DISCOUNTED LIKELIHOOD LINEAR REGRESSION FOR RAPID ADAPTATION

William Byrne and Asela Gunawardana

Center for Language and Speech Processing,
The Johns Hopkins University,
3400 N. Charles St.,
Baltimore, MD 21218, USA
{byrne,zilla}@jhu.edu

ABSTRACT

Rapid adaptation schemes that employ the EM algorithm may suffer from overtraining problems when used with small amounts of adaptation data. An algorithm to alleviate this problem is derived within the information geometric framework of Csiszár and Tusnády, and is used to improve MLLR adaptation on NAB and Switchboard adaptation tasks. It is shown how this algorithm approximately optimizes a *discounted likelihood* criterion.

1. INTRODUCTION

In speaker independent LVCSR systems, acoustic models are trained from a large amount of training data from many speakers. This is intended to yield robust models that work fairly well with a variety of channel conditions and speakers. As data becomes available for new speakers and channel conditions, model adaptation techniques can be used to adapt the models to the new conditions. We address here the difficult problem of rapid adaptation. A typical instance of this problem is a recognition task in which each speaker speaks only briefly so that only a small amount of speech is available for use in adaptation.

Many effective adaptation techniques estimate transformations of the acoustic model parameters to describe newly available data. Given a speaker independent (SI) model, choosing a model transformation is equivalent to choosing an adapted acoustic model. Thus, these transformations are constrained parameterizations of the acoustic models, and these adaptation techniques can be thought of as estimating constrained acoustic model parameters. For example, maximum likelihood linear regression (MLLR) [1] re-estimates constrained Gaussian means [2]. In fact, many adaptation schemes such as those of Padmanabhan *et al* [3] and of McDonough [4] which nominally transform the acoustic data to better match the SI models can be reformulated as transformations of the acoustic models, and as such can also be treated as model re-estimation schemes. Maximum *a posteriori* (MAP) adaptation schemes [5, 6] can also be viewed as re-estimating adapted model parameters from speaker independent parameters.

These adaptation schemes use the EM algorithm [7]. This well-known maximum likelihood iterative parameter estimation procedure may provide unreliable estimates when the amount of adaptation data is small. As an instance of this we present rapid adaptation examples in which MLLR overtrains in a single iteration. To address this problem we propose in Section 3 the *moment interpolation* variant of the EM algorithm. Our analysis of this variant makes use of the information geometric interpretation of the EM algorithm, presented by Csiszár and Tusnády [8] and reviewed in Section 2. In Section 5, it is shown that it approximately optimizes a *discounted likelihood* criterion [9] intended for use in estimation from small data sets.

This is a general modification to EM that can be used in many applications. In particular, it can be used to alleviate overtraining in the adaptation schemes discussed above. Since it optimizes an approximation of the discounted likelihood criterion, we call our procedure *Discounted Likelihood Linear Regression* (DLLR) when it is used for robust MLLR adaptation. In Section 4 we compare the performance of the DLLR procedure to that of the usual EM based MLLR procedure on supervised adaptation on the North American Business News (NAB) corpus and on unsupervised adaptation on the Switchboard corpus. Finally, we summarize our conclusions in Section 6.

2. AN INFORMATION GEOMETRIC VIEW OF THE EM ALGORITHM

Csiszár and Tusnády show that the EM algorithm can be viewed as alternating minimization between a parameter set Θ determined by the models and their parameterization, and a family of *desired distributions* $\mathcal{D}$ defined by

$$\mathcal{D} = \{P_X | P_Y = \delta_{\hat{y}}\},$$

where X is the complete variable, $Y = g(X)$ is the incomplete variable, P_Y is the Y-marginal of P_X, and $\delta_{\hat{y}}$ is a point mass at an observation $\hat{y}$ of Y. In the case of MLLR [1], the parameter family Θ is a set of linear transformations W_j that will be applied to the means of the HMM Gaussian observation distributions. The complete variable is (o_1^T, s_1^T) where $o_1^T = (o_1, \cdots, o_T)$ is the observed acoustic sequence and $s_1^T = (s_1, \cdots, s_T)$ is the hidden state sequence; the incomplete variable consists of only the acoustic sequence o_1^T. Note that $\mathcal{D}$ is the set of *all* probability densities (HMM or not), that put all their mass on the observed output sequence o_1^T.

The alternating minimization is done under the information divergence between a desired distribution $P_X \in \mathcal{D}$ and a parameter $\theta \in \Theta$, which is given by

$$D(P_X, \theta) = \sum_x P_X(x) \log \frac{P_X(x)}{Q_X(x;\, \theta)}.$$

Given a parameter iterate $\theta^{(p)}$, the alternating minimization procedure, which is shown schematically in Figure 1, consists of the following two steps:

1. **Forward Projection (E Step)**: Find a desired distribution $P_X^{(p+1)}$ such that

   $$P_X^{(p+1)} \in \arg\min_{P_X \in \mathcal{D}} D(P_X, \theta^{(p)}).$$

   Csiszár and Tusnády show that the unique minimizer is given by $P_X^{(p+1)}(x) = Q_{X|Y}(x|\hat{y};\, \theta^{(p)})$. This unique minimizer is known as the *I-projection* of $\theta^{(p)}$.

2. **Backward Projection (M Step)**: Find a parameter $\theta^{(p+1)}$ such that

   $$\begin{aligned} \theta^{(p+1)} &\in \arg\min_{\theta \in \Theta} D(P_X^{(p+1)}, \theta) \\ &\in \arg\min_{\theta \in \Theta} D(Q_{X|Y=\hat{y};\, \theta^{(p)}}, \theta). \end{aligned}$$

   It is easy to show that this minimization is equivalent to maximizing the EM auxiliary function, and thus to show

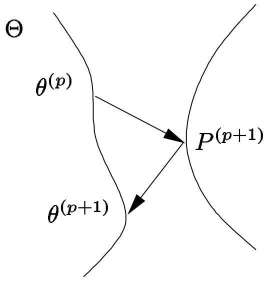

Figure 1. A schematic illustration of the alternating minimization procedure.

> that the alternating minimization procedure is equivalent to the EM algorithm.

When using the the EM algorithm for iterative MLLR with multiple regression classes, the transformations $W^{(p+1)} = \{W_1^{(p+1)}, \cdots, W_J^{(p+1)}\}$ are re-estimated from the previous transformations $W^{(p)} = \{W_1^{(p)}, \cdots, W_J^{(p)}\}$ by choosing the transformation $W_j^{(p+1)}$ for regression class $\mathcal{C}_j$ to satisfy the following condition [1]:

$$\sum_{s\in\mathcal{C}_j}\sum_{t=1}^{T}\tilde{\gamma}_s^{(p+1)}(t)\Sigma_s^{-1}o_t\nu_s' = \sum_{s\in\mathcal{C}_j}\sum_{t=1}^{T}\tilde{\gamma}_s^{(p+1)}(t)\Sigma_s^{-1}W_j\nu_s\nu_s' \tag{1}$$

where $\tilde{\gamma}_s^{(p+1)}(t)$ is the occupancy probability of state s at time t evaluated under the desired distribution $P_X^{(p+1)} = Q_{X|Y=\hat{y};\,W^{(p)}}$, ν_s and Σ_s are the extended mean vector and covariance matrix of that state's output density, and o_t is the observation vector at time t.

In practice, the re-estimation is done in two steps. First, the forward-backward algorithm is used to evaluate the accumulators

$$\begin{aligned}\tilde{c}_s^{(p+1)} &= \sum_{t=1}^{T}\tilde{\gamma}_s^{(p+1)}(t)\\ \tilde{d}_s^{(p+1)} &= \sum_{t=1}^{T}\tilde{\gamma}_s^{(p+1)}(t)o_t.\end{aligned} \tag{2}$$

Then, the re-estimation equations

$$\sum_{s\in\mathcal{C}_j}\Sigma_s^{-1}\tilde{d}_s^{(p+1)}\nu_s' = \sum_{s\in\mathcal{C}_j}\Sigma_s^{-1}\tilde{c}_s^{(p+1)}W_jx\nu_s\nu_s' \tag{3}$$

are solved for each transformation $W_j^{(p+1)}$. Note that these formulae are easily extended to HMMs with Gaussian mixture densities [1].

3. THE MOMENT INTERPOLATION ALGORITHM

The moment interpolation algorithm replaces the forward projection (E step) of the EM algorithm with the following:

$$P_X^{(p+1)} = \lambda Q_{X|Y=\hat{y};\,\theta^{(p)}} + (1-\lambda)P_X^{(p)}.$$

That is, it interpolates between the previous desired distribution and the I-projection of the previous parameter to get the next desired distribution. This is shown schematically in Figure 2.

By using the convexity of the divergence, it can be shown that this choice of $P_X^{(p+1)}$ guarantees that

$$D(P_X^{(p+1)}, \theta^{(p)}) \leq D(P_X^{(p)}, \theta^{(p)})$$

with equality only if $P_X^{(p)} = Q_{X|Y=\hat{y};\,\theta^{(p)}}$ [10]. This can be seen to be a GAM procedure as described by Gunawardana and Byrne [10, 11]; the algorithm converges to the same points as the EM algorithm does. However, this procedure slows down the convergence of the algorithm, allowing the opportunity to terminate the algorithm before overtraining occurs.

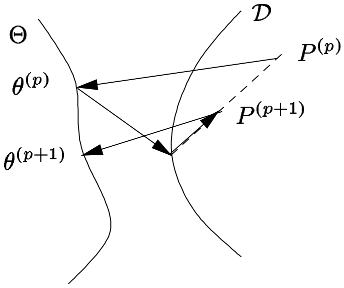

Figure 2. A schematic illustration of the moment interpolation algorithm.

Using moment interpolation for MLLR corresponds to replacing the occupancies $\tilde{\gamma}_s^{(p+1)}(t)$ in equations (1) and (2) by $\gamma_s^{(p+1)}(t)$, where

$$\gamma_s^{(p+1)}(t) = \lambda\tilde{\gamma}_s^{(p+1)}(t) + (1-\lambda)\gamma_s^{(p)}(t). \tag{4}$$

This is because the desired distribution $P_X^{(p+1)}$ (or $Q_{X|Y=\hat{y};\,\theta^{(p)}}$ in the case of EM) only enters the re-estimation equations through $\gamma_s^{(p+1)}(t)$, which are its moments, and therefore linear in it. In practice, the accumulators $\tilde{c}_s^{(p+1)}$ and $\tilde{d}_s^{(p+1)}$ in equation (3) are replaced by

$$\begin{aligned}c_s^{(p+1)} &= \lambda\tilde{c}_s^{(p+1)} + (1-\lambda)c_s^{(p)}\\ d_s^{(p+1)} &= \lambda\tilde{d}_s^{(p+1)} + (1-\lambda)d_s^{(p)},\end{aligned} \tag{5}$$

since the accumulators are also linear in the desired distribution $P_X^{(p+1)}$.

3.1. Initializing the Algorithm

The success of MLLR using moment decay depends strongly on how the distribution $P_X^{(0)}$ is initialized, which in turn determines how the accumulators $c_s^{(0)}$ and $d_s^{(0)}$ are initialized. Since the goal of moment interpolation is to make the re-estimation more robust in the small data regime, we use the SI training data to initialize these accumulators with robust counts. We do this by estimating a single global MLLR transform $W^{(0)}$ *using the entire speaker independent acoustic training corpus.* This procedure uses EM-based MLLR and and produces the accumulators $\sum_{t=1}^{T_{SI}}\gamma_{SI}{}_s^{(0)}(t)$ and $\sum_{t=1}^{T_{SI}}\gamma_{SI}{}_s^{(0)}(t)o_{SI}{}_t$, where T_{SI}, $\gamma_{SI}{}_s^{(0)}(t)$ and $o_{SI}{}_t$ are the length of the SI data, state occupancy on the SI data, and SI observation vector respectively. We cannot use these saved accumulators directly because they are moments of a distribution which puts all its mass on the T_{SI} length SI training data; we instead desire moments of a distribution which puts all its mass on the T length adaptation data. We therefore normalize the saved accumulators as follows to get our initial accumulator counts:

$$c_s^{(0)} = \frac{T}{T_{SI}}\sum_{t=1}^{T_{SI}}\gamma_{SI}{}_s^{(0)}(t)$$

$$d_s^{(0)} = \frac{T}{T_{SI}}\sum_{t=1}^{T_{SI}}\gamma_{SI}{}_s^{(0)}(t)o_{SI}{}_t$$

Notice that since equation (3) is multiplied by a constant, the re-estimation procedure is unchanged, and these normalized counts will generate the speaker independent transformation $W^{(0)}$. However, normalization does change the effect of interpolating these initial accumulator counts with those generated from the

	Iteration				
	SI	1	2	3	4
MLLR	10.6	10.7			
DLLR	10.6	10.3	10.1	9.9	10.0

Table 1. Word error rates for supervised adaptation on NAB using two sentences of adaptation data. The MLLR and DLLR procedures are used to estimate a global transformation.

	SI	global	3 class	10 class
MLLR	10.6	10.1	10.5	11.8
DLLR	10.6	9.9	9.6	9.4

Table 2. Word error rates for supervised adaptation on the NAB corpus using eight sentences of adaptation data. Performance the MLLR and DLLR procedures when estimating one, three, and ten transformations is compared.

adaptation data – without it, the SI counts would overwhelm the adaptation counts.

In practice, this initialization gives a transformation $W^{(0)}$ that is almost identity. It is not exactly identity since SI training of the original acoustic models is not done to convergence.

4. RESULTS

MLLR speaker adaptation using the moment interpolation algorithm was performed on both the NAB and Switchboard LVCSR corpora. The results are presented below.

4.1. Supervised Adaptation on the NAB Corpus

Lattice rescoring experiments were run on the Aug. 1994 development test set of the NAB corpus. The SI system had 6700 unique states, each with a mixture of 12 Gaussians, and was trained on 80 hours of speech from the Wall Street Journal corpus (the SI-284 set). Supervised adaptation was carried out on adaptation sets of 2 and 8 utterances. Results for global MLLR based on two utterances given in Table 1 show that the EM based MLLR procedure overtrains in the first iteration, while the moment interpolation based DLLR procedure improves the word error rate (WER) over multiple iterations.

Using eight adaptation utterances we find that the EM algorithm gives a small gain when estimating a global transformation, that this gain becomes smaller when estimating transformations for three regression classes, and that the algorithm overtrains when estimating ten transformations. The moment interpolation algorithm gives better performance at all these tasks, and does not overtrain. Table 2 summarizes these results.

When estimating transformations for multiple regression classes, we find that MLLR often overtrains unless class specific transformations are initialized by a global MLLR transformation. The MLLR results here were obtained in this manner. However, DLLR is more robust. The multiple regression class DLLR transformations used here were estimated directly.

In Figure 3, the two algorithms are compared in estimating a ten regression class transformation from eight utterances of adaptation data. It can clearly be seen that although the EM algorithm increases adaptation set likelihood faster than the moment interpolation algorithm, the test set likelihood does not track this increase under the EM algorithm. The results of this overtraining is reflected in the word error rate. However the moment interpolation algorithm increases both adaptation and test set likelihoods, and improves the word error rate.

4.2. Unsupervised Adaptation on the Switchboard Corpus

Lattice rescoring experiments were run on a development test set of the Switchboard corpus. The baseline system was trained on 60 hours of speech using PLP cepstral features with per-utterance cepstral mean subtraction. Triphone state clustering used word-boundary information and yielded an SI system with 8000 unique states, each with a mixture of 12 Gaussians. We note that this system can be further refined using conversation-side cepstral mean subtraction (CMS), vocal tract normalization (VTN), global MLLR and a trigram language model to yield a baseline performance of 36.6% WER. However, the baseline for our experiments is 43.1% – the global MLLR used in the refined system is based on the entire test set, and VTN and CMS require more than the five seconds of data we use for adaptation. Adaptation was done in an unsupervised fashion, and the adapted models were tested on data that did not overlap with the adaptation data according to the protocol defined in the Rapid Speaker Adaptation project at the 1999 JHU LVCSR workshop [12]. This protocol determines how well transformations learned on the small adaptation data generalize to unseen data, and also provides a test set large enough to obtain reliable measurements of performance.

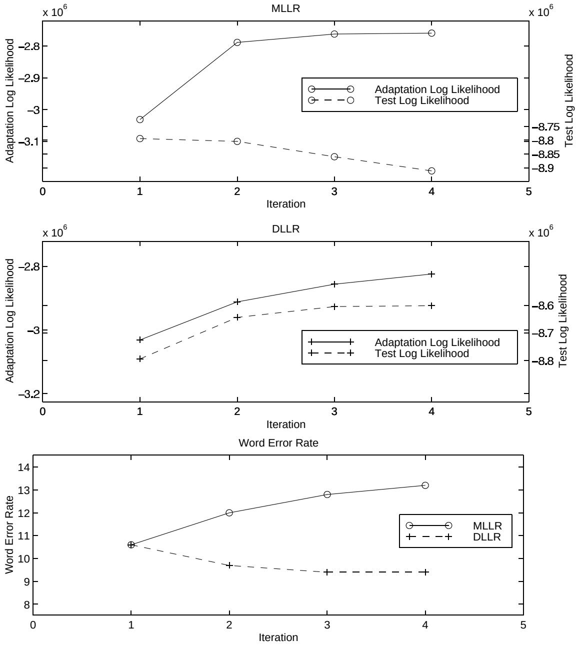

Figure 3. A comparison of adaptation and test likelihood as well as WER between the MLLR and DLLR procedures when used to estimate a 10 regression class transformation on the NAB corpus.

	Iteration					
	SI	1	2	3	4	5
MLLR-1	43.1	46.8				
DLLR-1	43.1	42.3	42.0	42.0	41.9	41.7
DLLR-3	43.1	42.6	42.2	42.5		
DLLR-10	43.1	43.0	42.8	42.9		

Table 3. Word error rates for unsupervised adaptation on the Switchboard corpus using five seconds of adaptation data. The performance of the MLLR procedure for estimating one transformation (MLLR-1) and that of the DLLR procedure for estimating one, three, and ten transformations (DLLR-1, DLLR-3, DLLR-10) are compared.

Table 3 compares the EM based MLLR and the moment interpolation based DLLR procedures when used to estimate one, three, and ten transformations based on five seconds of adaptation data. The EM algorithm overtrains serverly in one iteration, while the DLLR procedure provides a gain of 1.4% in word error rate. However, this gain is reduced when more regression classes are used. This shows that although the moment interpolation algorithm slows down overtraining, it cannot completely prevent it.

5. DISCUSSION

The results in Section 4 show that the moment interpolation algorithm can alleviate overtraining encountered in rapid adaptation. It is necessary at this point to discuss how the initialization of the algorithm affects its convergence behavior. The initialization described in Section 3.1 selects $P_X^{(0)}$ so that $P_Y^{(0)}$ puts all its mass on the SI training data $\hat{y}_{SI}$ rather than on the adaptation data $\hat{y}$. Thus, $P_X^{(0)}$ is not contained in the desired family $\mathcal{D}$, and we are no longer guaranteed that this algorithm will converge to the same points as the EM algorithm.

In this light, it is instructive to see how the moment interpolation algorithm relates to the discounted likelihood criterion [9], which is formulated specifically for the case where there is little data, and to the MAP estimation procedures to which it bears some resemblance.

5.1. Discounted Likelihood and Moment Interpolation

The discounted likelihood criterion finds a parameter θ which is at minimum divergence from a discounted desired family $\mathcal{D}_\lambda$ defined as

$$\mathcal{D}_\lambda = \{P_X | P_Y = \lambda \delta_{\hat{y}}\},$$

where $0 < \lambda < 1$. Define the divergence between the desired family $\mathcal{D}_\lambda$ and a parameter $\theta \in \Theta$ as

$$D(\mathcal{D}_\lambda, \theta) = \min_{P_X \in \mathcal{D}_\lambda} D(P_X, \theta).$$

The discounted likelihood criterion minimizes $D(\mathcal{D}_\lambda, \theta)$. The discounted desired family $\mathcal{D}_\lambda$ differs from the maximum likelihood desired family $\mathcal{D}$ in that it puts weight λ on the observation $\hat{y}$ rather than weight 1. The weight λ can be interpreted as a confidence in the observation $\hat{y}$. It can be argued that moment interpolation with weight λ optimizes an approximation of $D(\mathcal{D}_\lambda, \theta)$ [9].

Our case differs slightly from this ideal case in that the initialization of $P_X^{(0)}$ is such that it is not in $\mathcal{D}_\lambda$. In fact, it is easy to see that $P_X^{(0)}$ belongs to the set $\mathcal{D}_0$, since it puts no weight on the adaptation data (assuming the SI training data and the adaptation data are different). However, moment interpolation with interpolation weight λ guarantees that $P_X^{(p)}$ belongs to $\mathcal{D}_{\lambda^{(p)}}$ where $\lambda^{(p)}$ is given by

$$\lambda^{(p)} = 1 - (1 - \lambda)^p.$$

Thus, although the moment interpolation algorithm converges to maximum likelihood transformations, its iterates improve discounted likelihood criteria.

5.2. Moment Interpolation and MAP

Although there is a superficial similarity between the re-estimation formulas of the moment interpolation and MAP [5, 13, 6], the procedures differ in that the moment interpolation algorithm presented here optimizes a maximum likelihood criterion rather than a maximum *a posteriori* criterion. In fact, the moment interpolation procedure presented above has no dependence on a prior. As is often observed, MAP procedures converge to maximum likelihood solutions asymptotically in the amount of training data. The moment interpolation EM variant described here converges to maximum likelihood solutions for training sets of all sizes.

Moment interpolation can be incorporated into a MAP adaptation procedure to make it more robust in the small data case. An EM based MAP estimation procedure [13, 14] can be formulated as alternating minimization under the penalized divergence

$$D_{MAP}(P_X, \theta) = D(P_X, \theta) - \log \pi(\theta),$$

where $\pi(\theta)$ is the prior. Since the penalty term depends only on the parameter θ, it does not affect the forward projection (E step). One could replace the forward projection with a moment interpolation step and still converge to MAP estimates. Thus, not only is the moment interpolation procedure different from MAP estimation, it complements it.

6. CONCLUSIONS

The moment interpolation procedure has been shown to be robust in the small data case, providing a 1.4% improvement in accuracy on the Switchboard corpus, using only five seconds of adaptation data. The usual MLLR procedure overtrains in this case, *decreasing* the accuracy by 3.7%. The moment interpolation procedure is derived as a general extension of EM, and is therefore applicable to a large class of adaptation procedures.

Although it can be shown that the moment interpolation algorithm approximately optimizes the discounted likelihood criterion, it is essentially a maximum likelihood procedure (though it can be extended to MAP estimation). Therefore, if allowed to proceed to convergence, the algorithm will still be susceptible to overtraining. Further research into procedures that exactly optimize the discounted likelihood criterion is ongoing.

7. ACKNOWLEDGMENTS

We thank Entropic Cambridge Research Laboratory, Cambridge, UK for lattice generation and rescoring tools.

REFERENCES

[1] C. J. Leggetter and P. C. Woodland, “Speaker adaptation of continuous density HMMs using multivariate linear regression,” *Comp. Spch. & Lang.*, 1995.

[2] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, “Speaker adaptation using constrained estimation of gaussian mixtures,” *IEEE Trans. Spch. & Aud. Proc.*, vol. 3, pp. 357–366, Sept. 1995.

[3] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny, “Speaker clustering and transformation for speaker adaptation in speech recognition systems,” *IEEE Trans. Spch. & Aud. Proc.*, vol. 6, pp. 71–77, Jan. 1998.

[4] J. McDonough, W. Byrne, and X. Luo, “Speaker normalization with all-pass transforms,” in *ICSLP*, 1998.

[5] J.-L. Gauvain and C.-H. Lee, “Bayesian learning for hidden Markov model with Gaussian mixture state observation densities,” *Speech Communication*, vol. 11, pp. 205–213, 1992.

[6] K. Shinoda and C.-H. Lee, “Structural MAP speaker adaptation using hierarchical priors,” in *IEEE Wkshp. Spch. Recog. & Und.* (S. Furui, B.-H. Juang, and W. Chou, eds.), pp. 381–387, 1997.

[7] A. P. Dempster, A. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data,” *J. Roy. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[8] I. Csiszár and G. Tusnády, “Information geometry and alternating minimization procedures,” *Stat. & Dec., Supp. Iss. No. 1*, pp. 205–237, 1984.

[9] W. J. Byrne, “Generalization and maximum likelihood from small data sets,” in *IEEE-SP Wkshp. Neur. Net. Sig. Proc.*, 1993.

[10] A. Gunawardana and W. Byrne, “Convergence of EM variants,” tech. rep., CLSP, Johns Hopkins University, 1998. CLSP Research Note No. 32.

[11] W. Byrne and A. Gunawardana, “Convergence of EM variants,” in *IEEE Inf. Thry. Wkshp. Det. Est. Class. & Imag.*, p. 64, 1999.

[12] V. Digalakis, S. Berkowitz, E. Bocchieri, C. Boulis, W. Byrne, H. Collier, A. Corduneanu, A. Kannan, S. Khudanpur, and A. Sankar, “Rapid speech recognizer adaptation to new speakers,” in *ICASSP*, 1999.

[13] J.-L. Gauvain and C.-H. Lee, “Maximum *a posteriori* estimation for multivariate gaussian mixture observations of Markov chains,” *IEEE Trans. Spch. & Aud. Proc.*, vol. 2, no. 2, pp. 291–298, 1994.

[14] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions.* Wiley, 1997.