ON-LINE INCREMENTAL ADAPTATION FOR SPEAKER VERIFICATION USING MAXIMUM LIKELIHOOD ESTIMATES OF CDHMM PARAMETERS

Kin Yu & John Mason

Speech Research Group, Department of Electrical Engineering
University of Wales Swansea SA2 8PP, UK

e-mail: k.yu@swansea.ac.uk & j.s.d.mason@swansea.ac.uk

Phone: +44 1792 205678 ext. 4564, Fax: +44 1792 295686

ABSTRACT

This papers investigates two approaches to on-line incremental adaptation of CDHMM parameters.

First the popular MAP approach is examined, highlighting difficulties in automatically setting the adaptation rate. To overcome these problems we introduce a new approach based on the multi-observation estimation equations of the forward-backward algorithm called a cumulative likelihood estimate (CLE).

Experimental results using these two approaches are compared with and without the use of a speech model for enrolment on isolated word speaker models. In both enrolment procedures, the CLE approach can achieve approximately an EER of 1% for six adaptation sequences using a single digit test token.

1. Introduction

Incremental adaptation implies the incorporation of newly acquired data into existing models. In *speech* recognition this is known as speaker adaptation. Instead of performing the task of speech recognition after speaker adaptation, in this paper we apply speaker verification on these newly formed client models. Once enrolment has been performed with a small amount of speaker-specific data, the problem becomes one of continual update of the client model over time to improve recognition performance.

In applying an on-line incremental adaptation strategy we eliminate the requirement for storage of speaker-specific speech data and allow the model to change. Here the tasks of speech and speaker recognition present fundamental differences in class discrimination: initially only a small amount of data is likely to be available to define the new class, presenting a sparse training problem in speaker verification [1].

The two approaches discussed in this paper are (i) the maximum *a posteriori* approach developed for the CDHMM by Gauvain and Lee [2], and (ii) a new approach based on the multi-observation equations of the CDHMM, which applies a cumulative likelihood estimate (CLE) of the parameters of the model, storing the posteriori probability estimates.

2. Maximum *a posteriori* (MAP)

Many current adaptation approaches use a Bayesian learning framework to derive maximum *a posteriori* estimates of the parameters of a model. This method has been initially studied for single-mixture CDHMMs by Lee *et al.* [3], and extended to the multi-mixture case by Gauvain and Lee [2].

Application of the MAP approach requires the specification of the prior density parameters or hyper-parameters for each set of parameters we wish to update. In this paper the hyper-parameters are restricted to the posteriori density as if no prior information is available as described in [2]. Consequently, we can constrain the hyper-parameter estimation in such a way as to leave a single adaptation control parameter, referred to as τ [2]. Tying the hyper-parameters allows for simple application of MAP to incremental adaptation.

Similar instances of this MAP approach have been used in experiments by Gauvain [4] where $\tau = 2$, Ahadi where $\tau = 10$ and Matsui and Furui [5] where τ takes a range of values from 0.5 to 1.5.

For application to on-line incremental adaptation using the above procedure, we recursively update the hyper-parameters using the current model. A similar approach is described by Zavaliagkos [6] in his incremental adaptation procedure.

The specification of τ in this MAP approach is critical. When applied to speaker verification, the value of τ may vary according to the amount of available speaker-specific data as postulated by Matsui and Furui [5], and speaker-dependent control parameters in the speaker verification context may prove useful.

3. Cumulative likelihood estimates (CLE)

The method of cumulative likelihood estimates (CLE) of the CDHMM parameters attempts to address the problems as-

sociated with the choice of τ in the afore mentioned MAP approach. This is done by introducing the use of speaker-dependent parameters which control the rate of adaptation automatically.

The CLE method is derived directly from the well known bi-observation forward-backward (FB) estimates of the CDHMM parameters [7]. For the incremental training procedure of interest it is necessary to assume that given two sets of observations, one exists in model form. Rearranging the equations to accommodate for this assumption we can arrive at the following equations for means ($\hat{\boldsymbol{\mu}}_{ih}$), mixture weights ($\hat{w}_{ih}$), and covariances ($\hat{\boldsymbol{\Sigma}}_{ih}$) of a CDHMM:

$$\hat{\boldsymbol{\mu}}_{ih} = \frac{\tilde{\boldsymbol{\mu}}_{ih}\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t)\mathbf{x}_t}{\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t)} \quad (1)$$

$$\hat{w}_{ih} = \frac{\tilde{w}_{ih}\tilde{\gamma}_i + \sum_t \gamma_{ih}(t)}{\tilde{\gamma}_i + \sum_t \gamma_i(t)} \quad (2)$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{ih} = & \frac{(\tilde{\boldsymbol{\Sigma}}_{ih} + \tilde{\boldsymbol{\mu}}_{ih}\tilde{\boldsymbol{\mu}}'_{ih})\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t)\mathbf{x}_t\mathbf{x}'_t}{\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t)} \\ & - \frac{(\tilde{\boldsymbol{\mu}}_{ih}\tilde{\gamma}_{ih})(\tilde{\boldsymbol{\mu}}_{ih}\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t)\mathbf{x}_t)'}{(\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t))^2} \\ & - \frac{(\sum_t \gamma_{ih}(t)\mathbf{x}_t)(\tilde{\boldsymbol{\mu}}_{ih}\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t)\mathbf{x}_t)'}{(\tilde{\gamma}_{ih} + \sum_t \gamma_{ih}(t))^2}\end{aligned} \quad (3)$$

where $\tilde{\boldsymbol{\mu}}_{ih}$, $\tilde{\boldsymbol{\Sigma}}_{ih}$, $\tilde{w}_{ih}$, $\tilde{\gamma}_{ih}(t)$ are the mean, covariance, mixture weight and posteriori probability parameters of the current model respectively for state i, mixture h and observation vector time index t; and $\mathbf{x}_t$ and $\gamma_{ih}(t)$ are the observation vector and posteriori probability values from the current observation set. The notation is also simplified by defining $\tilde{\gamma}_{ih} = \sum_t \tilde{\gamma}_{ih}(t)$. The above CLE equations can be extended readily to combine multiple models of the same structure with multiple observations if required.

In applying the CLE method for incremental adaptation, we require extra storage only for the $\sum_t \tilde{\gamma}_{ih}(t)$ parameters which are calculated in the FB algorithm and are speaker-dependent. Update of these parameters then follows a simple recursive assignment of the posteriori probability values which are carried over to the next stage of training.

4. Experiments

Experiments conducted consist of comparing FB, CLE and MAP training of the client speaker models using the incremental training procedure and isolated word text-dependent models. Two types of experiments are considered here. They are enrolment without a speech model where the client CDHMMs are created using the first available client training set, and enrolment with a speech model where the structure of the client speaker model is dictated by the structure of the speech model.

Figure 4. illustrates how each training approach is applied, using the available training sets. Normal FB training is memoryless, so two conditions can be applied to this approach.

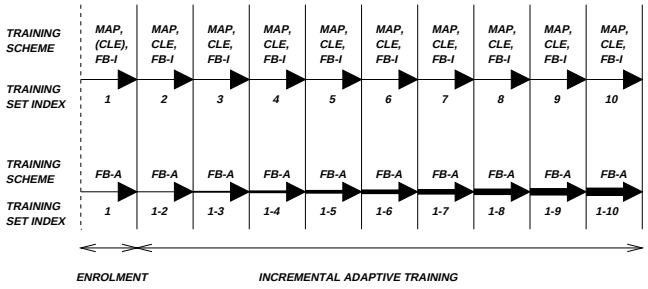

Figure 1: Illustration of how each training approach is applied using the 10 available training sets. The top diagram shows the training sets used in isolation, which is applicable to MAP, CLE and FB training (labelled FB-I). The lower diagram shows the training sets being accumulated, which is only applicable to FB training (labelled FB-A).

The first is a cumulative data store condition where training sets are carried over from one training sessions to the next, this is labelled FB-A. This is an unlikely scenario as it incurs the overhead of storing all the speech data but is used here as the target for the MAP and CLE. The second condition is when the training sets are used in isolation. For FB training this is likely to give the poorest result, since the model has client data from only a single utterance, which is labelled FB-I.

For the two memory retentive training approaches CLE and MAP, only the isolated data store condition applies. This is because the cumulative data store condition implies growing emphasis on the previous observation sets when applied recursively, a condition which is undesirable.

In all experiments the CDHMMs are of the form used in HTK [8]. This toolkit is used throughout the experiments with adjustments made to accommodate both MAP and CLE estimation.

16-state single-mixture Markov models with diagonal covariance Gaussians are used throughout. The topology for the word model is constrained to be left-to-right with self loops and no skips.

For verification the output score from the true speaker is normalised by the sum of the scores from all 20 client speakers. The equal error rates are then calculated using this normalised value [4].

4.1. Speech database and preprocessing

The verification experiment is performed using the BT Millar digit database (*one* to *nine* and *zero*), which has been collected in a quiet environment using a high quality microphone comprising 25 repetitions of each of the vocabulary items from each speaker. The sessions take place over a period of approximately three months with speakers encour-

aged to divide sessions evenly across this period. The speech is recorded at 20kHz using 16 bits (linear) per sample. In these experiments the data is bandpass filtered to telephone bandwidth and down-sampled to 8kHz prior to feature extraction. The feature is mel-scale 14th order cepstra over a Hamming window of 32ms, with 50% overlap.

This database is divided into training and testing sets. The first ten repetitions, i.e. the first two collection sessions, are reserved for training, with the remaining fifteen repetitions reserved for testing. This results in 3000 tests.

In addition to the 20 client speakers, the data from a separate 20 speaker set is used to train speaker-independent whole-word digit models. In this case, all 25 repetitions are used to create the speaker-independent isolated word speech models. These are in turn used as the starting point for each of the 20 client speakers in the verification experiment.

4.2. Enrolment without a speech model

In this experiment we compare the three training approaches, namely FB, CLE and MAP. Enrolment of the client speaker consists of uniform segmentation of the first available training set to estimate the values of the parameters, followed by an iterative Viterbi alignment of the training set to re-estimate the parameters. The last stage of this enrolment consists of FB, CLE or MAP training. Subsequent incremental training on the newly acquired training data set uses the three named training procedures on the current model. This experiment highlights the applicability of each training scheme to the task of incremental adaptation.

Only the means are updated in training with all variance elements set to unity. This is because on enrolment we do not have sufficient data for reliable estimates of the variance parameters [1].

For MAP estimation, the τ value is tested, and the results for $\tau = 2$ and 10 are shown in Figure 2. The plot illustrates that for a value of $\tau = 2$ the adaptation is fast, but when applied to the incremental framework, the estimation of the client model is less appropriate for discrimination between speakers, this is also true for smaller values of τ although not illustrated. A value of $\tau = 10$ causes a slower adaptation but is preferred because it is produces a more appropriate model given enough adaptation data. For values larger than $\tau = 10$, adaptation is slower.

Results from the ten training sets comparing FB, CLE and MAP ($\tau = 10$) training procedures used as described above are shown in Figure 3. As predicted, the FB-I training illustrates the poorest of the results. There is a general slow left to right trend, which is caused by the training data getting chronologically closer to the test data. This trend line highlights the lack of memory retention in typical FB training of a CDHMM. The lowest trend line highlights FB-A when the training sets are carried over, or accumulated. This, as

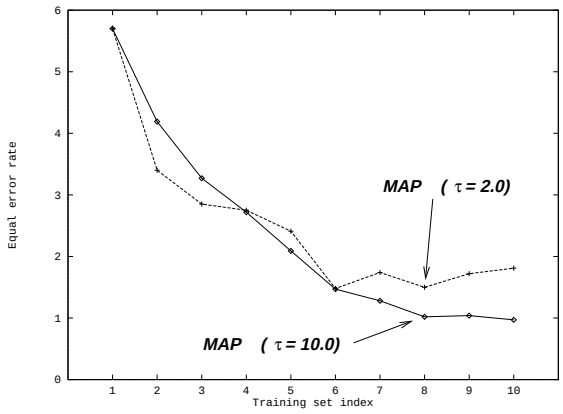

Figure 2: Equal error against training set index for MAP estimation of CDHMM parameters, using $\tau = 2$ and 10.

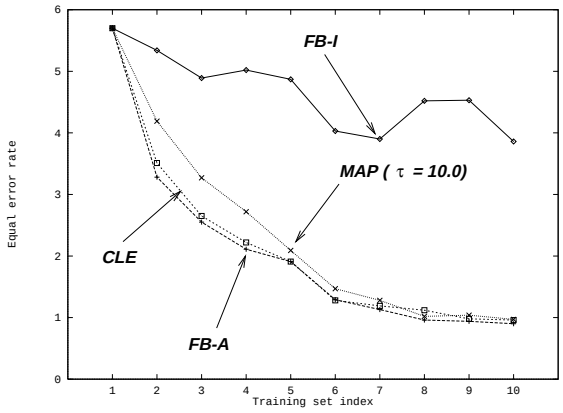

Figure 3: Equal error against the training set index for (a) FB estimation of parameters when the training sets are accumulated over time, (b) FB estimation of parameters when the training sets are used in isolation, (c) MAP ($\tau = 10$) and (d) CLE.

expected, gives the best result. The lowest line forms the target result we wish the CLE and MAP to approach.

The MAP trend line for $\tau = 10$ continually improves the model smoothly through the training sets to the target trend line, but the adaptation is a little slow. The CLE adaptation method follows the target line closely, and proves the usefulness of this approach by automatically adjusting the adaptation rate. The differences also highlight the gain that can be attained when speaker-dependent parameters are used (CLE), over speaker-independent parameters such as those used in the MAP framework.

4.3. Enrolment with a speech model

In this experiment, we extend the process of incremental adaptation by including speaker adaptation to create the

client speaker models at enrolment.

CLE cannot be used to adapt the speaker-independent speech model to a speaker-dependent client model for enrolment because the value of $\tilde{\gamma}_{ih}$ is now very much larger than the current posteriori probability estimates for a single observation set. This is due to the large amount of speech data used to train the model. Hence for enrolment we apply MAP with $\tau = 2$ to the seed model. Once the client model has been created, the speaker model can then be updated using CLE. Figure 4 illustrates the performance of this method and compares it against FB training using a speaker-independent speech model as the seed model, and a pure MAP approach using a τ value of 10. Again only the mean parameters are updated, and the structure of the speaker models are dictated by a single mixture 16 state diagonal covariance isolated word speech models.

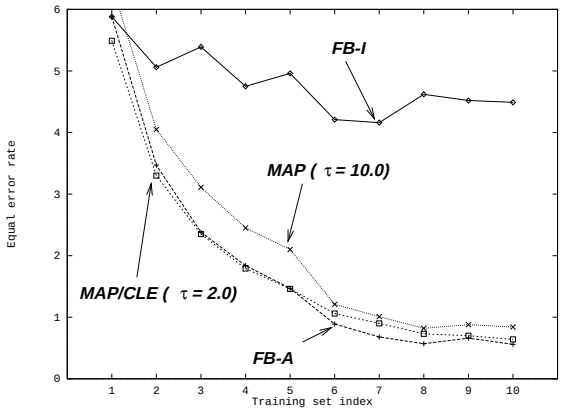

Figure 4: Equal error against the training set index for (a) FB estimation of parameters when then training sets are accumulated, (b) MAP/CLE ($\tau = 2.0$), and (c) MAP ($\tau = 10.0$).

The results for the first training set in this experiment are now different. This is because we use speaker adaptation to enrol the client speakers. FB-I training, is again the poorest of the results. The lowest trend line can be considered as a reasonable target performance. This is FB-A training when the training sets are accumulated. The pure MAP based approach follows the same trend as shown in the previous experiment except for this adaptation at enrolment. Using a combination of MAP for enrolment and CLE for continual update of the client speaker model, it is shown that this method follows the target line very closely.

5. Conclusions

In this paper we have discussed two methods for applying incremental adaptation to speaker models for the task of speaker recognition. One merit of CLE is storage of posteriori probability values of the model and carrying it over to the next stage of training. CLE has the advantage over MAP by using speaker-dependent parameters that are auto-setting, against empirically chosen speaker-independent parameters which might be dependent on the conditions imposed on the system. The use of speaker-dependent parameters is supported by the results in Figure 3 where CLE outperforms the chosen MAP approach consistently, and significantly when the amount of speaker-specific data is small. Although the CLE framework cannot be applied directly to the enrolment of a client speaker using a seed word model, the MAP/CLE combination proves powerful enough to attain discrimination between speakers. Further work is required to improve the CLE approach to accommodate enrolment with a speech model.

In both enrolment procedures, the CLE incremental adaptive approach can achieve an equal error rate of approximately 1% with six adaptations using a single digit test token.

6. Acknowledgment

The authors thank BT Labs for the use of the Millar database and continuing financial support for this work.

7. REFERENCES

1. Kin Yu, John Mason, and John Oglesby. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *IEE proc. vision, image and signal processing*, 142:313–318, 1995.
2. J. L. Gauvain and C. H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on speech and audio processing*, 2:291–298, 1994.
3. 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*, 39:806–814, 1991.
4. J. L. Gauvain, L. F. Lamel, and B. Prouts. Experiments with speaker verification over the telephone. In *Proc. Eurospeech-95*, volume 1, pages 651–654, 1995.
5. T. Matsui and S. Furui. A study of speaker adaptation based on minimum classification error training. In *Proc. Eurospeech-95*, volume 1, pages 81–84, 1995.
6. G. Zavaliagkos. Maximum a posteriori adaptation techniques for speech recognition. *Ph.D. thesis, Northeastern University*, 1995.
7. B. Juang. Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. *ATT Bell Laboratories Technical Journal*, pages 1235–1249, 1985.
8. S. J. Young and P. C. Woodland. *HTK: Hidden Markov model toolkit V1.4 User manual*. Cambridge University Engineering Department, Speech Group, 1992.