MAXIMUM LIKELIHOOD SEQUENTIAL ADAPTATION

Beth Logan

Compaq Computer Corporation, Cambridge Research Laboratory,
One Kendall Square Building 700, Cambridge MA 02139, USA.
Beth.Logan@compaq.com

ABSTRACT

We develop a new sequential adaptation technique for HMMs based on an incremental variant of the EM algorithm. The approach has little impact on the speed of normal Viterbi decoding and in the case of mean adaptation only, is equivalent to incremental MAP adaptation for a certain choice of priors.

We apply the technique to the ARPA HUB4 broadcast news task. Here since the acoustic conditions change frequently, it is advisable to 'reset' the adaptation process periodically. However, for this task, the acoustic conditions change so rapidly that it is difficult to obtain enough information for adaptation between model resets. Many existing adaptation schemes tackle this problem of data sparsity by cleverly updating unseen mixture components. We investigate an orthogonal strategy in which a set of models, each representing a different acoustic condition, is maintained and adapted. We show that small improvements in performance are possible using this approach.

Keywords: sequential adaptation, online adaptation, speech recognition, HMM.

1. INTRODUCTION

A major challenge currently facing developers of large vocabulary speech recognition systems is how to quickly adapt generic speech models to changing environments and speakers. While numerous techniques exist (e.g. [1], [2], [3]), many require additional training examples to adapt to new conditions. This limits how quickly new situations can be compensated for, as well as hindering the development of real time systems.

Recently, there has been interest in sequential or online adaptation techniques (e.g. [1]). These algorithms update hidden Markov models (HMMs) using a very small number of adaptation utterances (of the order of one). They are thus able to adapt quickly to changing conditions.

Since reliable estimates of model parameters cannot be made from so few utterances, a method of including prior information is necessary. For example, [1] includes additional statistics in a Bayesian framework to form maximum *a posteriori* (MAP) estimates of the parameters. The exact solution of the MAP incremental learning problem is intractable however leading to approximations.

In the next section we develop an alternative scheme based on a sequential version of the Estimate-Maximize (EM) algorithm. This leads to an approximate maximum likelihood (ML) solution with prior information included as accumulated sufficient statistics. It should be noted that this algorithm is equivalent to the approximate MAP scheme in some cases.

We investigate the application of our technique to the broadcast news task. This task poses particular challenges for online adaptation as the acoustic conditions change very rapidly. Our aim is to improve recognition performance with minimal impact on the time taken.

2. SEQUENTIAL ADAPTATION

We model speech using HMMs with model parameters λ. Given a sequence of speech observations $\mathbf{O}$, the standard procedure is to train the model parameters according to a ML criteria. That is we choose λ^* to satisfy

$$\lambda^* = \arg\max_{\lambda} \left[p(\lambda|\mathbf{O}) \right]. \tag{1}$$

For HMMs no closed-form solution exists to solve Equation 1 for λ^* so the EM algorithm is typically employed. This process consists of two repeated steps as follows:

E Step: Given λ and $\mathbf{O}$ accumulate sufficient statistics S;

M Step: Choose λ' to maximize $E\{\log p(\mathbf{O}|\lambda')|\lambda, S\}$.

Note that here this algorithm has been expressed in a slightly non-standard way so as to better introduce the next section.

We now consider adaptation as an extension of this training process and adjust the model parameters to more closely match (in the ML sense) new data as it is presented. Specifically we consider an incremental version of the EM algorithm presented in [4].

E Step: For the utterance at time t, accumulate sufficient statistics S. Combine these new statistics with previously accumulated sufficient statistics $S^{(t-1)}$ to give $S^{(t)}$ using

$$S^{(t)} = S + \gamma S^{(t-1)} \tag{2}$$

where γ is a 'forgetting factor';

M Step: $\lambda^{(t)^*} = \arg\max_{\lambda^{(t)}} E\{\log p(\mathbf{O}|\lambda^{(t)})|\lambda^{(t-1)}, S^{(t)}\}$.

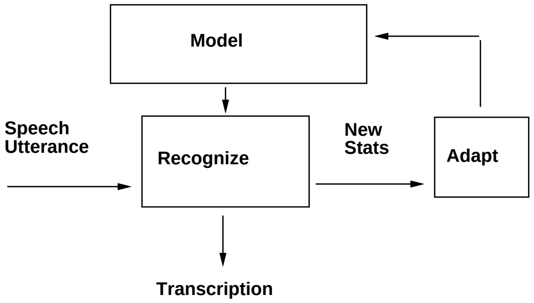

Figure 1: Basic online adaptation algorithm

Here the forgetting factor γ controls the speed of adaptation. This algorithm is exact if γ is the ratio of the number of frames in the newly presented utterance to the total number of frames used to train the model. For the case of sequential adaptation of a large vocabulary system however, this ratio approaches zero. Therefore we use an inexact form of the algorithm.

A standard HMM models the posterior state observation probabilities using mixtures of Gaussian densities with diagonal covariance matrices. In this case, the sufficient statistics consist of the mixture component occupation counts, state transition counts, and for each mixture component the sum and sum of squares.

Figure 1 shows the operation of the algorithm. As each new speech utterance is presented, the current model is used to decode it and to generate sufficient statistics. These statistics are then used in conjunction with the existing sufficient statistics to adapt the existing speech model. This has only minimal impact on the speed of the usual speech recognition process. We show unsupervised adaptation. Supervised adaptation is also possible if reference transcriptions are available. Unless stated, all our experiments perform unsupervised adaptation.

It should be noted that if the HMM states are modeled by mixtures of Gaussians and if only the means of these Gaussians are adapted, then the ML algorithm presented here is equivalent to approximate MAP adaptation [1] for a particular choice of priors. In this case γ is analogous to the 'precision factor' τ in the previous work.

3. EXPERIMENTS

We examine the performance of our algorithm on the ARPA HUB4 broadcast news task [5]. In nearly all cases, we consider only mean adaptation so our technique is equivalent to approximate MAP adaptation as mentioned. This approach has been applied with success to small to medium vocabulary tasks with substantial mismatch between the training and testing databases (e.g. [1], [6]). For the large vocabulary system considered here however, we use HMMs with many states to model the data. These cover the acoustic space reasonably well so since our test set is acoustically similar, we do not expect dramatic improvements in performance using our technique.

Our main concern in this paper is with the issues that arise when running an online recognition system. We look specifically at the problem of when to 'reset' the model parameters. Although we could continue to update the acoustic models indefinitely, this is inadvisable for two reasons.

The first is that the the algorithm is not exact. It is therefore possible for the models to diverge, giving increasingly worse performance. The second is that if the acoustic conditions change dramatically, it may be preferable to start from the initial models rather than those tuned to a more specific condition. We describe our experiments in the following sections.

3.1. Experimental Setup

We use the CRL Calista Large Vocabulary speech recognition system for experiments. We train generic HMMs on the HUB4e96 training set, modeling 13 dimensional mel-cepstrum features augmented by the first and second derivatives. We use 3-state triphone clustered continuous Gaussian density HMMs. There are a total of 6000 states. A standard trigram language model is used during recognition. Our experiments study a one-pass system where the adaptation step is considered to occur at the end of the pass. We use the HUB4e97 evaluation test set.

We consider two different segmentations of the test set. The first is based on the given transcriptions. The second is generated using the CMU segmenter [7]. This chooses segment boundaries based on differences in entropy between adjacent windows of the test set combined with a silence threshold. We shall refer to these segmentations as 'perfect' and 'imperfect' respectively. We also consider 'perfect' and 'imperfect' classification of the utterances into speaker and environment classes. In the latter case, we again use the tools provided by CMU.

3.2. Simple Resetting of Models

We first investigate a very simple strategy for resetting the models during online adaptation. Here, we reset the models after a set number of utterances have been observed. Table 1 shows the results of this experiment for models with 8 Gaussian mixture components per state. Only the first 200 utterances of the perfect segmentation of the test set are used. The baseline result, equivalent to using a standard Viterbi decoder, is obtained when γ is unity. In this experiment, only the mean of each Gaussian is adapted. We investigated adapting other parameters in this and other sections of the work, and found either no improvement or a degradation in performance.

These results indicate that the simple resetting strategy is inadequate for this data set. We note also that the choice of forgetting factor and resetting frequency are related: the longer the period between model resets, the less aggressive forgetting factor can be tolerated.

3.3. Resetting on Acoustic Changes

If we imagine online adaptation as slowly tuning the model parameters as more and more information is gathered about a particular condition, then a reasonable strategy for resetting the models is to reinitialize them when-

Reset Frequency (utts)	γ	% Error
-	1.00	26.2
100	0.95	26.1
	0.90	27.6*
50	0.95	26.3
	0.90	26.2
	0.85	27.0*
25	0.90	26.1
	0.80	26.1
	0.70	26.4

Table 1: Recognition results for online mean adaptation resetting the models periodically. Tested on the first 200 utterances of HUB4e97 using perfect segmentation. * indicates a significant difference to the baseline result with 95% confidence.

ever the environment or speaker changes. As a first step in this direction then, we conduct an artificial experiment in which we use perfectly segmented and labeled utterances. Every time the speaker or environmental condition changes, the models are reset. The recognition results on the full HUB4e97 test using models with 8 Gaussian mixture components per state are shown in Table 2. Again we only adapt the mean of each Gaussian.

From this table it is evident that the proposed adaptation scheme gives no improvement whatsoever. We also explored supervised adaptation and found it gave similar results. Thus the lack of adaptation is not due to erroneous transcriptions.

The reason lies in the nature of the data. For this task, the acoustic conditions change every few utterances. There is therefore insufficient information to adapt the model before it is reset since many triphones are not seen. Severely reducing γ so that less prior information is included does not solve this problem and in fact has a slight detrimental effect. We also explored resetting the model less often, such as on a speaker sex or environment change and saw no improvement.

To counteract the problem of insufficient information several approaches are possible. The first is to adapt more aggressively. Our algorithm only adapts the parameters of *observed* mixture components. However, since the problem of data sparsity is common to most adaptation schemes, much work has focussed on adapting the parameters of the *unobserved* mixture components (e.g. [2], [3], [8]). In these and other approaches, the general technique is to adapt the parameters of unseen mixture components according to the statistics of 'similar' observed mixture components. There are many variations on the way to identify similar components and how to use the information from these.

A second and in many respects orthogonal way to include more information is to maintain a set of models in parallel and adapt each of these sequentially as appropriate. Conceptually this is similar to dividing the testing data into sets with similar acoustics (such as speaker sets) and running online adaptation on each of these. In an actual online speech recognition system though, it is necessary to switch quickly between these models as the acoustic conditions change. This scheme is shown in Figure 2. We describe experiments with this system in the next section.

γ	% Error
1.00	26.1
0.90	26.1
0.80	26.1
0.50	26.1
0.30	26.1
0.05	26.2
0.01	26.3*

Table 2: Recognition results for online mean adaptation resetting the models on an environment or speaker change. Tested on the full HUB4e97 test set with perfect segmentation. * indicates a significant difference to the baseline result with 95% confidence.

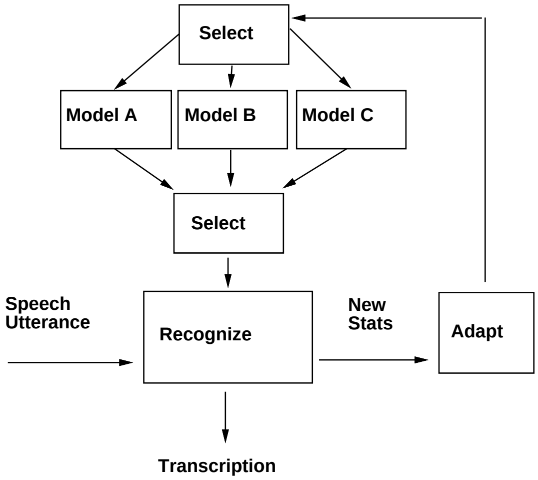

Figure 2: Online adaptation using parallel models

3.4. Parallel Model Adaptation

We first conduct an artificial experiment to investigate the feasibility of this scheme. We use perfectly segmented test data and assume that we can perfectly detect and classify speaker and environment changes. We maintain models for all classes with greater than two utterances in the test set (about 60 models) and perform online mean adaptation for these. The results of this experiment for models with varying number of Gaussian mixture components are shown in Table 3.

We see from these results that using more than one model gives a significant improvement in performance. Although the improvement is very minor, this must be balanced against the fact that almost no extra computation time is required to achieve it.

We now consider a less artificial experiment in which we use the imperfect segmentation of the database and classify each segment as telephone or non-telephone speech, then cluster these classifications further into speaker classes [7]. We use a different parallel model for each speaker class. This experiment is still artificial since the

Nr. Mix Components	γ	% Error
8	1.00	26.1
8	0.90	25.9*
8	0.85	26.0*
16	1.00	24.7
16	0.90	24.5*
16	0.85	24.6

Table 3: Recognition results for online mean adaptation using parallel models. Tested on the full HUB4e97 test set with perfect segmentation. * indicates a significant difference to the baseline result with 95% confidence.

Nr. Mix Comp.	Parallel Models	γ	% Error
16	-	1.00	25.4
	speakers from clusters	0.90	25.3
	male/female/telephone	0.90	25.3

Table 4: Recognition results for online mean adaptation using parallel models. Tested on the full HUB4e97 test set with imperfect segmentation.

clustering is performed in batch mode.

The second line of Table 4 shows the results of this experiment and we see a similar improvement in performance to the results in Table 3. We are unable to test the significance of these results as we do not have the correct transcription for each segment. (We are using the matched pairs test (e.g. [9]) which uses the difference in error rates on independent utterances as the test statistic.)

For the algorithm to be fully online, the choice of which model to use for recognition and adaptation must be made 'on the fly'. We therefore now consider using parallel models based on classes determined from the training data, using a Gaussian classifier to quickly classify each utterance. The third line of Table 4 shows the results of such an experiment using 3 models: a telephone model and non-telephone male and female models. These models are reset every 50 utterances. We see that even though the segmentation and classification are imperfect, a small improvement is still achieved.

We could also base the choice of parallel models on more specific classes such as speaker or environment clusters from the training set. Here an issue would be guaranteeing sufficient 'coverage' of all speakers and conditions. Additionally, it may be desirable to start the adaptation of each model from a less generic model. The exploration of these issues is the subject of ongoing work.

4. CONCLUSIONS

We have developed a new sequential adaptation technique for HMMs based on an incremental variant of the EM algorithm. If only the mean of each Gaussian model is adapted, the approach is equivalent to incremental MAP adaptation for a certain choice of priors. We applied the technique to online adaptation of the HUB4 broadcast news task, investigating strategies for resetting the models. We found that for this task, the frequent changes in acoustic conditions meant that it was difficult to obtain enough information for adaptation between model resets. Previous work has investigated schemes which sensibly update unseen mixture components given limited data. We investigated an orthogonal strategy in which a parallel set of models, each representing a different acoustic condition was maintained and adapted. We showed that small improvements in performance were possible using this approach, even if the models were predetermined and the model to use was selected imperfectly.

Although the improvements seen on this task were very minor, we stress that our scheme has little impact on the speed of a standard Viterbi decoder. Additionally, the use of parallel models is a general technique that can be applied to other online adaptation schemes, including those which update more than just the mixture components seen in the adaptation data.

5. ACKNOWLEDGMENTS

The author would like to thank the rest of the Speech Group at CRL for their help with running experiments and in particular Pedro Moreno for many fruitful discussions.

6. REFERENCES

[1] Q. Huo, C. Chan, and C. H. Lee, "On-line adaptation of the SCHMM parameters based on the segmental quasi-Bayes learning for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 2, pp. 141–144, 1996.

[2] G. Zavaliagko, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP*, pp. 676–679, 1995.

[3] C. Leggetter, *Improved acoustic modelling for HMMS using linear transformations.* PhD thesis, University of Cambridge, 1995.

[4] R. M. Neal and G. E. Hinton, "A new view of the EM algorithm that justifies incremental and other variants," *Learning in Graphical Models*, pp. 355–368, 1998.

[5] R. M. Stern, "Specification of the 1996 Hub 4 broadcast news evaluation," in *DARPA Speech Recognition Workshop*, 1997.

[6] K. Laurila, M. Vasilache, and O. Viikki, "A combination of discriminative and maximum likelihood techniques for noise robust speech recognition," in *Proc. ICASSP*, 1998.

[7] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news data," in *DARPA Speech Recognition Workshop*, pp. 97–99, 1997.

[8] V. Digalakis *et al.*, "Rapid speech recognizer adaptation to new speakers," in *Proc. ICASSP*, 1999.

[9] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*, pp. 532–535, 1989.