SIZE MATTERS: AN EMPIRICAL STUDY OF NEURAL NETWORK TRAINING FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

Dan Ellis[†] *and Nelson Morgan*[†,‡]

†International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
‡University of California at Berkeley, EECS Department, Berkeley, CA 94720
Tel: (510) 643-9153, FAX: (510) 643-7684, Email: {dpwe, morgan}@icsi.berkeley.edu

ABSTRACT

We have trained and tested a number of large neural networks for the purpose of emission probability estimation in large vocabulary continuous speech recognition. In particular, the problem under test is the DARPA Broadcast News task. Our goal here was to determine the relationship between training time, word error rate, size of the training set, and size of the neural network. In all cases, the network architecture was quite simple, comprising a single large hidden layer with an input window consisting of feature vectors from 9 frames around the current time, with a single output for each of 54 phonetic categories. Thus far, simultaneous increases to the size of the training set and the neural network improve performance; in other words, more data helps, as does the training of more parameters. We continue to be surprised that such a simple system works as well as it does for complex tasks. Given a limitation in training time, however, there appears to be an optimal ratio of training patterns to parameters of around 25:1 in these circumstances. Additionally, doubling the training data and system size appears to provide diminishing returns of error rate reduction for the largest systems.

1. INTRODUCTION

For about 10 years, we and others have trained large neural networks to estimate posterior probabilities of context-independent phonetic classes for use in speech recognition systems based on Hidden Markov Models (HMMs) [7]. For small tasks, moderate amounts of training data, and when simple models were used, we consistently found that we could provide competitive and often improved recognition performance in comparison with systems that used more standard architectures and training methods (e.g., Gaussian mixtures trained with a Maximum Likelihood criterion) [9]. However, for large tasks for which a great deal of training data was available, we have had difficulty achieving the performance of likelihood-based HMM systems. Some of this difference is undoubtedly due to scientifically uninteresting factors, such as the resources required to correct faulty transcription. We also have wondered whether some of the observed difficulty might be a straightforward trade-off between computation and performance. As we have usually designed it, hybrid HMM/ANN training requires the update of all network parameters in response to every input frame. On the other hand, direct training of state-conditional feature densities in HMM systems only requires the update of parameters corresponding to the state or states associated with each feature vector. Furthermore, at least in principle, likelihood-based HMM systems can always benefit from more acoustic data by improving the estimates for ever-finer categories (e.g., from triphones to quinphones), since with more data these rarer categories will become more populated.

Of course, there are analogous procedures available for connectionist systems; for instance, the CDNN described in [1], with variants explored in [2] and [5], can yield density estimates for context-dependent classes as a product of network outputs. The ACID/HNN system of [3] goes even further, resulting in an extensive set of polyphone probabilities that can be used for a fully context-dependent system in the spirit of the large HMMs. In experiments on Switchboard, for instance, this latter system appears to be quite comparable in performance to the best likelihood-based HMMs. However, to achieve this result the simplicity of the large single network is sacrificed, leaving us with the question: can we extract greater recognition accuracy from an increase in training data without complicating the structure?

In previous work, we typically did not incorporate large amounts of training data (e.g., much more than 10 hours of speech). We also did not have sufficient computational resources to explore the simplest approach: namely, to keep the simple architecture constant and merely increase the size of the network for training on larger training sets. This year, we developed a baseline recognition system for the Broadcast News task, for which we had 74 hours of acoustic data.[1] While using all of this data was preferable, systems trained on subsets were good enough to generate the comparative results for this experiment. For the parts of the experiment in which we used larger networks and larger fractions of the data, the amount of computation would have previously been prohibitive. However, we recently completed the development of a multiprocessor machine incorporating VLSI developed in our group, and this permitted trainings that required on the order of 10^{15} arithmetic operations for the larger cases.

[1] Components of a variant of this system are currently being developed for the 1998 DARPA Hub 4 evaluation, in collaboration with the connectionist groups at Cambridge University, Sheffield University, and Faculté Polytechnique de Mons.

Given the availability of acoustic materials and computational resources, we decided to push our simple system to its limit, and also to test it for a range of training set and neural net sizes.

2. METHODS

The basic procedure was to train neural networks with a range of sizes on acoustic training data from different amounts of large vocabulary continuous speech. Each network was then used in a hybrid HMM/ANN recognizer, and was evaluated with word error rate on a large vocabulary task using a 65K word lexicon.

2.1. Corpus

For these experiments, we used the Broadcast News corpus. This is a collection of speech from American radio and television news broadcasts, such as the National Public Radio program *All Things Considered* or *Nightline*, televised in the U.S. on the ABC network. These shows comprise a whole range of speaking conditions, from planned speech in studio environments to spontaneous speech in noisy field conditions over telephone lines. The (possibly multi-sentence) segments are divided into 7 different focus conditions representing different acoustic/speaking environments; the majority conditions are planned studio speech and spontaneous studio speech.

2.2. System Architecture

As in many of our previous papers [7], the underlying statistical model was an extremely simple HMM. For each of 54 phonetic categories, we had an HMM consisting of a strictly left-to-right model with multiple states tied to a single distribution; multiple repeated states were used to establish a minimum duration for each phone. Transition probabilities were set to .5. The emission probabilities of the HMM were scaled likelihoods estimated by dividing the network outputs by the priors for each class. The network was a Multi-Layer Perceptron (MLP) with a single sigmoidal hidden layer, whose size for these experiments was varied from 500 to 4000 by factors of two. For each choice of hidden layer size, a training was done using $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, and all of the 74 hours of acoustic training material that was available to us for this study. Note that the largest training incorporated about 700,000 parameters and 16 million acoustic frames. 54 outputs associated with the phonetic classes were generated by softmax functions computed from the weighted hidden unit outputs. For the main set of experiments reported here, the input consisted of feature vectors from the frame associated with the target label, as well as from 4 vectors into the past and 4 into the future.

To generate the features used in these experiments, the speech was filtered and downsampled to 8 kHz. PLP-12 features [4], including the PLP gain term to give a 13-element feature vector, were computed every 16 ms, and normalized according to the mean and variance of each segment in the training data. These segments were provided to us by our colleagues in the Cambridge University connectionist group, and roughly corresponded to an utterance or a group of utterances by a single speaker. In practice a good segmentation system (such as the one developed by the HTK group at Cambridge) does not degrade performance over that achieved by manually segmenting chunks associated with a single signal source [11]. We used the Noway large vocabulary decoder [8], and co-developed a large vocabulary pronunciation lexicon with our partners at Cambridge and Sheffield. The backoff trigram grammar from the Cambridge BN 97 system was used, incorporating 7M bigrams and 24M trigrams. It had been trained using both text sources (broadcast news and newswire texts) and broadcast news acoustic transcripts.

2.3. Training Hardware and Software

Connectionist training of large networks is quite computationally demanding, as noted above. To aid in this task, we developed a vector microprocessor described in [10], and vectorized software that incorporated efficient assembler routines for this processor while providing a C++ structure that permitted a moderate range of experimentation for our trainings. The board-level system (called the Spert-II) also includes 8 MB of fast (zero wait state) SRAM so that memory accesses are not a bottleneck for the neural computation of large networks. Current high end PCs and moderate level workstations are now fast enough to compete with this system (when highly optimized software is used), but we have also developed 2-processor and 4-processor systems which are significantly faster for sufficiently large networks. Using a commercial bus expansion chassis, these permit the connection of four Spert-IIs to a single Sparc host. Although the bus bandwidth is necessarily shared between the four processors, accumulating error gradients over 16 to 32 patterns permits a near linear speedup for the larger networks trained in this study. The Spert-II boards can also be used independently within the multiprocessor system for those problems with smaller networks.

3. RESULTS

	Training set size, hours			
	9.25	18.5	37	74
# Hidden units				
500	42.8%	41.0%	40.2%	39.2%
1000	41.8%	38.8%	36.5%	36.9%
2000	40.4%	37.2%	35.6%	34.4%
4000	40.3%	37.4%	33.9%	33.7%

Table 1: Word Error Rate percentages for the overall hybrid recognition system incorporating classifier networks with different-sized hidden layers, trained on varying amounts of acoustic data. In each case, the input consisted of 9 vectors of length 13, and the output layer was 54 units. The number of parameters for each layered network was ((9x13)+54)x(# Hidden units) weights, plus (# Hidden units +54) biases.

Table 1 details the effect of varying the size of the training set and the number of hidden layer units on the error rate of the overall hybrid system, all other parameters being held constant. These results are plotted as a 3-dimensional

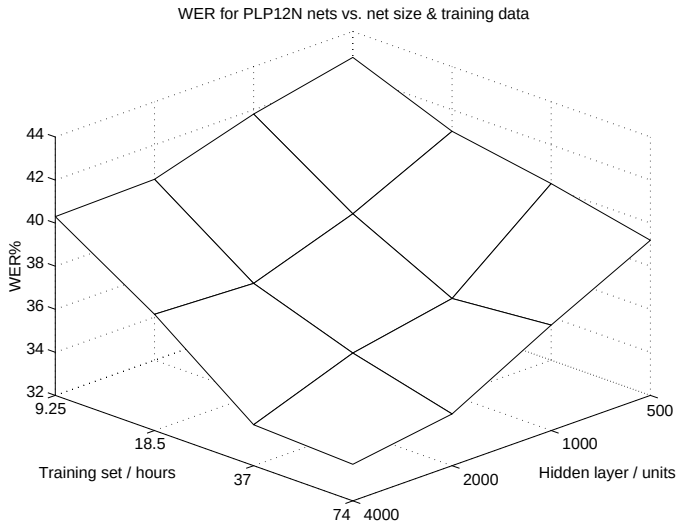

Figure 1: Surface plot of system word error rate as a function of the amount of training data and the hidden layer size.

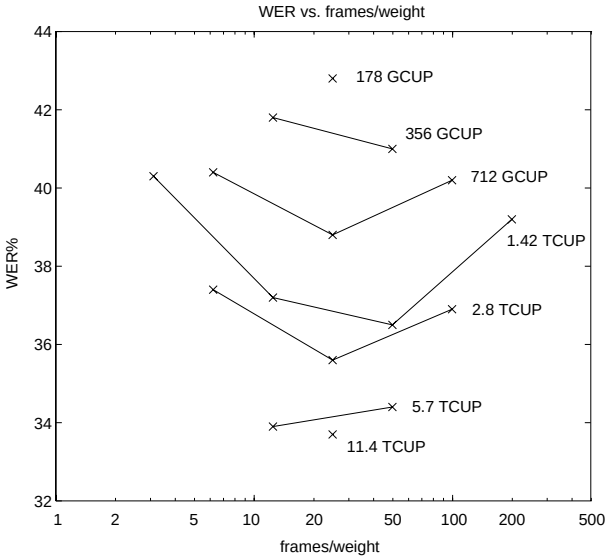

Figure 2: Slices through the surface of the previous figure, showing the variation of error rate with the ratio of training patterns to network weights for a fixed number of connection updates per training epoch.

surface, i.e. error rate as a function of training set and hidden layer size, in Figure 1. The most obvious trend is that increasing either parameter will improve the overall system performance in virtually every condition.

There are some caveats to be borne in mind when considering the word error rate figures. These results were obtained on a separate test set of 32 minutes containing 5938 words; by our reckoning, to be significant at the 5% level, error rates must differ by at least 1.5%.[2] There is additional variability introduced by the randomization of pattern presentation used in the network training.

Training followed a standard 'simulated annealing' process with repeated passes or 'epochs' over the entire training set; after initial stabilization, the learning rate was halved on each successive iteration until the frame classification accuracy on a separate cross-validation set improved by less than 0.5%. The interaction between this criterion and other variables meant that the different networks trained for different numbers of epochs, between 7 and 10.

The acoustic data for these experiments was limited to 4 kHz bandwidth before feature extraction.[3] While this processing succeeded in its intention of improving the relative performance on the telephone-channel speech which forms some 15% of the corpus, it appears to increase the error for the remaining full-bandwidth data. Finally, the decoder pruning for these tests was chosen to be fairly aggressive, giving a typical recognition speed of about 2x real-time; slower but more exhaustive decoding would yield a relative error rate improvement of 5-10%.

Figure 2 shows a succession of slices through the error-rate surface, taken in planes parallel to the view plane. Each slice corresponds to a constant product of training set size and hidden layer size, or equivalently the number of connection updates per complete training epoch. Each line is tagged with this number, with the maximum value of 11.4 TCUP (11.4 x 10^{12} updates) for the case of the 4000 hidden units by (117+54) weights per unit by 16.7 million training patterns. These slices confirm the central 'dip' visible in the error rate surface, indicating that for a given amount of training computation, there is an optimal ratio of training frames per network weight in the range 10 to 40.

4. DISCUSSION

Our primary observation, that improvements are almost always obtained by increasing either or both of the amount of training data or the number of network parameters, is not surprising. It is encouraging, however, that these increases continue to be significant out to the practical limits of our current resources, at least when considering simultaneous increases of training set size and network size (i.e. the leading diagonal of Table 1). In fact, the 1998 Broadcast News evaluations provide a second nominal 100 hour training set, so we are now planning to train an 8000 hidden unit net on 150 hours of data (using 28 features per frame). Even using our custom multi-processor hardware, this training will require over three weeks of computation. Were we to use our 300 MHz Sun Ultra-30, we project that this training would take several months to complete.

Given our earlier experience with training networks for speech recognition, our test points for this study straddled a minimum in the patterns-per-parameter dimension. The

[2]This test set, used internally within our collaboration with Cambridge and Sheffield, is a subset of the BN 97 evaluation set. Previous experience at Cambridge suggests that this subset is slightly "harder" than the larger set, typically by about a factor of 10% in relative error.

[3]This bandlimiting was done for its complementarity with the Cambridge system, with whom we would be merging models; in practice the overall loss in performance due to this bandwidth reduction appeared to be minor.

size of the available training set and the practical limits on network size coincided at about this ratio, using PLP-12 as the input feature. As part of our Broadcast News effort, we are also employing a different set of 28 features based on the modulation-spectrum, using a modified form of the approach described in [6]. While we have too few results to see if this ratio changes when evaluated over a different-sized vector of different features, it is clear from the experiments we have done that we do continue to derive improvements from increasing the network and training size.

Finally, although the error rate does continue to fall as we move to larger data sets and more parameters, examination of the leading diagonal for Table 1 shows that there does appear to be a diminishing of returns for this strategy. The error reduction for each doubling of both training set size and parameters goes from 9.3% for the first doubling down to 5.3% for the last. It may be that we are nearing the limits of potential improvements of this system without incorporating more structure. In fact, as previously noted, we are currently engaged in developing a joint system with our European partners in which we are merging estimators that often lead to different errors. Ultimately, this is likely to be the way in which we will incorporate an even larger number of parameters for improved recognition accuracy.

5. CONCLUSION

As stated in the title, it appears that over the range of parameters we investigated, size does matter, and the most obvious route to improving speech recognition, that of increasing the amount of training data and the number of classifier parameters, is still a viable course for the hybrid connectionist architecture. While our absolute system performance is not as good as some other more complex systems, it is notable how much can be achieved by this baseline. Routine refinements such as context-dependence, gender-dependence, feature adaptation (e.g. vocal-tract length normalization) and higher-order grammars can all be employed to improve performance. Also, simple model merging techniques using multiple hybrid HMM/ANN estimators form part of the overall Broadcast News evaluation effort we are conducting in collaboration with Cambridge University and our other partners.

Acknowledgments

We thank Eric Fosler and Adam Janin for their part in this work, and Jim Beck and David Johnson for the computational systems that we used. Additionally, we thank Steve Renals, Gethin Williams, Tony Robinson, and especially Gary Cook for a range of support that was essential for getting up to speed on the Broadcast News large vocabulary recognition task. This study was conducted with primary support from the European Community basic research grant SPRACH, under a subcontract from the Faculté Polytechnique de Mons in Belgium.

6. REFERENCES

[1] Bourlard, H., Morgan, N., Wooters, C., and Renals, S., "CDNN: A Context Dependent Neural Network for Continuous Speech Recognition" *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing,* San Francisco, II-349-352, 1992.

[2] Cohen, M., Franco, H., Morgan, N., Rumelhart, D., and Abrash, V., "Context-Dependent Multiple Distribution Phonetic Modeling", *Advances in Neural Information Processing Systems V*, pp. 649-657, 1993.

[3] Fritsch, J., "ACID/HNN: A Framework for Hierarchical Connectionist Acoustic Modeling," *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, eds. S. Furui, B.-H. Juang, and W. Chou, pp. 164-171, 1997.

[4] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," *Journal Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738-1752, 1990.

[5] Kershaw, D., Robinson, A., and Hochberg, M., "Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition Systems," *Advances in Neural Information Processing Systems VIII*, pp. 750-756, 1996.

[6] Kingsbury, B., Morgan, N., and Greenberg, S., "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25(1-2), August 1998.

[7] Morgan, N., and Bourlard, H., "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach." *Signal Processing Magazine*, pp 25-42, May 1995.

[8] Renals, S., and Hochberg, M., "Efficient Search Using Posterior Phone Probability Estimates," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing,* Detroit, Vol. 1, pp. 596-599, 1995.

[9] Steeneken, J.M. and Van Leeuwen, D.A., "Multi-lingual assessment of speaker independent large vocabulary speech-recognition systems: the SQALE project (speech recognition quality assessment for language engineering)," *Proceedings of EUROSPEECH'95* (Madrid, Spain), 1995.

[10] Wawrzynek, J., Asanović, K., Kingsbury, B., Beck, J., Johnson, D., Morgan, N., "SPERT-II: A Vector Microprocessor System," *IEEE Computer*, vol. 29, no. 3, pp 79-86, March 1996.

[11] Woodland, P., Hain, T., Johnson, S., Niesler, T., Tuerk, A., Whittaker, E., and Young, S., "The 1997 HTK Broadcast News Transcription System," *Proc. of the Broadcast News Transcription and Understanding Workshop,* Landsdowne, Virginia, pp 41-48, 1998.