CONNECTIONIST PROBABILITY ESTIMATION IN THE DECIPHER SPEECH RECOGNITION SYSTEM

Steve Renals, Nelson Morgan*, Michael Cohen† and Horacio Franco†*

*International Computer Science Institute, 1947 Center St., Berkeley CA 94704, USA
†SRI International, Menlo Park CA 94025, USA

ABSTRACT

Previously, we have demonstrated that feed-forward networks may be used to estimate local output probabilities in hidden Markov model (HMM) speech recognition systems. Here these connectionist techniques are integrated into the DECIPHER system, with experiments being performed using the speaker independent DARPA RM database. Our results indicate that:

- connectionist probability estimation can improve performance of a context independent maximum likelihood trained HMM system,
- performance of the connectionist system is close to what can be achieved using (context dependent) HMM systems of much higher complexity, and
- mixing connectionist and maximum likelihood estimates can improve the performance of a state-of-the-art context dependent HMM system.

1 INTRODUCTION

Previous investigations, both theoretical and experimental, have indicated that feed-forward networks (typically, multilayer perceptrons, MLPs) may be used to estimate local HMM output probabilities [1, 6]. Our previous published results have generally concentrated on speaker-dependent databases using an unsophisticated recognition system. In this paper, we extend our experiments to the speaker independent DARPA Resource Management (RM) task, incorporating our connectionist methods into SRI's DECIPHER [3], a state of the art continuous speech recognition system.

There are several reasons why probability estimation using MLPs is an attractive approach:

- MLPs are well matched to discriminative objective functions.
- Although an MLP is a parametric model, a large network defines an extremely flexible set of functions. Thus only weak assumptions are made about the input statistics. As a result of this they can combine multiple sources of evidence. For example a single MLP may be trained using input data that mixes samples drawn from several distributions, discrete or continuous.
- Maximum likelihood estimation of HMM parameters requires the assumption of conditional independence of outputs. MLPs can model correlations across an input window of adjacent frames.
- Since the recognition time computations are extremely regular, it is possible to have a simple, efficient implementation in parallel hardware.

2 TRAINING AND RECOGNITION ISSUES

In connectionist training, the posterior probability of an output class (HMM state)[1] given the acoustic data, $p(q_j|\mathbf{x})$ is estimated; (scaled) likelihoods, needed for HMM recognition may be obtained from these posteriors via Bayes' rule [7]. This is in contrast to the maximum likelihood training usually employed, where the likelihood of the data given the class $p(\mathbf{x}|q_j)$ is estimated directly, by fitting parameters to a PDF for each class. Maximum likelihood is the optimal training criterion if the true model is known to be in the space of models under investigation. This is not the case in continuous speech recognition. Thus, it makes sense to use discriminative methods which lower the posterior probability of incorrect classes in addition to increasing the posterior probability of the correct class.

We have used discriminatively trained feed-forward networks—MLPs—trained according to a relative entropy criterion (the Kullback-Liebler divergence) to perform these estimations. These networks were used in a '1-from-N' classification mode: each output unit corresponded to a particular class, with the target output vector being binary with just one unit on. If we constrain the output units to sum to one (e.g., by using some normalisation in the output units' transfer function), then we may regard the output units as representing a probability distribution. Note that the common sigmoid transfer function is not normalised. However it may be shown that at the minimum of a least squares or relative entropy objective function, the output units of a 1-from-N network will sum to 1. In practice, we find no significant difference between using a sigmoid transfer function or a normalised exponential ('softmax') transfer function on the output units of a MLP. More stringently, it has been proved that networks trained as classifiers (i.e. a 'hard' class labelling for each frame) output posterior probability estimates at the minimum of a relative entropy or least squares objective function [1, 4].

[1]For convenience, we shall consider 'class' and 'HMM state' to be synonymous.

'Failure is an opportunity to learn.' In 1988 N. Morgan and H. Bourlard [unsurprisingly unpublished] first used these methods in the DECIPHER system. On the speaker independent Resource Management task, without a grammar, a word accuracy of -30% was recorded. Now, using essentially the same approach, we have improved our recognition accuracy to about 70% on the same task. What changes were necessary to make the connectionist approach effective?

1. (Scaled) likelihoods must be used in the Viterbi search, not posteriors. These may be most simply obtained by dividing each network output by the relative frequency of that class. Although the equations used in the Viterbi search hold for both posteriors and likelihoods, posteriors (which incorporate the prior probabilities estimated from the data) should not be used: when a language model and phone-structured lexicon are defined (i.e. the overall HMM topology), the priors for each class are implicitly set. Thus we must factor out the data estimates of these priors [7].

2. To train large networks efficiently a stochastic gradient descent procedure should be adopted (i.e. 'per pattern' or 'online' weight update). This is the case because the training set is large and redundant. More sophisticated methods, such as conjugate gradient or quasi-Newton methods, rely on an exact computation of the gradient (i.e. batch weight update). Additionally methods that maintain a (diagonal) approximation to the Hessian are extremely expensive in terms of both computation and memory for large networks.

3. Cross-validation training is essential for good generalisation and preventing over-training, especially when using large networks. In our training schedule we cross-validate by withholding a certain proportion of the training data (typically 10–20%) and using this to validate the training after each epoch. When the classification performance on the validation set first fails to improve by a certain amount (typically 0.5%) the gradient descent step-size is reduced, typically by a factor of 2. This time-dependent reduction in stochastic gradient descent step-size (gain) may be understood in terms of the constraints on the gain sequence given by stochastic approximation theory [8][2]. After each succeeding epoch the step size is further reduced, until once again there is no improvement on the validation set. Training is then halted.

4. Input representation is important. In particular dynamic features (obtained via linear regression estimate of the temporal derivative) should be used in addition to static ones, and a multi-frame input (typically we use ± 4 frames of context), offers an improvement over single frame input [6].

5. The word transition penalty used in the Viterbi search should be increased in the case of multi-frame input. This empirical result may be explained in terms of a scaling relationship between the likelihood of a single frame and the joint likelihood of several frames, given the class.

[2]The conditions given by stochastic approximation theory ($\sum_n \alpha_n = \infty$ and $\sum_n \alpha_n^2 < \infty$) are not, in fact, met by our gain sequence, $\alpha_n \propto 1/2^n$, since $\sum_n 1/2^n < \infty$. A gain sequence such as $\alpha \propto 1/n$ would meet these constraints. This first constraint, that is violated by our gain sequence, may be regarded as ensuring that the gradient descent can in fact reach the minimum. Since we use a cross-validation training scheme, it may be that this condition is not necessary for us. Certainly a $1/2^n$ gain sequence results in faster training than a $1/n$ sequence.

3 THE DECIPHER SYSTEM

The systems into which we have previously integrated connectionist probability estimators were very simple: context independent phone models, single density models (with duration modelling) and single pronunciations of each vocabulary item. This paper continues this research by integrating such connectionist probability estimators into a large HMM continuous speech recognition system, SRI's DECIPHER. DECIPHER is a much richer system than the previous baseline systems we have used. It includes multiple probabilistic word pronunciations, cross-word phonological and acoustic modelling, context dependent phone models, and models with multiple densities.

Words are represented as probabilistic networks of phone models, specifying multiple pronunciations. These networks are generated by the application of phonological rules to baseform pronunciations for each word. In order to limit the number of parameters that must be estimated, phonological rules are chosen based on measures of coverage and overcoverage of a database of pronunciations. This results in networks which maximise the coverage of observed pronunciations while minimising network size. Probabilities of pronunciations are estimated by the forward-backward algorithm, after tying together instances of the same phonological process in different words. Phonological rules can be specified to apply across words, adding initial or final arcs which are constrained to connect only to arcs fulfilling the context of the rule [2, 3].

Context dependent phone models include word-specific phone, triphone, generalised triphone, cross-word triphone (constrained to connect to appropriate contexts), and left and right biphone (and generalised biphone). All these models are smoothed together, along with context independent models, using the deleted interpolation algorithm.

Most phone models have three states, each state having a self transition and a transition to the following state. A small number of phone models have two states, to allow for short realisations.

4 EXPERIMENTS

Experiments were performed on the speaker-independent DARPA Resource Management database. This database used a vocabulary of 998 words and no grammar (perplexity = 998) or a word pair grammar (perplexity = 60).

A 12th order mel cepstrum front end was used, producing 26 coefficients per frame: energy, 12 cepstral coefficients and derivatives of each static feature computed over a 4 frame window. The inputs to the MLP consisted of a frame in ± 4 frames of context, a feature vector length of 234. The MLPs that we used contained 512 hidden units (a number determined by empirical experiments, trading off representational power with computation) and 69 output units (corresponding to 69 monophone categories), giving a total of around 150,000 weights. Stochastic gradient descent training typically required about 10 passes through the training database of 1.3 million frames. This required less than 24 hours compute time, using a 5-board RAP (Ring Array Processor) [5], containing 20 TI TMS320C30 DSPs, each with 256kB of SRAM and 16MB of DRAM.

To train an MLP we require a bootstrap model to produce time-aligned phonetic labels. In this case we used the context independent DECIPHER system to perform the forced alignment between the training data and word sequence.

The baseline DECIPHER system models the output distributions using tied Gaussian mixtures. Training used the forward-backward algorithm to optimise a maximum likelihood criterion.

We used two sets of test sentences for evaluation. A 300 sentence development set (the June 1988 RM speaker independent test set) was used to tune the HMM recognition parameters, such as the word transition penalty. The results reported here were obtained from a 600 sentence test set (the February 1989 and October 1989 RM speaker independent test sets); no tuning of parameters was performed using this set.

4.1 Context Independent Models

We first experimented using context independent models. The baseline context independent DECIPHER system incorporated multiple pronunciations, cross-word phonological modelling, etc., but had only 69 two or three state phone models (200 distributions in all).

The baseline connectionist system had 69 single distribution phone models; the lexicon consisted of a single pronunciation for each word. Each phone model was a left-to-right model (with self-loops) with $N/2$ states, where N was the average duration of the phone. Transition probabilities were all tied to be 0.5. The connectionist probability estimator was integrated into DECIPHER in two ways:

- The usual {2,3}-state DECIPHER models were used, but each model had only a single output distribution (from the MLP). Thus the 2 or 3 states in a model shared a distribution.
- A new MLP was trained with 200 outputs, corresponding to the 200 states in the 69 context independent DECIPHER models.

The maximum likelihood transition probabilities (which basically encoded duration information) were retained.

Two heuristics were tried for combining the MLP and standard estimates of the state output probabilities. In the first weighted logs of the MLP and Gaussian mixture likelihood estimations were used:

$$\log(P(\mathbf{x}|q_j)) = \lambda_1 \log\left(\frac{P_{mlp}(q_j|\mathbf{x})}{P(q_j)}\right) + \lambda_2 \log(P_{gm}(\mathbf{x}|q_j))$$

where P_{mlp} denotes the MLP estimate of a probability and P_{gm} the Gaussian mixture estimate. A single set of λs was used over all the states: they were optimised for minimum recognition error over the 300 sentence development set.

	Parameters	% error Perplexity 998	60
Baseline MLP-69	155,717	36.1	12.8
CI-DECIPHER	125,762	44.7	14.0
MLP-69	155,717	30.1	8.3
MLP-200	222,920	34.9	11.4
MIX-69	281,548	29.5	7.9

Table 1: Results using 69 context independent phone models. The baseline MLP system uses 69 single distribution models with a single pronunciation for each word in the vocabulary. The DECIPHER system also uses 69 phone models, each with two or three states 200 independent distributions in total. The MLP-69 and MLP-200 systems use DECIPHER's multiple pronunciation and cross-word modelling. MLP-200 differs from the other MLP systems in that it has 200 outputs corresponding to DECIPHER's 200 states. The MIX-69 system is an a system interpolating the probabilities produced by the MLP and DECIPHER (rather than replacing the DECIPHER probabilities by MLP probabilities).

In the second heuristic, the log of a weighted average of the state ouput probabilities estimated by the MLP and the tied Gaussian mixtures was used:

$$\log(P(\mathbf{x}|q_j)) = \log\left(\lambda_1 \frac{P_{mlp}(q_j|\mathbf{x})P_{gm}(\mathbf{x})}{P(q_j)} + \lambda_2 P_{gm}(\mathbf{x}|q_j)\right).$$

In this approximation, the probability of the data $P(\mathbf{x})$ was required to ensure that the 2 likelihood estimates are scaled similarly. This cannot be obtained from the MLP, and was approximated by summing over the state conditional tied Gaussian likelihoods:

$$P_{gm}(\mathbf{x}) = \sum_i P_{gm}(\mathbf{x}|q_i)P(q_i).$$

The best results were obtained using the first method, which resulted in an 8.0% error on the development set, compared with an error of 9.1% using the second method. Thus the first approach was used in evaluating over the 600 sentence test set.

Results for these context independent systems are shown in table 1. There are several notable aspects to these results:

- The MLP system using single pronunciations and single distribution phone models has a lower error rate than the context independent DECIPHER system, which uses multiple pronunciations and cross-word phonological modelling.
- Incorporating the MLP estimator into the context independent DECIPHER system results in still higher performance, lowering the error rate substantially from 12.8% to 8.3%.
- The DECIPHER system uses multiple state, multiple distribution HMMs. Typically each phone model consists of 3 independent states. An MLP can be used to estimate these probabilities, simply by increasing the size of the output layer and using the maximum likelihood state segmentation as output targets. In

	Parameters	% error Perplexity 998	60
DECIPHER	5,541,844	21.9	4.9
MLP-DECIPHER	5,697,726	20.2	4.3

Table 2: Results using 3428 context dependent phone models. The MLP-DECIPHER hybrid system interpolates the MLP context independent probabilities with the DECIPHER context dependent probabilities.

maximum likelihood trained systems moving from single distribution to multi-distribution models is usually beneficial, since it enables acoustically different parts of a phone (e.g. onset, centre and offset of a vowel) to be modelled independently. However, such a change produced a performance degradation when using the MLP. We hypothesise that this was due to discriminative training. In many cases, different states of a phone are acoustically very similar. Thus, forcing an MLP to discriminate between such states could be counterproductive.

- Interpolation of MLP probabilities with context-independent DECIPHER maximum likelihood probabilities gives a small, but not significant (at the 0.95 level), improvement. The parameters for the interpolation were tuned on the development set. There was a more substantial improvement on the development set (around 2% error reduction), indicating that the interpolation parameters are being overfitted to the dataset used for tuning.

4.2 Context Dependent Models

Our experiments using context dependent models involved interpolating connectionist estimates of context independent output probabilities (as used above) together with the maximum likelihood tied mixture estimates of the context dependent probabilities used in DECIPHER. The first heuristic described above was used for these interpolations. Results for this hybrid system are shown in table 2. These results indicate that the MLP context independent probabilities may be used to afford a small improvement in recognition, which is not significant at the 0.95 level for the word-pair grammar case, but is significant at that level for the no grammar case.

We are currently researching the use of context dependent neural networks (CDNNs) to model context dependent phones. This approach and initial results are presented in a second paper to be presented at this conference.

5 CONCLUSIONS

The results presented here indicate that connectionist probability estimation is able to improve the performance of a state of the art recognition system. Perhaps the three most important conclusions are:

1. Comparing like with like, a discriminatively trained connectionist context independent system performs considerably better than the corresponding maximum likelihood tied mixture system.
2. The context independent MLP-DECIPHER system has an error of 8.3% compared with a 4.9% error produced by context dependent DECIPHER. However the latter system has 50 times the number of models and 35 times the number of parameters compared with the MLP system.
3. Interpolating MLP context independent probabilities into the context dependent DECIPHER system produces an increase in word accuracy.

ACKNOWLEDGEMENTS

Many thanks to Hervé Bourlard, Chuck Wooters and Phil Kohn, all of whom provided valuable input to this work. Support from the International Computer Science Institute is gratefully acknowledged. This work was partially funded by DARPA contract MDA904-90-C-5253.

REFERENCES

[1] H. Bourlard and C. J. Wellekens. Links between Markov models and multi-layer perceptrons. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 502–510. Morgan Kaufmann, San Mateo CA, 1989.

[2] M. Cohen. *Phonological Structures for Speech Recognition*. PhD thesis, University of California at Berkeley, 1989.

[3] Michael Cohen, Hy Murveit, Jared Bernstein, Patti Price, and Mitch Weintraub. The DECIPHER speech recognition system. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 77–80, Albuquerque, 1990.

[4] Herbert Gish. A probabilistic approach to the understanding and training of neural network classifiers. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1361–1364, Albuquerque, 1990.

[5] N. Morgan, J. Beck, P. Kohn, J. Bilmes, E. Allman, and J. Beer. The ring array processor (rap): A multiprocessing peripheral for connectionist applications. *Journal of Parallel and Distributed Computing*, page In Press, 1992.

[6] N. Morgan, H. Hermansky, H. Bourlard, C. Wooters, and P. Kohn. Continuous speech recognition using PLP analysis with multi-layer perceptrons. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 49–52, Toronto, 1991.

[7] Steve Renals, Nelson Morgan, and Hervé Bourlard. Probability estimation by feed-forward networks in continuous speech recognition. In *Proceedings IEEE Workshop on Neural Networks for Signal Processing*, pages 309–318, Princeton NJ, 1991.

[8] H. Robbins and S. Munro. A stochastic approximation method. *Annals of Mathematical Statisitics*, 29:400–407, 1951.