T-61.182 Special course on Information Technology

Robustness in Language and Speech Processing

Bayesian Speaker Adaptation Methods

Ramūnas Girdziušas
email: ramunasg@cc.hut.fi

11th June 2003

1 Introduction

The performance of speaker independent automatic speech recognizers (ASR) nowadays is sufficient to be applied to many realistic tasks. However, it is very common that an ASR system works well for most of its users, but for some speakers its performance is significantly degraded. To circumvent this problem one modifies all or some of the ASR parameters so that the ASR system gets matched to the particular speaker. Such a process, which generally is called *speaker adaptation*, can be performed by retraining ASR system in the so-called *batch adaptation* way after a sufficient amount of speaker data is available, or the speaker can be adapted in an *online* manner, when one updates parameters each time the new data arrives [12]. The most challenging is that ASR system involves many thousands of parameters, which, in general, require very large amount of speaker data to be re-estimated reliably. A great success has been achieved by applying the so-called *Bayesian adaptation* methods and a brief overview of them is presented here. First, a general issues are briefly mentioned concerning basics of the ASR giving emphasis on the phonetic level modelling. Then a hierarchical Bayesian estimation paradigm is stated. Finally, a short assessment of some of the selected results available in the ASR literature is given to show what improvement one might expect when applying various speaker adaptation techniques.

2 HMM-based speech recognition model

Here I present the basic levels of a continuous automatic speech recognition system so that a reader later will understand 'what and where' is actually adapted.

2.1 Speech recognition system

A general model of an ASR system is shown in Fig. 1. It comprises several different levels of hierarchy:

Acoustic level modelling processes raw speech data and produces the likelihood that a given speech segment belongs to a particular phoneme class.

Lexical modelling receives a sequence of phonemes, each given the importance weight by its corresponding likelihood value. Given different pronunciation models and lexical rules of phoneme composition, lexical modelling comprises a search engine that produces the likelihood of a word given a phoneme sequence.

Language modelling. Given as an input a set of hypothesized word sequences, it allows to select the best word sequence (sentence) by using the so-called bigram or trigram word sequence likelihoods $P(w_t|w_{t-1}, w_{t-2})$, which are usually estimated from a large text corpus.

The adapation of the ASR system to particular speaker is usually performed on the phoneme modelling level. Thus from now on we will focus on this part of the ASR process. As for the speaker adaptation in the higher levels of the ASR system hierarchy, the interested reader is referred to [2].

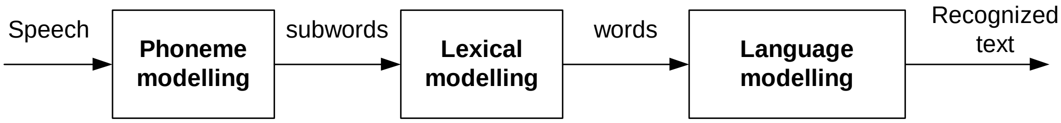

Figure 1: A generic speech recognition model.

More detailed overview of the acoustic modelling is shown in Fig. 2. Predominant number of the phonetic level modelling systems is implemented by using a feature extractor followed by a statistical pattern recognizer, which is normally a hidden Markov model (HMM) [25]. In this report I discuss the Bayesian speaker adaptation in the sense how the parameters of such a phoneme level system are adapted to a speaker, and what improvement one might expect to achieve. Below a short overview of an HMM is presented.

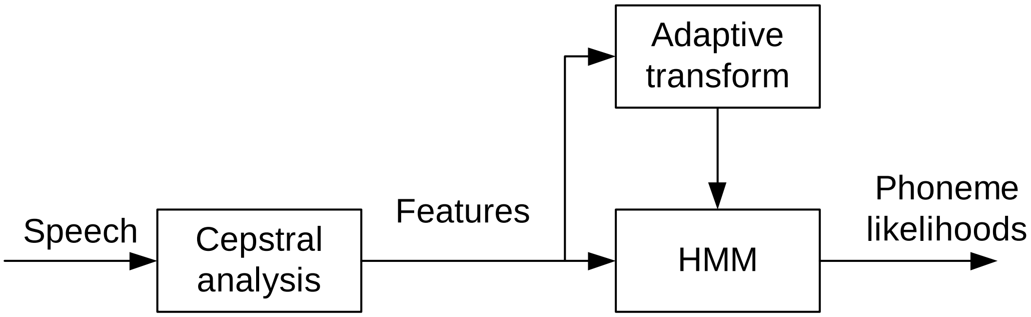

Figure 2: Acoustic level model.

2.2 Hidden Markov Model

2.2.1 Maximum likelihood sequence recognition

Hidden Markov model [25] is a stochastic approach to generate sequences. It starts by first producing a state sequence according to some random process specified by the so-called *state transition matrix* $A = a_{ij}$ which defines the probability that state $s_t = j$ will follow state $s_{t-1} = i$. Then, the sequence of the observed vectors $\mathbf{O} = \{\mathbf{o}_t, t = 1 \ldots T\}$ can be generated according to the probability distribution $b_{s_t}(o_t)$ attached to each state s_t. An HMM will shortly be denoted as $\theta = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$, where $\mathbf{B}$ denotes all the parameters of the state observation probability densities. In our case they are Gaussian mean and covariance matrices of the gaussian state observation probability distributions

$$b_i(\mathbf{o}_t) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_i|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_i)], \tag{1}$$

An HMM classifier provides the unknown observation sequence $\mathbf{O}_m$ with the likelihood $P(\mathbf{O}_m|\theta_s)$ that a particular model θ_s produced a given phoneme $\mathbf{O}_m$. Then one could make a decision about which class the unknown phoneme belongs to by choosing the maximum likelihood model $\theta_{l_m} = \arg\max_{\theta_s} P(\mathbf{O}_m|\theta_s)$.

The HMM likelihood can be written in a more detailed way as:

$$\begin{aligned} p(\mathbf{O}|\theta) &= \sum_{s_1,s_2,\cdots,s_T} p(\mathbf{O}|\mathbf{S},\theta)p(\mathbf{S}|\theta) \qquad (2) \\ &= \sum_{s_1,s_2,\cdots,s_T} \pi_{s_1} b_{s_1}(\mathbf{o}_1)\cdots b_{s_T}(\mathbf{o}_T) a_{s_1 s_2}(\mathbf{o}_1)\cdots a_{s_{T-1}s_T}(\mathbf{o}_T) \end{aligned}$$

The above-stated equation cannot be evaluated directly, but this problem is solved by means of dynamic programming [25] resulting in either the forward-backward recursions to evaluate the exact value of the likelihood or the so called Viterbi matching, when instead of performing a complete summation one approximates the likelihood by a single term given by the state sequence that provides the maximum joint probability $P(\mathbf{O},\mathbf{S}|\theta)$.

2.2.2 Maximum likelihood HMM parameter estimation

The HMMs are estimated by minimizing the negative log-likelihood of the training patterns generated by the models of their corresponding labels:

$$\begin{aligned} E_{ML} &= -\ln P(\mathbf{O}_1,\ldots,\mathbf{O}_M|\theta_1,\ldots,\theta_K) \\ &= -\sum_{m=1}^{M} \ln P(\mathbf{O}_m|\theta_{l_m}). \qquad (3) \end{aligned}$$

By a virtue of the independence assumption used in the second line, the maximum likelihood is greatly simplified: each model can now be trained separately by using only those training sequences that match the model. This can be performed by using the so called Baum-Welch algorithm which maximizes the likelihood of the observed sequence by iteratively performing two consecutive steps. First, it averages the logarithm of the joint probability of the observation sequence and the state sequence $P(\mathbf{O},\mathbf{S}|\theta)$ over all states. The auxiliary function $Q(\theta',\theta)$ of the unknown model parameters θ is obtained (E-step). Then the maximization over θ is performed and θ' is updated by the optimal solution $\hat{\theta}$.

- **E-step**: average logarithm of joint probability function:

  $$Q(\theta',\theta) = \sum_{\mathbf{S}} P(\mathbf{O},\mathbf{S}|\theta') \ln P(\mathbf{O},\mathbf{S}|\theta), \qquad (4)$$

  where θ' are the current model parameters.

- **M-step**: maximize $Q(\theta',\theta)$ over θ to obtain $\hat{\theta}$ and set $\theta' = \hat{\theta}$.

Such an algorithm can be intuitively understood by considering its simplified form, which is called segmental K-means clustering [14]. According to such an approach, one replaces the expectation step by determining only the best state sequence $\hat{\mathbf{S}}$ and considering only its corresponding term $\ln P(\mathbf{O}, \hat{\mathbf{S}}|\theta)$ instead of the summation in Eq. (4). The maximization step is then performed by assigning each observation to its corresponding state and then re-estimating the parameters of the state observation densities based on the newly attached to each state observation data. For example, the mean vector μ_i of each state s_i can be updated by using an averaging

$$\mu_i = \frac{1}{H} \sum_h \mathbf{o}_h, \tag{5}$$

where the index h runs over those observations that are attached to the state i after selecting the best state sequence $\hat{\mathbf{S}}$.

3 Bayesian adaptation methodology

Bayesian speaker adaptation is based on the estimation of the model parameters by using Bayes theorem:

$$p(\theta_{l_m}|\mathbf{O}_m, \beta_{l_m}, \alpha_{l_m}) = \frac{p(\mathbf{O}_m|\theta_{l_m}, \beta_{l_m})p(\theta_{l_m}|\alpha_{l_m})}{p(\mathbf{O})}, \tag{6}$$

If we compare (6) with (3), we can see, that such an equation utilizes the prior probability distribution $p(\theta_{l_m}|\alpha_{l_m})$ of the unknown phoneme model θ_{l_m} best representing a phoneme speech data $\mathbf{O}_m$. Thus, Bayesian approach can be viewed as a balance law between the term specified by our data, and the term that represents our a priori knowledge about the parameter values:

$$E_{BS} = \sum_{m=1}^{M} \ln P(\mathbf{O}_m|\theta_{l_m}, \beta_{l_m}) + p(\theta_{l_m}|\alpha_{l_m}). \tag{7}$$

In addition, the Bayesian approach also allows to consider the set of hyperparameters α_{l_m} and β_{l_m} that control the prior and likelihood distributions. The prior distribution $p(\theta|\alpha)$ is often chosen to be gaussian function

$$p(\theta|\alpha) = \left(\frac{\alpha}{2\pi}\right) \exp\left(-\frac{\alpha}{2}||\theta||^2\right), \tag{8}$$

so that α represents the inverse variance of the prior parameter values, and β might be the inverse variance of the noise in the likelihood function, such as the one shown in Eq. (1).

In summary, a hierarchical Bayesian parameter adaptation is to

- estimate the model parameters θ:

$$P(\theta|\mathbf{O}, \alpha, \beta, \mathcal{H}_i) = \frac{P(\mathbf{O}|\theta, \beta, \mathcal{H}_i)P(\theta|\alpha, \mathcal{H}_i)}{P(\mathbf{O}|\alpha, \beta, \mathcal{H}_i)} \tag{9}$$

- set hyperparameters

$$P(\alpha, \beta|\mathbf{O}, \mathcal{H}_i) = \frac{P(\mathbf{O}|\alpha, \beta, \mathcal{H}_i)P(\alpha, \beta|\mathcal{H}_i)}{P(\mathbf{O}|\mathcal{H}_i)} \tag{10}$$

- compare models

$$P(\mathcal{H}_i|\mathbf{O}) \propto P(O|\mathcal{H}_i)P(\mathcal{H}_i). \tag{11}$$

Hyperparameter estimation is usually performed by implementing the steepest descent on the negative log-likelihood of the Bayesian evidence $P(\mathbf{O}|\alpha, \beta, \mathcal{H}_i)$ [21]. Such a criterion usually is obtained by integrating $P(\mathbf{O}, \theta|\alpha, \beta, \mathcal{H}_i)$ w.r.t. model parameters θ, resulting in what is known as the evidence approach [20]. This method is often applied in many of the modern hierarchical Bayesian speaker adaptation techniques, see for example [27, 16, 35, 24]. However, other criteria has also been used to determine the hyperparameters of the speaker adaptive ASR system [28].

4 Brief review of selected Bayesian techniques

In this section, a brief overview of some of the ideas pertaining to Bayesian approaches to speaker adaptation is presented. The techniques to be discussed below have been successfully applied in speaker adaptation field, dating back, perhaps, already to 1977 [19].

4.1 Maximum a Posteriori adaptation

Below a simple example is presented to illustrate the basic ideas of the online Bayesian adaptation rules used to re-estimate the state observation gaussian density mean vector. It is notable even in this case the hyperparameter estimation requires nonlinear optimization procedures, and therefore we only state the first level of Bayesian inference Eq. (9). For the hyperparameter optimization one can be referred to [3].

Assume the one dimensional additive noise model

$$o_i = \theta + n_i, \tag{12}$$

where parameter θ is to be estimated from the data set $\mathbf{O} = \{o_i, i = 1, 2, \ldots, M\}$. Each random variable n_i is supposed to follow the gaussian distribution $N_{\mu,\sigma}(n_i)$

$$N_{\mu,\sigma}(n_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{n_i - \mu}{\sigma}\right)^2\right\} \tag{13}$$

whose mean $\mu = 0$, and variance $\sigma = \sigma_d$. Bayesian parameter estimation can be performed by using the following steps:

1. Prior specification

$$p(\theta|m_\theta, \sigma_\theta^2) = N_{m_\theta,\sigma_\theta}(\theta). \tag{14}$$

2. Likelihood

$$p(\mathbf{O}|\theta, \sigma_o^2) = \prod_{i=1}^{M} N_{\theta,\sigma_o}(o_i) \tag{15}$$

3. Posterior

$$\begin{aligned} p(\theta|\mathbf{O}) &= N_{m_o,\sigma_o}(\theta), & (16)\\ \mu_p &= \frac{\sigma_o^2/M}{\sigma_\theta^2 + \sigma_o^2/M} m_\theta + \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_o^2/M} \frac{1}{M} \sum_{i=1}^{N} o_i, & (17)\\ \sigma_p^2 &= \frac{\sigma_o^2 \sigma_\theta^2}{\sigma_o^2 + M\sigma_\theta^2}. & (18) \end{aligned}$$

Eq. (17) is one of the most fundamental mean estimation rules in the Bayesian speaker adaptation [23]. It shows that when the number of data points M is small, the posterior mean μ_p is approximately equal to the prior mean m_θ, whereas when M is large, the posterior mean is determined by the data. Such an approach indicates how one could overcome the problem of a small data available during the speaker adaptation by specifying the prior parameter values and then correcting it online by using the the newly acquired data. Two simplest adaptation rules explained above, i.e. Eq. (17) and (18) are often called *maximum a posteriori* (MAP) estimation. They have been applied in many works on Bayesian speaker adaptation, see especially [30, 17, 32]. It has been shown that the MAP approach outperforms the maximum likelihood gaussian mean adaptation [10, 11]. The rules Eq. (17) and (18)have been tested on word recognition task by using the HMM-based ASR system on 1000 word DARPA Resource Management Continuous Speech Corpus [23]. The data had twelve speakers, eleven of them were used to create a speaker independent system and one of them was used each time as a new unseen speaker to turn the ASR system into speaker dependent one. The best results were obtained by using only the adaptation of gaussian state observation means, which reduces the word error rate from 12% down to 9.5%, and as the authors state:

'*...with roughly three minutes of adaptation speech, the reduction in word error is approximately 16% which is nearly half the reduction achieved training a speaker-dependent model with 10 times more data.*'

Adaptation of variance according to Eq. (18) did not improve results. The conclusion that a speaker adaptation is superior to speaker-dependent training is also drawn in [9, 18]. These works use NIST Naval Resource Management corpus, and put more emphasis on the initial estimation of prior means m_θ.

4.2 Extended Maximum a Posteriori methods

One difficulty with the MAP approach is that only those models can be adapted which have their corresponding phonemes in a new speaker's data. Such a data sparsity problem is circumvented in the extended MAP (EMAP) approaches. Here one models the correlations between the observation mean vectors of the different phoneme models, rather than applying a set of separate different priors. One way to implement the idea of correlation modelling is to apply tree-structured HMMs, where the phoneme HMMs are considered as leaves of the HMMs that tie the states of different phoneme models [22, 27, 24].

The EMAP methods are often given names of vector field posterior smoothing [31] or the phone cross-correlation prediction methods [30, 1, 26, 36, 13]. A general approach to these ideas utilizes the random Markov Field concepts (MRF) concepts [26, 36], in particular, one considers a matrix, whose columns are phoneme mean vectors, as a spatial structure, where each element depends only on the elements located at its surrounding neighborhood. By postulating a particular dependency structure, one estimates the cross-correlations between different phoneme models from the training data set and then uses the MRF priors during the speaker adaptation. See also [7] for the general overview of the dependency modelling for the mean vectors of the prior state observation gaussian distributions.

4.3 Bayesian Principal Component Analysis

Bayesian principal component analysis (BPCA) speaker adaptation [16] technique aims to adapt the gaussian state observation means by stacking them into a *supervector*, and then considers reduced representations of such a vector so that much less parameters should be re-estimated during the adaptation. More precisely, the idea behind the approach can be explained as follows:

- obtain a set of speaker dependent HMM models, each having the mean vector $\mathbf{m}_{k,i,j}$ representing the mean vector of the jth gaussian distribution in the ith state of the kth speaker;
- construct a set $\mathbf{M}$ of the supervectors $\mathbf{m}_k$: $\mathbf{M} = \{\mathbf{m}_k = [m_{k,0,0}^T, \ldots, m_{k,i,j}^T, \ldots, m_{k,N,J}]^T\}$,
- estimate the model for the supervector set by using the Bayesian PCA [16]:

$$\mathbf{m} = \mathbf{W}\mathbf{x} + \mu_m + \epsilon, \tag{19}$$

  where $\mathbf{x}$ is a latent variable modelled as a gaussian random vector with unit variance and zero mean, gaussian noise $\epsilon \propto N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$ is assumed to be independent of $\mathbf{x}$. The estimated model comprises three sets of parameters $\hat{\theta} = \{W, \mu_m, \sigma_\epsilon^2\}$.
- An ASR system is adapted to a new speaker by using an on-line estimation formulae for the model parameters θ shown explicitly in Eq. (19).

Such an approach is a probabilistic generalization of the clustering-based speaker adaptation technique known as 'eigen-voice' approach and similar ideas can be found in maximum likelihood linear regression (MLLR) speaker transformation method [15]. According to these transformation methods, one first estimates canonical eigen-voice matrix $\mathbf{W}$, whose each column is an eigenvector of the covariance matrix estimated from a training set of speakers' supervectors. Speaker-dependent system is then obtained as a linear combination of the eigen-voices. This can be implemented to produce a very fast adaptation as one has to estimate only the linear weights that show the contribution of each eigen-voice to the adapted speaker's observation mean supervector, rather than estimating the whole supervector itself. The main difference between the BPCA and the 'eigen-voice' approach is that BPCA technique provides an on-line re-estimation formulae for the model θ based on the prior distributions of the model parameters [16]:

$$\theta_{MP} = \arg\max_{\theta} p(\mathbf{O}|\theta)p(\theta|\alpha,\beta), \tag{20}$$

whereas more classical transformation-based methods exploit only the knowledge of the likelihood term $p(\mathbf{O}|\theta)$.

Experiments performed with the Korean digits database, whose 105 speakers were used to obtain 'eigen-voices' and the other 35 speakers were used to evaluate the adaptation algorithms. The dimensionality of supervector was equal to approximately 3840 parameters. The BPCA method was shown to outperform the MLLR method, cl. 90.21% vs. 88.60% average recognition results [16].

4.4 Combined MAP and transformation-based approaches

It is known that the transformation-based techniques are good in rapid adaptation situations when one has at its disposal only small amounts of new data to adapt the ASR system. The MAP techniques usually show a considerable improvement only on the longer adaptation runs. A combination of MAP update rules for the re-estimation of the state observation gaussian distribution means and the transformation-based methods has been shown to improve the Bayesian speaker adaptation [8]. In particular, tests were run on the 'spoke 3' task of the phase-1 large vocabulary Wall Street Journal Corpus, which had about 140 non-native speakers of American English. The results show that after 20 adaptation sentences the transformation approach is superior to the MAP technique, however, after the 40 sentences the MAP method gives better results, i.e. 17% vs. 18% total word error rate is achieved compared to the initial 30% error rate given by the speaker-independent ASR system. The combination of both approaches decreases the word error rate down to 14.9%.

A comparison between the MLLR and MAP linear regression (MAPLR) technique fully utilizing Eq. (20) with the hyperparameter estimation shows that Bayesian approach achieves more robust speech recognition under the small amounts of adaptation data [6]. Similar observations are drwan in other works on joint adaptation of gaussian means and the linear transform $\mathbf{W}$ parameters [8, 33, 5, 34, 4, 29].

5 Conclusions

Bayesian speaker adaptation provides a variety of different ways to make an ASR systems more robust to inter-speaker variations. The simplest online adaptation rules for the mean vectors of an HMM state observation gaussian densities produce better results than their classical ML counterparts in case there is a small adaptation data available. Further improvements are made when modelling the correlations between different phoneme models, and henceforth building the joint all-phoneme priors, rather than independent priors of each phoneme. Rapid speaker adaptation in the presence of sparse data is achieved by introducing lower representations of the observation mean vectors via linear transforms. Such transforms can be better utilized by knowing their parameter uncertainty bars and this is the way how Bayesian PCA technique makes improvements on the more classical transformation-based speaker adaptation methods. It is clear that the determination of the model noise levels β and the regularization constants α by maximizing Bayesian evidence allows one to achieve model complexity control, prune the tree-structured HMMs and select simpler HMM structures. What is much less clear is the actual impact the estimation of the hyperparameters causes on the ASR system performance in a large vocabulary speech recognition.

References

[1] S.M. Ahadi and P.C. Woodland. Rapid speaker adaptation using model prediction. *IEEE ICASSP*, 1:684–687, May 1995.

[2] A. Berger and R. Miller. Just-in-time language modelling. *IEEE ICASSP*, 2:705–708, May 1998.

[3] Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[4] K.T. Chen and H.M. Wang. Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation. *IEEE ICASSP*, 1:317–320, 2001.

[5] J.T. Chien. Online hierarchical transformation of hidden markov models for speech recognition. *IEEE Trans. on Speech and Audio Processing*, 7(6):656–667, November 1999.

[6] J.T. Chien. Quasi-bayes linear regression for sequential learning of hidden markov models. *IEEE Trans. on Speech and Audio Processing*, 10(5):268–278, July 2002.

[7] V. Digalakis, H. Collier, S. Berkowitz, A. Corduneanu, E. Bocchieri, A. Kannan, C. Boulis, S. Khudanpur, W. Byrne, and A. Sankar. Rapid speech recognizer adaptation to new speakers. *IEEE ICASSP*, 2:765–768, March 1999.

[8] V. Digalakis and L. Neumeyer. Speaker adaptation using combined transformation and bayesian methods. *IEEE ICASSP*, 1:680–683, May 1995.

[9] J.L. Gauvain and C.H. Lee. Improved acoustic modeling with bayesian learning. *IEEE ICASSP*, 1:481–484, March 1992.

[10] Y Goto, M.M. Hochberg, D.J. Mashao, and H.F. Silverman. Incremental map estimation of hmms for efficient training and improved performance. *IEEE ICASSP*, 1:457–460, May 1995.

[11] S. Homma, J. Takahashi, and S. Sagayama. Iterative unsupervised speaker adaptation for batch dictation. *IEEE ICSLP*, 2:1141–1144, October 1996.

[12] X. Huang and K.F. Lee. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. *IEEE Trans. on Speech and Audio Processing*, 1(2):150–157, Alpril 1993.

[13] E. Jon, D.K. Kim, and N.S. Kim. Robust correlation estimation for emap-based speaker adaptation. *IEEE Signal Processing Letters*, 8(6):184–186, June 2001.

[14] Biing Hwang Juang and L. R. Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. on ASSP*, 38(9):1639–1641, 1990.

[15] Jean-Claude Junqua and Gertjan van Noord. *Robustness in Language and Speech Processing*, chapter D.T. Merino, Speaker Compensation in ASR. Kluwer Academic Publishers, 2001.

[16] D.K. Kim and N.S. Kim. Rapid speaker adaptation using probabilistic principal component analysis. *IEEE Signal Processing Letters*, 8(6):180–183, 2001.

[17] C. H. Lee, C. H. Lin, and B. H. Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Trans. on Signal Processing*, 39(4):806–814, April 1991.

[18] C.H. Lee and J.L. Gauvain. Speaker adaptation based on MAP estimation of hmm parameters. *IEEE ICASSP*, 2:558–561, 1993.

[19] B. Lowerre. Dynamic speaker adaptation in the harpy speech recognition system. *IEEE ICASSP*, 2:788–790, May 1977.

[20] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[21] D. J. C. MacKay. Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge, 1997.

[22] Yasunaga Miyazawa. An all-phoneme ergodic hmm for unsupervised speaker adaptation. *IEEE ICASSP*, 2:574–577, 1993.

[23] B.F. Necioglu, M. Ostendorf, and J.R. Rohlicek. A bayesian approach to speaker adaptation for the stochastic segment model. *IEEE ICASSP*, 1:437–440, March 1992.

[24] o. Siohan and A.C. Surendran. Structural bayesian predictive adaptation of hidden markov models. *Workshop on Adaptation Methods for Speech Recognition, Sophia-Antinopolis*, 2001.

[25] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[26] B.M. Shahshahani. A markov random field approach to bayesian speaker adaptation. *IEEE Trans. on Speech and Audio Processing*, 5(2):183–191, March 1997.

[27] K. Shinoda and C.H. Lee. Structural map speaker adaptation using hierarchical priors. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 381–388, December 1997.

[28] K. Shinoda and T. Watanabe. Speaker adaptation with autonomous model complexity control by mdl principle. *IEEE ICASSP*, 2:717–720, May 1996.

[29] O. Siohan, C. Chesta, and C.H. Lee. Joint maximum a posteriori adaptation of transformation and hmm parameters. *IEEE Trans. on Speech and Audio Processing*, 9(4):417–428, May 2001.

[30] R. Stern and M. Lasry. Dynamic speaker adaptation for isolated letter recognition using map estimation. *IEEE ICASSP*, 8:734–737, April 1983.

[31] J. Takahashi and S. Sagayama. Vector-field-smoothed bayesian learning for incremental speaker adaptation. *IEEE ICASSP*, 1:696–699, May 1995.

[32] E. Thelen. Long term on-line speaker adaptation for large vocabulary dictation. *IEEE ICSPL*, 4:2139–2142, October 1996.

[33] v. Digalakis, D. Rtishev, and L.G. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Trans. on Speech and Audio Processing*, 3(5):357–366, September 1995.

[34] N.J.-C. Wang, S.S.-M Lee, F. Seide, and L.-S. Lee. Rapid speaker adaptation using a priori knowledge by eigenspace analysis of mllr parameters. *IEEE ICASSP*, 1:354–348, 2001.

[35] S. Wang and Y. Zhao. Online bayesian tree-structured transformation of hmms with optimal model selection for speaker adaptation. *IEEE Trans. on Speech and Audio Processing*, 9(6):663–677, Spetember 2001.

[36] X. Wu and Y. Yan. Linear regression under maximum a posteriori criterion with markov random field prior. *IEEE ICASSP*, 2:997–1000, 2000.