References

[1] Bourlard, H., & Morgan, N. (1992). "CDNN: A Context Dependent Neural Network for Continuous Speech Recognition", *IEEE Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing*, pp. II.349-352, San Francisco, CA.

[2] Bourlard, H., & Morgan, N. (1994). *Connectionist Speech Recognition,* Kluwer Academic Press

[3] Hampshire, J.B., & Waibel, A. (1990). "Connectionist Architectures for Multi-Speaker Phoneme Recognition", *Advances in Neural Information Processing Systems 2*, D. Touretzky (ed.), Morgan Kaufmann, CA.

[4] Hermansky, H. (1990). "Perceptual Linear Predictive (PLP) Analysis of Speech", *Journal of the Acoustical Society of America,* 87:1738-1752.

[5] Jacobs, R., and Jordan M. (1993). "Linear Piecewize Control Strategies in a Modular Neural Network Architecture", *IEEE Trans. on Systems, Man, and Cybernetics,* March/April 1993, vol. 23, nr. 2, pp. 337-345.

[6] Konig, Y., and Morgan, N. (1993). "Supervised and Unsupervised Clustering of the Speaker Space For Connectionist Speech Recognition" *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing,*

[7] Krishnaiah, P.R., and Kanal, L.N., eds. (1982). *Classification, Pattern Recognition, and Reduction of Dimensionality.* Handbook of Statistics, vol. 2. Amsterdam: North Holland.

[8] Morgan, N,. Beck, J., Kohn, P., Bilmes, J., Allman, E., and Beer, J., "The Ring Array Processor (RAP): a multiprocessing peripheral for connectionist applications," *Journal of Parallel and Distributed Computing,* Special Issue on Neural Networks, vol. 14, pp.248-259, 1992

[9] Price, P., Fisher, W., Bernstein, J., and Pallett, D. (1988). "A Database for Continuous Speech Recognition in a 1000-Word Domain", *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing,* pp. 651.

[10] Robinson, T., Almeida, L., Boite, J.M., Bourlard, H., Fallside, F., Hochberg, M., Kershaw, D., Kohn, P., Konig, Y., Morgan, N., Neto, J.P., Renals, S., Saerens, M., & Wooters, C. (1993). "A Neural Network Based, Speaker Independent, Large Vocabulary, Continuous Speech Recognition System: The WERNICKE Project", Proceedings of *EUROSPEECH'93,* September 21-23, Berlin, Germany.

[11] Schwartz, R. (1993). Oral Presentation, *Speech Research Symposium XIII,* Johns Hopkins

four-cluster case is 7.5 hours for the rate-of-speech nets and 9 hours for the k-means nets (an average of two hours per net). The total training time for the two-cluster case is 18 hours for the rate-of-speech nets, and 11.5 for the k-means nets (average of 7.5 hours per net).

5 Discussion

In this paper, we have proposed a Parallel Net architecture for reducing the training time of the hybrid HMM/MLP system. Each of the experts in the system are trained on one region of the speaker space, hence making each net a *quasi*-speaker-dependent probability estimator. In our initial pilot experiments, we observed a strong gender effect. Also, there was strong evidence of over-tuning to the same category data. These two observations motivated us to restructure our experiments to reduce over-fitting, and to factor in gender effects.

We retrained the experts using same-gender SI cross-validation to avoid over-tuning. Also to further reduce over-fitting to the SD data, we cut the number of parameters by a factor of four and used a smaller learning rate. We experimented with different averaging schemes: weighted vs. equal, and average of scaled-likelihoods vs. sum of posteriors divided by sum of priors. The theory [see also Jacobs & Jordan, 1993] suggest that a non-uniform weighting mechanism is desirable. However, in our experiments, the weighted average was similar, if not worse, than an equal weighted average. This may only mean that we did not develop the correct method for determining the best weights in these examples; but in any event the evidence we have so far does not support computed weights, and equal weights in any event seem to work well enough to support a parallel approach. Also, we consistently got better results from averaging scaled likelihoods (equation 2 vs. 1).

The average error rate of the Parallel Net architecture was better than *both* the best SD system *and* the average error rate of all the female-SD systems, and the four-cluster male systems. Furthermore, the performance of the Parallel Nets was comparable to a single net trained on the aggregate training data. Given the shorter training time and the potential for taking advantage of parallel architectures, we believe that the Parallel Net architecture is the preferable architecture.

6 Acknowledgments

We would like to thank Eric Fosler, Yochai Konig, and Gary Tajchman for their help and advice. We also thank the National Science Foundation for its support through grant MIP-9311980, as well as general support from ICSI.

Word Recog Percent Error – Male Set				
System	Rate-of-Speech		K-means	
	2 CL	4 CL	2 CL	4 CL
PN, Eqn (2), eql wgts	8.1	7.0	8.8	8.0
PN, Eqn (2), gating, +smth	7.7	8.1	8.6	8.1
PN, Eqn (1), eql wgt	11.0	11.1	13.5	11.5
Best Net	7.6	7.7	8.2	9.8
Avg of Nets	9.4	9.1	10.6	11.9

Table 2: Word Recognition Percent Error for each of the systems tested on RM February 1991 SI male evaluation data. The "Best Net" column represents the error rate of the best single net. The "Avg of Nets" is the average of word recognition error of the nets.

4.2 Results

We used the male RM SI data for training, as mentioned above. Each of the nets in the two-cluster experiments were 512 hidden units each, and in the four-cluster experiments were 256 hidden units, making the number of parameters to be the same across all systems. Each net was trained on a partition of the training data with error back-propagation, started with a learning rate of 0.008, and was cross-validated on male data from February 1989 RM SI evaluation data to determine the stopping point for the training. The same data was also used for development purposes, for example setting the word transition penalty.

In order to perform a fair comparison between the Parallel Net architecture and monolithic net, we trained a 1000 hidden units net on all the male RM SI data. We tested all the systems on the male speakers of the February 1991 RM SI evaluation data. The word recognition error rate of our standard baseline system (which is trained on all genders of RM SI training data) was 8.9% for the males only. In comparison, the error rate of the monolithic all-male system was 7.3%, which is significantly better at $p < .05$ level.

The results of the Parallel Net architecture, presented in Table 2, are similar to that of the all-male monolithic net for the four-cluster cases, and the difference in error rates are not significant (7.0% and 8.0% for the four-clusters versus 7.3% for the male monolithic net). Weighted averaging gives worse results compared to the equal weighted averaging approach if the weights of the gating network are used directly. By introducing speaker continuity constraint and averaging the weights over a sentence (+smth row), the results of the weighting scheme improve and approach that of the equal weighting one.

The total training time for the monolithic net on our special purpose hardware [Morgan et al, 1992] is 18 hours. However, the total training time for the section of the speaker space. We experimented with two splitting criteria: rate-of-speech, and k-means clustering.

4.1 Splitting the Speakers

First, we used a priori knowledge about the domain and allocated the speakers to groups based on their rate-of-speech, where (inverse) rate-of-speech is measured as average number of seconds per word. In the second method, we use the k-means clustering [e.g., Krishnaiah & Kanal, 1982] algorithm.

4.1.1 Dividing the Space Based on Rate-of-Speech

Two observations motivated us to experiment with this split of the data. In the most recent ARPA WSJ evaluation, researchers reported significantly higher error rates on two fast speakers in the evaluation test set. The second motivation comes from our earlier results (section 3.2). We analyzed the relationship between the rate-of-speech of the female test speakers & the SD system's training data and word error rate. In order to have sufficient training data for each net, we chose to experiment with two and four clusters.

4.1.2 K-means Clustering

For the k-means clustering algorithm, we use a distance measure explained below. Let $X = \{X_1, X_2, ..., X_n, ..., X_N\}$ be the feature vector sequence corresponding to the speech of speaker S_x, where each X_n is a vector, $X_n = (x_{n1}, x_{n2}, ..., x_{nd}, ..., x_{nD})^t$. For each speaker S_k, we calculate a mean vector $\mu^j_k = (\mu^j_{k1}, ..., \mu^j_{kd}, ..., \mu^j_{kD},)^t$ and a covariance vector $\sigma^j_k = (\sigma^j_{k1}, ..., \sigma^j_{kd}, ..., \sigma^j_{kD})^t$[4] for each broad phonetic category $j = \{1...J\}$[5]. Define the distance between speaker S_x and speakers S_k as:

$$D(S_x, S_k) = \frac{1}{N} \sum_{n=1}^{N} \min_j \sum_{d=1}^{D} \left[\log \sigma^j_{kd} + \left(\frac{x_{nd} - \mu^j_{kd}}{\sigma^j_{kd}} \right)^2 \right] \tag{3}$$

So, for calculating the distance between two speakers, we use the μ's and σ's of one speaker, and the feature vectors of the other. Except for the distance measure, we follow the standard k-means clustering algorithm.

We can replace the gating network by using this distance measure. In order to determine the weights to use for each of the scaled likelihoods, we measure the distance of the unknown test speaker to the cluster centers and use an estimated probability (computed assuming a Gaussian distribution with a diagonal covariance matrix), the normalized $e^{-distance}$, as weight.

[4] Covariance matrix assumed to be diagonal.

[5] The five broad phonetic categories are based on the phonetic classes in the TIMIT set; they are fricatives, liquids, nasals, stops, and vowels.

speakers (from RM February 1989 SI evaluation set, as mentioned above).

For a fair comparison between the parallel architecture and a single-net system, we trained *one* net on the aggregate training data of the five SD systems, and cross-validated the training using female SI data (as explained above). We chose a 1000 hidden unit net that has about an equivalent number of parameters as the five female nets altogether. We bootstrapped this net from a TIMIT net in order to reduce the total training time. The initial learning rate was 0.008.

A net that was trained on the aggregate data of the SD nets, the female SI net, has an error rate of 7.4%. All of our experimental nets performed worse than this, but given a small data set of only 5 speakers, the results were not considered conclusive. However, the average error rate for the Parallel Nets using eqn (2) with an equal weight for each SD system is fairly close to the female SI net, with the Parallel Net having 13.5% more relative word error than female SI (8.4% versus 7.4%). This is not a statistically significant difference at the $p < .05$ level for this test set, so that in some sense there is no demonstrable difference in performance. The average performance of the Parallel Net architecture is better than *both* the average error of the SD systems' (13.0%) *and* the best SD system (9.7%) (significantly so for the average case).

Comparing the performance of our Female SI net with our baseline hybrid HMM/MLP system, we observe that Female SI has about 40% more relative error (7.4% versus 5.3%, which is significant at the $p < .05$ level) than the gender-independent SI net. This is unsurprising, since the baseline SI net is trained on over 30 speakers and is trained longer. The Female SI net, in contrast, is only trained on five speakers and goes through half as many epochs of training. The obvious remedy would be to train the Female SI net on more female speakers. In other words, train more SD systems to get a better representation of the sample space. Another possibility is to train each SD net on two or more same-gender speakers that are in the same region of the sample space, creating *quasi*-SD nets and increasing the coverage of the sample space that way. This conjecture was the basis for the main set of experiments.

4 Experiments and Results

In order to get a better representation of the speaker space, we increased the number of training speakers in the next experiment. We used the male speakers' data from the RM SI training set (November 1989), consisting of 49 male speakers, each uttering 40 sentences. Since there is little training data for each speaker, training 49 SD nets was not feasible. Instead, we can divide the speaker space based on some criterion and train one net on each

Gender Effects on Percent Word Error		
Training Speaker	Test Speaker	
	Male	Female
Male	45.4	111.7
Female	106.4	38.1

Table 1: Gender Effects on Word Error. This table shows the average error rate of SD Female (Male) nets when tested on SI Female (Male) data. Error rates of higher than 100% are due to counting insertions, deletions, and substitutions as errors.

There was an interesting unexpected result: the SD system with the worst recognition score on its own data generalized best to the speech of unknown speakers. On the other hand, a system which was almost perfectly tuned to speech from the same speaker generalized the worst. While not all the systems obeyed this rule, it was a general trend. This suggested that we should use SD nets that are not as fully tuned to the same speaker's data.

3.2 Retraining the Experts

In the next group of pilot experiments we examined the effect of using speaker-independent cross-validation to avoid overfitting to the speaker-dependent training data. We also reduced the number of parameters in contrast to the first experiment (again to combat over-fitting). We changed the size of the hidden layer from 1000 hidden units to 256 hidden units for each net. In order to reduce the training time, we bootstrapped each of the nets from a 256 hidden units net that was previously trained on the hand-labeled SI TIMIT database. Our training data for each net was 600 SD sentences. Same-gender SI data for cross-validation was chosen from the RM November 1989 SI training set: 460 sentences with 23 speakers for the female set, and 490 sentences with 49 speakers for the male set. We used a lower learning rate (α = 0.004) than for the *pretrained* nets. Each net went through only 1-2 epochs of training before cross-validation performance indicated that training should be stopped.

To estimate $P(M_i|x)$, we trained a *gating* network [Hampshire & Waibel, 1990]. We used a net with 10 hidden units, 234 input units, and n output units (where n is the number of nets). It was trained with back-prop to associate each feature vector with the label of the training speaker. In each of the experiments below, we have run Viterbi decoding on the output of each SD net and reported the results.

Based on the strong gender dependencies that we observed, we chose one gender (the female set) for the following experiments. The female set comprises five female SD systems (RM SD training data), and four female unknown test (nets) should be trained on subsets of data with different statistical properties. In all the experiments reported here, we use a speaker-dependent split in the training data.

3 Pilot Experiments

3.1 Using Pretrained Nets

In our first pilot experiment, we used twelve (five female and seven male systems) *pretrained* speaker dependent (SD) estimators, each of which were trained on data from one speaker of the RM SD November 1989. By *pretrained* we mean that the nets were previously trained to maximize the accuracy of SD recognition. Each net had 1000 hidden units, 61 outputs, 234 input units (= 26 PLP and delta PLP features[2] × 9 frame window size), and trained with phonetic labels which had gone through two iterations of forced viterbi realignment. 500 of the SD sentences were used for training, and 100 were held out for cross-validation. The nets were trained starting with a learning rate of 0.008, and the training took 4-6 epochs. The word recognition error rate of each system on the same speaker's test data (RM January 1990 SD evaluation — 25 sentences per speaker) ranges from 1.8% to 11.3%[3], while the error rate on the RM February 1989 SI evaluation test set (300 sentences, 10 speakers, 4 of which are female and 6 male) ranges from 64.6% to 82.0%.

We averaged (equal weighting) the scaled likelihoods of each of the SD subsystems using equation (2) and got 22.6% word recognition error, which is better than the performance of *both* the best SD sub-system *and* the average of all the SD sub-systems. However, this error rate is not comparable to that of a monolithic SI system (a net trained on RM November 1989 SI data), which has an error rate of about 5.1% for the same SI recognition task.

Upon analyzing the results, we came across striking gender effects, as shown in Table 1. Sub-systems trained on male speech generalized better to male speech than to female speech; vice versa for female nets. This provided motivation for another experiment: if the test speaker's gender is female (male), we only allowed the probabilities generated by female (male) systems to take part in calculating the average scale likelihood. Since this was a pilot experiment, the gender of the test speakers were *known* to the system. It is possible, however, to build a gender detector which reliably (approx. 98% accuracy) detects the gender of the test speaker [Konig & Morgan, 1993]. Table 1 shows the strength of this effect. Averaging in a gender-based way further decreases the overall word recognition error to 16%.

[2] PLP stands for Perceptual Linear Predictive analysis [Hermansky, 1990]

[3] It should also be noted that these *pretrained* nets were trained approximately two years ago so that the raw error numbers are probably somewhat higher than our current systems would achieve.

highly-dimensioned joint probabilities, such as the scaled likelihood of the data including a large acoustic context. Additionally, the MLP training is inherently discriminant, making effective use of parameters for limited training data, and estimating relatively detailed densities without strong parametric assumptions. Over the last few years, we have observed in a number of instances that the direct substitution of such an estimator for a tied-mixture module has resulted in significant improvements.

We are currently using a recognizer called Y0 (described in [Robinson et al., 1993]), which uses a single density per phone with repeated states for a simple durational model. The densities are trained with no explicit incorporation of phonemic context (e.g., triphones). Our current results on DARPA Resource Management (RM)[1] test sets show a performance that is comparable to that of much more complex context-dependent systems; the recognition word error on the February 1989 test set of the baseline hybrid HMM/MLP system, used here for comparison, was equal to 5.1% (including insertions, deletions, and substitutions).

In the current experiments, we train multiple networks on separate partitions of the training set. If M_i represents cluster i, let $P(M_i)$ be the probability that M_i is a better match than $M_k, \forall k \neq i$. If all M_i's are mutually exclusive, and cover all possible cases, $\sum_{i=1}^{n} P(M_i) = 1$, we can calculate the likelihoods (within a constant factor $P(x)$) by:

$$P(x|q_k) = \frac{\sum_{i=1}^{n} P(M_i|x)P(q_k|x, M_i)}{\sum_{i=1}^{n} P(M_i)P(q_k|M_i)} \tag{1}$$

where q_k is an HMM state, M_i represent each of the n MLPs, and $P(M_i|x)$ is the probability that M_i is the "correct" estimator of the sound class, given the data x. For instance, in the case of a male/female partition, this probability would be the probability that the speaker is male or female (2 probabilities that sum to 1). As an alternative, we calculate a weighted average of the scaled likelihoods:

$$P(x|q_k) = \sum_{i=1}^{n} P(M_i|x)\frac{P(q_k|x, M_i)}{P(q_k|M_i)} \tag{2}$$

In some of our experiments, we have used a simplified form of the above formula and inserted an equal weight averaging factor of $1/n$ instead of $P(M_i|x)$. This amounts to the following two assumptions: that $P(M_i|x)$ is independent of the data, and that all priors of M_i are equal (i.e., male prior equal to female prior).

Mixture of experts approaches are most effective when each expert has different statistical properties and biases. Therefore, each of our sub-systems

[1] This is a speaker-independent continuous speech recognition task that has a vocabulary of roughly 1000 words and uses a word-pair grammar with a perplexity 60; it is described in many places in the speech literature, including [Price et al, 1988].

on disjoint elements in the training set, and then combined in some manner, communication is minimized. Additionally, there is some hope that the right partitioning and weighting of the separate estimates could provide some improvement in performance; for instance, in the gender case, separating male and female training data has some demonstrable advantages.

There is some evidence that data splitting should at least provide equivalent performance. In one report at a recent SRS meeting, R. Schwartz of BB&N described an experiment in which Hidden Markov Model (HMM) Gaussian mixture parameters were separately estimated for individual speakers and then averaged [Schwartz, 1993]. The resulting system was comparable in performance to a more standard estimator that was trained on the pooled data from all speakers. Of course, this experiment reported the estimation of data likelihoods and the averaging of Gaussian parameters, and this does not necessarily show that a posterior estimator like a Multi-layer Perceptron (MLP) will permit a similar parallelization. However, it suggests that a test is worthwhile. Recent results in applying the split net strategy in control theory [Jacobs & Jordan, 1993] are another indication that such approach may be advantageous.

Another related effort was that of the Meta-Pi network [Hampshire & Waibel, 1990]. In this approach, speaker-dependent estimators for voiced stop consonant probabilities were weighted and summed with gating elements trained with error back-propagation. For a source dependent speaker (i.e., one of the six training speakers), the performance of the Meta-Pi architecture on a six speaker three phone (/b,d,g/) task was comparable to a speaker dependent system. For a source independent (i.e., unknown) speaker, however, the error rate was almost tripled.

In the work described here, we also are using an MLP trained with back-propagation; however, these estimators are trained to be discriminant for the 61 phone set of TIMIT. We have focused our efforts on the speaker-independent case. In other words, we would like our mixture of estimators to perform at least as well as a monolithic estimator (which was trained on all of the data) when tested on an unknown speaker (which was not present in the training data).

2 Approach

In our experiments we use estimators that are based on the hybrid HMM/MLP method as explained in [Bourlard & Morgan, 1994]. The main idea in this method is to train an MLP (using a squared error or relative entropy criterion) for phonemic classification; such a net can be used as an estimator of posterior class probabilities, and when divided by class priors can estimate scaled likelihoods. The MLP estimator has the potential advantage (over standard Gaussian or Gaussian mixture estimators) of the ability to estimate

Parallel Training of MLP Probability Estimators for Speech Recognition: A Gender-Based Approach

Nikki Mirghafori, Nelson Morgan, and Hervé Bourlard
International Computer Science Institute, Berkeley, California
1947 Center St, Suite 600
Berkeley, CA. 94704
Tel: (510) 642-4274, FAX: (510) 643-7684
{nikki, morgan, bourlard}@icsi.berkeley.edu

Abstract

In this paper we explore the averaging of mixtures of multiple neural network probability estimators in speech recognition. We experiment with different ways of dividing up the speaker space. A division based on gender seems to be the most important. The division based on a priori knowledge (in our case, rate of speech) resulted in lower error rates than the use of k-means clustering. The overall accuracy of the Parallel Net architecture is about the same as the monolithic probability estimator, but communication costs on parallel machines can be expected to be significantly lower. Additionally, the overall product of patterns times parameters is lower with such a partitioning, resulting in reduced training time even on serial machines.

1 Introduction

In previous work, we have examined the factorization of Multi-layer Perceptrons (MLPs) that are viewed as probabilistic estimators. In two particular cases, we partitioned out the influence of phonemic context [Bourlard & Morgan, 1992], and of speaker gender [Konig & Morgan, 1993]. These partitionings permitted the evaluation of the posterior probabilities of a large number of classes without the explicit computation of a huge output layer.

In our current work, we are interested in partitioning not only the network estimators, but also the training data. This is of increasing relevance as we move to larger and larger data sets. Cycling through these data requires more than a linear increase in computation, as the estimators themselves should (ideally) be expanded to a greater number of parameters in order to take advantage of the increased coverage in the training materials.

Parallelism is a potential remedy for this increased computational burden, but depending on the machine and the algorithm, communications costs can overwhelm any advantage due to numerical parallelization. Training set parallelism is a potential cure for this difficulty. If multiple estimators are trained