Speech Recognition with a Cortex Model: Preliminary Results and Outlook

Todd K. Leen, Ronald Cole,
Dan Hammerstrom, Jon Inouye

Oregon Graduate Institute
Department of Computer Science
and Engineering
19600 N.W. von Neumann Drive
Beaverton, OR 97006-1999 USA

Technical Report No. CS/E 90-022

June, 1990

Speech Recognition with a Cortex Model: Preliminary Results and Outlook[1]

Todd K. Leen, Ronald Cole, Dan Hammerstrom, Jon Inouye
Oregon Graduate Institute, 19600 N.W. von Neumann Dr.,
Beaverton, OR, 97006-1999

June 6, 1990

tleen@cse.ogi.edu

This is a brief summary of the results obtained in applying the model of Ambros-Ingerson, Granger and Lynch [1] to phonetic discrimination tasks.

Training the Model for Phonetic Discrimination

The model of reference [1] performs top-down, hierarchical clustering of a distribution of input patterns by successive partitionings of the input space. At each level in the hierarchy a devoted group of cells performs a metric clustering of the input patterns presented. The clustering is arrived at by competition and Hebbian learning.

A network of n groups of cells performs an $n-$level partitioning which is conveniently represented by a dendrogram. For low-dimensional ($D \leq 3$) data representations, the partitioning can be represented directly by plotting the input patterns as points, shaded to indicate which cells in a group are maximally activated by the pattern (see Fig. 1). We have used such plots to track the convergence of the algorithm and otherwise verify the simulation software.

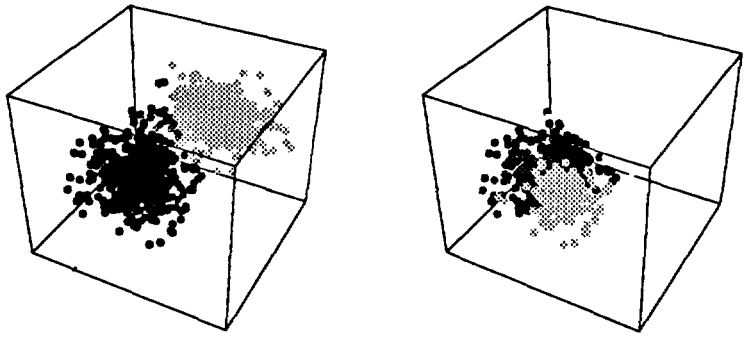

Fig. 1 Partitioning of 3-D data in the first two levels of the hierarchy

[1]This work was supported by the Office of Naval Research under contract N00014-90-J-1349, and by DARPA through a grant to the Department of Computer Science and Engineering.

Following training, the network response to an input pattern is a set of activated cells; one cell in each group is maximally active. Each input pattern is thus mapped to a binary signature. We describe below how the signatures are used for classification.

For our preliminary experiments we are using vowels extracted from spoken letters [2]. The data base consists of 52 utterances from each of 120 native English speakers. The data base is divided into four equal parts, each corresponding to 30 speakers. Subsets of the first three are used for training, the fourth is used for testing the trained network. (Note that the utterances in the test set are from speakers *not* included in the training set. This tests the model for *speaker-independent* recognition.)

Each utterance was digitized to 16 bit accuracy at 16kHz sampling rate. A DFT was computed over 10 ms sampling windows, at 3 ms time increments. Our final pattern vectors are the lowest 32 DFT coefficients, time-averaged over the center 1/3 of the vowel. These coefficients span the frequency range from 0 to 4 kHz.

We describe results of experiments conducted on the vowels in the letters **A**, **E** and **F** . This set was chosen as it presents a fairly difficult phonetic discrimination task. Figure 2 is a dendrogram of the signatures produced in response to the training data. Each leaf of the dendrogram corresponds to a binary signature. The leaves are labeled according to the letter from which the vowel was extracted ($1 \leftrightarrow$ **A**, $2 \leftrightarrow$ **E**, $3 \leftrightarrow$ **F**). Each leaf is also labeled (in parenthesis) by the number of training examples with the corresponding signature.

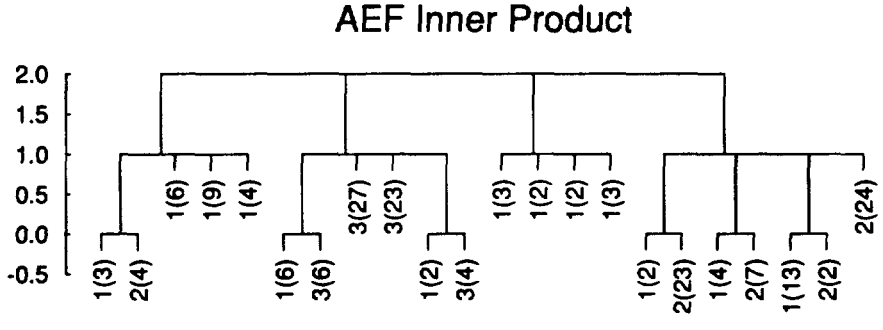

Fig. 2 Dendrogram of signatures produced by the training set.

Pairs of leaves that are joined at the lowest level ("0" on the vertical scale) correspond to signatures that are common to two different vowels. In such cases,

the model incorrectly clusters both vowels in the same class. Presumably, adding hierarchical levels would help eliminate such confusions.

We display the statistical properties of the distribution of input patterns by scatter plots of the data. To identify the important degrees of freedom, principal components of the training data were computed. Scatter plots of the first two principal components of the data are shown in Fig. 3. The classes are fairly separable, though not linearly. The figure also shows a fair amount of overlap between the classes.

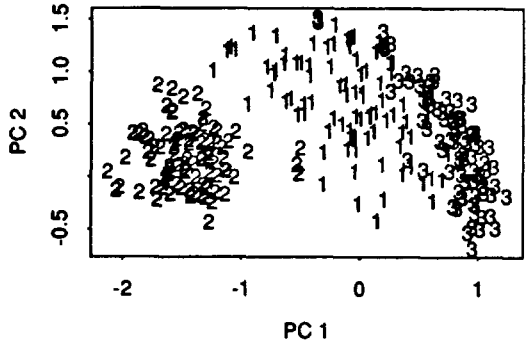

Fig. 3 Scatter plots of the first two principal components of the vowel data.

Classification Performance

To evaluate the model as a classifier, each pattern from the test set is fed to the trained network. The signature (activation pattern) produced is then compared to the set of signatures produced by the training set. If the test pattern signature matches one of the training signatures, then the test pattern is assigned to the same vowel that produced the training signature. If the test pattern matches a signature common to two different vowels (two leaves of the training dendrogram joined at level "0"), then that test pattern is assigned to the vowel most frequently represented in that signature.

Occasionally, test patterns produce signatures *not* generated during the training. We have employed two techniques to disambiguate this situation. In the first method, the offending winning cells are disregarded and the cells with the next highest activation are chosen to contribute to the signature. This process is repeated until we arrive at a signature that matches one generated during the training. The test pattern is then classified as described above. In the second

method, the leading bits of the signature that match signatures from the training set are retained, and the bits beyond the mismatch are discarded. The bits retained correspond to the grossest level of partitioning. The resulting truncated signatures reach part-way down the training set dendrogram. The test pattern is then assigned to the vowel most frequently represented in the sub-tree below the mismatch point. Both of these techniques yield similar classification performance.

Figure 4 shows the classifier performance as a function of the number of training epochs. The performance varies somewhat with changes in the network configuration (number of cells in each hierarchy). We have explored the use of both inner-product and Euclidean-distance activation rules. The plot in Fig. 4 shows performance with the inner-product activation. The mean performance is similar for both, but the inner-product activation rule produces a wider variation in performance with changes in the amount of training data used. The three curves give the test set results for networks trained on 1/3, 2/3 and 3/3 of the training corpus. The peak classification performance for this experiment is 95.6%.

In comparison, feed-forward networks trained on the same data offer somewhat better performance. Networks with 6 and 9 hidden nodes, trained by a conjugate gradient algorithm scored at 97.78% on the test data. This performance advantage is not surprising, given the ability of such networks to reproduce arbitrary mappings. We are encouraged by our results, particularly in consideration of the computational simplicity and rapid training of this algorithm relative to backpropagation.

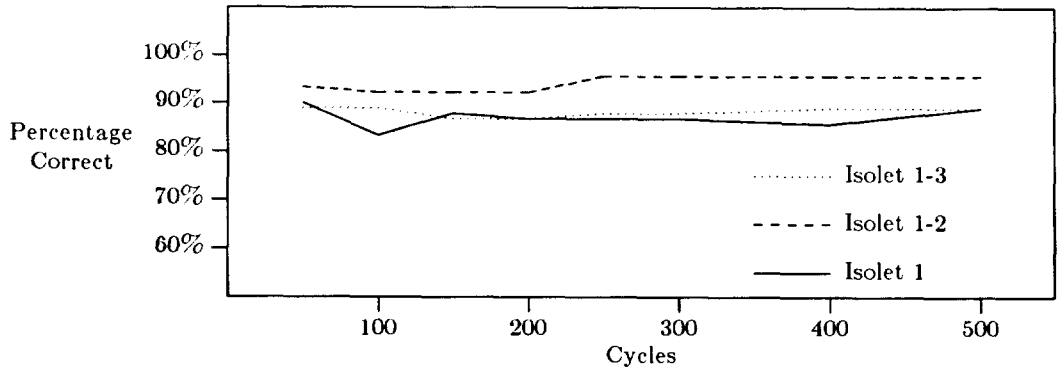

Fig. 4 Classification performance for A-E-F.

Future Directions and Outlook

We are currently conducting experiments to discriminate more phonemes, and evaluate the performance under the constraints imposed by limited computational precision. In addition, we will evaluate the performance of the model when trained on cochleagrams of the signal, rather than DFT's. Finally, we are planning to extend the model to deal naturally with time-dependence, making use of the temporal cues in natural speech.

Acknowledgements – We thank Richard Granger and Jose Ambros-Ingerson for stimulating conversation, and Yeshwant Muthusamy for preparing the data and performing the backpropagation experiments.

References

[1] Jose Ambros-Ingerson, Richard Granger, and Gary Lynch. Simulation of paleocortex performs hierarchical clustering. *Science*, 247, 16 March 1990.

[2] Ron Cole, Yeshwant Muthusamy, and Mark Fanty. The ISOLET spoken letter database. Technical Report CSE 90-004, Oregon Graduate Institute of Science & Technology, March 1990.