IMPROVING STATISTICAL SPEECH RECOGNITION

Steve Renals*, Nelson Morgan*, Michael Cohen†, Horacio Franco† and Herve Bourlard‡ *
*** International Computer Science Institute, Berkeley CA 94704, USA**
† SRI International, Menlo Park CA 94025, USA
‡ L&H Speechproducts, Ieper B-8900, Belgium

ABSTRACT

In this paper we summarise the theory underlying our hybrid connectionist/HMM continuous speech recognition systems and report some recent experimental results and further theoretical developments. Highlights include:

1. Experimental results indicating that our connectionist methods can significantly improve the performance of a context independent HMM system (to a performance close to that of state of the art context dependent systems of much higher complexity);
2. Experimental results demonstrating that a state of the art context dependent HMM system (SRI's DECIPHER) can be significantly improved by interpolating context independent connectionist probability estimates;
3. The development of a principled network decomposition method that allows the efficient and parsimonious modelling of context-dependent phones, with no independence assumptions.

1 INTRODUCTION

Since 1988, we have been investigating the use of feed-forward connectionist networks to improve hidden Markov model (HMM) speech recognition systems. The basis of this approach has been a proof that feed-forward networks can be used as probability estimators [2]. We have been using such networks to estimate the output probabilities of HMMs [5, 6].

In this paper, we summarise the theory underlying our hybrid connectionist–HMM systems, and present the issues (both theoretical refinements and practical necessities) that we have had to address in order to use these methods successfully. Our recent experiments have used connectionist methods to augment the DECIPHER system developed at SRI International [3].

The connectionist systems described here estimate probabilities for context independent phone models. However, advanced continuous speech recognition systems utilise context dependent phone models [7, 4]: multiple models for each phone depending on the surrounding phonetic context. In section 5, we present a new network decomposition method that allows feed-forward networks to model phones in context efficiently and parsimoniously. We show this technique to be theoretically justified—no independence assumptions are necessary—and initial experiments have demonstrated that the decomposition method does indeed produce similar probability estimates to the much larger original networks.

2 HMM—CONNECTIONIST HYBRIDS

While neural networks have shown themselves to be powerful classifiers, it is still the case that there are no architectures and associated algorithms to model time-dependence that can compete with the HMM approach for continuous speech recognition. In this section we summarise the basic features of HMM speech recognition and explain how neural network methods may be integrated within this framework, to produce hybrid systems.

2.1 HIDDEN MARKOV MODELS

In continuous speech recognition our basic units of speech are typically at sub-word level—here we use phones. Each phone is modelled by a HMM, a stochastic automaton defined by a set of states q_i, a topology specifying allowed state transitions and a set of local output probability density functions $P(\mathbf{x}_{t+1}, q_i|q_j, \mathbf{X}_1^T)$. If we assume that the output at any time is independent of previous outputs and depends only on the current state, then we may separate a local probability into the product of a transition probability $P(q_i|q_j)$ and a state output probability $P(\mathbf{x}_{t+1}|q_i)$. The state of a HMM at any given time t is *hidden*: only the acoustic vector $\mathbf{x}_t$ is observed. We perform speech recognition by computing the most likely state sequence that could have generated the data. This is usually accomplished by a dynamic programming procedure referred to as the Viterbi algorithm.

Note that the HMM is a powerful hierarchical representation, since words are composed of sequences of phone models (constrained by a lexicon of pronunciations) and sentences are sequences of word models (constrained by a grammar).

We wish to estimate $P(\mathbf{Q}_1^N|\mathbf{X}_1^T, \Theta)$, the probability of a sequence of HMM states $\mathbf{Q}_1^N = q_1, \ldots, q_n$, given the acoustic input $\mathbf{X}_1^T$ and the parameters of our models Θ. Using Bayes' rule we may rewrite this as:

$$P(\mathbf{Q}_1^N|\mathbf{X}_1^T, \Theta) = \frac{P(\mathbf{X}_1^T|\mathbf{Q}_1^N, \Theta)P(\mathbf{Q}_1^N|\Theta)}{P(\mathbf{X}_1^T|\Theta)} \tag{1}$$

We call $P(\mathbf{X}_1^T|\mathbf{Q}_1^N, \Theta)/P(\mathbf{X}_1^T|\Theta)$ the acoustic model and $P(\mathbf{Q}_1^N|\Theta)$ the language model. The language model encodes the constraints given by the model topology: i.e. word pronunciations and grammar. This is a prior, perhaps estimated from some larger text database. In this paper we are concerned with the acoustic model.

The acoustic model is often represented by some parametric density estimator (e.g. a tied mixture of Gaussians): the parameters of such models are usually set using a maximum likelihood procedure, the EM (forward-backward) algorithm, which maximises the likelihood $P(\mathbf{X}_1^T|\mathbf{Q}_1^N, \Theta)$. Note that this is a different acoustic model to the one used in the previous paragraph: the $P(\mathbf{X}_1^T|\Theta)$ denominator is not included. Whilst this denominator is constant at recognition training (no model dependency), it is not constant at training time since the parameters Θ are being adapted. Since

$$P(\mathbf{X}_1^T|\Theta) = \sum_{\mathbf{R}_1^N} P(\mathbf{X}_1^T|\mathbf{R}_1^N, \Theta)P(\mathbf{R}_1^N|\Theta) \ , \tag{2}$$

maximising $P(\mathbf{X}_1^T|\mathbf{Q}_1^N, \Theta)/P(\mathbf{X}_1^T|\Theta)$ corresponds to maximising the likelihood of the correct state sequence generating the data, and minimising the probability of the data being generated by all incorrect state sequences. This is *discriminative training*.

2.2 CONNECTIONIST PROBABILITY ESTIMATION

In 1989 Bourlard and Wellekens [2] proved that a feed-forward network (e.g. a multi-layer perceptron, MLP) trained to minimise a least squares or relative entropy objective function in a 1–from–N classification task will output estimates of the class conditional posterior probability, $P(q_i|\mathbf{x}, \Theta)$ of class q_i given input $\mathbf{x}$. Now,

$$P(q_i|\mathbf{x}, \Theta) \propto \frac{P(\mathbf{x}|q_i, \Theta)}{P(\mathbf{x}|\Theta)} \ ,$$

where the proportionality constant is $P(q_i|\Theta)$, which is the relative frequency of class i in the training set. Thus if we divide the outputs of a trained MLP by these relative frequencies, we have probabilities that define an acoustic model for HMM, derived in a discriminative fashion.

3 RECENT SPEECH RECOGNITION EXPERIMENTS

We have applied these methods to problems in continuous speech recognition. Most of our experiments have been performed using the DARPA Resource Management (RM) speaker independent continuous speech database. This is a very well studied database with a vocabulary of 998 words. The standard training set contains around 4000 sentences from over 100 speakers: about 1.5 million frames of data. Evaluation is performed over various test sets, typically containing 300 sentences from 12 new speakers.

The difficulty of a speech recognition task is measured by the perplexity, the geometric average number of words that can follow any other word. When no grammar is used, the RN task has a perplexity of 998—any word can follow any other. With a simple word-pair grammar listing allowable pairs of words, the perplexity is reduced to 60.

The front end used in these experiments was a mel cepstral analysis, producing 12 coefficients, plus energy for each 10ms frame of speech. In addition to these 13 coefficients, we estimate their derivatives (via a regression), giving 26 coefficients per frame.

We have trained feed-forward networks to perform phonetic classification, using this data. Thus, when trained, these networks output estimates to the class conditional posterior probabilities[1].

The networks we trained to perform these estimations were typically large; in addition to 26 inputs, we typically appended $\pm$ 4 frames of context, giving a total of 234 inputs. We usually used networks with about 512 hidden units, and with 69 output classes, giving a total of over 150,000 weights. Training a network with this many weights is not trivial (we used the RAP, a ring array processor with 4–20 TMS320C30 DSPs, see Kohn et al., these proceedings): in particular we have used a cross-validation procedure combined with stochastic gradient descent (per-pattern update). 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. 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.

The systems into which we have previously integrated connectionist probability estimators were very simple, using context independent phone models[2] and single pronunciations of each vocabulary item. This paper continues this research by integrating our connectionist probability estimators into a large HMM continuous speech recognition system, SRI's DECIPHER [3]. DECIPHER is a much richer system than the previous baseline systems we have used. It includes both multiple probabilistic word pronunciations and context dependent phone modelling.

4 RESULTS

We have performed continuous speech recognition experiments using the database, networks and system described above. Training a network of around 150,000 weights required around 10 passes through the training database of 1.3 million training patterns (200,000 patterns were held out for cross-validation)

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 modelled the output distributions using tied Gaussian mixtures. Training used the forward-backward algorithm to optimise a maximum likelihood criterion.

1. The theorem of Bourlard and Wellekens[2] only applies when networks are trained to a global minimum; however, numerical simulations have indicated that networks trained to a stopping point above the minimum also output good estimates of the posteriors.
2. A *context independent* phone model, is one which models a phone in whatever context it may occur. *Context dependent* phone models, however, model phones in specific phonetic contexts, to take account of the acoustic effect of context. Thus there may be several different models for a single phone, depending on the surrounding context.

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

A context independent form of the DECIPHER system was used as a baseline. It was trained using the maximum likelihood forward-backward procedure gave a word error of 14.0%, using the word-pair grammar (perplexity 60). When the usual HMM output probability estimators were replaced with a MLP trained to classify its input acoustic vectors into 1 of 69 classes (and outputting posterior probability estimates) the word recognition error improved to 7.8%. In the no grammar case, the MLP improved the error rate of DECIPHER from 44.7% to 30.1%.

The baseline CI-DECIPHER system uses multiple densities to model each phone: that is, the beginning of a phone may be modelled with a different distribution, compared with the beginning or end of that phone. This multiple density modelling has proven effective for "pure" HMM systems. However, in our MLP variant of DECIPHER, although each phone is modelled by multiple states, all the states in a phone shared the same density. We have experimented with multiple densities in the MLP system—they worsened the performance. We hypothesise that this was due to discriminative training. In many cases, different states of a phone are acoustically similar. Forcing an MLP to discriminate between them could be counter-productive.

4.2 CONTEXT DEPENDENT MODELS

We have also used MLP context independent output probability estimates to improve the full, context-dependent DECIPHER system. This system had 3428 models with over 5.5 million parameters; the context independent system had 69 models with about 150,000 parameters. Context dependent DECIPHER is a state of the art continuous speech recogniser, producing an error of 4.9% on our test set. By interpolating the MLP estimates with the DECIPHER output probability estimates, we have improved the word error rate from 4.9% to 4.1% using a word-pair grammar. This is a statistically significant improvement with 95% confidence. This interpolation between the scaled likelihoods of the MLP and the likelihood estimates of the pure HMM was performed in the log domain, so that the MLP $P(\mathbf{x}|\Theta)$ factor became additive. In the no grammar case, word error improved from 21.9% to 19.5%, which is also statistically significant at the 95% level.

5 CDNN: A CONTEXT DEPENDENT NEURAL NETWORK

In the experiments reported above, connectionist methods were used only to estimate context independent probabilities of the form $P(q_i|\mathbf{x})$. Such probabilities are considered "context independent" since the left hand side of the posterior, contains no term involving neighbouring phones[3]. However, state of the art recognisers, such as DECIPHER, are of greater complexity: in particular they use context dependent phone models. Will the consistent improvements offered by our connectionist methods be washed out in systems with more detailed models?

One difficulty with more complex models is that many more parameters must be estimated with the same limited amount of data. Brute-force application of our earlier techniques to context dependent phones would result in an output layer with many thousands of units, and a network with many millions of connections. This network would be impractical to train, both in terms of computation and learnability, using current-sized public data bases. Here, we present a method for estimating likelihoods for context dependent phone models, using networks that are not substantially bigger than our context independent networks, and that require only a small increase in computation. This method is a general technique for decomposing a larger classification network into several smaller networks without any simplifying assumptions, such as independence of outputs.

3. Note, that these probabilities do include an acoustic context, since we typically use a 9 frame window. However this acoustic context is usually on a different timescale to phonetic context.

Consider a system with K phone classes. A context independent network trained to estimate class posterior probabilities would then have K outputs. For a context dependent model, we may wish to estimate the joint probability of a current HMM state with a particular neighbouring phonetic category. Using $\mathcal{C}$ to represent the set of possible contexts, we wish to estimate $p(q_k, c_j|\mathbf{x})$, where $c_j \in \mathcal{C} = \{c_1, ..., c_L\}$, where q_k are HMM states and $\mathbf{x}$ the acoustic data, as usual. If there are L context classes, this will require $K \times L$ output units for an MLP estimator. However, if we use the conditional probability axiom, the desired expression can be broken down as follows:

$$p(q_k, c_j|\mathbf{x}) = p(q_k|\mathbf{x}) \times p(c_j|q_k, \mathbf{x}) \tag{3}$$

Thus, the desired probability is the product of the monophone posterior probability and a new conditional. The former can be realised with the usual monophone network. Viewing an MLP as an estimator of the left side of a conditional given the right side as input, the second term can be estimated by an MLP trained to generate the correct context class given inputs of the current class and the speech input frame. The latter network only has as many outputs as there are context classes.

This procedure reduces the training of a single network with $K \times C$ outputs to the training of two smaller networks with K and C outputs respectively, and represents a generic way of splitting large MLPs used in classification mode into several smaller ones. It has the potential, however, of requiring much greater computation during the recognition phase. Solutions to these and other practical problems are discussed in [1].

We compared biphone probabilities generated by the original and split MLP for the speaker independent Resource Management database. Each biphone consisted of a phone with a right context of one of eight broad speech categories, based on place of articulation. The number of hidden units in each MLP was chosen such that the number of parameters was approximately the same in both cases. After having trained both cases on 4,000 sentences, biphone probabilities were computed on a test set of 100 sentences pronounced by 4 different speakers, yielding a total of 17,012,088 probabilities. To compare both sets of probabilities we computed the correlation coefficient to be 0.65, and the mean absolute difference that was equal to 0.0017. Thus, the two sets of probabilities are significantly correlated.

6 CONCLUSIONS

Two substantial advances in connectionist methods for continuous speech recognition have been reported:

- We have demonstrated that connectionist probability estimation may be used to improve the performance of a state of the art continuous speech recognition system. In particular:
  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 7.8% 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 a significant increase in word accuracy.
- We have presented a general method for factoring a multi-layered classification network with a large output layer into a number of smaller networks, with no statistical independence assumptions. We have demonstrated that this technique may be used estimate likelihoods for context dependent phone models.

ACKNOWLEDGEMENTS

Thanks to Phil Kohn for software support and Chuck Wooters for experimental assistance. This work was partially funded by DARPA contract MDA904-90-C-5253.

REFERENCES

[1] H. Bourlard, N. Morgan, C. Wooters, and S. Renals. CDNN: A context dependent neural network for continuous speech recognition. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, 1992. In press.

[2] H. Bourlard and C. J. Wellekens. Links between Markov models and multilayer perceptrons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-12:1167–1178, 1990.

[3] M. Cohen, H. Murveit, J. Bernstein, P. Price, and M. Weintraub. The DECIPHER speech recognition system. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 77–80, Albuquerque, 1990.

[4] K. F. Lee. *Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1988.

[5] 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.

[6] S. Renals, N. Morgan, M. Cohen, and H. Franco. Connectionist probability estimation in the DECIPHER speech recognition system. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, 1992. In press.

[7] R. Schwartz, Y. Chow, O. Kimball, S. Roucous, M. Krasner, and J. Makhoul. Context-dependent modelling for acoustic-phonetic recognition of continuous speech. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1205–1208, Tampa FL, 1985.