ON-LINE SPEAKER ADAPTATION FOR HMM BASED SPEECH RECOGNISERS

B. Watson

Department of Electrical and Computer Engineering
University of Queensland

ABSTRACT – An investigation of a gradient-descent based training technique was performed for the on-line adaptation of hidden Markov models to new speakers in a speech recognition system. It was found to be successful for supervised speaker adaptation, improving the recognition performance on a 46 word task (alphabet, digits and control words) from 88.0% to 93.2% after adaptation with nine repetitions of each word. Unsupervised adaptation on the same task was unsuccessful. However, for an easier 20 word vocabulary, unsupervised adaptation improved the recognition performance from 97.7% to 99.0%.

INTRODUCTION

The Baum-Welch re-estimation procedure for hidden Markov models (HMMs) directly maximises the likelihood of the training observations given the model. It is suited to the batch estimation of model parameters. In order to update models as new data becomes available, a smooth on-line learning algorithm is desirable. Baldi and Chauvin have proposed such an algorithm, and have experimented with its use for HMM training in a molecular biology problem (Baldi et al., 1993). Baldi and Chauvin's approach to HMM training is based on the use of a gradient descent algorithm. Suppose we wish to maximise some function $\mathcal{L}$, that is dependent on a model parameter x. We can do this by calculating the gradient of the objective function $\mathcal{L}$, and making a small change (Δx) to the value of the parameter based on this. The value of the parameter x_n after training step n is:

$$x^n = x^{n-1} + \Delta x^n$$

where Δx^n is calculated using the partial derivative of the objective function with respect to the parameter, $\frac{\partial \mathcal{L}}{\partial x}$, evaluated at value x^n. That is:

$$\Delta x^n = \mu \left.\frac{\partial \mathcal{L}}{\partial x}\right|_{x^{n-1}} + \eta \Delta x^{n-1}$$

μ is a learning rate parameter, which controls the size of the parameter changes between training steps. It will affect both the speed of convergence, and the stability of the model parameter estimates. A momentum parameter, η, can also be introduced into the update calculations, in order to allow the use of a larger learning rate parameter, while still maintaining consistent updates, as it can provide partial averaging over the training observations (Hertz et al., 1991).

Consider an N state discrete output HMM with M possible outputs. Let the state of the model at time t be denoted as s_t, and the output at time t be denoted as o_t. The model parameters are: initial state probabilities, $\pi_i = Pr(s_1 = i)$; transition probabilities, $a_{ij} = Pr(s_{t+1} = j | s_t = i)$; and discrete output probabilities, $b_{ik} = Pr(o_t = k | s_t = i)$. The parameters of the HMM will be referred to collectively as λ. That is, $\lambda = \{A, B, \pi\}$, where A, B, and π, are the sets of transition, output, and initial state probabilities, respectively. Because all of these parameters are probabilities, when summed over the appropriate range, they will sum to 1.

For the HMM, we wish to update the model corresponding to a particular utterance, O, so that it maximises the likelihood of the model producing that utterance, $L = Pr(O|\lambda)$. Gradient descent on the log-likelihood is numerically better conditioned than gradient descent on the likelihood (Levinson et al., 1983) so the log-likelihood was used as the objective function for model training.

For both the transition and output probabilities, a normalised-exponential representation is introduced, to ensure that as probabilities they will have values between 0 and 1, and sum over the appropriate ranges to 1. The transition and output probabilities are written in terms of new parameters, w_{ij} and v_{ij} respectively, which are the values before normalisation:

$$a_{ij} = \frac{e^{\varphi w_{ij}}}{\sum_k e^{\varphi w_{ik}}}$$
$$b_{ij} = \frac{e^{\varphi v_{ij}}}{\sum_k e^{\varphi v_{ik}}}$$

φ is a parameter of the normalisation function which, can be absorbed into the learning rate.

The derivative of the log-likelihood with respect to w_{ij} and v_{ij} can then be derived (see (Baldi and Chauvin, 1994)) so that updates can be calculated. The derivatives are:

$$\frac{\partial \log L}{\partial w_{ij}} = \varphi[\sum_t \gamma_{t,i,j} - \sum_t \gamma_{t,i} a_{ij}] \quad (1)$$
$$\frac{\partial \log L}{\partial v_{ij}} = \varphi[\sum_t \xi_{t,i,j} - \sum_t \gamma_{t,i} b_{ij}] \quad (2)$$

where $\gamma_{t,i,j} = Pr(s_t = i, s_{t+1} = j|O, \lambda)$, $\gamma_{t,i} = Pr(s_t = i|O, \lambda)$, and $\xi_{t,i,j} = Pr(s_t = i, o_t = j|O, \lambda)$. These values can be computed using the forward and backward algorithms.

SUPERVISED ADAPTATION

Although our main interest in the proposed gradient descent technique is for on-line unsupervised learning, we began by analysing its behaviour for supervised adaptation of speech recognition models in order to determine: what models constitute the best prototype models, which are to be used as the starting point for speaker adaptation; how the performance of the adaptation algorithm depends on the amount of adaptation data; and what values of the learning rate and momentum parameters give the best results.

Experiments were performed on the Texas Instruments 46 word speech database. This database contains recordings from sixteen speakers - eight male and eight female. Each speaker has uttered the letters of the alphabet, the digits from zero to nine, and ten "control" type words (enter, erase, go, help, no, rubout, repeat, stop, start, yes). The data from each speaker has been divided into a training set of nine or ten utterances, and a test set of sixteen utterances of each word.

The TI-46 database was selected for experimentation because it represents a small but challenging task. Accurate alphabet recognition is difficult because the vocabulary contains a number of highly confusable words such as the e-set (B, C, D, E, G, P, T, V, Z (for American English)).

The speech waveforms from the TI-46 database were processed in a standard manner for input to the HMM recogniser. The speech waveforms in the database were sampled at 12.5 kHz with 12-bit quantisation. These waveforms were segmented into frames of 512 samples (40 milliseconds) with a new frame starting every 128 samples (10 milliseconds). The samples were pre-emphasised with a filter whose transform function is $1 - 0.95z^{-1}$, and Hamming windowed. For each frame, twelfth order LPC cepstral coefficients were calculated, along with first and second order delta coefficients.

Left-to-right hidden Markov models with ten states were used to model each of the vocabulary words. Supervised adaptation of each of the models, for each speaker, was performed by using each training utterance to update the relevant model, using values computed according to equations (1) and (2). In figure 1, the recognition performance following supervised adaptation is presented as a function of the learning rate and the amount of adaptation data used. The results presented were obtained when the starting models were speaker independent, but had only been trained using the data from the other speakers of the same gender as the target speaker. Genderless speaker independent models were observed to give recognition performances approximately 1% lower than those obtained using gender specific models. However, in both cases the adaptation improved the recognition performance substantially.

Adapting from the gender based speaker independent models, with a learning rate of between 0.18 and 0.26, the recognition performance improved from 88.0% to 93.2% following adaptation using nine

samples of each word. By way of comparison, Baum-Welch training using the set of nine adaptation samples yielded a recognition rate of only 90.8%. The adapted recognition performance increased as more and more speaker specific data was made available from the target speaker, but quite small amounts of adaptation data were required to realise a significant improvement in performance. For example, after only two samples of each utterance, the performance had risen to 90.5%, a reduction in error rate of over 20%.

The value of the learning rate parameter, μ, is critical. If too small a learning rate is used, the information present in the adaptation samples is not fully utilised. When too great a learning rate is used, the updates become unreliable, and the knowledge contained in the speaker independent models is under-utilised. In practice, however, we found that the word models were reliably updated over a range of values for the learning rate parameter.

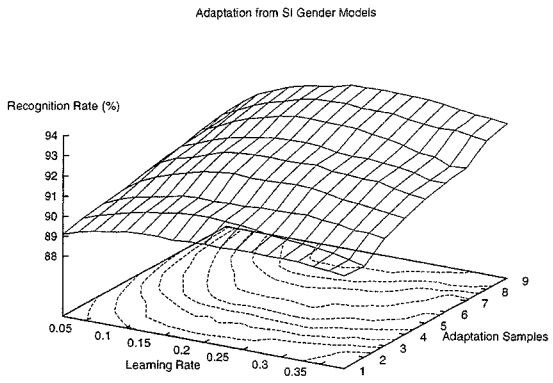

Figure 1: Recognition performance with adaptation starting from speaker independent gender specific models, as a function of the number adaptation samples and the learning rate.

It should be noted that in these (and subsequent) experiments a *single* update step was performed using each piece of adaptation data. A second update step with each utterance improved the final mean adapted recognition rate (to 93.4% for a learning rate of 0.22), at the cost of double the computation. Further update iterations were found not to be beneficial, as the multiple updates placed too much emphasis on the new adaptation utterance at the expense of losing information already contained in the model.

The observed best results were achieved when using a small learning rate. For gradient descent algorithms, the use of a momentum term (Hertz et al., 1991) frequently allows the use of a larger learning rate parameter, while still maintaining consistent updates, as it provides partial averaging over the training observations. An investigation of the effect of momentum on this task was performed, but it was found that it did not yield an improvement in the best adapted recognition performance.

It should be noted that it is also common practice to modify the learning rate (and possibly the momentum) over time for gradient descent algorithms. It is a normal procedure in gradient descent learning to start with a large learning rate, and gradually decrease it to a small value, as the training set is used up. Such techniques were not pursued here, because the amount of adaptation data available was quite minimal, and it was clear that the models were still improving significantly with each new piece of adaptation data.

The reliability of an adaptation strategy is an important consideration in determining its usefulness. Although an adaptation technique may improve the average recognition rate, measured across a set

of speakers, its utility will be greatly reduced if, for certain individuals, the recognition performance actually degrades. In this investigation, however, it was found that the on-line supervised adaptation substantially improved the correct recognition rate for all speakers. The reductions in error rates for individual speakers ranged between 27.8% and 58.8% after adaptation with nine samples of each word, using a learning rate of 0.22.

MINIMISING ADAPTATION COMPUTATION

The modelling power of the HMM transition probabilities can be limited, with the state occupancy distributions reducing to simple exponentials (Rabiner, 1989). This has motivated some researchers to propose extensions to HMMs, including explicit duration modelling (Russell and Cook, 1985)(Levinson, 1986). Observing that the exact values of the HMM transition probabilities can be relatively unimportant in determining the effectiveness of the models for recognition, researchers at AT&T, chose not to use transition probabilities in their system, and instead assumed that forward and self transitions in the left-to-right HMMs which they used were equally likely (Lee et al., 1992). If the values of the transition probabilities are relatively unimportant, then adapting them introduces unnecessary processing. The results of three adaptation experiments in which (1) all model parameters were adapted, (2) only the transition probabilities were adapted, and (3) only the output probabilities were adapted, are presented in figure 2. The adaptation of the transition probabilities does indeed appear to be superfluous, with their values having a very minor impact on recognition performance.

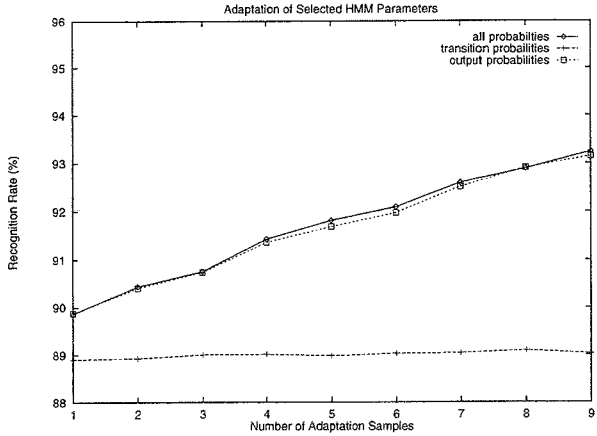

Figure 2: Recognition performance following adaptation of selected HMM parameters.

The on-line learning algorithm, as implemented for the preceding studies, required the computation of both the forward and backward probabilities for each utterance, in order to calculate the changes to each of the model parameters. It is possible to produce either the forward or backward probabilities as a by-product of a recognition stage, but in practice, recognition for HMM systems is usually performed using algorithms which require less computation, such as the Viterbi algorithm. The Viterbi algorithm finds the most likely path through a model (for an observation sequence), and its likelihood, while the forward and backward algorithms find the total likelihood that the observation sequence was produced by the model. In practice, the recognition results obtained using the Viterbi algorithm are often sufficiently close to the results obtained using the forward algorithm, that the Viterbi algorithm is preferred because of its lower computation requirements.

In a similar manner the Viterbi algorithm is often used as the basis for HMM parameter re-estimation, instead of the Baum-Welch expectation maximisation based algorithm, by using the segmental k-means algorithm. The model parameters can be updated to make the most likely path for an observation more likely. Counts of the frequencies of occupation of a state in the most likely path replace the expected

number of times the state was occupied in the update formulae, for example.

The gradient descent training algorithm can also be approximated using Viterbi counts (Baldi and Chauvin, 1994). A comparison of the gradient descent adaptation algorithm, using the original formulation (based on the forward and backward probabilities), and using the Viterbi counts, using a learning rate of $\mu = 0.22$, revealed no significant difference in recognition performance following adaptation. However, the adaptation using the Viterbi approximation required significantly less computation - for the experiments described here, it was 25% faster.

UNSUPERVISED ADAPTATION

The preceding studies were all concerned with analysing the performance of the gradient descent algorithm for supervised adaptation of the HMMs. However, the algorithm has properties which suggest its suitability for application to unsupervised adaptation of a speaker's models as they speak: it can update models using a single utterance immediately after the utterance has been obtained by the system; and it does not involve prohibitively large amounts of processing.

In contrast to the supervised adaptation approach discussed previously, implementation of unsupervised adaptation using the gradient descent training algorithm requires that the input speech first be recognised, and the recognised samples then be used for adaptation. It is suited to applications where a speaker immediately commences using a speech recognition system (using the speaker independent models), and as the speaker talks, the models are gradually adapted to be more suitable for recognising the speaker's speech.

As with the supervised adaptation study, the recognition performance for the speaker adapted models was examined as a function of the learning rate parameter, and the amount of speaker specific adaptation data available from each speaker. No matter how small the learning rate, the result of unsupervised adaptation in our study was a decline, rather than an improvement, in recognition performance. We expected that the recognition performance might decline if the recognition error rate was high - leading to numerous erroneous adaptation steps. However, in this case, it was not the overall error rate that caused problems, but the unfortunate distribution of errors. Most of the errors were confined to a few words. In particular, e-set words contributed many of the errors. For words such as /c/, the probability of recognition error was initially over 50%, and hence we must expect many bad adaptation steps for these words. Consequently the recognition performance after several adaptation steps for these words was actually worse than prior to adaptation. For the e-set words the mean recognition error rate was initially 15.1%. This increased substantially to 20.8% following adaptation (with $\mu = 0.20$), while for the other words in the vocabulary the initial recognition error rate of 11.3% increased only slightly to 12.0% following unsupervised adaptation.

The alphabet (and particularly the e-set) is hardly a typical vocabulary. The possibility of confusion is extremely high. This makes it difficult to perform reliable unsupervised adaptation. In order to show that unsupervised adaptation can be successful, an easier task was considered - recognising just the digits and control words present in the TI-46 word database. The effect of adaptation for a range of learning rates and increasing amounts of adaptation data (from one to nine repetitions of each word) is presented in figure 3. The recognition rate was high prior to adaptation - 97.7%. Following adaptation with nine samples of each word the best recognition rate was 99.0% (for a learning rate of 0.36 or 0.40).

While the best recognition rates were obtained with quite large learning rates, the improvement in performance following adaptation for large learning rates was not as uniform as that obtained with smaller rates. With smaller learning rates (less than 0.20), each set of adaptations with a single repetition of each word in the vocabulary yielded an improvement in recognition performance. With larger adaptation rates the recognition performance following adaptation steps actually declined occasionally when a number of wrongly recognised words resulted in a bad adaptation.

CONCLUSIONS

Gradient descent algorithms can be used to incrementally adapt hidden Markov models. Using supervised adaptation on a system for recognising a 46 word alphabet, digits and control words vocabulary, recognition performance was improved from 88.0% to 93.2% after adaptation using nine samples of each vocabulary word. It was found that the computation required to adapt the models could be min-

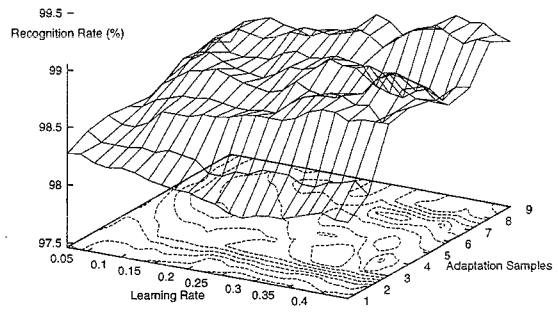

Figure 3: Recognition performance following unsupervised model adaptation for a twenty word vocabulary consisting of the digits and ten control words.

imised by not adapting the transition probabilities, and by using Viterbi counts instead of forward and backward probabilities, without significantly impacting on the recognition improvements obtained.

Unsupervised adaptation was also found to be possible when the recognition rate for all words in the vocabulary was initially excellent, but was not successful for a vocabulary containing a number of poorly recognised words.

REFERENCES

Baldi, P. and Chauvin, Y. (1994). Smooth on-line learning algorithms for hidden Markov models. *Neural Computation*, 6(2):307–318.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. (1993). Hidden Markov models in molecular biology: new algorithms and applications. In *Advances In Neural Information Processing Systems 5*, pages 747–754. Morgan Kaufmann Publishers, USA.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley Publishing Company, USA.

Lee, C. H., Gauvain, J. L., Pieraccini, R., and Rabiner, L. R. (1992). Large vocabulary speech recognition using subword units. In *Proceedings of the Fourth Australian International Conference on Speech Science and Technology*, pages 342–353.

Levinson, S. E. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1(1):29–45.

Levinson, S. E., Rabiner, L. R., and Sondhi, M. M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286. Published as Proceedings of the IEEE, volume 77, number 2.

Russell, M. J. and Cook, A. E. (1985). Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. In *International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 5–8.