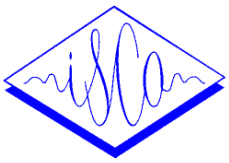

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

Phonetic Context in Hybrid HMM/MLP Continuous Speech Recognition

N. Morgan[†], H. Bourlard[†,‡], C.Wooters[†], P.Kohn[†], & M. Cohen[*]

[†]International Computer Science Institute, Berkeley, CA 94704, USA
[‡]Lernout&Hauspie Speechproducts, 1780 Wemmel, Belgium
[*]SRI International, Menlo Park, CA, USA

ABSTRACT

Earlier work has shown the ability of Multilayer Perceptrons (MLPs) to estimate emission probabilities for a Hidden Markov Model (HMM) [1][2][3]. In these reports, we have shown that these estimates have led to improved performance over counting estimation techniques in the case where a fairly simple HMM was used. However, current state-of-the-art continuous speech recognizers require HMMs with greater complexity, e.g. multiple densities per phone and/or context-dependent phone models.

Brute-force application of our earlier techniques to triphones (the standard approach to context-dependent HMMs) would result in an output layer with many thousands of units, and many millions of connections to train. In this report we describe another approach to the application of MLPs to context-dependent probability density estimation, as well as some practical aspects of efficient implementation of the method.

INTRODUCTION

As has been shown in [1] and demonstrated practically in [2,3], networks trained for phonetic classification produce values that are very close to posterior probabilities, that is, the probability of a phonetic class given some speech representation for a time segment. This is true for any trained classification system with the following characteristics:

1) The error criterion is sum of squared errors or relative entropy

2) The targets are 1.0 for only one of the classes and 0.0 for all of the others

3) The trained system reaches a global minimum.

The first two conditions are easily met. In general, trained systems do not meet the last condition. However, we have shown experimentally that MLP outputs trained to local minima function very well as probabilities; they sum, on the average, to roughly 1.0, and the usual axioms of probability appear to hold. In particular, we have used Bayes' rule to transform the outputs to likelihoods for use in an HMM continuous speech recognizer. These probability estimates lead to respectable performance on speaker-dependent training and recognition (currently about 8% word error on the Resource Management task using the perplexity-60 wordpair grammar). In essence, our recognition system consists of a standard and simple HMM recognizer in which the emission probabilities are computed by the MLP. Direct comparisons between simpler training approaches (e.g., co-occurrence counting for VQ features) and the MLP have consistently shown an advantage for this approach to probability estimation.

On the other hand, the best HMM systems still perform substantially better than our experimental systems. For instance, on the same task, scores of 2-3% have been reported [4]. While there are many design differences between any two real-world systems, the major obvious difference between the best HMM systems and our experimental programs is that the former incorporate context-dependent phonetic models, and the latter do not. Speech researchers using HMMs have consistently found context-dependent phonetic models to be very important in reducing recognition errors.

Context dependent modeling poses problems to both HMM and ANN systems in that many more parameters must be estimated with the same limited amount of data. The standard HMM approach is to use triphone models. A triphone is a model of a phone given a particular left and right phonetic context. The direct application of this technique to our system would pose problems. For the Resource Management task, our system uses 61 context independent phone models (with a single output unit for each phone). If we were to extend the system to triphone models, we would have 61^3 or about 220,000 different context-dependent models. With a typical hidden layer of 500 units, our

system would have about 10^8 connections, which is far too many for a practical system.

As suggested above, HMMs have a similar problem when modeling triphones. In order to deal with this problem, HMM systems use a variety of techniques either to reduce the number of context-dependent models, or to smooth poorly trained context-dependent models with well trained context-independent models. For example, SRI's DECIPHER system [5] uses a form of generalized context based on broad phonetic classes. A hierarchy of models, from general to specific, is defined. These include, from most specific to most general, word-specific, triphone, generalized triphone, biphone, generalized biphone, and context independent models. All of these models are trained simultaneously using the forward-backward algorithm. The actual phonetic models used when building word models are a combination of the models trained at all levels of specificity. They are combined using the deleted interpolation algorithm [6].

Similar approaches must be designed in order to make it possible to use context dependent models in ANN systems. Even if enough training data were available, networks with million of parameters can be expected to take impractical amounts of time to train using back-propagation approaches, even with fast special-purpose machines such as our Ring Array Processor (RAP) [7].

In the approach reported here, we are able to estimate likelihoods for context-dependent phonetic models with nets that are not substantially larger than our context-independent MLPs, and that require only a small increase in computation. The remainder of this paper describes this approach, as well as some practical aspects of efficient implementation.

CONTEXT-DEPENDENT MLPs

As described above, with a few assumptions an MLP may be viewed as estimating the probability $p(q|x_n)$ where q is a speech class, and x_n is the input data (speech features) for frame n. If there are K such classes, then K outputs are required in the MLP. This probability may be considered "context-independent" in the sense that the left-hand side of the conditional probability contains no term involving the neighboring phones.

For a context-dependent model, we may wish to estimate the joint probability of a current phone with a particular neighboring phone. Using c to represent the class of context. we wish to estimate $p(q,c|x_n)$. As noted previously, if there are C context classes, this will require $K \times C$ output units for an MLP estimator. However, if we use the definition of conditional probability, the desired expression can be broken down as follows:

$$p(q,c|x_n) = p(q|x_n) \times p(c|q,x_n)$$

Thus, the desired probability is the product of the monophone posterior probability and a new conditional. The former can be realized 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 networks with K and C outputs respectively, a potentially huge savings in time and in parameters. It has the potential, however, of requiring much greater computation during the recognition phase. If one implements this method naively, the second network must be computed K times for each frame during recognition, since the output probabilities depend on an assumption of the current class (corresponding to a monophone model in a hypothesized word sequence at that point in the dynamic programming). However, this expense can largely be circumvented. Assuming no hidden layers shared between inputs q and x_n for the second net, the contribution to the output vector (pre-sigmoid) coming from q can be pre-computed for all possible values of q and c. Look-ups to this table can be added to pre-sigmoid outputs due to x_n at each required point in the dynamic programming. This look-up is a minor increase for the computation of each local distance. The most significant increase in computation is then due to the computation or look-up of the sigmoid, which now must be done for each local distance calculation in the dynamic programming, rather than once each frame and each class. In practice this appears to at most double the recognition time. Alternatively, a second table with all biphone or triphone probabilities can be computed from the first by adding in the pre-sigmoid output due to x_n, and calculating the sigmoids once per frame. Assuming that the number of Markov states examined per frame is much larger than the size of this table, this would be a further computational savings.

This recognition procedure is illustrated in Figure 1 for the biphone case. The lower left of the figure shows the q inputs which are all zero except for a one corresponding to the current phoneme (either known, as in the training case, or hypothesized, as in the recognition case). The pre-sigmoid weighted sum of hidden unit outputs corresponding to context class l is called Y, and is added to the pre-sigmoid output Z, which is due to the data vector x. Note that Y is data-independent, and can be pre-computed prior to recognition, while Z is only computed once per frame.

Using this approach, one can estimate biphones or generalized biphones. Extending this in the obvious way, one can estimate triphone probabilities using three networks. Preliminary tests with generalized biphones appear to show some small improvement when the biphone probabilities are smoothed with monophone probabilities using a simple heuristic. This will be formalized in upcoming months by applying smoothing techniques such as the common approach of deleted interpolation [6]. For the MLP-based implementation of this algorithm, the network would be trained N times in a jacknife procedure, using $\frac{N-1}{N}$ of the data. Each choice of the remaining $\frac{1}{N}$ of the data is then used both for cross-validation in the neural network training and for the deleted-interpolation estimation of the smoothing parameters.

DISCUSSION

As has been discussed in our previous papers, networks can generate the probabilities $p(Y|X)$, where Y is the target vector and X is the input vector. Using this fact and a few simple probabilistic axioms, highly-dimensioned models can be factored into several low-dimensioned models, without any added assumptions of independence. While our implementations of these ideas are at too early a stage to know if they are of practical importance, they are at least in principle a way of strongly reducing the number of parameters required for context-dependent models. An additional issue for future consideration is the smoothing of context-dependent densities with context-independent densities; should this be done using Maximum Likelihood techniques such as deleted interpolation, when our density estimation techniques are discriminant?

These and other practical smoothing considerations are the major topics of our current investigations.

ACKNOWLEDGEMENTS

We note and appreciate the continued support from ICSI and DARPA contract MDA904-90-C-5253 for this work.

REFERENCES

[1] H. Bourlard, and N. Morgan. "Merging Multilayer Perceptrons & Hidden Markov Models: Some Experiments in Continuous Speech Recognition" in *Artificial Neural Networks: Advances and Applications*, North Holland Press, 1990, E. Gelenbe editor (In Press)

[2] N. Morgan and H. Bourlard. "Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models", *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, pp. 413-416, Albuquerque, New Mexico, 1990.

[3] N. Morgan, C. Wooters, H. Bourlard, and M. Cohen. "Continuous Speech Recognition on the Resource Management Database using Connectionist Probability Estimation", Proceedings of ICSLP-90, 1337-1340, Kobe, Japan, 1990

[4] D.S. Pallett, J.G. Fiscus, and J.S. Garofolo. "DARAPA Resource Management Benchmark test Results, June 1990", Proceedings of the DARPA Speech and Natural Language Workshop, Hidden Valley, Pa., June 1990

[5] M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, D. Bell. "Linguistic Constraints in Hidden Markov Model Based Speech Recognition" *Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing*, pp. 699-702, Galsgow, Scotland, 1989.

[6] L.R. Bahl, F. Jelinek, and R. Mercer. "A Maximum Likelihood Approach to Continuous Speech Recognition", IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-5(2):179-190, March, 1983

[7] Morgan, N., Beck, J., Kohn, P., Bilmes, J., Allman, E., and. Beer, J."'The RAP: a Ring Array Processor for Layered Network Calculations," *Proc. of Intl. Conf. on Application Specific Array Processors*, pp. 296-308. IEEE Computer Society Press, Princeton, N.J., 1990.

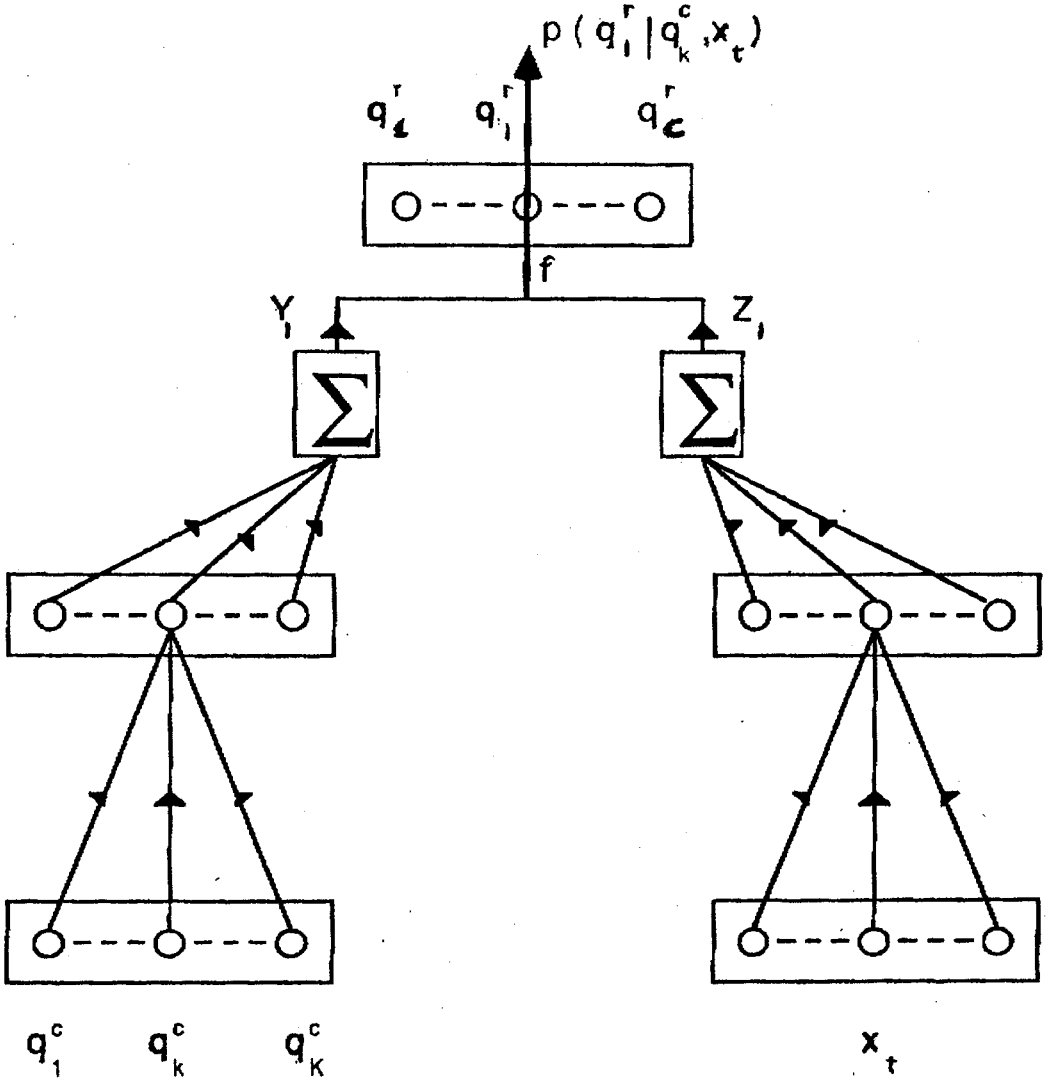

Fig. 1