CONNECTIONIST MODEL COMBINATION FOR LARGE VOCABULARY SPEECH RECOGNITION

M. M. Hochberg G. D. Cook S. J. Renals A. J. Robinson
Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, England

Abstract–Recent reports in the statistics and neural networks literature have expounded the benefits of merging multiple models to improve classification and prediction performance. The Cambridge University connectionist speech group has developed a hybrid connectionist-hidden Markov model system for large vocabulary, talker independent speech recognition. The performance of this system has been greatly enhanced through the merging of connectionist acoustic models. This paper presents and compares a number of different approaches to connectionist model merging and evaluates them on the TIMIT phone recognition and ARPA Wall Street Journal word recognition tasks.

INTRODUCTION

An acoustic pre-processor or *front-end* is a common feature of all large vocabulary speech recognition systems. The front-end maps the sampled waveform onto a lower-dimensional representation of the acoustic signal. Typically, the specific mapping is selected as the front-end which performs best on some development test set. Since different front-ends may provide better representations for different acoustic events (e.g., phoneme class, talker, etc.), it would seem advantageous to sensibly merge multiple front-ends and their associated models.

There has been speech recognition research into merging multiple sources of information. For example, work at BBN has addressed merging the parameters of speaker-dependent hidden Markov models (HMMs) to obtain a speaker-independent system [1] and Cohen and Franco at SRI have merged a conventional HMM and multi-layer perceptron [2]. Recently, model combination has been shown to be a promising area of neural network research. Techniques such as *Generalized Stacking* [3] and Bayesian approaches [4] have been explored as a means to most effectively utilize all the available information. This paper presents an application of connectionist model merging to speech recognition. Multiple acoustic representations are merged resulting in a significant reduction in the recognition error rate.

THE HYBRID CONNECTIONIST-HMM

The hybrid connectionist-HMM employs the same basic framework as described in [5], but utilizes a different connectionist component. The speech recognition sys-

0-7803-2026-3/94 $4.00 © 1994 IEEE

tem uses a recurrent network to map a sequence of acoustic vectors to a sequence of posterior phone probabilities. The network outputs are used as estimates of the observation probabilities within an HMM framework, i.e., the observations are considered as a stochastic process on a non-observable, first-order Markov chain. Given new acoustic data and the connectionist-HMM framework, the maximum *a posteriori* phone or word sequence is then extracted using standard Viterbi decoding techniques.

The basic acoustic modeling system is illustrated in Figure 1. At each 16ms frame, an acoustic vector, $\mathbf{u}(t)$, is presented at the input to the network along with the previous state vector, $\mathbf{x}(t-1)$. These two vectors are passed through a single-layer, fully-connected, feed-forward network to give the output vector, $\mathbf{y}(t)$, and the next state vector, $\mathbf{x}(t)$. Forward acoustic context is modeled by expanding the input vector to cover additional frames and by delaying the target. The state vector provides the mechanism for modeling the dynamics of the acoustic signal in various contexts.

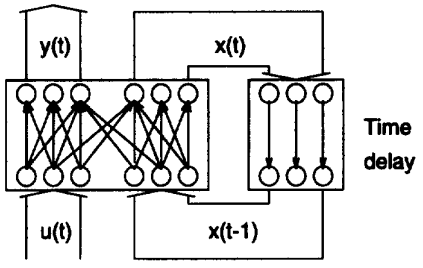

Figure 1: The recurrent net used for phone probability estimation.

Each output channel corresponds to a particular phone in the phone set. The use of the *softmax* nonlinearity for the output nodes with the cross-entropy training criterion implies that the outputs can be considered estimates of the posterior probability of the phones given the (local in time) acoustic data. This network is trained by back-propagation through time. (A more complete description of the network may be found in [6].)

THE MODELS

Because the goal of this work is to reduce the recognition error rate through merging multiple recurrent networks, it is important that each portion of the speech signal can be modeled by at least one of the individual networks. In the experiments presented here, the parameters for each network are estimated on the same speech data, but processed with different front-ends. Two successful spectral representations have been found to be a 20 channel mel-scaled filter bank with voicing features [7] and 12th order cepstral coefficients derived from perceptual linear prediction [8]. The filter bank and cepstra are referred to in this paper as MEL+ and PLP, respectively. In addition, because the recurrent network is time asymmetric, training the network to classify

forward in time will result in different dynamics than training to classify backwards in time. Based on the above considerations, four networks were constructed from the possible representations; FORWARD MEL+, BACKWARD MEL+, FORWARD PLP, and BACKWARD PLP.

MODEL COMBINATION

Probability Domain Merging

The most straightforward approach to merging the recurrent networks is through a linear combination of the model outputs. In the most general framework, the merged estimate of the posterior probability of phone i given the acoustic data up to time t is given by

$$y_i(t) = \sum_{k=1}^{K} \beta_{ik} y_i^{(k)}(t) \tag{1}$$

where $y_i^{(k)}(t)$ is the estimate of the kth model and β_{ik} are the merging weights. Note that the weights can be dependent on the input data, e.g., $\beta_{ik} = \beta_{ik}(\mathrm{u}(t))$. Sufficient conditions on the βs to guarantee a statistical interpretation of the output are that they are *tied* across phones (i.e., $\beta_{ik} = \beta_k$), *sum-to-one* (i.e., $\sum_k \beta_{ik} = 1$), and are *non-negative*. With these conditions, the merged output will meet the constraints needed for interpretation of the output as the posterior phone probabilities. As is seen in the results section, relaxing these constraints does not necessarily lead to poorer performance.

Log-Probability Domain Merging

For computational reasons, the mapping of the phone probabilities into recognized word strings is usually performed in the log-probability domain. This has led to experiments evaluating merges performed after the conversion of the network output to the log domain, i.e.,

$$\log y_i(t) = \sum_{k=1}^{K} \beta_{ik} \log y_i^{(k)}(t). \tag{2}$$

With this approach, it is difficult to assign a probabilistic interpretation to the merged outputs. However, if the models are assumed to be independent, then the estimated joint likelihood of the different data is proportional to the product (or sum in the log-domain) of the network outputs.

Merge Criteria

Given the connectionist-HMM framework, there are number of different approaches to determine the βs. In all cases where training data was required to learn the merge

parameters, the data was taken from an independent development set. Although the amounts of data in the training set was quite large, this approach was taken to further reduce the chance of obtaining a merge with substantial bias.

Uniform. The first attempt at combining networks assumed the merge weights were independent of the data with uniform probabilities, i.e., $\beta_{ik} = 1 / K$. This approach maintains the probabilistic interpretation of the merged output in the probability domain. Good initial results using this simple merging approach [9] has led to the evaluation of more complex merging techniques.

Linear Regression. Recent work has shown that merging regression predictors through linear regression (referred to as *Stacked Regressions*) produce an estimator that is better than any of the individual estimates [10]. The regression approach determines the βs through minimizing the sum-squared error

$$\sum_t \sum_i \left(\hat{y}_i(t) - \sum_{k=1}^{K} \beta_{ik} y_i^{(k)}(t) \right)^2 \tag{3}$$

on a development set. Here, $\hat{y}$ is the desired target and the regression parameters, β_{ik}, are assumed to be fixed after training. In [10], Breiman found that constraints on the βs improved performance. In this paper, the regression merging is evaluated with and without constraints such that the merge weights are tied across models and/or sum-to-one. It was rarely found that any of the merge parameters were ever less than zero.

Mixture of Experts. This framework (see Figure 2) employs a gating network to determine data-dependent merge parameters. The approach is equivalent to Jordan and Jacob's *mixture of experts* [11] with fixed experts. The data-dependent merging coefficients can be determined by maintaining a probabilistic interpretation and employing the expectation-maximization (EM) algorithm [12]. Let $U = \{\mathbf{u}(t)\}$ be the set of acoustic training data for each frame and let $C = \{c(t)\}$ be the corresponding phone. Assuming each frame is independent results in the likelihood $L(U)$ given as

$$L(U) = p(U|C, Y) = \prod_{t=1}^{T} p(\mathbf{u}(t)|c(t), \mathbf{y}^{\mathcal{M}}(t)) \tag{4}$$

where $Y = \{\mathbf{y}^{\mathcal{M}}(t)\}$ represents the outputs of all the models, i.e, $\mathbf{y}^{\mathcal{M}}(t) = \{\mathbf{y}^{(m)}(t)\}$. The merging comes about by assuming that $p(\mathbf{u}(t)|c(t), \mathbf{y}^{\mathcal{M}}(t))$ is a mixture density of the form

$$p(\mathbf{u}(t)|c(t), \mathbf{y}^{\mathcal{M}}(t)) = \sum_{k=1}^{K} p(\mathcal{M}_k|c(t), \mathbf{y}^{\mathcal{M}}(t)) p(\mathbf{u}(t)|\mathcal{M}_k, c(t), \mathbf{y}^{\mathcal{M}}(t)) \tag{5}$$

where $\mathcal{M}_k$ represents the kth model. Here, the mixing coefficients $p(\mathcal{M}_k|i, \mathbf{y}^{\mathcal{M}}(t)) = \beta_{ik}(\mathbf{u}(t))$. As in [11], a generalized linear model is used as the gating network to compute $\beta_{ik}(\mathbf{u}(t))$.

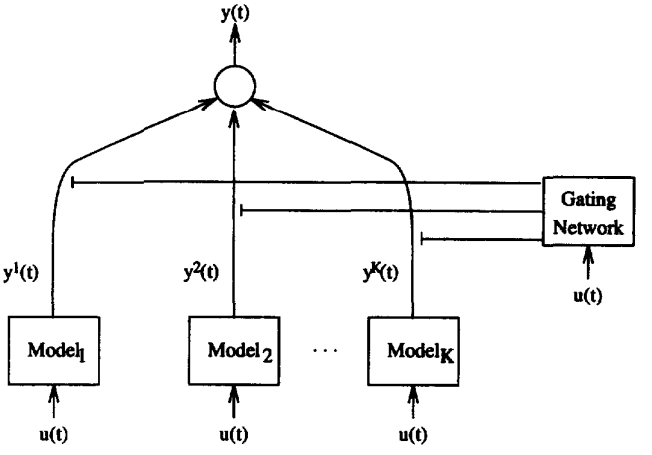

Figure 2: Mixture-of-experts framework.

The generalized EM algorithm [12] is an iterative approach used to compute the maximum likelihood estimates of the gating network parameters. Each iteration applies two conceptual steps. The E-step computes the posterior probability of each model

$$p(\mathcal{M}_k|\mathbf{u}(t), c(t)) = \frac{\beta_{ik}(\mathbf{u}(t))y^{(k)}_{c(t)}(t)}{\sum_{n=1}^{K}\beta_{in}(\mathbf{u}(t))y^{(n)}_{c(t)}(t)} \tag{6}$$

for each pair $\{\mathbf{u}(t), c(t)\}$ in the development set. The M-step estimates the parameters of the generalized linear model using the Iteratively Re-weighted Least Squares procedure (IRLS) [11] with $\mathbf{u}(t)$ as the inputs and $p(\mathcal{M}_k|\mathbf{u}(t), c(t))$ as the desired outputs. This procedure results in a method for learning the parameters of the gating network for each phone. The procedure insures that the merging weights do sum-to-one.

The standard mixture of experts approach has the weights tied across models. This is accomplished by assuming $p(\mathcal{M}_k|c(t), \mathbf{u}(t)) = p(\mathcal{M}_k|\mathbf{u}(t))$ and results in many fewer free parameters. A variation of this approach is to replace the input of the gating network with the output of one of the networks. In the experiments described later in the paper, the gating network inputs were either three contiguous frames of the acoustic feature vector or a single frame of a network output.

In addition to the above variations, the case where there are no inputs was also considered. In this case, the gating network outputs constant values and the EM algorithm [12] specifies an iterative solution for the maximum likelihood coefficients. The parameter update equation becomes simply

$$\hat{\beta}_{ik} = \frac{1}{T}\sum_{t=1}^{T}\frac{\beta_{ik}y_i^{(k)}(t)}{\sum_{n=1}^{K}\beta_{in}y_i^{(n)}(t)}\delta_{c(t),i} \tag{7}$$

where $\hat{\beta}$ represents the updated estimate and δ is the Kronecker delta function.

EXPERIMENTAL RESULTS

Recognition Tasks

TIMIT. TIMIT is one of the standard speech corpora for the evaluation of phone recognition systems. It is divided into 462 training speakers and 168 test speakers. Each speaker utters two calibration sentences and eight sentences that are used in these evaluations, giving a training set of 3696 sentences and 1344 test sentences. In the experiments described here, 1152 of the test sentences were used for cross-validation estimation of the merging parameters and 192 (the core test) sentences were used for testing.

Wall Street Journal. The Wall Street Journal (WSJ) is the current ARPA large-vocabulary recognition task. The training data used was the short-term speakers from the WSJ0 corpus consisting of 84 speakers uttering a total of 7,200 sentences. The November 1993 spoke 5 development test data was used for estimation of the merging parameters. This data was collected from 10 talkers and 216 sentences using a Sennheiser microphone. Results are reported for the November 1993 spoke 6 development test. This test has 202 sentences from the same 10 talkers as spoke 5. The test are from a closed 5,000 word, non-verbalized punctuation vocabulary using the standard bigram language model [13].

Results and Analysis

Tables 1 and 2 show the TIMIT and WSJ results for the various approaches to model merging. In the tables, *frame rate* is the classification rate of the merged system on the development data, *error rate* is the phone or word recognition error rate on the test set computed as

$$100 \times \frac{\#\text{ insertions} + \#\text{ deletions} + \#\text{ substitutions}}{\#\text{ phones}} \tag{8}$$

and *improvement* is measured relative to the average error rate. For the EXPERTS merges, ACOUS., PROB., MEL+, and PLP indicate the type of inputs to the gating network. For the TIMIT experiment, only the FORWARD AND BACKWARD MEL+ front-ends were merged.

The tables clearly show the benefits of model merging. Each of the networks trained on different front-ends have similar performance, but the frame rate is substantially improved by merging the network outputs. This improvement is reflected in the error rate by a reduction of 9% and 27% for the TIMIT and WSJ tasks, respectively. For both tasks, the simple uniform merging accounts for most of the improvement and the best results were achieved by merging in the log-probability domain.

For the regression merge approach, not much variation in either the frame rate or the recognition error rate is observed across the different types of constraints.

Merge Type	Constraints	Frame Rate %	Error Rate %	Improv. %
FORWARD ONLY	-	65.9	31.7	-
BACKWARD ONLY	-	65.7	31.8	-
AVERAGE	-	65.8	31.8	-
UNIFORM	Tied, Sum	69.3	29.4	7.5
UNIFORM (LOG)	Tied	69.2	29.0	8.8
REGRESSION	Tied, Sum	69.3	29.3	7.9
REGRESSION	Sum	69.3	29.3	7.9
REGRESSION	Tied	69.3	29.1	8.5
REGRESSION		69.7	29.3	7.9
EXPERTS (ACOUS.)	Tied, Sum	69.3	29.2	8.2
EXPERTS (ACOUS.)	Sum	69.5	29.4	7.5
EXPERTS (PROB.)	Tied, Sum	69.4	29.1	8.5
EXPERTS (PROB.)	Sum	69.0	29.5	7.2

Table 1: TIMIT phone recognition results for different merge approaches. Frame rate is computed on development data and error rate is computed on test data.

Merge Type	Constraints	Frame Rate %	Error Rate %	Improv. %
FORWARD MEL+	-	78.1	15.0	-
FORWARD PLP	-	76.6	15.1	-
BACKWARD MEL+	-	73.8	15.5	-
BACKWARD PLP	-	76.1	14.4	-
AVERAGE	-	76.2	15.0	-
UNIFORM	Tied, Sum	82.5	11.4	24.0
UNIFORM (LOG)	Tied	82.8	11.0	26.7
REGRESSION	Tied, Sum	82.5	11.5	23.3
REGRESSION	Sum	82.8	11.3	24.7
REGRESSION	Tied	82.6	11.7	22.0
REGRESSION		83.1	11.4	24.0
EXPERTS (MEL+)	Tied, Sum	82.7	11.4	24.0
EXPERTS (PLP)	Tied, Sum	82.7	11.4	24.0

Table 2: WSJ word recognition results for different merge approaches. Frame rate is computed on development data and error rate is computed on test data.

This indicates that over-fitting of the training data does not seem to be a problem. Examination of the sum-squared error obtained from (3) in the merge process also shows little variation for the different constraints or from the uniform case. This implies that – at least for these networks – little improvement over the uniform merge can be expected.

TIMIT results obtained with the mixture of experts approach show that a single gating network achieves better performance than a set of separate gating networks for each phone. This is most likely due to insufficient training data to estimate the multiple gating network parameters. Even with large amounts of training data, some phones occur very infrequently which makes it difficult to estimate the parameters of a gating network. Conditioning the mixture of experts gating network on the acoustic signal or network output achieved similar performance on TIMIT. For WSJ, using MEL+ or PLP features as inputs to the gating network had no effect on the recognition results.

As indicated in Tables 1 and 2, simple model merging improves performance but the use of more complex merging strategies does not significantly improve the recognition results. Analysis of the TIMIT task indicates that the different merge types are all reasonably close to the optimal merge. Figure 3 shows the results of a line search on the merge parameter with the tied and sum-to-one constraints. It

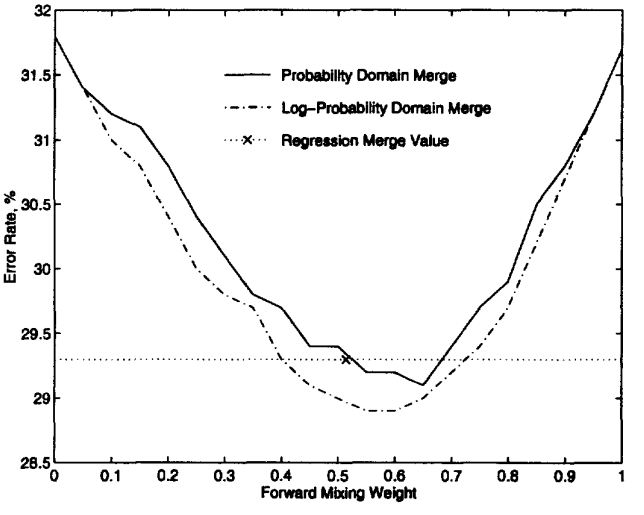

Figure 3: Error rate versus forward network mixing coefficient for probability and log-probability domain mixing on the TIMIT task.

is easy to see that the best performance is certainly in the region around 0.5 (the uniform merge). The regression estimate of the merge parameter shown in the figure is 0.51 and the mixture of experts has a mean value of 0.52 with variance 0.005. This implies that better/additional acoustic models are necessary to greatly improve the

TIMIT results.

To determine which front-end merge provides the most improvement, forward and backward models with the same spectral representation were merged. Similar merges were performed across spectral representations with the same time indexing. The results are shown in Table 3 and indicate that both the variation in spectral representation and processing of different data are important to the merging process. In addition, merging all front-ends resulted in better performance than any of the subset of the front-end merges.

Merge Type	ins. %	sub. %	del. %	errors %
AVERAGE FRONT-END MERGE	1.0	8.5	3.4	12.8
AVERAGE TIME MERGE	0.8	8.3	3.6	12.7

Table 3: Connectionist model subset merging results on the WSJ word recognition task.

DISCUSSION

This paper investigated various approaches to merging multiple, different acoustic models within the hybrid connectionist-HMM framework. Given the chosen acoustic models (recurrent networks), it was found that

- merging results in a significant reduction in error rate,
- the uniform, linear regression, and mixture of experts approaches all had similar performance, and
- the log-probability domain merging gave consistently better results.

The results presented here indicate the potential of this model merging approach. The fact that the linear regression and mixture of experts approaches did not do much better than the uniform merge may be a result of the selected networks. These techniques should show more significant gains when merging networks with different performance levels. As Figure 3 shows, the uniform merge of the log-domain probabilities may not be the best choice and research is planned in this area. In conclusion, this work shows model merging within the hybrid connectionist-HMM framework to be a very powerful mechanism for improving speech recognition performance. TIMIT results obtained with the merged system are the best known to the authors. Even with orders of magnitude fewer parameters, the merged system is competitive with state-of-the-art HMM systems on the WSJ task.

ACKNOWLEDGEMENTS

This work was partially funded by ESPRIT project 6487 (WERNICKE). Two of the authors (T.R. and S.R.) are supported by SERC fellowships. The authors would like to acknowledge MIT Lincoln Laboratory for providing the language model and Dragon Systems for providing the pronunciation lexicon for the WSJ task.

REFERENCES

[1] F. Kubala and R. Schwartz, "A new paradigm for speaker-independent training," in *1991 International Conference on Acoustics, Speech, and Signal Processing*, (Toronto, Canada), pp. 833–836, IEEE, May 1991.

[2] S. Renals, N. Morgan, M. Cohen, and H. Franco, "Connectionist probability estimation in the Decipher speech recognition system," in *1992 International Conference on Acoustics, Speech, and Signal Processing*, (San Francisco, California), pp. 601–604, IEEE, Mar. 1992. Volume 1.

[3] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[4] W. Buntine, "Learning classification trees," in *Artificial Intelligence Frontiers in Statistics III* (D. J. Hand, ed.), pp. 182–201, Chapman & Hall, 1993.

[5] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. The Kluwer International Series in Engineering and Computer Science. VLSI, Computer Architecture, and Digital Signal Processing, Boston, Massachusetts: Kluwer Academic Publishers, 1994.

[6] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, pp. 298–305, Mar. 1994.

[7] T. Robinson, "Several improvements to a recurrent error propagation network phone recognition system," Tech. Rep. CUED/F-INFENG/TR.82, Cambridge University Engineering Department, Sept. 1991.

[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[9] M. M. Hochberg, S. J. Renals, and A. J. Robinson, "ABBOT: The CUED hybrid connectionist-HMM large-vocabulary recognition system," in *Proc. of Spoken Language Systems Technology Workshop*, ARPA, Mar. 1994.

[10] L. Breiman, "Stacked regressions," Tech. Rep. 367, Department of Statistics, University of California, Berkeley, August 1992.

[11] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, Mar. 1994.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *J. Roy. Statist. Soc.*, vol. B39, pp. 1–38, 1977.

[13] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Fifth DARPA Speech and Natural Language Workshop*, (Harriman, New York), pp. 357–362, DARPA, Morgan Kaufman Publishers, Inc., Feb. 1992.