ELSEVIER

Speech Communication 34 (2001) 159–174

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Transformation-based Bayesian prediction for adaptation of HMMs

Arun C. Surendran *, Chin-Hui Lee

Multimedia Communications Research Laboratory, Bell Labs, Lucent Technologies, 600 Mountain Ave, Murray Hill NJ 07974, USA

Abstract

Due to inaccuracies in the modeling procedure, estimation errors, and poor data to parameter ratios, adaptation techniques can perform poorly when only a limited amount of data is available. Modeling inflexibility, on the other hand, limits their potential when large amounts of data are present. In this paper, we present a transformation-based Bayesian predictive approach to hidden Markov model (HMM) adaptation that addresses the above problems. The new technique, called Bayesian predictive adaptation (BPA), treats adaptation as *model evolution* arising from *attempted transformation* of the model parameters. The transformation is a structural representation of the assumed mismatch between the trained models and the adaptation data. Instead of estimating the transformation parameters directly, and blindly treating the estimates as if they are the true values, BPA averages over the variation of the parameters to generate a new model that can be used in the decoding process. By combining the power of Bayesian prediction to take into consideration the errors in estimation and modeling, with the power of transformation based techniques to use fewer parameters for adaptation, the proposed approach creates a new family of techniques that tend to be robust to estimation and modeling errors when only limited data are available, and to modeling inflexibility when large amounts of data are present. We present adaptation results under channel and speaker mismatches, and compare the performance of BPA to other adaptation techniques to demonstrate its effectiveness.

Keywords: Speech recognition; Model adaptation; Bayesian prediction

1. Introduction

Powerful statistical learning techniques are used to train models for speech recognition using data collected from a limited number of acoustic conditions. The model typically consists of a large number of hidden Markov models (HMMs) each describing a particular phone unit, possibly a triphone. The nature of these techniques is such that, assuming that the model structure is accurate, the parameters are estimated to optimize some pre-determined criterion computed on the given data. Popular parameter estimation approaches are *maximum likelihood* (ML) (Rabiner, 1989) and *maximum a posteriori* (MAP) (Gauvain and Lee, 1994). As a result of training the models specifically on the given data, the systems do not perform as well when data from different acoustic conditions are presented during testing.

In order to bridge the acoustic mismatch between the new and the training environments, researchers have used *model adaptation*, statistical

* Corresponding author. Tel.: +1-908-5826351; fax: +1-908-5827308.

E-mail addresses: acs@research.bell-labs.com (A.C. Surendran), chl@research.bell-labs.com (C.-H. Lee).

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00052-2

learning techniques that adjust the parameters of the models of speech given some data that describe the new environment. Adaptation techniques are geared towards situations where complete retraining should be avoided, and when the amount of adaptation data available is much less than that during full-scale training. It is desirable to adapt as many HMM parameters as possible, even if the data corresponding to some of the phone models are not available. This is a challenging issue in adaptation because the algorithm is asked to meet very conflicting requirements – generalize from limited data, yet at the same time learn the data very well. Under such tough requirements, adaptation algorithms meet some unique challenges that traditional training algorithms do not face. The following questions should be addressed to determine how well an adaptation scheme performs:

- Modeling errors – is the model general enough to suit different acoustic conditions? In transformation based adaptation techniques, does the structure truly capture the nature of the mismatch between the training and testing environments?
- Is the model flexible enough so that limited adjustments using the new data can capture the mismatch sufficiently?
- Estimation errors – how well can the adaptation data represent the new environment and how close are the computed estimates to the actual values?
- Is the adaptation algorithm capable of capturing all the information the data set has to provide, while being robust to errors and outliers?

These are very important issues that are not addressed as much in adaptation techniques. We will now briefly review some adaptation techniques and discuss how they deal with the issues of modeling and estimation errors. There are two broad approaches to adaptation: (1) direct HMM parameter adaptation and (2) indirect transformation based parametric adaptation. Direct adaptation techniques usually employ some Bayesian schemes where the original HMM is considered as prior information and the new parameters are estimated as interpolations of the original HMM parameters and the empirical estimates computed from the given adaptation data. The so-called maximum a priori (MAP) adaptation of HMMs (Gauvain and Lee, 1994) for batch adaptation and the quasi-Bayesian adaptation schemes (Huo and Lee, 1997) for incremental adaptation fall under this category. There are many transformation based adaptation schemes (e.g. stochastic matching (Sankar and Lee, 1996), linear regression – maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) and MAP–LR (Chesta et al., 1999), non-linear transformation of model means (Surendran et al., 1996), etc.) that model the mismatch as structural transformations and estimate the HMM parameters through them. The functional forms of these transformations can be based on prior knowledge about the distortion. These transformation parameters can be shared among different models. Such a sharing can be performed in a hierarchical fashion so that tying is extensive when data available are limited, and can be gradually reduced as more data become available, e.g. structural MAP (Shinoda and Lee, 1998) and hierarchical on-line transformation (Chien, 1999). Other techniques like cluster adaptive training (CAT) (Gales, 1998) and Eigenvoice (Kuhn et al., 1998) express the model means of the new speaker as linear combinations of some basis vectors representing ''prototypical'' speakers. In CAT, the basis vectors are the mean values of speaker clusters, while in the Eigenvoice method, they are orthogonal vectors computed from the means of the models of many speakers using some dimensionality reduction technique like principal component analysis. The direct and indirect approaches mentioned above can also be used together, either one after another, or optimized jointly (Siohan et al., 2000; Digalakis and Neumeyer, 1996).

Most techniques assume that the modeling is accurate and the mismatch can be learned given sufficient data. They consider the estimates to be true and accurate instances of the desired parameter and use them as such in a decoder. MAP adaptation schemes perform very well when large amounts of data are present. However, when only a few utterances are available for adaptation, they do not do as well. The transformation based techniques model the mismatch

using fewer parameters and hence perform better than MAP when limited data are available. But they do not perform well when large amounts of data are present. This is because the structure used for transformation is unable to capture the mismatch between the two environments sufficiently, e.g. the bias transformation in stochastic matching (Sankar and Lee, 1996) assumes that the distortion is a slowly varying linear channel whereas channels encountered in practice, like telephone channels, are mostly non-linear. Hence when large amounts of data are available, the limited structure is unable to learn all the information in the data. We will demonstrate this later in our experiments. The techniques that use tying (e.g. SMAP) assume that their structure for tying captures the correlations accurately, and they do not take into consideration the estimation errors. Shinoda and Lee (1998) reported that transforming the variance parameters with limited amount of data is so unreliable that it degrades the performance. In CAT and eigenvoice approaches, only the linear combination weights need to be estimated for adaptation, hence they work well when limited data are available. But these approaches also assume that the basis vectors adequately model the important variations of the speakers, and the prototypical speakers represent the entire speaker space.

We can see that most techniques face problems due to modeling and estimation, but they do not address these problems directly. In this paper, we have tried to address these problems specifically.

Recently, a Bayesian predictive approach (Berger, 1980; Ripley, 1996) called Bayesian predictive classification (BPC) was used to compensate for modeling and estimation errors during the classification phase (Jiang et al., 1999; Huo et al., 1997). To understand BPC, let us briefly look at MAP adaptation. Given some adaptation data $\mathscr{X} = \{\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_T\}$, with its corresponding transcription W, the adapted model Λ_{MAP} is estimated as the value that maximizes the *posterior distribution* of Λ given $\mathscr{X}$,

$$\begin{aligned}\Lambda_{\text{MAP}} &= \operatorname*{argmax}_{\Lambda} P(\Lambda|\mathscr{X}, W) \\ &= \operatorname*{argmax}_{\Lambda} P(\mathscr{X}|\Lambda, W)P(\Lambda),\end{aligned} \tag{1}$$

where $P(\Lambda)$ is an assumed prior. During the recognition stage, with a given test utterance Y, Λ_{MAP} is incorporated directly in the MAP decoder to obtain the recognized sentence

$$\begin{aligned}\hat{W} &= \operatorname*{argmax}_{W} P(W|Y) \\ &= \operatorname*{argmax}_{W} P(Y|\Lambda_{\text{MAP}}, W)P(W).\end{aligned} \tag{2}$$

Here $P(Y|\Lambda, W)$ and $P(W)$ are the assumed parametric distributions of the acoustic and language models, respectively. Coming back to our discussion on BPC, to account for the uncertainty in the parameters, the acoustic model probability $P(Y|\Lambda_{\text{MAP}}, W)$ in Eq. (2) is replaced with a *predictive distribution*

$$\hat{P}(Y|W) = \int_{\Omega} P(Y|\Lambda, W)P(\Lambda)\,\mathrm{d}\Lambda, \tag{3}$$

where Ω is the space of the model Λ. $P(\Lambda)$, the prior of the model Λ, is assumed to be known. Thus the BPC decoder can be written as

$$\hat{W} = \operatorname*{argmax}_{W} \hat{P}(Y|W)P(W). \tag{4}$$

The hope here is that $P(\Lambda)$ can adequately capture the variation of the model due to errors, and hence the integral performed in Eq. (3) can effectively average over these errors. This method can be shown to minimize the overall recognition error, given the uncertainty (Huo et al., 1997). The quality of the prior can be improved using some training data $\mathscr{X}$ by replacing $P(\Lambda)$ with a posterior distribution of the model

$$P(\Lambda|\mathscr{X}, W) = \frac{P(\mathscr{X}|\Lambda, W)P(\Lambda)}{\int_{\Omega} P(\mathscr{X}|\Lambda, W)P(\Lambda)\,\mathrm{d}\Lambda}, \tag{5}$$

where W is the transcription of the training data. Thus the BPC decoder can be re-written as

$$\hat{W} = \operatorname*{argmax}_{W} \hat{P}(Y|\mathscr{X}, W)P(W), \tag{6}$$

where $\hat{P}(Y|\mathscr{X}, W)$ is another predictive density which is obtained by replacing $P(\Lambda)$ with $P(\Lambda|\mathscr{X}, W)$ in Eq. (3). Optionally, a parametric form $P(\Lambda|\phi)$ can be adopted to represent the posterior, and then its hyperparameters can be computed empirically from some training data

(Huo et al., 1997). Since the model has a large number of parameters, formulating the prior $P(\Lambda)$ and estimating its hyperparameters can be difficult especially when only limited data are available.

It would be easier to apply Bayesian prediction to a small number of parameters instead of all the model parameters. In addition, if information regarding the nature of the distortion becomes available, a better predictive density can be computed. In this paper, we use this idea to formulate a new scheme which applies Bayesian prediction to a limited number of transformation parameters that represent the structural form of the distortion. The proposed technique views adaptation as *evolution of the model* arising from attempted transformations of the model parameters. The goal of the new approach is to develop a technique that overcomes the estimation and modeling shortcomings faced by conventional adaptation techniques like MAP when only a small amount of adaptation data is present, and to account for the modeling insufficiency of transformation based techniques when large amounts of data are present. This goal is realized by combining the power of Bayesian prediction to normalize for variation due to errors, with the power of structural transformation techniques to capture the mismatch using a limited number of parameters. We call this approach Bayesian predictive adaptation (BPA). BPA is unique in the following ways:

- BPA uses Bayesian prediction on a small set of *transformation parameters* unlike BPC which applies prediction on the entire set of model parameters.
- Unlike MAP or ML based estimation techniques, BPA does not use the estimates of the parameters as-is in a plug-in MAP decoder. It averages over the uncertainty in the parameters to generate a new, predictive model of the data which is then used in decoding.

Although BPA is applied to simple transformations in this study, it is a general approach and can be extended to other transformations. We will analyze BPA in detail in Section 2.

The rest of the paper is organized as follows. In Section 2, we explain the idea of applying Bayesian prediction to model adaptation via transformation parameters. In Section 3, we discuss the problems of prior selection, posterior computation and predictive distribution calculation. In Section 4, we discuss how BPA can be used to enhance the adaptation of delta coefficients, which have a small dynamic range and hence can be more difficult to adapt than the cepstra. In Section 5, we present experimental results and finish with a summary in Section 6.

2. Bayesian predictive adaptation

As mentioned earlier, BPA views adaptation as *evolution of the model* arising from attempted transformations of model parameters. What this means is that BPA models the adaptation procedure as the transformation of the trained models Λ_X using a small number of transformation parameters; but then, instead of estimating these parameters, BPA averages over their variation using Bayesian prediction. Thus BPA yields a model that need not be equivalent to the model obtained by actually applying the transformation, i.e., the new model can have a different structure with a different family of distributions than the initial model or the model resulting from applying the transformation. In this sense, the model can be said to have evolved rather than adapted, to better suit the new condition. The meaning of this will become clearer from the following analysis.

Transformation based adaptation techniques assume that the mismatch between the training and testing environments can be modeled adequately by a function that transforms the parameters of the trained models,

$$\Lambda_Y = G_\theta(\Lambda_X), \tag{7}$$

where θ are the parameters of the transformation whose admissible range is assumed to be Θ. When the test data Y is available, it is decoded using the new model Λ_Y as

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(Y|\Lambda_Y, W)P(W) \tag{8}$$

$$= \underset{W}{\operatorname{argmax}} P(Y|\theta, \Lambda_X, W)P(W). \tag{9}$$

Since θ is unknown, it is often estimated using an optimization metric. For example, an ML

estimator (Sankar and Lee, 1996; Leggetter and Woodland, 1995) computes the value such that

$$\theta_{\mathrm{ML}} = \operatorname*{argmax}_{\theta} P(Y|\theta, \Lambda_X, W) \tag{10}$$

and the MAP estimator (Chien et al., 1997; Chesta et al., 1999) maximizes the posterior distribution $P(\theta|Y, \Lambda_X, W)$ such that

$$\begin{aligned}\theta_{\mathrm{MAP}} &= \operatorname*{argmax}_{\theta} P(\theta|Y, \Lambda_X, W) \\ &= \operatorname*{argmax}_{\theta} P(Y|\theta, \Lambda_X, W)P(\theta),\end{aligned} \tag{11}$$

where $P(\theta)$ is the prior distribution of the parameter θ. These point estimates are used directly in the likelihood function $P(Y|\theta, \Lambda_X, W)$ to be used in the decoder. If the distortion does not vary much, and if the distortion model is accurate, then the parameters of the transformation are well represented by a prior $P(\theta)$ with a sharp mode. In such cases, the point estimate of θ can adequately capture the distortion. In practice, distortion channels are non-linear and can be difficult to model with simple functions. Further, they can vary over time and also depend on the speech units of the utterance, e.g. a speaker with an accent can distort each phone in a different way. Such variations cannot be captured by a single parameter.

In BPA, instead of plugging the point estimate θ_{MAP} (or θ_{ML}) into $P(Y|\theta, \Lambda_X, W)$, and using the plugged-in version $P(Y|\theta_{\mathrm{MAP}}, \Lambda_X, W)$ (or $P(Y|\theta_{\mathrm{MAP}}, \Lambda_X, W)$) directly in the decoder, we compute a *predictive distribution* $\hat{P}(Y|\Lambda_X, \mathscr{X}, W)$ for each word sequence W,

$$\hat{P}(Y|\Lambda_X, \mathscr{X}, W) = \int_{\theta\in\Theta} P(Y|\theta, \Lambda_X, W)P(\theta|\mathscr{X})\,\mathrm{d}\theta, \tag{12}$$

where $P(\theta|\mathscr{X})$ is the posterior distribution of θ given some adaptation data $\mathscr{X}$. $\hat{P}(Y|\Lambda_X, \mathscr{X}, W)$ is called a "predictive distribution" since it can "predict" Y given Λ_X independent of the intervening unknown parameter θ (Berger, 1980). The decoder used in BPA can be written as

$$\hat{W} = \operatorname*{argmax}_{W} \hat{P}(Y|\Lambda_X, \mathscr{X}, W)P(W). \tag{13}$$

Comparing Eqs. (12) and (13) to those for BPC (Eqs. (3) and (6)) we can see that unlike BPC, BPA has not removed the dependence on the model – only the dependence on the distortion parameters. Since the transformation can have fewer parameters than the model, it is easier to calculate the prior of the parameters of the transformation rather than quantify the prior of the model parameters directly. Potentially another predictive distribution $\tilde{P}(Y|\mathscr{X}, W)$ can be further computed from $\hat{P}(Y|\Lambda_X, \mathscr{X}, W)$ using the prior distribution $P(\Lambda_X)$ of the model Λ_X.

Now, comparing the decoder for BPA (Eq. (13)) to that using MAP or ML estimate (Eq. (2)), we can see that BPA uses a new model that has been adapted to take into account the variations due to estimation and modeling errors. It can be shown that because of this averaging, the predictive approach is more robust, and minimizes the overall error given the prior distribution (Berger, 1980; Ripley, 1996). It is hoped that the predictive approach will account for estimation and modeling errors when limited data are available, and for modeling errors and modeling inflexibility when large amounts of data are available. If the prior has a sharp mode, the results of the predictive method and MAP approaches should be similar. The predictive technique is a conservative strategy and hence the focus is more on robustness (Berger, 1980; Ripley, 1996). It assumes that errors are present and tries to compensate for them. In situations where the parameters can be accurately estimated, i.e., when modeling errors are minimal and enough data is available to accurately estimate the parameters, MAP and ML approaches can outperform the predictive approach (Surendran and Lee, 1998). Otherwise, the predictive approach is more robust to changes in the environment (Surendran and Lee, 1998, 1999).

Similar to the MAP formulation, the key issues to be addressed in the new technique are the choice of prior family and estimation of the parameters of the posterior distribution. The additional problem in BPA is the evaluation of the predictive distribution given in Eq. (12). Frequently approximations have to be used to represent the posterior and the predictive distributions (Berger, 1980; Ripley, 1996). In speech recognition, since the underlying sequence of states is hidden, this makes the problem even more difficult.

The idea of Bayesian prediction can also be used for model compensation during testing in an unsupervised adaptation manner. This idea was used in (Surendran and Lee, 1998) to adapt the model means. We will discuss Bayesian predictive compensation in more detail in Section 6.

BPA can also be used in conjunction with hierarchical tying structures similar to those in (Shinoda and Lee, 1998; Chien, 1999; Furui, 1989) to further improve performance.

2.1. Choice of transformation functions

The first step in BPA is to choose functions that transform the model parameters. The choice of these functions depends on our idea of the mismatch. Further, the complexity of these transformations dictate the complexity of computing a posterior distribution and the difficulty of computing an accurate predictive density. The reader is referred to (Chesta et al., 1999) for a discussion of the difficulty of choosing priors and computing posteriors for linear regression of mean vectors. Since BPA can inherently compensate for modeling errors, the choice of the transformation can be simple, yet be effective.

In this paper, we assume additive biases to transform the means and multiplicative factors to transform the coefficients of the precision matrix. We assume that the probability distribution of an observation vector Y in state s of the HMM Λ_X is given by a mixture of K multivariate normal density functions,

$$P(Y|\Lambda_X, s) = \sum_{k=1}^{K} w_{sk} P(Y|\Lambda_X, s, k) = \sum_{k=1}^{K} w_{sk} \mathcal{N}(Y|\mu_{sk}, R_{sk}), \quad (14)$$

where w_{sk} is the weight of the mixture component k in state s such that $\sum_{k=1}^{K} w_{sk} = 1$, and

$$\mathcal{N}(Y|\mu, R) \propto |R|^{1/2} \exp\left\{ -\frac{1}{2}(Y-\mu)^t R (Y-\mu) \right\} \quad (15)$$

is a normal distribution with mean μ and precision matrix R. If each mixture component k of state s has a mean vector μ_{sk} and a diagonal precision matrix R_{sk} whose diagonal entries are r_{ski}, $i = 1, \ldots, D$, then we assume that the means are transformed using an additive bias vector ℓ_{sk}, and the ith coefficient of the precision matrix is transformed using a multiplicative factor η_{ski},

$$\mu^*_{sk} = \mu_{sk} + \ell_{sk}, \quad (16)$$

and

$$r^*_{ski} = \eta_{ski} r_{ski}, \quad i = 1, \ldots, D, \quad (17)$$

where D is the number of components in the feature vector and $\theta = (\ell, \eta) \in (\mathscr{B} \times \Xi)$ is the parameter set of the transformation. $(\mathscr{B} \times \Xi)$ is the admissible space of the parameter $\theta = (\ell, \eta)$. Instead of using different sets of parameters for each mixture component, it is possible to use *parameter tying*, i.e. divide the mixture components into groups and share the transformation functions among the group. In such a case, the parameters ℓ_{sk} and η_{ski} in Eqs. (16) and (17) can be replaced by $\ell_{f(s,k)}$ and $\eta_{f(s,k)i}$, where $f(s,k)$ indicates a tying function that maps the mixture components into clusters based on the state s and mixture index k.

Although BPA is used in this study with the simple transformations mentioned above, it is a powerful approach that in general can be applied to any transformation function. For example, we can apply it to linear regression parameters.

3. Prior, posterior and predictive distributions

3.1. Prior selection

An important element of the predictive approach is the selection of a prior probability distribution of $\theta = (\ell, \eta)$ over $(\mathscr{B} \times \Xi)$. Care must be taken to see that the prior/posterior reflects the variability in the transformation parameters. At the same time, the form of the prior should also be such that the posterior and the predictive densities in Eq. (12) are mathematically tractable, computationally manageable and useful. The form of the prior can be chosen using empirical evidence. In some cases, the final predictive density arising from such a choice might not be suitable for the

given decoder (Surendran and Lee, 1998). In such cases, it may be better to choose an approximate prior that may not model the data exactly but may suit the decoding process well. In (Surendran and Lee, 1998) we chose a predetermined form of the prior density and computed its parameters directly from the data using empirical techniques.

Another popular approach is to choose the prior such that the prior and the posterior belong to the same family of distributions. Such a choice of prior is called a *conjugate prior* (DeGroot, 1970). We choose the conjugate prior in this paper since it not only makes the computation easier, but also makes it easy to extend the adaptation procedure to incremental and hierarchical approaches. For a Gaussian likelihood function with unknown mean and unknown multiplicative factor for the precision matrix, the conjugate prior is a normal-gamma density (DeGroot, 1970), i.e., the conditional distribution of the mean given the multiplicative factor is a normal distribution, and the marginal of the multiplicative factor is a gamma distribution (DeGroot, 1970). In speech recognition, since each state in the HMM is assumed to be a mixture of Gaussians, no sufficient statistic, and hence no conjugate prior density exists. A good approximation is to use the product of the conjugate prior densities of each component of the mixture (Gauvain and Lee, 1994). Assuming a diagonal precision matrix and assuming each component of the feature vector is transformed independent of the others, the joint conjugate prior $P(\theta)$ can be written as a product of the conjugate priors of each coefficient i,

$$P(\theta) = P(\ell|\eta)P(\eta), \tag{18}$$

with

$$P(\ell|\eta) \propto \Pi_{i=1}^{D} |\eta_i \tau_i|^{1/2} \exp\left[-\frac{\eta_i \tau_i}{2} (b_i - m_i)^2 \right], \tag{19}$$

$$P(\eta) \propto \Pi_{i=1}^{D} \eta_i^{(\alpha_i - 1)} \exp(-\beta_i \eta_i), \quad i = 1, \ldots, D, \tag{20}$$

where $\alpha_i > 0$, $\beta_i > 0$ are the hyperparameters of the gamma distribution and $\tau_i > 0$ and m_i are the hyperparameters of the conditional normal distribution.

3.2. Initial estimation of hyperparameters

The initial estimate of the hyperparameters should sufficiently reflect the mismatch present between the data and the model. A good initial value would be one that reflects an "average" mismatch – a global value that is computed using all the data.

In the gamma distribution, the initial values were assumed to be $\alpha = 1.0$ and $\beta = 1.0$. This implies that the initial mean of the multiplicative factor is 1.0 ($E(\eta) = \alpha/\beta$). This also makes the variance ($\mathrm{var}(\eta) = \alpha/\beta^2$) equal to 1. This has no physical meaning but is mathematically convenient. The mean and precision of ℓ in the conditional normal distribution are estimated using empirical Bayes techniques (Berger, 1980). Given the decoded state sequence $S = \{s_1, s_2, \ldots, s_T\}$ and the data $\mathscr{X} = \{\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_T\}$, a global mean vector m is estimated as

$$m = \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} (\mathscr{X}_t - \mu_{s_t k}), \tag{21}$$

and the ith coefficient of τ can be computed as

$$\frac{1}{\tau_i} = \frac{1}{T-1} \sum_{t=1}^{T} \sum_{k=1}^{K} (\mathscr{X}_{ti} - \mu_{s_t k i} - m_i)^2. \tag{22}$$

When only a limited amount of data is available, the above estimates can be unreliable especially for the delta and delta–delta coefficients. To overcome this problem, we use a transformation based hierarchical approach to compute the hyperparameters of the delta coefficients. This procedure is detailed in Section 4.

3.3. Computing the posterior distribution

Computing the posterior distribution can be the most difficult part of the entire procedure. This is where most often approximations need to be used that can determine the overall performance.

Given the prior $P(\ell, \eta)$, the adaptation data $\mathscr{X} = \{\mathscr{X}_1, \mathscr{X}_2, \ldots, \mathscr{X}_T\}$ and the HMM Λ_X, we need to compute the posterior distribution $P(\ell, \eta|\mathscr{X}, \Lambda_X)$. Since the underlying state sequence $S = \{s_1, s_2, \ldots, s_T\}$ and the mixture component

sequence $C = \{c_1, c_2, \ldots, c_T\}$ are missing, the posterior can be computed as

$$P(\ell, \eta|\mathscr{X}, \Lambda_X) = \sum_S \sum_C P(\mathscr{X}|\ell, \eta, S, C, \Lambda_X) P(S, C|\Lambda_X, \ell, \eta) \times P(\ell, \eta)/P(\mathscr{X}|\Lambda_X). \quad (23)$$

The missing data problem makes the above calculation cumbersome and difficult. Since it is not possible to directly compute the above distribution, there are many ways to approximate it. Bayesian data augmentation schemes (Tanner and Wong, 1990), Gibbs sampler (Gelfand and Smith, 1990) or quantile integration (Johnson, 1992) are just a few approaches that can be used. One approach would be to approximate the posterior using a different function which is much easier to compute than Eq. (23) and yet retains the major modes of the posterior. Hence the posterior can be replaced by

$$\Psi(\ell, \eta|\mathscr{X}, \Lambda_X) = \exp(R(\ell, \eta|\mathscr{X}, \Lambda_X)), \quad (24)$$

where

$$R(\ell, \eta|\mathscr{X}, \Lambda_X) = E\{\log P(\mathscr{X}, S, C|\ell, \eta) + \log P(\ell, \eta)|\mathscr{X}, \Lambda_X\} \quad (25)$$

$$= Q(\ell, \eta|\mathscr{X}, \Lambda_X) + \log P(\ell, \eta). \quad (26)$$

Readers familiar with the EM algorithm (Dempster et al., 1977; Gauvain and Lee, 1994) can identify this expression as the auxiliary function that is maximized in the MAP approach.

Following an analysis similar to (Gauvain and Lee, 1994) and (Shinoda and Lee, 1998), we can see that $\Psi(\ell, \eta|\mathscr{X}, \Lambda_X)$ is also a product of normal-gamma distributions. For a given state s, the ith coefficient of the new parameters of the posterior distribution can be computed as (Surendran and Lee, 1999)

$$m_i^* = \frac{\tau_i m_i + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(s,k) r_{ski} (\mathscr{X}_{ti} - \mu_{ski})}{\tau_i + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(s,k) r_{ski}}, \quad (27)$$

$$\tau_i^* = \tau_i + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(s,k) r_{ski}, \quad (28)$$

$$\alpha_i^* = \alpha_i + \frac{1}{2} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(s,k) \quad (29)$$

and

$$\beta_i^* = \beta + \frac{1}{2} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(s,k) r_{ski} (\mathscr{X}_{ti} - \mu_{ski} - \overline{\mathscr{X}}_i)^2 + \frac{1}{2} (m_i^* - m_i) \tau_i (\overline{\mathscr{X}}_i - m_i), \quad (30)$$

where

$$\overline{\mathscr{X}}_i = \frac{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(s,k) r_{ski} (\mathscr{X}_t i - \mu_{ski})}{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(s,k) r_{ski}}, \quad i = 1, \ldots, D. \quad (31)$$

where the subscript i indicates the index of the coefficient. $\gamma_t(s,k)$ is the probability of being in state s with mixture component label k at time t (Gauvain and Lee, 1994), and is calculated from the forward–backward algorithm. As mentioned before, parameter tying can be used to improve the quality of the estimates (Shinoda and Lee, 1998; Chien, 1999). In such a case, the summation over the mixture component index ($\sum_{k=1}^{K}$) in the above equations should be replaced by a summation over all components in the tying cluster.

The above expression can once again be computationally intensive. A useful approximation can be to use a segmental approach, where the state sequence S is computed using standard decoding techniques and then the posterior is computed based on that segmentation. The only change to the above equations would be in the way $\gamma_t(s,k)$ is computed,

$$\gamma_t(s,k) = \begin{cases} \dfrac{w_{sk} \mathscr{N}(y_t|\mu_{sk}, r_{sk})}{\sum_{j=1}^{K} w_{sj} \mathscr{N}(y_t|\mu_{sj}, r_{sj})} & \text{if } s_t = s, \\ 0 \quad \text{otherwise.} \end{cases} \quad (32)$$

The posteriors for the delta parameters are computed after the cepstra are adapted. Hence that discussion is postponed to Section 4.

This method can also be extended to an on-line adaptation scheme where the data are made available sequentially, either in chunks or in a stream. Using the posterior computed as each step as the prior for the next stream of data, BPA can then be implemented incrementally.

3.4. Estimating the predictive density

As mentioned before the predictive density in BPA is computed by deriving the marginal of the data given the model. This means that the influence of the nuisance or transform parameters is removed from the likelihood computation. In the process of removing their influence, what is left behind is an averaging of their effect over the entire range of their variation – a density that can "predict" how the data behaves given the model without the intervening nuisance parameters.

Whether the predictive density has a closed-form solution or not depends on the forms of the likelihood and the posterior distributions. Frequently approximations have to be used. Even if the predictive density has a closed-form solution, it may not be useful in a decoder – for example, if the mean adaptation produces a non-symmetric density whose mode is different from its mean, it will be of no use in an MAP decoder that prefers the mode (Surendran and Lee, 1998).

For a given component k of state s of the HMM, the predictive density can be computed as

$$P(Y|\Lambda_X,\mathscr{X},s,k) = \int_{\eta\in\Xi}\left[\int_{\ell\in\mathscr{B}} P(Y|\ell,\eta,\Lambda_X,\mathscr{X},s,k)P(\ell|m^*,\tau^*,\eta)\,d\ell\right] \times P(\eta|\alpha^*,\beta^*)\,d\eta, \quad (33)$$

where m^*,τ^*,α^* and β^* are computed from Eqs. (27)–(30). The integration over ℓ can be done easily since the likelihood function and the posterior are Gaussian. After integrating over ℓ the predictive density can be written as

$$P(Y|\Lambda_X,\mathscr{X},s,k) = \int_{\eta\in\Xi} P(Y|\eta,\Lambda_X,\mathscr{X},s,k,m^*,\tau^*)P(\eta|\alpha^*,\beta^*)\,d\eta, \quad (34)$$

where $P(Y|\eta,\Lambda_X,\mathscr{X},s,k,m^*,\tau^*)$ is a normal density with mean vector

$$\mu^*_{sk} = \mu_{sk} + m^*, \quad (35)$$

and precision matrix r^*_{sk} whose ith coefficient is

$$r^*_{ski} = \frac{r_{ski}\tau^*}{r_{ski}+\tau^*}, \quad i = 1,\ldots,D. \quad (36)$$

After integrating over the posterior distribution of the multiplicative factor, the predictive distribution is a t-distribution with $2\alpha^*$ degrees of freedom, location parameter μ^* and precision $\alpha^* r^*/\beta^*$,

$$P(Y|\Lambda_X,\mathscr{X},s,k) \propto \left[1+\frac{1}{2\alpha^*}\frac{\alpha^* r^*(y-\mu^*)^2}{\beta^*}\right]^{-(2\alpha^*+1)/2}. \quad (37)$$

Comparing Eq. (37) to Eq. (14) we can see that the form of the mixture component, and hence the state distribution has evolved.

In theory, this adaptation scheme can be extended to the mixture weights and transition probabilities in the HMM.

The predictive distribution (Eq. (37)) can be used as such in the decoder. In this experiment, a normal distribution which is a minimum divergence approximation to Eq. (37) was used. It can be shown that such an approximation will have parameters

$$\text{mean} = \mu^* \quad \text{and} \quad \text{var} = \frac{\beta^*}{(\alpha^*-1)r^*}. \quad (38)$$

4. Applying BPA to delta coefficients

Adapting the delta and delta–delta coefficients is particularly difficult when limited data is available. This is because the dynamic range of the deltas, especially for the higher order coefficients, is very small, and amount of data needed to adapt them increases as their dynamic range shrinks. One solution to this problem is to not adapt the delta parameters at all. This solution is attractive when limited data is available, but as the amount of data increases the delta coefficients can be adapted reliably, and not adapting them hurts performance (Surendran, 2000).

Recently, we proposed a novel transformation based hierarchical Bayesian scheme for adapting delta coefficients (Surendran, 2000). We use that technique in this paper in the framework of BPA to improve performance when limited data is available. The motivation for the new procedure is detailed in (Surendran, 2000) but we briefly review

it here: Since the delta coefficients can be considered as transforms of the cepstra (Sankar and Lee, 1996; Surendran, 2000), we can conclude that the transformations of delta-coefficients are themselves functions of the transforms of the cepstral coefficients. Since the estimate of the cepstral coefficient are much more reliable than the deltas when limited data is available, deriving the transforms of the deltas through the cepstra is more reliable. But as more data become available, it is better to estimate the transforms of the deltas from the data itself. Using these facts, we adopt a two-step hierarchical Bayes scheme to adapt the deltas. First, the cepstra are adapted using BPA. Knowing the transformation from the cepstra to the delta coefficients, we derive the prior distribution of the deltas based on the posterior distribution of the cepstral features. Once the new prior is determined, the deltas are adapted using BPA.

Thus the proposed hierarchical method helps the adaptation scheme perform close to the best of its ability given a specific amount of data.

The delta cepstra are computed by $\Delta C_{l,m} = \sum_{k=-K}^{K} GkC_{l-k,m}$, where $\Delta C_{l,m}$ and $C_{l,m}$ are the mth delta-cepstral and the cepstral coefficients for the lth frame of the signal. $G = 0.375$ is a gain term, and $K = 2$. The delta–delta cepstra are computed as $\Delta^2 C_{l,m} = \sum_{n=-N}^{N} Gn\Delta C_{l-n,m}$, where $\Delta^2 C_{l,m}$ is the mth delta–delta cepstral coefficients for the lth frame (Sankar and Lee, 1996). From the above expressions, we can determine that the means of the deltas are zero, and the variances can be computed as

$$\sigma^2_{\Delta C_{l,m}} = \sum_{k=-K}^{K} G^2 k^2 \sigma^2_{C_{l-k,m}}, \tag{39}$$

$$\sigma^2_{\Delta^2 C_{l,m}} = \sum_{n=-N}^{N} \sum_{k=-K}^{K} G^4 n^2 k^2 \sigma^2_{C_{l-k-n,m}}. \tag{40}$$

Assuming that the HMM parameters are transformed using Eqs. (16) and (17), the transformation parameters of the deltas can be computed as

$$b_\Delta = 0, \tag{41}$$

$$b_{\Delta^2} = 0, \tag{42}$$

$$\eta_\Delta = 1 \Big/ \left(\sum_{k=-K}^{K} G^2 k^2 \right) = 1/A, \tag{43}$$

$$\eta_{\Delta^2} = 1 \Big/ \left(\sum_{n=-N}^{N} \sum_{k=-K}^{K} G^4 n^2 k^2 \right) = 1/B, \tag{44}$$

where $A = \sum_{k=-K}^{K} G^2 k^2$ and $B = \sum_{n=-N}^{N} \sum_{k=-K}^{K} G^4 n^2 k^2$ are constant values computed from Eqs. (43) and (44). Here we assumed that the adjacent frames of data are independent and usually have similar (or the same) variance values.

Once the cepstras have been adapted, using Eqs. (38)–(40), (43) and (44) the new variances of the delta coefficients can be written as

$$(\sigma^*)^2_\Delta = (A/w^*)(\sigma^2 + s^*), \tag{45}$$

$$(\sigma^*)^2_{\Delta^2} = (B/w^*)(\sigma^2 + s^*), \tag{46}$$

where $w^* = (\alpha^* - 1)/\beta^*$ and $s^* = 1/\tau^*$.

The prior of the delta coefficients can now be computed to reflect the new values in Eqs. (45) and (46). Once again assuming a Normal-gamma distribution for the prior, the hyperparameters can be computed as

$$m_\Delta = 0, \tag{47}$$

$$\tau_\Delta = \tau^*, \tag{48}$$

$$\beta_\Delta = \frac{A}{w^*} \alpha_\Delta, \tag{49}$$

$$\alpha_\Delta = \alpha^*. \tag{50}$$

α_Δ can also be chosen heuristically and determines the learning rate of the system. The hyperparameters for delta–delta coefficients are calculated similarly, except A is replaced by B.

The adaptation of the deltas is carried out as described by Eqs. (27)–(30).

5. Experimental results

5.1. Database description

Sentences from the 991-word DARPA resource management (RM) task (Price et al., 1988) were spoken by five non-native male speakers and recorded simultaneously through two channels: (1) a close-talking microphone (MIC), and (2) a telephone handset over a dial-up line (TEL). The fact that MIC and TEL data were recorded

simultaneously was not exploited in the adaptation scheme.

The data consisted of 300 sentences for adaptation and 75 sentences for testing. The speech was down-sampled from 16 kHz to 8 kHz. For each 30 ms frame (with 20 ms overlap), a 38-dimensional feature vector (12 cepstra, 12 Δ-cepstra, 12 ΔΔ-cepstra, Δ-log energy, ΔΔ-log energy) was extracted based on 10th order LPC analysis. 1769 context dependent (CD) subword unit models were built, with a maximum of 16 mixture components per state (Lee et al., 1992). The RM word pair grammar which gives a perplexity of about 60 was used for the experiments.

A set of speaker independent (SI) HMMs (Λ_{SI}) were generated from the RM SI-109 training set consisting of 3990 utterances from 109 American talkers (31 females, 78 males). The SI model Λ_{SI} was further adapted using MAP adaptation (Gauvain and Lee, 1994) on the male data only to give a speaker-adaptive male speaker model (Λ_{SIM}). The word accuracy for the male speaker models on native speakers using high quality desktop microphones was 96% in most reported results, e.g., (Lee et al., 1993).

Recognition results are provided for two different cases:

- Λ_{SIM}, the speaker-independent male model used with MIC test data spoken by non-native speakers – speaker mismatch. The baseline word accuracy for this data set is 74.5% (Table 1).
- Λ_{SIM}, the speaker-independent male model used with TEL test data spoken by non-native speakers – channel, speaker and transducer mismatch with possible additive noise. The baseline word accuracy for this data is 42.6% (Table 2). This performance indicates that the degradation due to the telephone channel is quite severe.

We can see that this is a challenging database which provides different types of mismatches and is well suited to test the performance of our new technique.

We compare the performance of BPA with transformation based approaches as well as

Table 1
Word accuracy with change in amount of adaptation data when Λ_{SIM} tested on MIC data for five speakers

	Number of sentence					
	A	B	C	D	E	Avg.
Baseline	74.0	51.2	84.5	73.6	89.3	74.5
1	78.7	70.4	88.7	82.6	89.2	81.9
5	80.0	69.1	89.3	82.9	88.4	81.9
10	81.5	70.3	89.3	86.6	87.8	83.1
50	87.4	67.4	88.4	85.4	89.2	83.6
100	90.4	77.9	89.9	88.3	91.3	87.6
200	89.6	85.4	93.8	92.2	94.0	91.0
300	92.2	88.4	94.1	94.4	93.8	92.6

Table 2
Word accuracy with change in amount of adaptation data when Λ_{SIM} tested on TEL data for five non-native speakers

	Number of sentence					
	A	B	C	D	E	Avg.
Baseline	50.3	18.3	52.9	35.7	55.7	42.6
1	65.0	56.3	70.4	67.7	69.5	65.8
5	66.7	47.7	70.6	66.1	70.0	64.2
10	64.9	48.5	66.4	69.5	71.0	64.1
50	70.6	52.0	72.8	71.3	69.5	67.2
100	75.1	68.8	78.7	79.4	75.5	75.5
200	83.2	81.8	85.6	89.8	83.8	84.8
300	86.9	86.9	89.5	90.8	86.3	88.1

conventional MAP adaptation. Since, in this paper, bias transforms for the means and multiplicative transforms for the variances are used for BPA, we can directly compare this formulation to stochastic matching based adaptation of model parameters (SM) presented in (Sankar and Lee, 1996). SM uses bias transformations of both model means and model variances. Although SM was originally used for unsupervised compensation during testing (Sankar and Lee, 1996), in this paper we have used it for supervised adaptation. We have also compared BPA to MLLR. Since the transformations used in MLLR (linear regression) are more sophisticated than the ones used by BPA in this paper, a direct comparison is not entirely fair. Ideally, MLLR should be compared to Bayesian prediction applied to linear regression (BP–LR). Nevertheless, the comparison brings out some interesting observations. SM and MLLR experiments were done with different number of transformation at different data levels. Two (one for silence and one for speech), six (one for each broad phoneme classes), 47 (one for each phone) and 1769 (one for each model) transformations were used at each data level, and the number that gave the best performance was selected. This type of hand-tuning was not done for BPA or MAP. Results presented are averaged over the five speakers.

Table 1 shows the results of BPA performed on the MIC data for each one of the five speakers. Even when using only one sentence for adaptation, BPA improves the word accuracy from 75.4% to 81.9%. SM performs quite poorly with limited data, and shows improvement over the baseline only when a lot of data is available. It is important to note that BPA outperforms SM significantly. The performance of BPA is similar to MLLR when limited data are available. But while the performance of MLLR stays flat as more data becomes available, BPA improves its performance steadily and outperforms MLLR (85.1% versus 92.6% for BPA with 300 sentences). BPA is better than MAP with limited data. But as the amount of data increases, MAP outperforms the predictive approach. The comparisons with SM, MAP and MLLR are shown in Fig. 1.

Table 2 shows the performance of BPA for the five different speakers using TEL data. The initial average word accuracy is 42.6%. Even when only one sentence is used for adaptation, BPA improves the performance to 65.8%. Once again BPA outperforms SM everywhere, and outperforms MAP with limited data. But MLLR performs better than BPA when limited data are present, e.g. 73.3% WA versus 65.8% for BPA with one sentence. Similar to the results on the MIC data, as the amount of data increases, BPA and MAP both outperform MLLR. Fig. 2 shows the performance of the predictive approach when compared to SM, MAP and MLLR. We will discuss the results in the following section.

BPA was shown to be effective even when used for model compensation, i.e. model adaptation during testing done in an unsupervised manner (Surendran and Lee, 1998). In that paper, only the model means were adapted. A parametric form was assumed for the prior and the parameters were estimated using the test data only. This technique was shown to be more effective than a simple mean compensation using stochastic matching.

5.2. Discussion

5.2.1. BPA versus SM

Since BPA (as presented in this paper) and SM have similar complexity, comparison between them best highlights the power of BPA. It is clear from the above results that BPA significantly outperforms SM under all circumstances. When limited data are available, SM performs poorly due to its limited modeling power as well as errors in estimation, while BPA is robust to these errors. With the MIC data, SM degrades the performance below the baseline value. When a lot of data is available, the performance of SM improves but significantly underperforms BPA. This shows the shortcomings of its limited structure, while BPA, using a similar limited structure, is able to learn much more from the data. Thus we can conclude that BPA compensates for modeling inflexibility.

5.2.2. BPA versus MLLR

As mentioned before, the direct comparison between BPA of simple transformations and MLLR is not entirely fair. But some surprising observations emerge from the results. First of all,

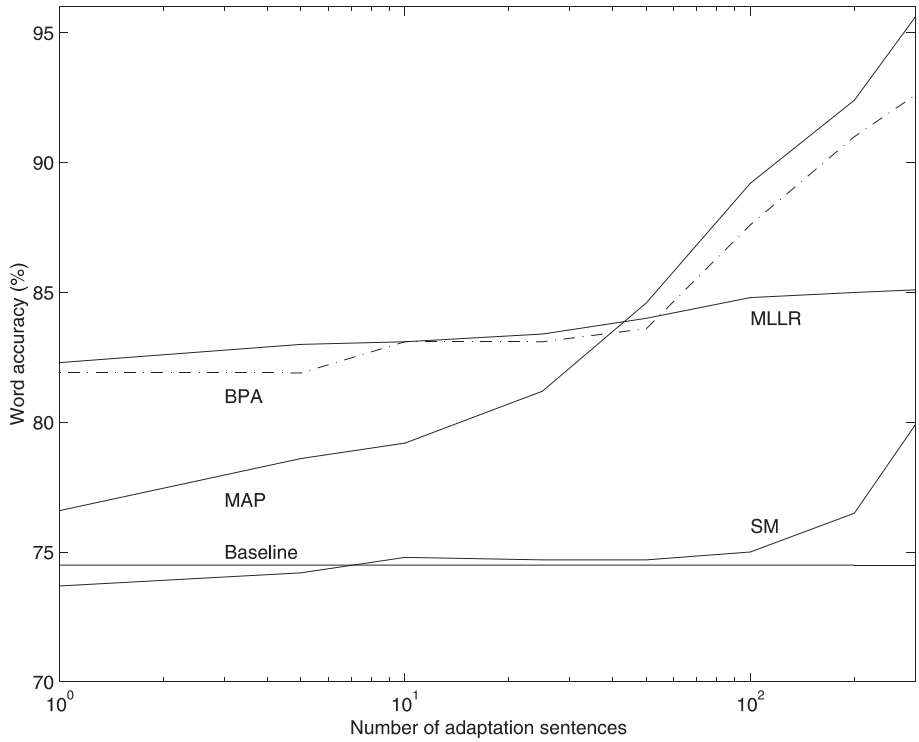

Fig. 1. Average word accuracy for five non-native speakers using BPA, SM, MLLR and MAP when Λ_{SIM} models were tested on MIC data.

we can observe that the performance of MLLR is good with limited data, but does not improve significantly when large amounts of data become available. This is a general limitation of all transform based approaches. When a limited amount of data is available, there are fewer parameters to estimate and hence they do well. But as more data become available, the transformation functions are restricted by their fixed structure, and fail to take advantage of all the information that is available. Since BPA in this paper uses a structure that is much less sophisticated than MLLR, we would expect it to be worse than MLLR for all amounts of data. Surprisingly, it performs better than MLLR for both MIC and TEL data when large amounts of data are present. The advantage in modeling power should help MLLR perform better than the simple BPA at lower data levels, and it does so for the TEL data. For the MIC data, BPA is almost as good as MLLR (most of the differences up to 50 sentences are statistically insignificant). This reinforces the point that BPA, to a certain extent, normalizes modeling inaccuracies and is able to go beyond the restrictions of its components.

Since Bayesian prediction can be applied to any transformation technique, it can be applied to linear regression also. But the mathematical formulation, computation of the posterior and the predictive distributions are more complicated and may not be tractable. More work is needed in this direction.

5.2.3. BPA versus MAP

BPA performs better than MAP when limited amounts of data are present. As discussed before, this is partly due to the fact that BPA is a transformation based technique and hence uses fewer parameters. But the same reason restricts its performance as compared to MAP when large

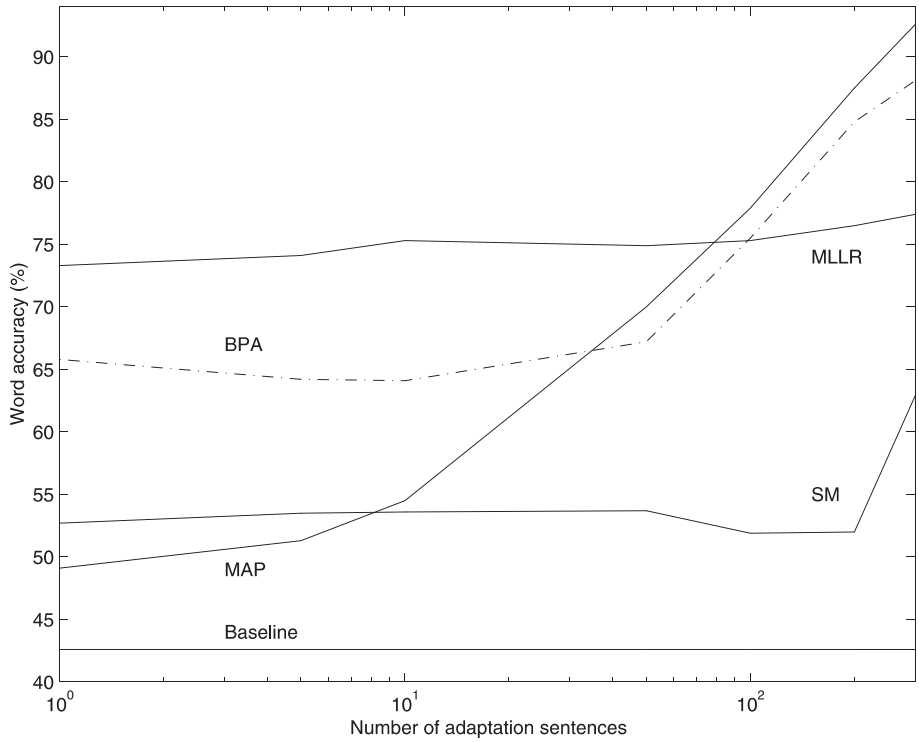

Fig. 2. Average word accuracy for five non-native speakers using BPA, SM, MLLR and MAP when Λ_{SIM} models were tested on TEL data.

amounts of data are available. Adaptation approaches like MAP are less restrictive in terms of their assumptions, so they are able to give better performance when more data are available. As discussed before, one way to improve the performance of BPA is to apply it to more elaborate transformation functions. Further improvement can be achieved by using it in hierarchical frameworks similar to SMAP (Shinoda and Lee, 1998).

Another reason why BPA may perform worse than MAP when a lot of data is available could be that the predictive approach is a conservative strategy – it assumes that there are errors in estimating the parameters and compensates for them. The MAP approach assumes that the estimated values are true. When a limited amount of data is available, the conservative strategy is more robust, but as more and more data become available, the MAP estimates are more accurate and hence do better for ASR.

6. Summary

In this paper, we have introduced a new adaptation paradigm that combines the powers of structure based transformation techniques and Bayesian prediction to account for two kinds of errors: (1) estimation and modeling errors, which occur when limited data are available, and (2) modeling inflexibility, which limits the ability of transformation based techniques when large amounts of data are available. Unlike MAP and BPC, which operate on the model parameters directly, BPA operates on the model indirectly using a limited number of transformation parameters. Instead of estimating these parameters and using them in a decoder directly like MLLR, MAP or SM, BPA creates a new model using Bayesian prediction that takes into account the variation in the parameters due to the uncertainty in estimation and inaccuracies in modeling. Since the

transformation parameters are limited in number, their distributions can be estimated more reliably than the distributions of the model parameters (which are needed in BPC). The transformations also allows us to incorporate into the adaptation any additional information that we may have regarding the nature of the distortion. In this paper, we have transformed the means and variances of the HMMs using additive biases and multiplicative factors, respectively. We have presented adaptation results to show that BPA is effective under different mismatch conditions. Our comparisons with SM, MLLR and MAP have shown that BPA indeed performs the tasks it has set out to do. It is significantly better than a transformation technique of similar complexity (SM) under all conditions. When the amount of adaptation data is limited, BPA is more effective than MAP. Even though, due to the restrictive modeling of transform based adaptation, BPA tends to perform slightly worse than MAP when large amounts of data are available, it is better than more complex transformation techniques like MLLR. We believe that this modeling restriction can be alleviated by using a more elaborate transformation technique and/or by using a hierarchical data tying techniques. In a related work, the predictive approach has been shown to be effective for compensation during testing in an unsupervised manner.

References

Berger, J.O., 1980. Statistical Decision Theory and Bayesian Analysis, second ed. Springer, New York.

Chesta, C., Siohan O., Lee, C.-H., 1999. Maximum a posteriori linear regression for hidden Markov model adaptation. In: Proc. Eurospeech 99, Budapest, September 1999, pp. 211–214.

Chien, J.T., 1999. On-line hierarchical transformation of hidden Markov models for speech recognition. IEEE Transactions on Speech and Audio Processing 7 (6), 656–667.

Chien, J.T., Wang, H.-C., Lee, C.-H., 1997. Bayesian affine transformation of HMM parameters for instantaneous and supervised adaptation in telephone speech recognition. In: Proc. Eurospeech, Rhodes, Greece, pp. 2575–2578.

DeGroot, M.H., 1970. Optimal Statistical Decisions. McGraw-Hill, New York.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series (B) 39, 1–38.

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

Furui, S., 1989. Unsupervised speaker adaptation based on hierarchical spectral clustering. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, Glasgow, pp. 286–289.

Gales, M., 1998. Cluster adaptive training for speech recognition. In: Proceedings of the International Conference on Spoken Language Processing '98, Sydney, Australia, Vol. 5, pp. 1783–1786.

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

Gelfand, A., Smith, A.F.M., 1990. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Huo, Q., Lee, C.-H., 1997. On-line adaptive learning of continuous density hidden Markov models based on approximate recursive Bayes estimate. IEEE Transactions on Speech and Audio Processing 5, 161–172.

Huo, Q., Jiang, H., Lee, C.-H., 1997. A Bayesian predictive classification approach to robust speech recognition. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing 97, pp. II.1547–II.1550.

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

Johnson, V.E., 1992. A technique for estimating marginal posterior densities in hierarchical models using mixtures of conditional densities. Journal of the American Statistical Association 87 (419), 852–860.

Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K., Contolini, M., 1998. Eigenvoices for speaker adaptation. In: Proceedings of the International Conference on Spoken Language Processing '98, Sydney, Australia, Vol. 5, pp. 1771–1774.

Lee, C.-H., Giachin, E., Rabiner, L.R., Pieraccini, R., Rosenberg, A., 1992. Improved acoustic modeling for large vocabulary continuous speech recognition. Computer Speech and Language 6, 103–127.

Lee, C.-H., Gauvain, J.-L., Pieraccini, R., Rabiner, L.R., 1993. Large vocabulary speech recognition using subword units. Speech Communication 13, 263–279.

Leggetter, C.J., Woodland, P.C., 1995. Speaker adaptation of HMMs using linear regression. Computer Speech and Language 9, 171–185.

Price, P., Fisher, W., Bernstein, J., Pallet, D., 1988. A database for continuous speech recognition in a 1000-word domain. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 651–654.

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

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

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

Shinoda, K., Lee, C.-H., 1997. Structural MAP speaker adaptation using hierarchical priors. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara.

Siohan, O., Chesta C., Lee, C.-H., 2000. Joint maximum a *posteriori* estimation of transformation and HMM parameters. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing.

Surendran, A.C., 2000. Hierarchical Bayes approach to adapting delta and delta-delta Cepstra. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing.

Surendran, A.C., Lee, C.-H., 1998. Predictive compensation and adaptation for robust speech recognition. In: Proceedings of the International Conference on Spoken Language Processing '98, Sydney, Australia.

Surendran, A.C., Lee, C.-H., 1999. Bayesian predictive approach to adaptation of HMMs. In: Proceedings of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Finland.

Surendran, A.C., Lee, C.-H., Rahim, M., 1996. Non-linear compensation for stochastic matching. IEEE Transactions on Speech and Audio Processing 7 (6), 643–655.

Tanner, M., Wong, W.H., 1990. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–550.