(12)

AD-A261 523

| | |
|---|---|
| 1. REPORT NUMBER #58 | ESSION NO. 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and) Supervised and unsupervised feature extraction from a cochlear model for speech recognition. | 5. TYPE OF REPORT & PERIOD COVERED Technical Report |
| | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s) N. Intrator and G. Tajchman | 8. CONTRACT OR GRANT NUMBER(s) N00014-91-J-1316 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Brain and Neural Systems Brown University Providence, Rhode Island 02912 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS N-201-484 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS Personnel & Training Research Program Office of Naval Research, Code 442PT Arlington, Virginia 32217 | 12. REPORT DATE 12/23/92 |
| | 13. NUMBER OF PAGES 10 pages |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | 15. SECURITY CLASS. (of this report) Unclassified |
| | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Publication in part or in whole is permitted for any purpose of the United States Government.

DTIC ELECTE JAN 25 1993 S B

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Published in B.H. Juang, S.Y. Kung, and C.A. Kamm, editors. Neural Networks for Signal Processing - Proceeding of the 1991 IEEE Workshop, pages 460-469. IEEE Press, New York, NY 1991.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Feature Extraction
Supervised Learning
Unsupervised Learning
Lyon's Cochlear Model
Speech Recognition

93 1 22 080

93-01198

11pf

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We explore the application of a novel classification method that combines supervised and unsupervised training, and compare its performance to various more classical methods. We first construct a detailed high dimensional representation of the speech signal using Lyon's cochlear model and then optimally reproduce its dimensionality. The resulting low dimensional projection retains the information needed for robust speech recognition.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

# Supervised and Unsupervised Feature Extraction From a Cochlear Model for Speech Recognition

Nathan Intrator*
Center for Neural Science
Box F, Brown University
Providence, RI 02912
nin@brownvm.brown.edu

Gary Tajchman†
Cognitive and Ling. Sciences
Box 1978, Brown University
Providence, RI 02912
tajchman@browncog.bitnet

**Abstract–We explore the application of a novel classification method that combines supervised and unsupervised training, and compare its performance to various more classical methods. We first construct a detailed high dimensional representation of the speech signal using Lyon's cochlear model and then optimally reduce its dimensionality. The resulting low dimensional projection retains the information needed for robust speech recognition.**

## INTRODUCTION – SPEECH PREPROCESSING METHODS

Many speech recognition systems, in particular, those based on HMMs, use LPC derived cepstral coefficients as the first step in preprocessing the speech data. These cepstra are then typically passed through vector quantization (VQ), or used directly as input to the HMM. The VQ step discretizes the multidimensional input vectors into a small set of possible inputs. This helps simplify training the system, but also introduces varying degrees of distortion [11]. This limitation is partially overcome by using methods to estimate output parameters for the continuous space defined by the cepstra. These techniques also run into problems when the dimensionality of the input vector gets large. In spite of these potential problems, LPC-based systems have performed well, especially when augmented with energy and time-differenced cepstra [11].

Speech recognition systems using ANNs have employed a much more heterogenous set of preprocessing techniques. Everything from raw speech to LPC-based cepstra has been tried [12]. However, most have used some form

*Research supported by NSF, the Army Research Office, and ONR.
†Research supported by NSF grant DIR-89-07769.

of preprocessing inspired by the representation produced by the mammalian peripheral auditory system. Examples include Mel scale and bark scale spectra. Other more sophisticated techniques exist that produce more detailed representations.

While there is a tendency for preprocessing based on auditory system constraints to be used with ANNs and preprocessing based on vocal tract constraints to be used with HMMs, this is not always the case. For instance, some current HMM systems include a Mel scale transformation when computing cepstra, and as mentioned above, LPC-based cepstra have been used with ANNs. The differences in preprocessing for HMMs and ANNs can be largely attributed to the fact that ANNs are good at integrating over large dimensional representations, while HMMs do best with much smaller dimensional input.

In this paper we focus on ANN techniques for processing the detailed, high dimensional auditory system representation of speech produced by Lyon's cochlear model [13]. We explore the application of a novel classification method that combines supervised and unsupervised training, and compare its performance to various methods. Our task is feature extraction and classification of voiceless stops extracted from the TIMIT corpus.

### What are features of recognition for speech data

When moving to a much larger representation of the speech data, many existing techniques such as classifiers, or vector quantizers fail to work, mainly because of the curse of dimensionality [1]. This problem is related to the sparsity of high dimensional spaces, and implies that the amount of training data has to grow exponentially with the dimensionality.

In many cases, it is conceivable to assume that the important information for speech recognition lies in a much smaller dimensional space, and the question becomes, how to find this low dimensional structure, or how to extract the relevant features from the data. This question can be put in a much broader statistical formulation, in which one has a data set that lies in high dimensional space, with a lower dimensional structure and tries to reduce the dimensionality of the data, without losing the important structure. These problems may be addressed using a recent statistical tool called Exploratory Projection Pursuit [3] which has an effective implementation with a biologically motivated neural network [6].

## LYON'S MODEL OF COCHLEAR PROCESSING

We chose to use a fairly sophisticated auditory model to preprocess the speech data for our neural network. One reason for doing this was to assess the feasibility of using such a model as front end for a recognizer. Auditory models typically produce very large output representations in order to retain much of the detail the higher centers in the brain receive from the cochlea.

The auditory model we used to preprocess the speech data was Lyon's cochlear model [13] as implemented by Slaney [18]. For each time slice, 84 channels of output were produced (for data sampled at 16kHz). Time slices were separated by 2 msecs. Therefore, for 56 msec of speech, the model produced 2352 bytes of data. While this is still orders of magnitude smaller than what is transmitted through the auditory nerve to higher centers of the brain, it is much larger than the data representations typically used for speech recognition tasks.

The channels in the model correspond to nerve fibers evenly spaced along the basilar membrane in the cochlea. The center frequencies of the set of channels are logarithmically spaced, giving the lower frequencies a more dense representation than the higher frequencies. Neighboring channels overlap to a large degree. This models the highly redundant representation used by the mammalian auditory nerve. The band pass regions of the channels increase linearly with frequency.

Each channel is implemented as a second order digital filter. The entire filter bank is implemented with a cascade design giving the representation realistic amplitude and group-delay response in addition to making the computation efficient. To model the effects of the inner and outer ear, the signal is passed through a pre-emphasis stage and then processed by the cascade of second order filters. The final stage of processing is preceded by half-wave rectification to model the unidirectional transduction of the basilar membrane movement by the inner hair cells.

The final phase of the cochlear model passes the output of each channel through a series of adaptive gain control (AGC) elements. These AGC elements attempt to keep the output levels of each filter within specific range. Each AGC is coupled with its nearest neighbors to each side. This helps model the masking effects found in real cochlear processing. The resulting rectangular frequency by time representation forms an image of auditory nerve activity and is called a cochleagram.

In sum, much of the detail and character of the representation used by the auditory nerve is retained in the cochleagram representation The task then becomes how to best use all of this information.

## FEATURE EXTRACTION IN HIGH DIMENSIONAL SPACE – THE BCM MODEL

From a mathematical view point, extracting features from the rectangular representation of the cochleagram is related to dimensionality reduction in high dimensional vector space, in which an $n \times k$ pixel image is considered to be a vector of length $n \times k$. In such high dimensional spaces the *curse of dimensionality* [1] says that it is impossible to base the recognition on the high dimensional vectors, because the number of training patterns needed for training a classifier should increase in an exponential order with the dimensionality, and therefore dimensionality reduction should take place before attempting

the classification. Due to the large number of parameters involved, a feature extraction method that uses the class labels of the data, will be biased to the training data [5], which translates to having features with poor generalization or invariance properties. Thus, the feature extraction should be unsupervised. A recent statistical method to address this problem of dimensionality reduction called exploratory projection pursuit (EPP) [3] assumes that features can be constructed from projections of the input space onto a small dimensional space. This method defines interesting features as those projections whose single dimensional projected distribution is far from Gaussian. Since high dimensional clusters translate to low dimensional multi-modal projected distributions, a plausible measure of deviation from normality can be based on a measure of multi-modality of the projected distribution. Intrator [6] has recently shown that a variation of the Bienenstock Cooper and Munro neuron [2] performs exploratory projection pursuit using a projection index that measures multi-modality. A network implementation which can find several projections in parallel is still computationally efficient and therefore may be applicable for extracting features from very high dimensional vector spaces of the type generated by the cochlear model.

The unsupervised feature extraction/classification method is presented in Figure 1. Similar approaches using the RCE and back-propagation network have been carried out by [15], and using the unsupervised charge clustering network by Scofield [17]. Huang and Lippmann [4] described a feature-map classifier for vowel recognition, in which internal nodes compute kernel functions related to the Euclidean distance between the input and cluster centers represented by these nodes. The unsupervised vector quantizer was trained to form the new representation which trained the supervised classifier. Kohonen et al. [10] used a similar approach with LVQ network. Review on various other unsupervised/supervised approaches appears in [12].

Although unsupervised feature extraction has the potential of being less biased to the training data, its result may be suboptimal since it ignores the information contained in the class labels. It is possible for example, that not all the information required for the classification is contained in those directions which are considered interesting by the feature extractor (some trivial examples are discussed in [8]). Therefore, it is possible that a hybrid of unsupervised/supervised feature extractor may yield better performance.

Another way to look at the problem is from the classification side; The performance of the classifier that reduces dimensionality based solely on the class labels, may be improved if an additional measure of the information carried in the projections is added. In the case of a back-propagation classification network, a local penalty term may be added to the energy functional minimized by error back propagation. This penalty which is added only to the hidden layer units, is the projection index defined by the BCM network [6, 9]. Therefore, the modification equations for the hidden layer units are affected by the delta rule [16] and by the BCM modification equations. This method is described in detail in [7].

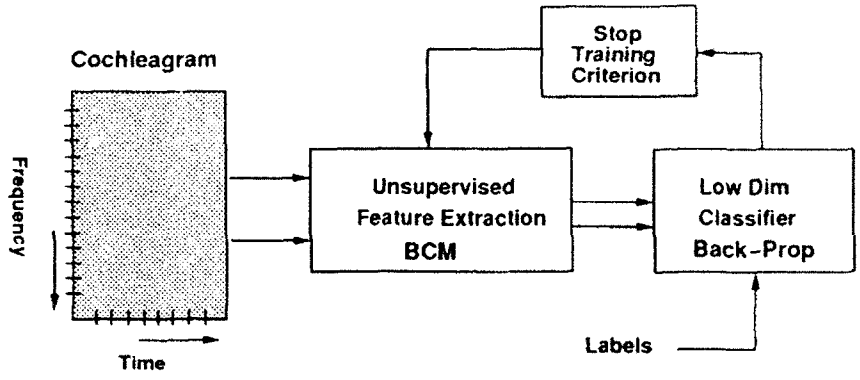


Figure 1: Low dimensional classifier is trained on features extracted from the high dimensional data. Training of the feature extraction network stops when the misclassification rate drops below a predetermined threshold on either the same training data (cross validatory test) or on different testing data.

## METHODS

### Data - Voiceless Stops from TIMIT

In this work we focused on feature extraction and classification of the voiceless stop consonants [p, t, k]. The source of our data was the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT). This database contains utterances from many talkers, with coverage of all the major dialect regions in the United States.

All tokens used in these experiments consisted of a stop followed by a vowel. We used only four vowel contexts [aa, ao, er, iy] in the training set. These vowels give a reasonable, but not complete coverage of the vowel space. This restricted set allowed us to test how well the feature extraction generalized to new vowel contexts.

These tokens were drawn from the utterances of 268 different talkers. Multiple talkers and various sentential contexts contribute to a fair degree of variability between tokens of the same CV type. The segment boundaries we used were exactly those provided with TIMIT. We made no attempt to sharpen or correct any misalignments that might exist in the data.

For each CV type, an average over the 25 tokens used for training is presented in the cochleagram matrix shown in Figure 2. The vertical axis is frequency, low to high from top to bottom, and the horizontal axis is time for each cochleagram. Looking at the lower left corner of the images, it can be seen that [p]s have low energy at the high frequencies, [t]s have a sharp burst in the high frequencies, and [k]s have diffuse energy in the high frequencies. These features tend to distinguish between the three voiceless stops for the cochleagram representation.

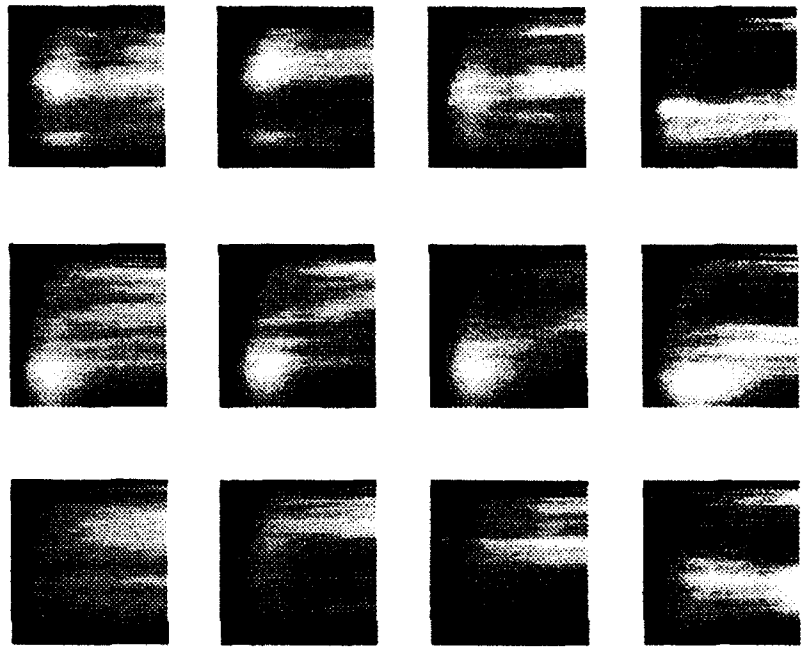

Figure 2: The output of Lyon's cochlear model for the 12 CV pairs. ¿From top to bottom [k, t, p], and from left to right [aa, ao, er, iy]. Each image is the average of 25 tokens from each CV type showing 75msec of speech aligned to burst release. White areas represent high energy.

## Training

In the first experiment features were extracted from the large representation of the speech segment using a BCM network. Here the BCM weights were only affected by the unsupervised modification rule. Classification was accomplished by training a small back-propagation network with the output of the BCM network as shown in Figure 1. An important issue of avoiding over fitting (in either of the nets) was addressed by testing (during training) on a third set of tokens (Pseudo test set).

In the second experiment the modification of the hidden units of a 3 layer back-propagation network, was a combination of the BCM synaptic modification equations, and the error propagated from the top layer. The performance of the networks in the first and second experiments were compared to the performance of a simple back-propagation network.

## Testing

| Training Method | 4 Vowels Training | 4 Vowels Testing | 7 Vowels Testing |
|---|---|---|---|
| BCM B-P | 81% | 73.8% | 72.7% |
| BCM/B-P | 92.6% | 83.8% | 81.5% |
| B-P | 98.7% | 84.8% | 78.2% |

Table 1: Comparison between classification using (1) projections from BCM unsupervised learning as input to back-propagation; (2) a hybrid of BCM unsupervised learning and supervised learning via error back-propagation; and (3) a plain back-propagation net.

We used two generalization paradigms to test the feature extraction and classification ability of the system. First, the standard type of generalization to new instances of the same class was carried out. For each of the 12 CV types, we tested with 25 novel instances[1]. This kind of generalization requires the system to categorize instances that fall within the region of the input space it has had experience with. Many recognition systems are specifically focused on this kind of generalization. However, the second kind of generalization, where a system trained with a limited set of contexts generalizes well in new contexts, is possibly more important. If a system can transfer to new contexts, or to a region of the input space it has not experienced, the set of abstract features it is using must be capturing highly relevant aspects of the input training space. The ability to discover such features strongly suggests the technique being used is well suited for robust speech recognition. We demonstrate this kind of generalization by training on four vowel contexts [aa, ao, er, iy], and testing with the seven vowel contexts [uh, ih, eh, ae, ah, uw, ow].

## RESULTS AND DISCUSSION

A comparison between the different training methods is shown in Table 1. The low dimensional projections of the cochleagrams discovered with BCM learning, served as input to a small back-propagation network to yield the first set of results. This training method yielded reasonable performance on the training set, and very nearly the same performance on the two test sets. The small difference in generalization to instances of the same-4-vowel-contexts test set and generalization to instances from the new-7-vowel-contexts test set implies the features discovered with this method are good abstractions, and robust. The weight matrices of the eight units used in the BCM network are shown in Figure 3.

Features distinguishing between the different bursts are evident. The synaptic weight image on the top row, furthest to the right shows a white area in the high frequencies which corresponds to a distinguishing feature between

[1] There were only 21 new tokens available for [pao]. All other CV groups had 25 tokens.

[t] and [k]. The image directly below is useful for distinguishing [p] from the other two stops.

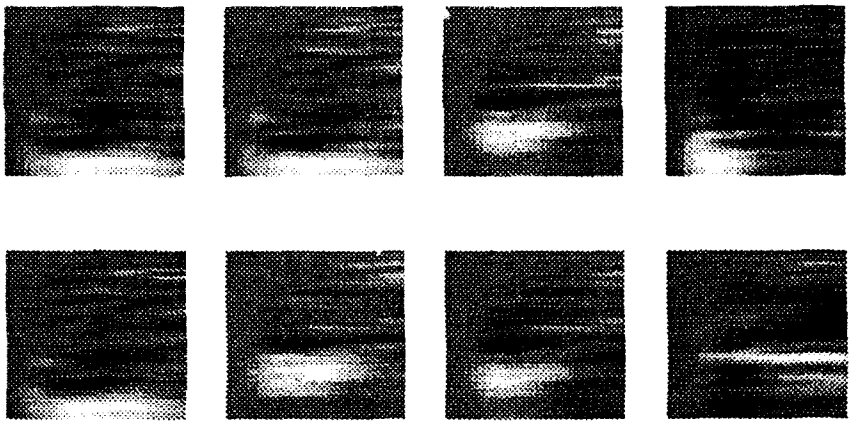

Figure 3: The synaptic weight matrices for 8 units after unsupervised training on 25 tokens of each CV type.

The results of the second training method, in which error back-propagation was modified to incorporate BCM-like constraints, are shown on the second line of Table 1. This novel integration of supervised and unsupervised techniques boosted the performance significantly over the previous training method. However, the pattern of results are very much the same; good and nearly equal performance with both types of generalization.

In contrast, this pattern was not found with the plain back-propagation net. While it did achieve the best performance of the three networks on the training set, it did not transfer its good generalization performance on the same-4-vowel-contexts to the new-7-vowel-contexts test set. Straight back-propagation training only attempts to minimize errors with the training set. It does not necessarily search for abstract features.

At this point, the only comparison we can make with HMM performance is very loose. Niles [14] constructed a baseline HMM system to classify the standard set of 39 phonetic classes in TIMIT. The speech was preprocessed using an order-18 LPC cepstral analysis, and then VQ codebooks for the cepstra, time-differenced cepstra, log energy, and delta log energy were used as input. A three state HMM was trained up for each phoneme. This system classified 82.0 percent correct when tested with just the voiceless stops. While this does give a ballpark indication that the systems we investigated here are doing reasonably well, any further comparison is precluded by methodological differences. For instance, Niles trained the HMMs for voiceless stops with all phonetic contexts, while our tokens always had a following vowel. Also, the HMM system was used as a baseline system, and was not fine tuned.

These preliminary results suggest that BCM training can be beneficially incorporated into a network architecture/training-paradigm for speech recog-

nition. Moreover, the cochleagram input representation produced by Lyon's cochlear model contains details about the speech events that are useful in classifying speech tokens. A set of experiments making specific, quantitative comparisons between the system we have proposed here and current HMM methods is planned.

# References


[1] R. E. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ, 1961.

[2] E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal Neuroscience*, 2:32–48, 1982.

[3] J. H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266, 1987.

[4] W. M. Huang and R. P. Lippmann. Neural net and traditional classifiers. In D. Anderson, editor, *Neural Information Processing Systems*, pages 387–396. American Institute of Physics, New York, 1988.

[5] P. J. Huber. Projection pursuit. (with discussion). *The Annals of Statistics*, 13:435–475, 1985.

[6] N. Intrator. Feature extraction using an unsupervised neural network. In D. S. Touretzky, J. L. Ellman, T. J. Sejnowski, and G. E. Hinton, editors, *Proceedings of the 1990 Connectionist Models Summer School*, pages 310–318. Morgan Kaufmann, San Mateo, CA, 1990.

[7] N. Intrator. Combining exploratory projection pursuit and projection pursuit regression with application to neural networks, 1991. Preprint.

[8] N. Intrator. Localized exploratory projection pursuit. In Ed Wegman, editor, *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 237–240. Amer. Statist. Assoc., Washington, DC., 1991.

[9] N. Intrator and L. N. Cooper. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17, 1992.

[10] T. Kohonen, G. Barna, and R. Chrisley. Statistical pattern recognition with neural networks: Benchmarking studies. In *IEEE International Conference on Neural Networks*, volume 1, pages 61–68, New York, 1988. (San Diego 1988), IEEE.

[11] K. F. Lee. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System.* PhD thesis, Carnigie Mellon University, 1988.

[12] R. P. Lippmann. Review of neural networks for speech recognition. *Neural Computation*, 1(1):1–38, 1989.

[13] R. F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, May 1982.

[14] L. T. Niles. Timit phoneme recognition using an hmm-derived recurrent neural network. In *Eurospeech91*, September 1991. Genoa, Italy.

[15] D. L. Reilly, C. L. Scofield, L. N. Cooper, and C. Elbaum. Gensep: a multiple neural network with modifiable network topology. In *INNS Conference on Neural Networks*, 1988.

[16] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, pages 318–362. MIT Press, Cambridge, MA, 1986.

[17] C. Scofield. Learning internal representations in the coulomb energy network. In *Proc. IEEE First Int'l Conf on Neural Networks, San Diego.* 1988.

[18] M. Slaney. Lyon's cochlear model. Technical report, Apple Corporate Library, Cupertino, CA 95014, 1988.

[19] Timit acoustic-phonetic continuous speech corpus. National Institute of Standards and Technology Speech Disc 1-1.1, October 1990. NTIS Order No. PB91-505065.