ROBUST ESTIMATION FOR RAPID SPEAKER ADAPTATION USING DISCOUNTED LIKELIHOOD TECHNIQUES

Asela Gunawardana and William Byrne

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

ABSTRACT

The discounted likelihood procedure, which is a robust extension of the usual EM procedure, is presented, and two approximations which lead to two different variants of the usual MLLR adaptation scheme are introduced. These schemes are shown to robustly estimate speaker adaptation transforms with very little data. The evaluation is carried out on the Switchboard corpus.

1. INTRODUCTION

In rapid speaker adaptation, the acoustic models of a speaker independent LVCSR system are modified to better match a given test speaker using a very small amount of adaptation data from that speaker. Standard adaptation methods such as maximum likelihood linear regression (MLLR) [1], constrained re-estimation of Gaussian means [2], and MAP adaptation [3] are based on the well known EM algorithm [4], which often provides unreliable estimates when the amount of adaptation data is small.

This problem has been addressed in the past by either severely restricting the space of model transforms [5, 2], or by using priors on the transforms [6, 7]. The first approach has the potential weakness of limiting the power of the adaptation, while the second requires assumptions on the form of the prior which may or may not hold in practice, and also requires estimating hyper-parameters.

We propose a more robust variant of the EM algorithm based on the *discounted likelihood* criterion [8]. Previously, we have presented initial experiments based on an approximate implementation of a discounted likelihood algorithm [9]. Here, we take a more principled approach, and discuss an alternate approximation that leads to a different implementation of the algorithm, and discuss the benefits and shortcomings of the two approaches.

The discounted likelihood procedure and its application to the estimation of MLLR-type speaker adaptation transforms is presented in Section 2, and two approximations leading to efficient implementations of the resulting discounted likelihood linear regression (DLLR) procedure are discussed in Section 3. In Section 4 we present results using these implementations for speaker adaptation on the Switchboard corpus. We discuss some of the relative merits of the algorithms presented in Section 5, and summarize our conclusions in Section 6.

2. DISCOUNTED LIKELIHOOD LINEAR REGRESSION

The discounted likelihood algorithm is derived as an alternating minimization procedure between a parameter set Θ determined by the models and their parameterization, and a family of *desired distributions* $\mathcal{D}_\lambda$ defined by

$$\mathcal{D}_\lambda = \{P_X | P_Y(\hat{y}) = \lambda\}\,,\ 0 < \lambda \leq 1$$

where X is the complete variable, $Y = g(X)$ is the incomplete variable, P_Y is the Y-marginal of P_X, and $\hat{y}$ is an observation of Y. Note that $\mathcal{D}_\lambda$ is the set of *all* probability measures that put mass λ on the observation $\hat{y}$. Csiszár and Tusnády [10] show that the case $\lambda = 1$ (i.e. when the desired distributions are concentrated on the observation $\hat{y}$) leads to the EM algorithm.

The procedure attempts to minimize the divergence between the desired distributions $P_X \in \mathcal{D}_\lambda$ and the parameters $\theta \in \Theta$. This divergence $D(P_X, \theta)$ is defined as follows. Given a parameter θ we define a measure $\tilde{Q}_{X;\,\theta}$ which assigns mass $q_Y(\hat{y};\ \theta)$ to the point $\hat{y}$, and is equivalent to $Q_{X;\,\theta}$ everywhere else (where $q_{Y;\,\theta}$ is the density or probability mass function of the marginal $Q_{Y;\,\theta}$). We then define $D(P_X, \theta)$ by

$$D(P_X, \theta) = D(P_X || \tilde{Q}_{X;\,\theta})$$

where $D(\cdot||\cdot)$ is the information divergence[1].

The alternating minimization procedure consists of the iterative application of the following two steps to produce a sequence of parameter iterates $\theta^{(p)}$:

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

$$P_X^{(p+1)} \in \operatorname*{arg\,min}_{P_X \in \mathcal{D}_\lambda} D(P_X, \theta^{(p)}).$$

[1] Technically, $D(P_X||\tilde{Q}_{X;\,\theta})$ should include a normalization term since $\tilde{Q}_{X;\,\theta}$ is not necessarily a probability measure. However, for our purposes, we do not include this term. The reasons for this follow from Csiszár and Tusnády [10].

0-7803-6293-4/00/$10.00 ©2000 IEEE.

It can easily be shown that the unique minimizer is given by the density (or probability mass function)

$$p_X^{(p+1)}(x) = \begin{cases} \lambda \, q_{X|Y}(x|\hat{y};\, \theta^{(p)}) & \text{for } y = \hat{y} \\ (1-\lambda)\, q_X(x;\, \theta^{(p)}) & \text{for } y \neq \hat{y}, \end{cases} \quad (1)$$

where $q_{X;\,\theta}$ and $q_{X|Y;\,\theta}$ are the joint and conditional densities of the model. This unique minimizer is known as the *I-projection* of $\theta^{(p)}$ on $\mathcal{D}_\lambda$.

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

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

□

Thus, the procedure attempts to find parameters θ such that the models $Q_{X;\,\theta}$ come close (under the divergence) to putting weight λ on the observation $\hat{y}$. Discounted likelihood differs from EM in that it allows this weight to be less than 1. We will return to this point in the discussion in Section 5.

In this paper, we discuss the use of the discounted likelihood procedure to estimate MLLR-type transforms [1] for adaptation of acoustic model means. In this case, the model $Q_{X;\,\theta}$ is composed of a language model $Q_{V_1^N}$ which is independent of θ, and acoustic HMMs $Q_{O_1^T, S_1^T | V_1^N;\, \theta}$, where the complete variable X is comprised of the acoustic sequence O_1^T, the HMM state sequence S_1^T, and the corresponding word sequence V_1^N. The HMM densities $q_{O_1^T, S_1^T | V_1^N;\, \theta}$ are specified through transition probabilities, mixture weights, and Gaussian means and covariances, all of which remain fixed, as well as the mean transforms $\{W_j\}$ which are to be estimated. Thus, in this case the parameters θ are sets of MLLR-type transforms $\{W_j\}$ which are applied to the means of Gaussians belonging to regression classes $\{C_j\}$. The estimation is based on an observation $\hat{y} = (\hat{o}_1^T, \hat{v}_1^N)$ of the incomplete variable $Y = (O_1^T, W_1^N)$. While $\hat{v}_1^N$ is provided in the case of supervised adaptation, in the case of unsupervised adaptation an initial recognition pass yields the words $\hat{v}_1^N$ corresponding to the adaptation acoustic data $\hat{o}_1^T$.

The alternating minimization procedure presented above results in the following equation for re-estimating new transforms $\{W_j^{(p+1)}\}$ from the previous transforms $\{W_j^{(p)}\}$ [1]:

$$\sum_{s \in C_j} \Sigma_s^{-1} \mathbf{E}_{P_X^{(p+1)}} \left[\sum_{t=1}^{T} \mathbf{1}(S_t = s) O_t \right] \mu_s' =$$
$$\sum_{s \in C_j} \Sigma_s^{-1} \mathbf{E}_{P_X^{(p+1)}} \left[\sum_{t=1}^{T} \mathbf{1}(S_t = s) \right] W_j \mu_s \mu_s' \quad (2)$$

where $\mathbf{1}(S_t = s)$ is an indicator function which takes value one when the hidden state S_t at time t takes the value s and is zero otherwise, and μ_s and Σ_s are the extended mean vector and the covariance matrix of that state's output density. Note that for $\lambda = 1$, the I-projection $P_X^{(p+1)}$ is given by $Q_{X|Y=\hat{y};\, \theta^{(p)}}$, and equation (2) becomes the familiar MLLR re-estimation equation [1].

In practice, the re-estimation is done in two steps. First, the accumulator counts

$$\begin{aligned} c_s^{(p+1)} &= \mathbf{E}_{P_X^{(p+1)}} \left[\sum_{t=1}^{T} \mathbf{1}(S_t = s) \right] \\ d_s^{(p+1)} &= \mathbf{E}_{P_X^{(p+1)}} \left[\sum_{t=1}^{T} \mathbf{1}(S_t = s) O_t \right] \end{aligned} \quad (3)$$

are evaluated, and the re-estimation equations

$$\sum_{s \in C_j} \Sigma_s^{-1} d_s^{(p+1)} \mu_s' = \sum_{s \in C_j} \Sigma_s^{-1} c_s^{(p+1)} W_j \mu_s \mu_s' \quad (4)$$

are solved for each transform $W_j^{(p+1)}$. Since this allows estimation of MLLR-type transforms under discounted likelihood criterion, we call this procedure discounted likelihood linear regression (DLLR). Note that these formulae are easily extended to HMMs with Gaussian mixture densities [1]; for simplicity, we present only the single mixture component case.

3. IMPLEMENTING DLLR

From equation (1) it can be seen that the I-projection of a parameter $\theta^{(p)}$ onto $\mathcal{D}_\lambda$ can be written as

$$P_X^{(p+1)} = \lambda \tilde{P}_X^{(p+1)} + (1-\lambda) Q_{X;\, \theta^{(p)}}$$

where $\tilde{P}_X^{(p+1)} = Q_{X|Y=\hat{y};\, \theta^{(p)}}$ is the I-projection that would be obtained under the EM algorithm (i.e., when $\lambda = 1$). Thus, the accumulator counts are given by

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

where $(\tilde{c}_s^{(p+1)}, \tilde{d}_s^{(p+1)})$ and $(\bar{c}_s^{(p+1)}, \bar{d}_s^{(p+1)})$ are expectations under $\tilde{P}_X^{(p+1)}$ and $Q_{X;\, \theta^{(p)}}$ respectively. Since $\tilde{P}_X^{(p+1)} = Q_{X|Y=\hat{y};\, \theta^{(p)}}$, it can be seen that $\tilde{c}_s^{(p+1)}$ and $\tilde{d}_s^{(p+1)}$ are the usual EM counts obtained from the forward-backward algorithm. However, since $\bar{c}^{(p+1)}$ and $\bar{d}^{(p+1)}$ are expectations under $Q_{X;\, \theta}$ instead of $Q_{X|Y=\hat{y};\, \theta}$, they are not conditioned on $\hat{v}_1^N$. Therefore, their computation involves a sum over all possible word sequences, which is intractable.

We now present two ways of approximating these counts that yield two different implementations of the DLLR procedure.

3.1. DLLR through Moment Interpolation

Given a distribution $P_X^{(p+1)}$, it can be shown that for model families with densities of the form

$$q_X(x;\, \theta) = \frac{1}{z(\theta)} \exp(\theta \cdot g(x)),$$

the M-step is given by choosing a parameter $\theta^{(p+1)}$ such that

$$\mathbf{E}_{Q_{X;\theta^{(p+1)}}} [g(X)] = \mathbf{E}_{P_X^{(p+1)}} [g(X)]. \quad (5)$$

However, HMMs with Gaussian observations are, in the terminology of Efron [11], curved exponential families of the form

$$q_X(x;\ \theta) = \frac{1}{z(\theta)} \exp(f(\theta) \cdot g(x)),$$

as illustrated by the following example.

Example: Suppose the acoustic features are one dimensional, that the transforms are simple scalings, and that $\sigma_s^2 = 1/2\pi$, $\mu_s = 1$ for all states s. Then

$$\begin{aligned} q_X(x;\theta) &= q_{V_1^N}(v_1^N)\ q_{S_1^T|V_1^N}(s_1^T|v_1^N)\ q_{O_1^T|S_1^T}(o_1^T|s_1^T;\ \{W_j\}) \\ &= e^{g_0(v_1^N, s_1^T)} \prod_j \prod_{s\in C_j} \prod_{t=1}^{T} \left(e^{\pi(o_t - W_j)^2}\right)^{\mathbf{1}(s_t=s)} \\ &= \exp\left\{ g_0(v_1^N, s_1^T) + \pi \sum_{t=1}^{T} o_t^2 \right. \\ &\qquad + \sum_j W_j \Big(2\pi \sum_{s\in C_j} \sum_{t=1}^{T} o_t \mathbf{1}(s_t = s)\Big) \\ &\qquad\qquad \left. + \sum_j W_j^2 \Big(\pi \sum_{s\in C_j} \sum_{t=1}^{T} \mathbf{1}(s_t = s)\Big) \right\} \end{aligned}$$

and since the density involves terms that depend on W_j^2, it is a curved exponential family [12].

Thus equation (5) applies only approximately when $f(\theta)$ is close to linear. Assuming that this is so, it follows that $\bar{c}_s^{(p+1)} \approx c_s^{(p)}$ and $\bar{d}_s^{(p+1)} \approx d_s^{(p)}$, giving

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

which we term the *moment interpolation* algorithm [9].

3.2. DLLR Through Speaker Independent Counts

As an alternative to the approximation discussed above, we attempt to find the needed counts directly. By expanding the definition of $\bar{d}_s^{(p+1)}$ and rearranging we get

$$\begin{aligned} \bar{d}_s^{(p+1)} &= \mathbf{E}_{Q_X}\left[\sum_{t=1}^{T} \mathbf{1}(S_t = s) O_t;\ \theta^{(p)}\right] \\ &= \mathbf{E}_{Q_{V_1^N, S_1^T}}\left[\sum_{t=1}^{T} \mathbf{1}(S_t = s) \mathbf{E}_{Q_{O_t|S_t}}\left[O_t | s; \theta^{(p)}\right];\ \theta^{(p)}\right] \\ &= \mathbf{E}_{Q_{V_1^N, S_1^T}}\left[\sum_{t=1}^{T} \mathbf{1}(S_t = s) W_j^{(p)} \mu_s;\ \theta^{(p)}\right] \\ &= \bar{c}_s^{(p+1)} W_j^{(p)} \mu_s, \end{aligned}$$

where C_j is the regression class to which the state s belongs. Thus, we need only estimate the state occupancy counts under the model, as opposed to the conditional state occupancy counts given the word sequence $\hat{v}_1^N$ and the acoustic sequence $\hat{o}_1^T$. Unfortunately, this still involves a sum over all word sequences. Currently, we approximate this by the speaker independent (SI) state occupancies obtained during acoustic model training: $\bar{c}_s^{(p+1)} \approx \frac{1}{K} c_s^{SI}$, where the constant K is chosen so that $\sum_s \bar{c}_s^{(p+1)}$ agrees with the number of frames of adaptation data. However, since these are conditional counts given the speaker independent acoustic data and corresponding word sequences, there is much room for improvement, as discussed in Section 6.

4. RESULTS

DLLR speaker adaptation using the moment interpolation algorithm as well as using the speaker independent counts was performed on the Switchboard LVCSR corpus. Results of lattice rescoring experiments on a development test set of the corpus are reported.

Adaptation was done in an unsupervised fashion based on approximately 5 seconds of adaptation data, 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 1998 JHU LVCSR workshop [13]. This protocol determines how well transforms learned on the small adaptation data generalize to unseen data, and also provides a test set large enough to obtain reliable measurements of performance.

The baseline system was trained on 60 hours of speech using 39-dimensional PLP cepstral features with per-utterance cepstral mean subtraction. Triphone state clustering used word-boundary information and yielded an SI system with about 8000 unique states, each with a mixture of 12 Gaussians. This system provides a baseline word error rate (WER) of 43.1% on bigram lattices.

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 global MLLR used in the refined system is based on the entire conversation side, and VTN and CMS require more than the five seconds of data we use for adaptation. Thus, these refinements cannot be used when only 5 seconds of adaptation data is provided.

Training a global speaker adaptation transform based on the 5 seconds of adaptation data using MLLR leads to overtraining in a single iteration, even when a block diagonal transform of three thirteen-element blocks is estimated. However, both DLLR methods are able to robustly estimate full transforms for two regression classes, and give a gain of 1.4% WER.

Directly interpolating MLLR transforms (again for two regression classes) with a robust transform obtained from the SI counts (i.e. interpolating the transforms themselves instead of the counts that give rise to them) was also investigated. This experiment is meant to be compared to the interpolation of MLLR counts with SI counts as described in Section 3.2. Note that estimation based directly on SI counts yields transforms that are almost identity, since the unadapted acoustic model means were trained from the SI data. This approach was not as effective, and only gave a 0.7% WER improvement.

These results are summarized in Table 1. The value of λ was chosen for these experiments was 0.1.

	Iteration					
	SI	1	2	3	4	5
MLLR-full	43.1	46.8				
MLLR-block	43.1	43.7				
MLLR-interp	43.1	42.4	43.0	43.0		
DLLR-MI	43.1	42.6	42.2	42.0	41.8	41.7
DLLR-SI	43.1	42.6	42.0	42.0	41.9	41.7

Table 1: The performance of the MLLR procedure for estimating a full transform (MLLR-full) and a block diagonal (MLLR-block) transform as well as that of the moment interpolation based and SI count based DLLR algorithms (DLLR-MI and DLLR-SI resp.) are compared. Also shown are the results of interpolating the results of MLLR with a robust transform (MLLR-interp).

5. DISCUSSION

The results in Section 4 show that the discounted likelihood algorithms can alleviate overtraining encountered in rapid adaptation. This can be understood in terms of the description of the alternating minimization procedure given in Section 2. The MLLR procedure ($\lambda = 1$) attempts to find transforms such that the models put all their probability mass on the adaptation data. Since this data is unreliable, an algorithm which attempts to put only a mass $\lambda < 1$ on the adaptation data leads to more robust behavior.

Even though it produces more robust estimates than the EM in its early stages, it can be shown that the discounted likelihood procedure converges to stationary points of the likelihood, regardless of the value of λ. Thus, the fixed points of the procedure do not depend on λ, and are also fixed points of the EM algorithm (since the EM algorithm results when $\lambda = 1$). However, the discounted likelihood criterion is less greedy, and the results presented show that overtraining can be avoided by early termination of the algorithm.

It can be seen that the effect of the discounted likelihood criterion on the usual EM based estimation procedure is to augment the counts gathered from unreliable data by reliable counts generated from the model to give more robust statistics on which to base the estimation. We believe that this is the natural way in which to do robust estimation. It can be seen that the more obvious method of interpolating the unreliable estimates obtained from the EM algorithm with more robust speaker independent transforms is not as effective.

6. CONCLUSIONS

The discounted likelihood criterion leads to estimation procedures which are robust in the face of small amounts of data. Two implementations of DLLR have been described that robustly estimate speaker adaptation transforms for the Switchboard corpus from 5 seconds of data. Although the two implementations of DLLR provided similar results, the moment interpolation algorithm makes assumptions on the form of the model family that may not hold. Also, the moment interpolation algorithm requires the retention of the interpolated forward-backward accumulator counts from iteration to iteration, which can be costly.

The second approximation, on the other hand, can be improved by using state occupancy counts that are not conditioned on the speaker independent acoustics or word sequences. Instead, the occupancies could be calculated directly from the language model and state transition probabilities. As mentioned earlier, computing these counts exactly would require a summation over all word sequences. However, we believe that the counts could be well approximated by using either the language model or the training data to judiciously sample the space of word sequences.

7. REFERENCES

[1] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of continuous density HMMs using multivariate linear regression," *Comp. Spch. & Lang.*, vol. 9, pp. 171–185, Apr. 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] J.-L. Gauvain and C.-H. Lee, "Bayesian learning for hidden Markov models with Gaussian mixture state observation densities," *Spch. Comm.*, vol. 11, pp. 205–213, 1992.

[4] 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.

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

[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] W. Chou, "Maximum a posterior linear regression with elliptically symetric matrix variate priors," in *Eurospeech*, vol. 1, pp. 1–4, 1999.

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

[9] W. Byrne and A. Gunawardana, "Discounted likelihood linear regression for rapid adaptation," in *Eurospeech*, 1999.

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

[11] B. Efron, "Defining the curvature of a statistical problem (with applications to second order efficiency)," *Ann. Stat.*, vol. 3, no. 6, pp. 1189–1242, 1975.

[12] S.-I. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, vol. 8, no. 9, pp. 1379–1408, 1995.

[13] 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.