RECURRENT INPUT TRANSFORMATIONS FOR HIDDEN MARKOV MODELS

V. Valtchev *S.Kapadia* *S.J.Young*

Cambridge University Engineering Department
Trumpington Street, Cambridge, England.

ABSTRACT

This paper presents a new architecture which integrates recurrent input transformations (RIT) and continuous density HMMs. The basic HMM structure is extended to accommodate recurrent neural networks which transform the input observations before they enter the Gaussian output distributions associated with the states of the HMM. During training the parameters of both HMM and RIT are simultaneously optimised according to the Maximum Mutual Information (MMI) criterion. Results are presented for the E-set recognition task which demonstrate the ability of recurrent input transformations to exploit longer term correlations in the speech signal and to give improved discrimination.

1. INTRODUCTION

Hidden Markov Models (HMM's) constitute the most successful and widely used approach to modelling acoustic events in speech recognition. However, the assumptions made by a first order HMM as to the underlying nature of speech are poor. A major flaw of HMM's when applied to automatic speech recognition (ASR) tasks is the output observation independence assumption

$$P(Y_t = y_t | S_1^T = s_1^T, Y_1^T = y_1^T) = P(Y_t = y_t | S_t = s_t) \quad (1)$$

i.e. the probability that an acoustic observation y will occur at time t depends only on the output distribution associated with the present state s_t of the Markov chain but not on the other observations. The problem with systems of this kind is that slowly varying articulatory processes introduce significantly larger amounts of long-term correlation which cannot be modelled adequately by the state transition probabilities alone. Modelling of acoustic signal dynamics can be improved by adding new dimensions to the observation vectors. Improved performance is normally achieved by introducing first order derivative information as as part of the observation vector. Recently reported results [8] indicate that using second order derivatives can also reduce recognition error rate. In most cases, the greater the number of parameters employed by the model, the greater will be the potential for modelling complex acoustic events. However, more parameters will require more training data and the associated computational requirements will be high. The obvious solution to this problem is to transform the output of the preprocessor so that the dimensionality is reduced while retaining as much information as possible. It is clearly desirable to rank the components of the observation vector according to their relative information content and perform selective pruning using some performance related criterion. Various input transformations of this kind have been investigated [7], [3]. However, unlike the work described here, once derived, these transformations remain unchanged throughout the training process. Furthermore, the criteria used to derive them are not directly related to the objective function used to optimise the HMM parameters.

Reconsidering the independence assumption problem we realise that using higher order derivative information with fixed input transformations to reduce dimensionality is not a complete solution to the problem. Ideally we would like to enhance the transformation by incorporating a recurrent mechanism which allows the present output y_t to be a function of $(y_0, y_1, ...y_{t-1})$. In this paper we introduce the concept of recurrent input transformations (RIT) in the HMM framework. The basic HMM structure is extended to accommodate recurrent neural networks which transform the input observations before they enter the state output of the HMM. During training the parameters of both HMM and RIT are simultaneously optimised according to the Maximum Mutual Information (MMI) criterion.

2. THE HMM/RIT ARCHITECTURE

In a conventional HMM, the probability of an observation y given a Gaussian output distribution p with parameters (C, m) is

$$p(y) = \frac{1}{(2\pi)^{d/2}|C|^{1/2}} e^{-1/2(y-m)^T C^{-1}(y-m)} \quad (2)$$

where m is the mean vector of the distribution, C is the covariance matrix, and d is the dimensionality. An input transformation is defined which transforms the output of the preprocessor x into the observation y seen by the Gaussian such that

$$y = f_N(U_N^T f_{N-1}(U_{N-1}^T \dots f_1(U_1^T x)\dots)) \quad (3)$$

where $U_1 \dots U_N$ are matrices and $f_1 \dots f_N$ are differentiable functions. The above transformation can be viewed as an N-layer neural network where the units in layer i are set to compute $f(.)$ and the weights of layer i are the values in U_i. A recurrent mechanism can be incorporated into 3 by allowing the argument of f_i to carry information about past states of the transform. In the following sections we consider single layer recurrent transformations with state units whose structure is shown in figure 1. The output of the transformation can be expressed as

$$\begin{aligned} y_t &= f_O(U_{IO}^T x_t + U_{RO}^T r_{t-1}) \quad (4) \\ r_t &= f_R(U_{IR}^T x_t + U_{RR}^T r_{t-1}) \quad (5) \end{aligned}$$

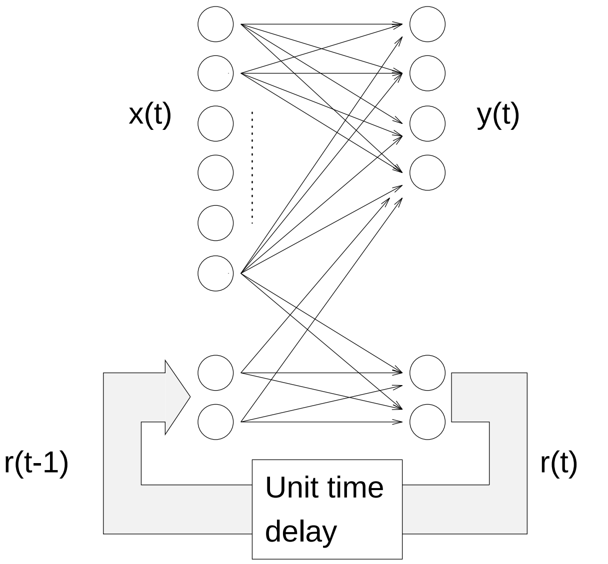

Figure 1. General structure of recurrent input transformation. x is the output of the preprocessor y is the output of the transform, r is the output of the state units and t is the time index. (note: not all connections are drawn for clarity of presentation)

where r is the output of the state units, t is the time index, U_{IO}, U_{IR}, U_{RR}, U_{RO} are the matrices describing the *input-to-output*, *input-to-recurrent*, *recurrent-to-recurrent* and *recurrent-to-output* connections respectively. The functions f_O and f_R are chosen to compute symmetric sigmoids with output ranging in the interval $[-1.0, 1.0]$.

Part of our objective in this work is to investigate the benefits of tying input transformations to specific states or set of states. The system we have implemented therefore allows input transformations to be *state based* (no sharing), *HMM based* (single transformation shared by all states of a model) or *global* (single transformation shared by all states of all models). In fact, the system has been developed as an extension of the HTK package described in [10], [9] and allows arbitrary tying of input transformations in common with all of the other HMM parameters.

3. INITIALISATION

Several different approaches can be taken to initialising the RIT's. Since we attempt to provide a better acoustic model with improved discrimination it seems plausible to initialise the RIT's to perform discriminative analysis on the feature vectors. Following Woodland [7], we initialise the transformations to increase the discrimination between in-class and potentially confusable out-of-class vectors. The in-class data is characterised by a mean vector m_i and covariance matrix C_i. These parameters are directly available from the Gaussian output distributions associated with the states of the HMM (eq. 2). The out-of-class (confusion) data can be similarly expressed in terms of a mean vector m_o and covariance matrix C_o. The covariance of the confusion data centred on the in-class mean, Q_o can be calculated by

$$Q_o = C_o + (m_o - m_i)(m_o - m_i)^T \qquad (6)$$

The required transformation matrix S is the one such that

$$S^T C_i S = I \text{ and } S^T Q_o S = D$$

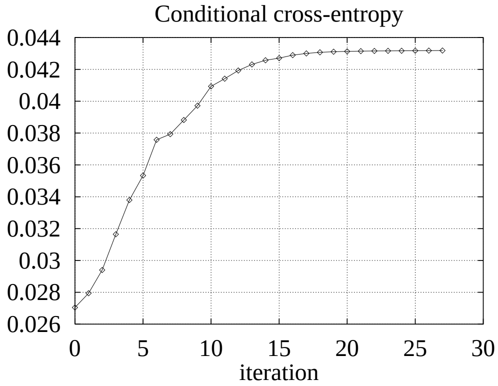

Figure 2. Plot of the MMI objective function for 15 state HMM's with 24x16x10 model-based input transformations.

where D is diagonal. Solving the general eigen problem, S can be expressed as

$$S = R_1 \Lambda^{-1/2} R_2 \qquad (7)$$

where R_1 is the matrix of eigen vectors of C_i, Λ is the diagonal matrix of corresponding eigen values and R_2 is the matrix of eigen vectors of $(R_1\Lambda^{-1/2})^T Q_o (R_1\Lambda^{-1/2})$. Various possible ways of gathering confusion data have been discussed in [7], however, in the work reported here we assume that Q_o is the average covariance over all classes. Improved discrimination in a subspace of the original observation space can be achieved by discarding dimensions in which the transformed variance of the confusable data has a smaller variance than the in-class data. In the following stage, we use S to initialise a single layer recurrent input transformation with parameters U_{IO}, U_{IR}, U_{RR} and U_{RO}. Following Bengio [2], the *input-to-output* connections can be initialised with $U_{IO} = \epsilon S$ where ϵ is a small positive number. Consequently, the total input to the output units of the network will be small and the sigmoid functions will operate within a linear range. The *recurrent-to-recurrent* and *recurrent-to-output* connections are set to small random numbers and all biases and *input-to-recurrent* connections are set to zero. The advantages of such deterministic initialisation are that the parameter estimation process does not depend entirely on random initial conditions and the recognition performance of the set of HMMs used to derive the transformations is partially or fully preserved.

4. TRAINING

We consider a speech recognition task where each class is represented by a single HMM. The set of HMM's with full covariance output distributions is trained using ML and the BW algorithm. Input transformations are initialised from the corresponding covariance matrices and shared covariance matrices will imply subsequent sharing of input transformations.

In the MMI approach the parameters of the model are reestimated by maximising

$$I_\lambda = \sum_n \log p_\lambda(y(n)|t(n)) - \log p_\lambda(y(n)|r) \qquad (8)$$

where $y(n)$ is the sequence of observations, $t(n)$ is the correct transcription of $y(n)$ and r represents the recognition-time HMM. Traditionally MMI optimisation of HMM parameters is carried out using some form of steepest descent. The partial derivative of the cost function is calculated with respect to each parameter in the system and using this information gradient descent is performed in the parameter space. Unfortunately, gradient descent type optimisation is extremely slow and it scales up poorly as tasks become larger and more complex.

Another approach to improving the speed of convergence is to make explicit use of high order derivatives. Let $\mathbf{a}(t-1)$ be the vector containing the present values of all parameters in a system. Given higher order derivative information, a new parameter vector $\mathbf{a}(t)$ can be computed by

$$\mathbf{a}(t) = \mathbf{a}(t-1) - \eta \mathbf{H}^{-1}\mathbf{g}(t) \tag{9}$$

where $\mathbf{a}(t-1)$ is the old parameter vector, $\mathbf{g}(t)$ is the gradient of the objective function with respect to the parameter vector and $\mathbf{H}$ is the Hessian matrix of second derivatives. In order to reduce the computational load due to the calculation, inversion and storage of the Hessian matrix most implementations of this method use some approximation to the Hessian. In the work presented here we adopt the assumption that all parameters are independent. We further simplify the computation by using a difference approximation to the second derivatives rather than exact values.

$$\mathbf{H} = [h_{ii}] \tag{10}$$

$$h_{ii} = \frac{\partial^2 I_\lambda}{\partial a_i^2} \approx \frac{\frac{\partial I_\lambda}{\partial a_i}(t) - \frac{\partial I_\lambda}{\partial a_i}(t-1)}{\Delta a_i(t-1)} \tag{11}$$

Using equations 11 and 9 gives

$$\Delta a_i(t) = -\eta \frac{1}{h_{ii}} g_i(t) \tag{12}$$

$$= \eta \frac{g_i(t)}{g_i(t-1) - g_i(t)} \Delta a_i(t-1) \tag{13}$$

For η set to 1.0 expression 13 transforms into the update strategy of Fahlman's QuickProp [4]. Subsequently, the special cases arising in 13 with regard to limiting the changes in parameter value are handled in a similar fashion to the method used by Fahlman in the original paper. No parameter change is allowed to be greater in magnitude than μ times the previous update of that parameter. If the change computed by the update formula is too large or in the opposite direction to the current gradient, we instead use μ times the previous change as the current change. At the start, one steepest descent iteration is used to bootstrap the process.

5. DATABASE AND GENERAL EXPERIMENTAL SETUP

The task chosen to evaluate the performance of the two training techniques was the speaker independent (SI) recognition of the members of the British English E-set ("B", "C", "D", "E", "G", "P", "T" & "V"). The E-set recognition is considered to be a particularly difficult task due to the high level of confusability between the different classes in the set. The data used for the experiments was collected and distributed by British Telecom Laboratories (BTL) and forms a subset of their spoken alphabet database. The same database has also been explored by Woodland [7] and McCulloch [5] which allows for a more realistic comparison of

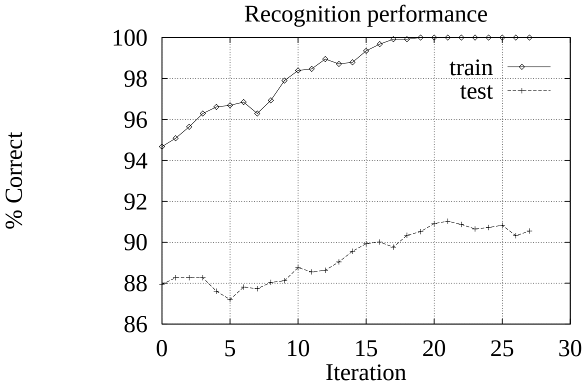

Figure 3. Train/Test set performance for 15 state HMM's with 24x16x10 model-based input transformations globally optimised using MMI.

the different training techniques employed. The experimental conditions are similar to those employed by Woodland in [7]. Each member of the E-set is represented by three utterances from each of the 104 different speakers (54 males, 50 females). The speakers are split into two halves to form a training set of 1239 utterances and a test set of 1219 utterances. The acoustic preprocessor uses the output of a 27 channel filterbank to produce 12 Mel Frequency cepstral coefficients (MFCCs) and their first order differentials (the twelve coefficients include the zeroeth coefficient which is the average value of the log power spectrum). The data was then further transformed using the principal components decomposition of the common covariance matrix of the training set. This resulted in a data set with common covariance matrix equal to the identity matrix. This transformation step further simplified the calculation of S (eq. 7 by setting Q_o (eq. 6) to the identity matrix.

6. RESULTS & DISCUSSION

Some preliminary experiments were carried out in order to establish the usefulness of the proposed architecture. In the first set of experiments, each class was modelled by a 3 state left to right HMM with no skips. The distributions of the final state were tied across all models to provide modelling of the vowel part of each utterance. By restricting the HMM's to just 3 states, the baseline performance is considerably reduced from the best achievable (for example with 15 state HMM's we obtained 93.7% on the same recognition task). However, the primary objective of our research is to investigate the ability of a globally optimised HMM/RIT configuration to exploit the available parameters. Subsequently, we wish to determine the advantages of the HMM/RIT architecture compared to the conventional Maximum Likelihood (ML) trained HMM systems.

The first row of results in table 1 established the baseline performance for Maximum Likelihood and Maximum Mutual Information trained models. A global input transformation was then introduced with output dimensionality reduced three times. The transformation was initialised to perform discriminative analysis and after the initialisation stage, the HMM parameters were reestimated via ML. The low performance figure obtained at this stage of the experiment (2nd line in table 1) indicates the loss of discrimination

Transformation	*ML*		*MMI*	
	train %	*test %*	*train %*	*test %*
none	73.61	72.44	94.51	76.54
24x8	63.84	59.23	85.47	75.47
24x8x20	63.84	59.23	89.34	80.75

Table 1. E-Set results for various input transformations using 3 state HMM's and single global transformation. The type of the transformation used is encoded as $IxOxR$ where I is the number of input units, O is the number of output units and R is the number of recurrent (state) units

Transformation	*ML*		*MMI*	
	train %	*test %*	*train %*	*test %*
24x24	95.00	90.65	100.00	91.88
24x16	94.35	87.94	100.00	88.27
24x8	91.69	81.46	98.55	82.12
24x16x10	94.35	87.94	100.00	91.03
24x8x12	91.69	81.46	100.00	88.36

Table 2. E-Set results for various input transformations using 15 state HMM's and model based transformations. The type of the transformation used is encoded as $IxOxR$ where I is the number of input units, O is the number of output units and R is the number of recurrent (state) units

information introduced with the vastly reduced dimensionality. In the next stage, the HMMs and the associated input transformation were simultaneously optimised according to the MMI criterion. The recognition results after this stage show that the global optimisation enables the MMI training to succeed in restoring the discrimination abilities of the models. The third line of table 1 shows the results obtained when the above experiment was repeated except that 20 recurrent units were added to the input transformation. In this case, the combined RIT and HMM, simultaneously trained using MMI show a significant improvement in performance. This confirms that there is extra information in the long term correlation which the model is able to exploit.

Table 2 gives results of using model-based transformations with 15 state HMM's. In this case, the final 9 states of each model were tied and shared a single input transformation, the remaining 6 states of each model shared a separate model-based transformation. Figure 2 shows the evolution of the MMI objective function for the case of model-based recurrent input transformations with 16 outputs and 10 recurrent units. The graph clearly demonstrates the fast convergence properties of QuickProp. Figure 3 plots the recognition performance achieved after each iteration of MMI training. As can be seen from table 2, the performance on the training data in four out of the five cases tested was 100.00% and this suggests that the system is undertrained. As a result the absolute performance level achieved is below that of our best full-covariance MMI-trained system. Nevertheless, the positive effect of adding recurrent state units is still clearly seen.

7. CONCLUSIONS

A system has been described which combines recurrent input transformations with conventional continuous density HMM's. Preliminary results have been presented which demonstrate that global optimisation allows the dimensionality of the feature vectors to be reduced without loss of performance. Furthermore, when recurrent states are introduced, performance increases significantly suggesting that there is further information to be extracted from the local context.

The use of recurrent input transformations on our best models has not yet achieved an improvement when compared to an MMI trained full covariance HMM's. However, the BTL E-set database is very small and we expect more substantial gains will be possible when more data is available.

ACKNOWLEDGEMENTS

V. Valtchev is a Benefactors' student at St. John's College, Cambridge. S. Kapadia is supported by a CASE Research Studentship.

REFERENCES

[1] A.J.Robinson. *Dynamic Error Propagation Networks.* PhD thesis, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2, June 1989.

[2] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe. Global Optimisation of a Neural Network - Hidden Markov Model Hybrid. Technical Report TR-SOCS-90.22, McGill University School of Computer Science, 3480 University street, H3A2A7, Montreal, Qc., Canada, December 1990.

[3] Peter F. Brown. The Acoustic-Modeling Problem in Automatic Speech Recognition. Technical report, IBM Thomas J. Watson Research Center, August 1987.

[4] Scott E. Fahlman. An Empirical Study of Learning Speed in Back-Propagation Networks. Technical report, CMU, September 1988.

[5] N. A. McCulloch. *Neural Network Approaches to Speech Recognition & Synthesis.* PhD thesis, Department of Communication and Neuroscience, Keele University, June 1990.

[6] S.Kapadia, V.Valtchev, and S.J.Young. Maximum Likelihood and Maximum Mutual Information Training of Continuous Density Hidden Markov Models - Experiments on the E-Set. In *Proc. 1992 Autumn Conference Speech and Hearing, Windermere, England.* IOA, November 1992.

[7] P.C. Woodland and D.R. Cole. Optimising Hidden Markov Models using Discriminative Output Distributions. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.* IEEE, April 1991.

[8] P.C. Woodland and S.J. Young. Benchmark DARPA RM results with the HTK portable HMM toolkit. To appear in *Proc. DARPA Continuous Speech Recognition Workshop*, September 1992.

[9] S. J. Young. The General Use of Tying in Phoneme-Based HMM Speech Recognisers. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.* IEEE, March 1992.

[10] S.J. Young. *HTK: Hidden Markov Model Toolkit V1.3 - Reference Manual.* Cambridge University Engineering Department, January 1992.

RECURRENT INPUT TRANSFORMATIONS FOR HIDDEN MARKOV MODELS

V. Valtchev , S.Kapadia and S.J. Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, England.

This paper presents a new architecture which integrates recurrent input transformations (RIT) and continuous density HMMs. The basic HMM structure is extended to accommodate recurrent neural networks which transform the input observations before they enter the Gaussian output distributions associated with the states of the HMM. During training the parameters of both HMM and RIT are simultaneously optimised according to the Maximum Mutual Information (MMI) criterion. Results are presented for the E-set recognition task which demonstrate the ability of recurrent input transformations to exploit longer term correlations in the speech signal and to give improved discrimination.