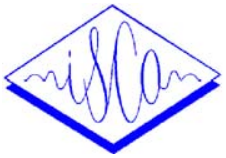

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MAXIMUM A POSTERIORI LINEAR REGRESSION FOR HIDDEN MARKOV MODEL ADAPTATION

Cristina Chesta[†] *Olivier Siohan*[‡] *Chin-Hui Lee*[‡]

[‡]Bell Laboratories – Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
[†]Dipartimento di Automatica e Informatica - Politecnico di Torino, Italy

ABSTRACT

In the past few years, transformation-based model adaptation techniques have been widely used to help reducing acoustic mismatch between training and testing conditions of automatic speech recognizers. The estimation of the transformation parameters is usually carried out using estimation paradigms based on classical statistics such as maximum likelihood, mainly because of their conceptual and computational simplicity. However, it appears necessary to introduce some constraints on the possible values of the transformation parameters to avoid getting unreasonable estimates that might perturb the underlying structure of the acoustic space. In this paper, we propose to introduce such constraints using Bayesian statistics, where a prior distribution of the transformation parameters is used. A Bayesian counterpart of the well known maximum likelihood linear regression (MLLR) adaption is formulated based on maximum *a posteriori* (MAP) estimation. Supervised, unsupervised and incremental non-native speaker adaptation experiments are carried out to compare the proposed MAPLR approach to MLLR. Experimental results show that MAPLR outperforms MLLR.

1. INTRODUCTION

Acoustic mismatches between training and testing conditions of an automatic speech recognizer can significantly upset the recognition accuracy, and in the past few years, acoustic model adaptation techniques have been shown as an efficient way to reduce these discrepancies, whether they may arise from changes in speaker characteristics, speaking environment or incorrect modeling assumptions [1].

Let Λ be a set of speaker-independent (SI) hidden Markov models (HMM). A transformation-based model adaptation consists of applying some transformations $F_\eta(\cdot)$ to various clusters of HMM parameters. Given some adaptation data, Y, the objective of the adaptation is first to derive the parameters η of the transformations, and then use the transformed models $F_\eta(\Lambda)$ to recognize the incoming speech. The estimation of η is traditionally carried out using classical statistics that assume that η are some fixed but unknown parameters. Because of its simplicity, the maximum likelihood (ML) criterion is usually chosen, which states that $\hat{\eta}_{ML}$ should maximise the likelihood of the adaptation data given the transformed model, $p(Y|\Lambda,\eta)$:

$$\hat{\eta}_{ML} = \underset{\eta}{\operatorname{argmax}}\, p(Y|\Lambda,\eta). \tag{1}$$

However, maximum likelihood estimation does not introduce any constraints on the possible values of η and relies only on the adaptation data Y and the original acoustic model Λ. It is therefore desirable to constraint the possible values of η to avoid getting a transformed model $F_\eta(\Lambda)$ that might corrupt the underlying structure of the acoustic space. Moreover, a poorly structured transformation can lead to a quick saturation of the performance improvement provided by the adaptation.

A possible solution to this problem is to introduce some constraints on the values of the transformation parameters. This can be achieved by introducing a constraint of the form $g(\eta) = 0$ during estimation, where $g(\cdot)$ is the constraint. It is however quite difficult to specify and justify the choice of $g(\cdot)$. Rather than relying on such an ad-hoc formulation, a mathematically attractive alternative can be formulated by carrying out the parameter estimation under a Bayesian statistical framework. Under a Bayesian formalism, η is no longer fixed but random, and is therefore described by its probability density function (pdf), $p(\eta)$. The density $p(\eta)$ represents the prior knowledge that we might have on the values that the transformation parameters can take. This constraint can be formally inserted in the estimation process by using a maximum *a posteriori* (MAP) criterion:

$$\begin{aligned}\hat{\eta}_{MAP} &= \underset{\eta}{\operatorname{argmax}}\, p(\eta|Y,\Lambda) \\ &\propto \underset{\eta}{\operatorname{argmax}}\, p(Y|\eta,\Lambda)\,p(\eta).\end{aligned} \tag{2}$$

The MAP criterion simply states that some values of η, described by the prior density $p(\eta)$, are more likely than others and this knowledge is used to constraint the estimation process. If the prior distribution reflects transformation parameters preserving the structure of the acoustic space, then the MAP-estimated transformation parameters will also incorporate that knowledge. Obviously, the performance of the algorithm is highly related to a proper choice of the prior distribution. However, this prior distribution can also be trained and the whole process is entirely data-driven.

In this paper, we focus on a simple transformation family which consists of applying an affine transformation to each cluster of HMM mean vectors, $\hat{m} = Am + b$, where m is an original mean vector, A is a square transformation matrix, b is a translation vector and $\hat{m}$ is the transformed mean vector. When the estimation criterion is maximum likelihood, this adaptation scenario corresponds to the well-known maximum likelihood linear regression, or MLLR technique [2]. Rather than carrying out the estimation using maximum likelihood, we derive an estimate of $\eta = (A,b)$ using MAP. The proposed MAPLR algorithm is described in section 2. Experimental results are given in section 3 where the performance of MLLR and MAPLR are compared on a non-native speaker adaptation task. Section 4 concludes the paper.

2. MAXIMUM A POSTERIORI LINEAR REGRESSION

2.1. Notations

In the following sections, we derive an estimate of η using MAP when Λ is a continuous density hidden Markov model (HMM). In a given state n, the pdf of an observation vector y is modeled by a mixture of M Gaussian distributions:

$$p(y|S=n) = \sum_{m=1}^{M} \omega_{n,m} N(y;\mu_{n,m},R_{n,m}), \tag{3}$$

where $N(y;\mu_{n,m},R_{n,m})$ is a Normal distribution of mean $\mu_{n,m}$ and precision matrix $R_{n,m}$ defined as:

$$N(y;\mu_{n,m},R_{n,m}) \propto |R_{n,m}|^{1/2} \cdot \exp\left\{-\frac{1}{2}\operatorname{tr}(y-\mu_{n,m})(y-\mu_{n,m})'R_{n,m}\right\}. \quad (4)$$

A mean vector $\mu_{n,m} \in \mathbb{R}^p$ is adapted using an affine transformation $\eta = \{A,b\}$, where $A \in \mathbb{R}^{p\times p}$ is the transformation matrix and $b \in \mathbb{R}^p$ is a bias vector:

$$\hat{\mu}_{n,m} = A\mu_{n,m} + b. \quad (5)$$

For notation simplicity and as commonly done in the MLLR formulation [2], the above transformation can be represented using a single $p\times(p+1)$ transformation matrix $W=(A,b)$ applied on an extended mean vector $\tilde{\mu}_{n,m}$ defined as $\tilde{\mu}'_{n,m} = (\mu'_{n,m},1)$. Equation (5) becomes:

$$\hat{\mu}_{n,m} = W\tilde{\mu}_{n,m}. \quad (6)$$

Clusters of mean vectors are also defined (based on phonetic or acoustic similarity) so that all mean vectors from the same cluster c share the same transformation W_c.

2.2. Selection of the prior density $p(W)$

As usual in MAP estimation, the choice of the prior density $p(W)$ can be based on some physical considerations regarding the parameter W or on some mathematical attractiveness like the existence of conjugate prior densities which can greatly simplify the maximization of (2). However, unlike MAP estimation of HMM parameters [3], no obvious conjugate prior densities could be found in our case. In our initial formulation [4], $p(W)$ was chosen as the product of a Normal-Wishart density with a Normal distribution. While the bias part of the transformation could then be derived under closed-form, no closed-form solution could be obtained for the square transformation matrix. In [5], Chou suggests to select $p(W)$ from a family of elliptically symmetric distributions. In [4], we chose a special case of elliptical distribution, namely a matrix variate normal prior density [6], which can be seen as a matrix version of a multivariate normal distribution:

$$p(W) \propto |\Sigma|^{-(p+1)/2}|\Phi|^{-p/2} \cdot \exp\left\{-\frac{1}{2}\operatorname{tr}(W-M)'\Sigma^{-1}(W-M)\Phi^{-1}\right\}, \quad (7)$$

where $W,M \in \mathbb{R}^{p\times(p+1)}$, $\Sigma \in \mathbb{R}^{p\times p}$, $\Sigma \geq 0$, $\Phi \in \mathbb{R}^{(p+1)\times(p+1)}$ and $\Phi \geq 0$. An additional advantage of this prior density is the existence of maximum likelihood estimates of its hyperparameters M and Σ when Φ is known (cf. section 2.4). We want to point out that $p(W)$ does not even have to be a probability density function since the MAP estimation is still valid with improper priors. The only constraints is that $p(W)$ should be a non-negative function.

2.3. Maximization of the posterior density $p(W|Y,\Lambda)$

Let $Y=\{y_t\}$ be an adaptation utterance used to derive W. Because of the missing data problem in Gaussian mixture HMMs, the maximization of (2) cannot be carried out directly. However, the EM algorithm [7] provides an efficient way to address this problem by defining an auxiliary function $Q(\cdot)$ having the same optimal solution as $p(W|Y,\Lambda)$ but whose maximization is usually simpler. The maximization is performed under an iterative fashion where $Q(\eta|\bar{\eta})$ is maximized w.r.t η given the previous estimate $\bar{\eta}$, until convergence. The auxiliary function is defined as:

$$\begin{aligned} Q(\eta_c|\bar{\eta}_c) &= E\{\log p(Y,S,L|\Lambda,\eta_c) + \log p(\eta_c)|Y,\Lambda,\bar{\eta}_c\} \\ &= \sum_S\sum_L p(S,L|Y,\Lambda,\bar{\eta}_c)\log p(Y,S,L|\Lambda,\eta_c) \\ &\quad + \log p(\eta_c), \end{aligned} \quad (8)$$

where $S=\{s_t\}$ represents the state sequence, $L=\{l_t\}$ is the mixture sequence, and can be rewritten as:

$$\begin{aligned} Q(\eta_c|\bar{\eta}_c) &= \sum_S\sum_L p(S,L|Y,\Lambda,\bar{\eta}_c)\sum_{t=1}^{T}\big[\log a_{s_{t-1},s_t} \\ &\quad + \log\omega_{s_t,l_t} + \log p(y_t|\eta_c,\mu_{s_t,l_t},R_{s_t,l_t})\big] \\ &\quad + \log p(\eta_c) \\ &= \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m)\log p(y_t|\eta_c,\mu_{n,m},R_{n,m}) \\ &\quad + \log p(\eta_c) + \Psi, \end{aligned} \quad (9)$$

where $\gamma_t(n,m) = P(s_t=n,l_t=m|Y,\Lambda,\bar{\eta}_c)$ is the probability of being in state n and mixture m at time t, given the sequence Y, the model Λ and the current transformation $\bar{\eta}_c$. Ψ represents all terms independent of η_c, including the transition probability $a_{i,j}$ and the mixture weights $\omega_{n,m}$.

After differentiating (9) w.r.t each element of W and equating to zero, the following system of $p\times(p+1)$ linear equation is obtained [4], where w_{ij} is the (i,j)-th component of the matrix W, r_{ij}, m_{ij}, σ_{ij} and ϕ_{ij} are the (i,j)-th components of the matrices $R_{n,m}$, M, Σ and Φ, and where $\tilde{\mu}_i$ is the i-th component of $\mu_{n,m}$[1]:

$$\sum_{k=1}^{p}\sum_{l=1}^{p+1} w_{kl}\left[\sum_{n=1}^{N}\sum_{m=1}^{M}\left(\sum_{t=1}^{T}\gamma_t(n,m)\right)r_{ik}\tilde{\mu}_l\tilde{\mu}_j + \frac{1}{2}\sigma_{ki}\phi_{jl} + \frac{1}{2}\sigma_{ik}\phi_{lj}\right] = z_{ij}, \quad \begin{matrix}1\leq i\leq p\\ 1\leq j\leq p+1\end{matrix} \quad (10)$$

where z_{ij} is defined as:

$$z_{ij} = \sum_{k=1}^{p}\sum_{l=1}^{p+1}\left[\sum_{n=1}^{N}\sum_{m=1}^{M}\left(\sum_{t=1}^{T}\gamma_t(n,m)y_k(t)\right)r_{ik}\tilde{\mu}_j + \frac{1}{2}\sigma_{ki}m_{kl}\phi_{jl} + \frac{1}{2}\sigma_{ik}m_{kl}\phi_{lj}\right]. \quad (11)$$

The matrix W can be obtained by solving the system of $p\times(p+1)$ linear equations described by (10) and (11). It is worth noting that this system of equations is very similar to the standard MLLR solution except for the additional terms related to the prior density. This system can be further simplified by assuming that $R_{n,m}$ and Σ are diagonal [4], in which case W can be obtained by solving p systems of $p+1$ linear equations which gives the same complexity as the standard MLLR formulation. In the experiments performed in this paper, such a simplification was not used.

2.4. Estimation of the hyperparameters of $p(W)$

One last difficulty related to MAP estimation is the choice or estimation of the hyperparameters $\{M,\Sigma,\Phi\}$ of the prior distribution $p(W;M,\Sigma,\Phi)$. The estimation is based on an empirical Bayesian approach where a set of transformation matrices $\{W_1,\ldots W_K\}$ is

[1] The subscripts n and m indicating the state and mixture index have been discarded for notation simplicity. Also note that the summations should be performed only over the mean vectors belonging to the cluster c.

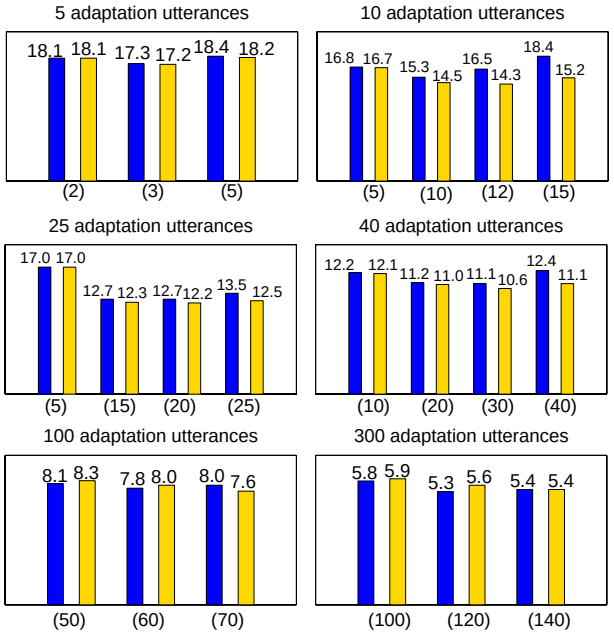

Figure 1: Word error rate (%) for batch supervised experiments for MLLR (black) and MAPLR (white) for various amount of adaptation data and transformation matrices (the number of transformations is in parenthesis). Microphone database (baseline system, Word Error rate = 18.5%).

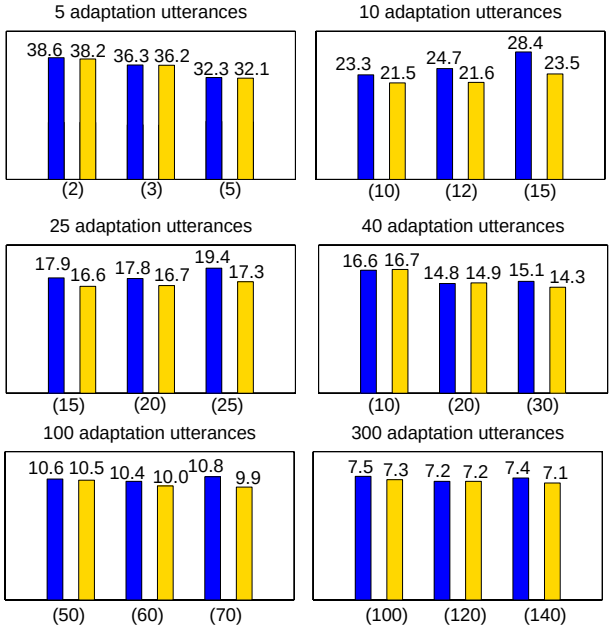

Figure 2: Word error rate (%) for batch supervised experiments for MLLR (black) and MAPLR (white) for various amount of adaptation data and transformation matrices (the number of transformations is in parenthesis). Telephone database (baseline system, Word Error rate = 34.9%).

first derived, and the hyperparameters are then obtained using maximum likelihood estimation. One advantage of the matrix variate prior density $p(W)$ is that the MLE estimation of M and Σ can be easily derived when Φ is given [6] and in this paper we assume that Φ is the identity matrix. In [4], we have proposed a technique to estimate the set of matrices $\{W_1, \ldots W_K\}$ directly from the original speaker independent (SI) model and therefore no extra training data is needed to derive the prior density. The basic idea is to use some of the SI model mean vectors as data and derive each W_k using a MLLR-like approach. In our experiments, the clusters of HMM mean vectors are derived from the decision tree used to train the context dependent SI models. The set of matrices $\{W_1, \ldots W_K\}$ used to estimate the hyperparameters are also organized in this tree, which means that different clusters of mean vectors can have different hyperparameters. Details can be found in [4].

3. EXPERIMENTS AND RESULTS

Experiments are performed on the resource management (RM) task where new recordings have been collected for 5 non-native speakers simultaneously through 2 channels: a close talking microphone and a telephone line. For each speaker, the data consists of 300 utterances used for adaptation and 75 test utterances. The recognizer is built on the official RM training set down-sampled at 8kHz and therefore our experiments reflect compensation of channel and speaker mismatches. Context-dependent acoustic models are built using the decision tree state tying algorithm described in [8]. The language model is the official RM word-pair grammar. We use a standard acoustic front-end. The signal is pre-emphasized using a first order difference and 10th order linear predictive coding (LPC) coefficients are derived every 10ms over 30ms Hamming windowed segments. The 10 LPC coefficients are converted to 12th order cepstral coefficients (LPCC) and a feature vector of 39 components, consisting of 12 LPCC plus the energy term and their first and second derivatives is produced at each frame. The objective of our experiments is to compare a standard MLLR formulation with our proposed MAPLR on various speaker adaptation tasks: batch supervised, batch unsupervised and incremental unsupervised.

3.1. Supervised experiments – Batch mode

In batch adaptation mode, the adaptation data is used to derive the transformation matrices. The acoustic models are then transformed and used to recognize the test data. The adaptation is supervised which means that the transcription of the adaptation utterances is given when estimating the transformation matrices. For each speaker and for each channel condition, various amount of adaptation data (from 5 utterances up to 300 utterances) is used to derive the transformation matrices. Full transformation matrices are used for both MLLR and MAPLR.

For a given amount of adaptation data, we first tune up the number of transformation matrices to get the best MLLR performance. MAPLR experiments are then performed for the same number of transformation matrices and this number is slightly modified to study the sensitivity of MLLR and MAPLR to this factor. Only 3 prior densities are derived from the SI independent models for the MAPLR adaptation. Results are given in Figure 1 and 2 in terms of word error rate averaged over the 5 speakers for the microphone and telephone database. For a given number of adaptation utterances, several configurations are evaluated depending on the number of transformation matrices (indicated in parenthesis under each plot). Both MLLR and MAPLR provide a significant improvement over the original SI models which have a word error rate of 18.5% and 34.9% for the microphone and telephone database.

For all configurations, MAPLR performs as well or better than MLLR. This is especially true for small amount of adaptation utterances (10 to 40) and appears more strongly on the telephone database. As usual in MAP learning [3], as the amount of adaptation data increases MAPLR converges asymptotically to MLLR. We also observe that MAPLR seems less sensitive than MLLR

	# Utter.	Cor	Sub	Del	Ins	W. Err.
No Adapt.	0	87.2	11.5	1.3	5.7	18.5
MLLR	5	86.7	11.7	1.6	4.8	18.1
MAPLR	5	86.7	11.8	1.6	4.7	18.0
MLLR	10	87.6	10.6	1.8	4.4	16.8
MAPLR	10	88.3	10.0	1.8	4.3	16.1
MLLR	25	89.7	9.0	1.4	4.1	14.5
MAPLR	25	89.9	8.8	1.3	3.9	14.0
MLLR	40	90.5	8.4	1.1	3.8	13.3
MAPLR	40	90.9	8.1	1.0	3.5	12.6
MLLR	100	92.5	6.7	0.8	3.1	10.6
MAPLR	100	92.8	6.4	0.8	3.1	10.2
MLLR	300	94.6	5.0	0.5	2.6	8.0
MAPLR	300	94.7	4.9	0.5	2.6	7.9

Table 1: Batch unsupervised experiments for MLLR and MAPLR for various amount of adaptation data and transformation matrices. Microphone database.

to the number of transformations, as shown for 10 adaptation utterances for the microphone and telephone database. Considering that the priors are derived from the SI models only and are therefore quite poor since they do not reflect very accurately non-native speaker variations nor telephone channel variation, we believe that these results illustrate the efficiency of the regularization provided by MAPLR over MLLR.

3.2. Unsupervised experiments – Batch mode

In batch unsupervised adaptation experiments, the transcription of the adaptation data is not known. It is therefore necessary to recognize the adaptation data using the original SI models and then use the obtained transcription to carry out the adaptation as in supervised experiments. These experiments are performed only on the microphone database and for the number of transformations that gave the best results in supervised experiments. Results are given in Table 1 in terms of word correct, substitution, deletion, insertion and word error rates. While there is a significant loss in accuracy compared to the supervised adaptation scenario, MAPLR still outperforms MLLR in all configurations.

3.3. Unsupervised experiments – Incremental mode

Unsupervised incremental adaptation, sometimes called compensation, means that the test data is used as adaptation data. As more and more test utterances are processed and recognized, the model adaptation is periodically carried out, every n utterances. As an example, if $n = 5$, utterances 1 to 5 are recognized using the original SI models. Once the fifth utterance has been recognized, transformation matrices are derived and the models are adapted. Utterances 6 to 10 are then recognized using the adapted models, and so on. MLLR as well as MAPLR adaptation can be performed in such an incremental way since, as indicated by (10) and (11), the time dependent quantities $\gamma_t(n,m)y_k(t)$ and $\gamma_t(n,m)$ needed to derive the transformation matrices can be accumulated utterance after utterance. Moreover, rather than setting the number of transformations to a fixed value, the number of transformations is dynamically derived according to the amount of data already recognized. As more and more test data is processed, the corresponding number of transformation matrices keeps increasing. In our experiments, the initial number of transformation matrices is set to 3, and the transformations are re-estimated every 5 utterances. Experimental results are given in Table 2. While only slight improvements have been obtained for batch adaptation, a significant reduction of the word error rate is now obtained when using MAPLR.

Adapt. method	Cor	Sub	Del	Ins	W. Err.
No Adapt.	87.2	11.5	1.3	5.7	18.5
MLLR	89.8	9.2	1.0	4.9	15.0
MAPLR	91.8	7.3	0.9	3.8	12.0

Table 2: Incremental unsupervised experiments for MLLR and MAPLR for various amount of adaptation data. Microphone database.Transformation matrices are updated every 5 utterances.

4. CONCLUSION

A Bayesian counterpart of the standard MLLR adaptation algorithm has been formulated based on MAP estimation. If the prior density is chosen appropriately, the MAPLR estimation can be carried out by solving a system of $p \times (p+1)$ linear equations (while MLLR solves independently p systems of $p+1$ equations). Under additional simplifying assumptions, the complexity of the MAPLR formulation can be reduced to be the same as MLLR but no experiments have been performed yet to evaluate the validity of that simplification. The prior densities are derived directly from the SI models and no additional data is needed. Experimental results performed on non-native speaker and channel adaptation tasks have shown that MAPLR slightly outperforms MLLR on supervised and unsupervised batch adaptation, while the improvement is larger on incremental adaptation. These results are particularly encouraging considering that the derivation of the prior density was not tuned in any of our experiments and that the prior density was somehow quite poor since the telephone channel variability was not represented in the SI models. Future works involve experiments on large vocabulary speech recognition and joint MAP estimation of HMM parameters and transformation parameters.

5. REFERENCES

[1] C.-H. Lee. On stochastic feature and model compensation approaches to robust speech recognition. *Speech Communication*, 25:29–47, 1998.

[2] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185, 1995.

[3] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori extimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, April 1994.

[4] O. Siohan, C. Chesta, and C.-H. Lee. Hidden Markov model adaptation using maximum a posteriori linear regression. In *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, 1999.

[5] W. Chou. Quasi-bayesian approach to MAP_MLLR. Private Communication.

[6] A. K. Gupta and T. Varga. *Elliptically Contoured Models in Statistics*. Kluver Academic Publishers, 1993.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society Ser. B*, 39:1–39, 1977.

[8] W. Reichl and W. Chou. A decision tree state tying based on segmental clustering for acoustic modeling. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 801–804, Seattle, Washington, USA, 1998.