Generalization in Neural Networks
Experiments in Speech Recognition

Elizabeth Lake Richards

CU-CS-538-91 August 1991

University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

ANY OPINIONS, FINDINGS, AND CONCLUSIONS OR RECOMMENDATIONS EXPRESSED IN THIS PUBLICATION ARE THOSE OF THE AUTHOR(S) AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE AGENCIES NAMED IN THE ACKNOWLEDGMENTS SECTION.

GENERALIZATION IN NEURAL NETWORKS: EXPERIMENTS IN SPEECH RECOGNITION

by

ELIZABETH LAKE RICHARDS

B. A., University of California at Los Angeles, 1956
M. A.,.University of California at Los Angeles, 1961
M. S., University of Colorado at Boulder, 1971

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science
1991

DEDICATION

To Carol, who makes all things possible.

Richards, Elizabeth Lake (Ph. D., Computer Science)

Generalization in Neural Networks:

Experiments in Speech Recognition

Thesis directed by Assistant Professor Gary L. Bradshaw

This research is an investigation of the problem of generalization in neural networks: how do the task which the network must learn, the architecture of the network, the training of the network, and the data representations used in that training, both individually and collectively, affect the ability of a network to learn the training data and to generalize well to novel data.

A psychological model of speech perception, Liberman and Mattingly's Motor Theory, provides the theoretical foundation for the tasks and architectures specified for the networks used in the research. Linguistic theories of vowel perception guided the preparation of speech data representations used in training the networks. Vowel data was collected across varying contexts and speakers to provide a broad test of the networks' ability to generalize to highly variable data.

Results of the research show that networks having different task requirements but trained with the same number and type of data representations form a family of networks which exhibit similar generalization across a broad range of hidden units. Contradicting commonly accepted guidelines, networks trained with larger data representations exhibit better generalization than networks trained with smaller representations, even though the larger networks have a significantly greater capacity. In addition, networks having the same training performance can exhibit different levels of generalization; researchers interested in generalization must track generalization directly. Finally, given an appropriate architecture, training algorithm, and sufficient training data, the data representation itself is the primary determiner of a network's ability to generalize well to new data.

ACKNOWLEDGEMENTS

How nice it would be to think that the results of one's research are solely the result of one's own efforts. Such is rarely the case, however. Were it not for the support and assistance of others this work would never have been completed. I owe so many thanks to -

- Barb Dyker and the csops folks, Taki Meghjee and the CLIPR support staff, and Barb Davis and the unixops group - for providing support for all the hardware which I so ravenously utilized.
- Scott Katz and the ICS Speech Lab for developing and maintaining the Speechbench software. To Hewlett-Packard Corporation for providing the equipment on which it runs.
- The secretaries and support staff at both ICS and the Computer Science department who contributed to making my Ph.D. years productive and enjoyable; thanks are due, in particular, to Dotty Foerst who was always there to help.
- The members of the Boulder Connectionist Research Group, especially to Kevin Markey and Rick Blumenthal for their helpful comments.
- Yoshiro Miyata for the high quality of the PlaNet simulation software he developed, and, most especially, for all his patient help and advice on how to use it.
- Thanks are due to Dirk Grunwald for his gracious offering of extra cpu cycles.
- The Colorado Institute for Artificial Intelligence for funding this research and U. S. West for its role as the industry partner.

- Dr. Hynek Hermansky of U. S. West for his ever thoughtful comments and advice.

To my friends, Molly Barden, Connie Delzell, Beth Parsons, Kate Puetz, Carolyn Schauble, and Sharon Smith, go many thanks for encouragement, assistance, and prayers.

To the members of my committee:

- Alan Bell, for all his help and advice in the linguistics aspect of this research;
- Andrzej Ehrenfeucht, for his thoughtful and considerate support;
- Mike Moser, for many stimulating technical discussions;
- Paul Smolensky, for his insightful and thought provoking comments;

Thank you all!

To my *Doktor Vater*, Gary Bradshaw, I owe more than words can ever express. For providing me the opportunity to enter the world of speech recognition, for patiently training me in the ways of research, for gently guiding me towards my goal and for friendly encouragement in the difficult moments - for being a Teacher among teachers - I will be forever grateful.

Last, but truly first, I give thanks to my family: to my father, for teaching me that a job worth doing is worth doing well - may light perpetual shine upon him; to my mother, who by her example has taught me the meaning of courage and endurance; to my daughter, Ingrid Anne, for keeping me laughing and being a shining light through the dark times; and to Carol, who has shared my life and been my ever constant support throughout the past nine years.

This work is dedicated to Carol and submitted *ad maiorem dei gloriam.*

CONTENTS

CHAPTER

TABLES

TABLE

APPENDIX

FIGURES

FIGURE

CHAPTER 1

INTRODUCTION: GENERALIZATION

1.1 Generalization

The essence of intelligence is the ability to learn. Learning includes both memorization of the concepts which are required to be learned and generalization to new concepts which have not yet been encountered. Memorization can be thought of as a search for a function that successfully relates or maps training inputs to their corresponding outputs. A function $f : X \rightarrow Y$, called a mapping, is a rule which assigns to each element of an input set X a unique element $f(X)$ of the output set Y. Generalization can be thought of as the ability to map novel inputs to the correct corresponding outputs based upon properties discovered in the training set. It should be noted that it is not always possible to generalize from items which have been memorized. For example, in a telephone book we find many people who have the same last name, while others have the same first name or initials. The ability to memorize names and the associated telephone numbers does not permit one to predict the phone number of a person who shares the same surname with one person and a given name with another. In other words, items can be memorized based upon unsystematic relationships between the input and output or upon idiosyncratic features of the items. Generalization requires that the memorized items have underlying similarities which are capable of being extracted and applied to new items from the same domain.

In a neural network the capacity of the network, or the number of mapping functions that can be represented in the search space of the network, is a function of the architecture of the network and can be measured by the number of connections in the network. Under appropriate conditions, a system having a large capacity has the potential to memorize a large training set or to represent a large number of functions. A system having small capacity is capable of memorizing only a small training set or representing only a small number of functions. It is possible for a system to be trained so that it finds a function that successfully relates or maps the training inputs to their corresponding outputs. That function may not correctly map novel inputs to their corresponding outputs. For example, a system which has a large capacity relative to the size of the training set it is required to learn may develop a function that is based upon idiosyncratic features rather than upon important features in the training set. Such a system will not generalize well to novel inputs. A number of researchers [22, 26] maintain that a large capacity system requires a large number of training examples to narrow the search down to the correct mapping function so that generalization will occur. Reducing the capacity of the system may reduce the number of training examples required to achieve a high level of generalization. It may also have the effect of eliminating the desired mapping function from the search space; the reduced system capacity may not even be able to memorize the training set.

Clearly there is a complex interaction between these four factors: memorization, generalization, system capacity, and number of training examples. Judd [13] states that "The business of finding regularities in data and generalizing from them depends totally on the embedded problem of simply remembering data." Memorization requires sufficient system capacity. A looser definition of learning does not require complete memorization. It permits a degree of deviation from the given data. Even in this case, a large training set size relative to the system capacity is required to find an appropriate mapping function for generalization to occur. To design and train neural networks that generalize well to new examples after having been trained on a sufficiently large set of training examples is a major goal of connectionist learning [12]. There is considerable current interest directed to understanding under what conditions good generalization occurs in neural networks.

I turn now to a review of existing research into the connectionist generalization problem and describe some limitations of that research. The unstated but implicit rationale underlying much of the research into the connectionist generalization problem is the labor and financial cost involved in the acquisition and preparation of training data. Since the acquisition and preparation of training data is in many instances the most labor intensive and least cost effective aspect of a project, there is an understandably strong desire to reduce the amount of time, labor, and money that must be spent on this portion of the project. One way to do that is to reduce the amount of data that is required for training a system and one obvious approach to accomplishing that is to reduce the capacity of the system.

Following the review of existing research, I will present a brief overview of the approach I take in investigating the generalization problem together with some definitions of key terms critical to an understanding of that approach. I will include some of the guidelines and hypotheses which guided my research.

1.2 Connectionist Generalization

Probably the most quoted guideline in connectionist literature regarding the amount of data required to train a network and achieve good generalization is attributed to Bernard Widrow [42]. Widrow suggested that the size of the training sample to be used in training a network should be at least 10 times the number of connection weights in the network. For networks intended to cope with real world problems, this suggestion immediately places a tremendous strain on the data acquisition resources available for the total research project. To illustrate, I have developed networks with between 615 and 22,050 connection weights. Using Widrow's dataset-size guideline these networks would required 6,150 examples to 220,500 training examples! Considering the resources and time required to collect such large amounts of data it is not surprising that researchers seeking to better understand under what conditions a network will exhibit good generalization would begin by exploring the relationship between the size of the network being trained, and the amount of training data required for the network to exhibit good generalization.

1.2.1 Generalization: architecture and training set size. By far the greatest amount of research into the connectionist generalization problem has been concentrated on exploring the impact of architecture and training set size on connectionist training and generalization. By network architecture I mean the set of input units, hidden units, and output units together with the connections between units and the unit groupings which combine to form a network of computing elements. Because of the difficulties involved in analyzing the behavior of networks using nonlinear activation functions, theoretical work on the relationship between network architecture and generalization has been limited.

Baum and Haussler [2] follow the theoretical approach to determining an appropriate sample size for a given network size for a classification task. Assume the error rate ϵ is defined as $0 < \epsilon \leq 1/8$. With N nodes and W weights, if $m \geq \mathcal{O}(\frac{W}{\epsilon} \log \frac{N}{\epsilon})$ random examples chosen from some arbitrary distribution can be stored in a feedforward network of linear threshold functions so that at least a fraction $1 - \epsilon/2$ of the training examples are correctly classified, then one can have confidence approaching certainty that the network will correctly classify $1 - \epsilon$ new testing examples selected from the same distribution. This estimate of m is an upper bound on the number of training samples required. A lower bound estimate is on the order of W/ϵ. A training set containing approximately 10 times as many training examples as there are weights in the network is required for a network having a generalization error rate of 10% or less. It is important to note that Baum and Haussler are concerned with linear threshold units only and that their results promise certainty regarding the generalization to be realized in the trained classifier network. Their results have not been extended to nonlinear functions such as the sigmoid function. Their results are not significantly different from Widrow's suggested value. The individual researcher may be willing to weigh the amount of effort required to collect and prepare a training set offering certainty of a specific generalization performance against the possibility of achieving that same degree of generalization performance with something less than absolute certainty and utilize a smaller training set size.

A number of empirical results have been reported concerning the appropriate training set size for a fixed size architecture or reducing the capacity of a network to accomodate a fixed size training set. Examples of the first approach, determining the appropriate training set size, are provided by Ahmad and Tesauro's work and also by that of Leung and Zue. Ahmad and Tesauro [1] trained a feed-forward network with no hidden units to learn a simple linearly-separable problem. Their task was to report the majority function, which returns a 1 if the majority of the inputs are on. They wanted to study the relationship between the size of the network, the number of training instances, and the generalization exhibited by the network. The size of the input layer was varied. Their results showed that for a given input size the failure rate decreased exponentially with the number of training patterns: $f = \alpha e^{-\beta S}$ where f is the failure rate and S is the number of training patterns. In addition, the number of training patterns required to achieve a fixed performance level was shown to increase linearly with the network size.

Leung and Zue [17] explored the impact of training set size on the training of a network having a fixed architecture. They used speech data to train a single hidden layer network having 100 input units representing three spectral frames of speech data, 32 hidden units, and 16 output units that corresponded to 16 American English vowels excised from 1,000 sentences spoken by 200 speakers. The network was trained with sets consisting of from 80 to over 8,000 training tokens. With an increase in training set size, training recognition results reached an asymptote of about 80%. Speaker-independent testing showed that the network generalization increased monotonically with training set size from 30% to 54% with the greatest increase in accuracy occuring between 80 and 800 training tokens. It is not possible to compare Leung and Zue's results against Baum and Haussler's theoretical estimate, since Leung and Zue do not achieve appropriate error rates on either the training or generalization error of the network. They state that their network containing 1392 connections does not exhibit significantly better performance with 8,000 training tokens than it did with 800 training tokens. This lack of improvement may indicate that their architecture had insufficient capacity. Further evidence is given by the 80% asymptotic training recognition. The network does not appear to have had sufficient capacity even to learn the training data. In a second experiment they sought to improve network generalization by increasing the information content of the input data representation. These efforts will be discussed in Section 1.2.2 below.

Examples of the second empirical approach, reducing the capacity of a network to accomodate a fixed size training set, involve varying the number of hidden units, varying the connectivity of the network, and varying the total capacity of the network. Several researchers have investigated systems which vary the number of hidden units in the network. For example, in a paper whose focus is primarily the analysis of hidden unit behavior in a network, Gorman and Sejnowski [10] report an increase in network generalization with an increase in number of hidden units. Neural networks were trained to perform the classification of sonar returns from two undersea targets, a metal cylinder and a similarly shaped rock. The networks had 60 input units and 2 output units. The number of hidden units was varied from 0 to

24. Table 1.1 illustrates their results. From 0 to 12 hidden units, both the training recognition and the testing generalization increase with the number of hidden units. The significance of observed increase in network performance with the number of hidden units was tested, and verified, by an analysis of variance. At 24 hidden units the training recognition continues to increase while the testing generalization exhibits a very slight decrease.

Table 1.1: Gorman and Sejnowski's Results

	Average Performance	
Hidden	Training	Testing
0	79.3	73.1
2	96.2	85.7
3	98.1	87.6
6	99.4	89.3
12	99.8	90.4
24	100.0	89.2

Gorman and Sejnowski noted the impact of number of hidden units on performance and network generalization: the improving recognition and generalization from 0 to 12 hidden units. They failed, however, to note the essentially similar generalization exhibited across a range of from 6 to 24 hidden units - a fourfold increase in the number of hidden units.

Morgan and Bourlard [26] performed an empirical study of the relationship between the number of weights in a feedforward network and the ability of the network to generalize well to new examples. They used both simulated data sets and speech data to train the networks. For the speech data they used German sentences spoken by a single speaker: 100 training sentences and 100 testing sentences. Input to the network was 9 frames of vector-quantized mel cepstra data for a total of 1188 input units. The output layer contained 50 units corresponding to the 50 phonemes to be recognized. The hidden layer varied from 20 to 200 units. Performance on the test set was monitored after each epoch and training was discontinued when test set performance showed no further improvement. As the number of hidden units increased from 20 to 200, performance on the training set increased from 75.7% to 86.7%. Corresponding performance on a novel test set decreased from 62.7% to 59.6%. Their results indicate:

> While both studies show the expected effects of overparameterization, (poor generalization, sensitivity to overtraining in the presence of noise), perhaps the most significant result is that it was possible to greatly reduce the sensitivity to the choice of network size by directly observing the network performance on an independent test set during the course of learning (cross-validation). ...Networks which require many more parameters than there are measurements will certainly reach lower levels of peak performance than simpler systems.

While Morgan and Bourlard investigated the impact of number of hidden units on generalization, their results may be limited because of the restricted range of hidden units they chose to investigate. For a network with 1188 input units and 50 output units they only explored a hidden layer having from 5 to 200 hidden units (see Table 1.2). From 5 to 200 hidden units, they show an increase in training recognition.

Table 1.2: Morgan and Bourlard's Experiments

	Phoneme Recognition	
Hidden	Training	Testing
5	62.8	54.2
20	75.7	62.7
50	73.7	60.6
200	86.7	59.6

An increase in testing generalization, however, is only exhibited from 5 to 20 hidden units. From 20 to 200 hidden units a small decrease in generalization can be observed.

The possibility that they may well not have been in the right ballpark is apparent from an examination of these results. Although training recognition was continuing to rise up to 200 hidden units at no point can it be claimed that the networks truly learned the training set; the highest training recognition exhibited is less than 87%. The slight decrease in testing generalization between 20 and 200 hidden units can be quite plausibly explained as a result of experimental variation. The fact that no networks having more than 50 but less than 200 hidden units were explored makes it extremely difficult to determine the stability of either training recognition or generalization within this range. The networks could exhibit non-linear performance results in this range, for example, performance could increase beyond 86.7% and then decrease to 86.7%. In addition, it is not at all clear why Morgan and Boulard did not study networks having more than 200 hidden units, since even the 200 hidden unit network was not succesful. The 86.7% recognition accuracy seems to indicate that this network had insufficient capacity to learn the training data. Additional experimentation might have revealed significantly increased training and testing recognition performance with an increase in the number of hidden units.

Mozer and Smolensky and also Sietsma and Dow report methods that automatically reduce the number of hidden units in a network. The rationale for Mozer and Smolensky's [27] approach is the observation that learning to criterion is faster for networks having many hidden units. Generalization is similarly assumed to be better with fewer hidden units. No evidence is offered in support of this position. Mozer and Smolensky use a relevance measure which computes for a given unit the approximate change in the difference between the network error with the unit and the network error without the unit. Here the error is a linear function rather than the usual quadratic function. The network is first trained to some specified margin around the target value, then the relevance is computed and the unit having the least relevance is removed from the network. Training then continues and the procedure is repeated a specified number of times. Examples of the approach are given for tasks such as the four-bit multiplier problem and the random mapping problem where a set of random 20-element input vectors is mapped to random 2-element output vectors. For the latter problem the network failed to reach criterion with two hidden units on 17% of the training runs. A network trimmed to two hidden units using the skeletonization process reached criterion on all but 8.3% of the training runs. In addition, the skeleton network reached criterion much sooner with six hidden units than did a network having only two hidden units. Its performance did not significantly decline as the network was trimmed. It is not clear how well the approach would scale-up for larger networks using large sets of training data collected from real-world problems.

Sietsma and Dow [33] developed a technique to test the hypothesis that multilayered feedforward networks with few hidden units on the first hidden layer generalize better than networks with many hidden units on the first hidden layer. They trained a network to classify sine waves of different frequencies. Two test sets were generated: one consisted of sine waves of different frequencies with different phase shifts from the training set, while the second consisted of similar sine waves corrupted with random noise added to the test patterns. The initial network architecture had 64 input units, 20 hidden units in the first hidden layer, eight hidden units in the second hidden layer, and three output units. Results were obtained for training the network with several different initial conditions. The network was then re-trained while unused or noncontributing units were pruned from the network. Here pruning was accomplished by programs outside of the network simulator; these programs read the network state and determined the units suitable for deletion. In general, Sietsma and Dow found that reducing networks to the smallest size capable of classifying the training set degraded the generalization capability of the networks "... indicating that in some circumstances networks with many hidden units generalize better than networks with few hidden units."

Examples of approaches which automatically vary the connectivity of the network are suggested by Hinton [12] and empirical results for such an approach are reported by Weigend [40]. Hinton suggests a cost term in the training algorithm error function which penalizes large weights. If the cost term is $\sum w^2$ then the derivative corresponds to a weight decay and decreases towards zero.

Weigend, Huberman, and Rumelhart trained a feedforward network with one hidden layer having non-linear activations and an output layer having linear activations to predict sunspot activities. Two training algorithm approaches were used to determine the optimum network size, one of which will be discussed in Section 1.2.2 below. In their approach, it was assumed that the best generalization occurs when the smallest network still able to fit the training data has been identified. They began with a network that was "too large" for the problem. No rationale is offered as to how this was determined. Each connection in the network has a cost associated with it:

$$\lambda \sum_{i,j} \frac{w_{ij}^2}{1 + w_{ij}^2}$$

The parameter λ is used to represent the relative importance of the weight cost function with respect to the standard performance error term. Use of this cost function in the training algorithm encourages the reduction and elimination of as many of the weights as possible. This approach is similar to that suggested by Hinton [12]. It has the disadvantage that network training must be monitored so that the value of λ can be changed as required: start with λ at 0, slowly increase λ until performance begins to decline and thereafter increase or decrease it as appropriate for a continued decrease in the training error. They used an average relative variance measure to show that prediction results from the weight elimination method are better than results from a network trained without weight elimination: 0.38 versus 0.42, approximately.

The most thorough experiment to date varying the capacity of the network is reported by Martin and Pittman. Martin and Pittman [22] used real-world data in their investigation of the neural network generalization problem. They designed and trained a network to recognize hand-printed letters and digits. Input patterns to the network consisted of $15X24$ bitmaps of pre-segmented size-normalized grayscale arrays for each character. The networks were fully connected and feedforward. The output units matched the number of categories to be identified. For the digits, they used training set sizes of 100 to 35,200 samples and a test set size of 4,000 samples. For the letters they used training set sizes of 500 to 6300 samples and a testing set of 2,368 samples. In exploring the effect of network capacity on generalization, they manipulated the architecture of the network in various ways: varying the number of hidden units, limiting the connectivity between layers, distributing the hidden units across both 1 and 2 layers, and sharing connection weights between hidden units. Each of these approaches had only a negligible effect on generalization. Using nets with fewer hidden units did not improve generalization, nor did limiting the connectivity between layers: "The fact that we find no advantage to reducing the number of connections conflicts with Baum & Haussler's estimates and the underlying assumption that capacity plays a strong role in determining generalization." Using local receptive fields and shared weights resulted in only a slight improvement in generalization and those occurred only in the case of relatively small training set sizes. The size of the training set, however, appeared to affect the ability of the networks to generalize to new examples:

> Given an architecture that enables relatively high training performance, we find only small effects of network capacity and topology on generalization performance....it is probably better to devote limited resources to collecting a very large, representative training set than to extensive experimentation with different net architectures. The variations in net capacity and topology we've examined do not substantially affect generalization performance for sufficiently large training sets. Sufficiently large should be interpreted as on the order of a thousand to tens of thousands of samples for hand-printed character recognition.

They report their numeric results in terms of performance versus training set size for a fixed capacity. They do not report numeric results for performance versus capacity for a fixed training set size. Consequently, it is not possible to compare the results obtained in my investigation with the results of their investigation.

Finally, Cheung, Lustig, and Kornhauser [6] took yet another approach to investigating the relationship between training set size and generalization. They reasoned that back-propagation treats all training patterns as being equally important during training. Nevertheless, some training patterns are more difficult to master than others. Based on this rationale they developed a process to select out at each training cycle the most poorly-trained patterns. They then enlarged the training set by including additional copies of these selected patterns for use in the continued training of the network. The dynamic training set that was thus created contained the original pattern set plus copies of the most recent poorly-trained patterns. Using two toy domains, such as counting the number of 1's in a 6 bit pattern, they tested this approach and showed an increase in generalization with a reduction in error rate from a 24% error rate to a 0% error rate. No attempt was made to determine if these results would scale up to larger networks using larger data sets collected from actual problems.

In summary, reported results for the relationship between network architecture and generalization indicate that either (a) network capacity has no apparent effect on generalization [22], [17]; (b) networks with more hidden units generalize better [10, 33]; or (c) networks with fewer hidden units generalize better [26]. There is an obvious disparity of results reported with respect to the relationship between the architecture of a network and the ability of that network to generalize to new data. With regard to the relationship between training set size and generalization, reports indicate that (a) more training examples results in better generalization [1, 6], or (b) an increase in training set size results in an increase in generalization only up to a certain point, beyond which additional training examples do not increase generalization [17]. The latter results appear to indicate that quantity of training data alone is not sufficient to guarantee good generalization.

1.2.2 Generalization: other factors.

By far the greatest amount of research into the connectionist generalization problem has been concentrated on exploring the impact of the relationship between the capacity of the network being trained and the size of the training data set on the network's ability to exhibit good generalization. Several researchers have reported on other factors which appear to affect network generalization: the amount of training a network has received, the data representation used in that training, and the task the network was required to learn.

Reports that the amount of training a network receives has a bearing on generalization have been made by both Weigend, Huberman, and Rumelhart [40] and also by Morgan and Bourlard [26]. Weigend, Huberman, and Rumelhart trained a feedforward network with one hidden layer having non-linear activations and an output layer having linear activations to predict sunspot frequency. A large number of hidden units was allocated to the network architecture. In order to ensure that overfitting did not occur with this approach, network performance on a cross-validation set was monitored and training was discontinued when performance on the cross-validation set ceased to improve. The rationale for their approach is that a network will not be trained to the point where it begins to learn noise present in the training set. A figure comparing the results of training by this method with the previously discussed training by weight elimination indicates that "...the fitting of the noise of this training set happens [*sic*] to have no effect on the error of this cross-validation set." The authors further indicate that the results depended strongly on the specific training and cross-validation data used. No numerical results were presented to compare the approach with networks which were supposedly "overtrained".

Morgan and Bourlard studied the ability of a network to recognize speech data. In their investigation performance on the test set was monitored after each epoch and training was discontinued when the test set performance showed no further improvement. They report that their results indicate:

> ...that it was possible to greatly reduce the sensitivity to the choice of network

> size by directly observing the network performance on an independent test set during the course of learning (cross-validation). If iterations are not continued past this point, fewer measures are required....

Of studies on other factors which appear to affect network generalization, reports of the relationship between data representation and the generalization exhibited by a network are also of interest. These fall into two general categories. The first category consists of reports regarding the effect of representational changes designed to improve the information content or quality of the data prior to creating the training patterns. An example is the smoothed spectral representation which results from the convolution of a two dimensional Gaussian filter with an original spectral representation of that data. The second category consists of reports regarding changes in data representation which are really changes in the training patterns, for example, using −1, +1 rather than 0, 1 to represent a binary pattern.

Reports regarding the relationship between data representation and network generalization are provided by Leung and Zue [17] and also by Kamm and Singhal [14]. Recall that Leung and Zue used speech data to train a single hidden layer network having 100 input units representing three spectral frames of speech data to recognize 16 American English vowels. Training reached an asymptote of about 80% correct. Speaker-independent testing of the trained networks showed generalization of approximately 54%. In an attempt to improve the generalization exhibited by the network, Leung and Zue decided to improve the information content of the input data by providing the network with information about the 122 possible contexts for the original speech input. This improvement in information content necessitated an increase in the size of the input representation by an additional 122 units. These additional units require additional weights that serve to increase the capacity of the network. Nevertheless, training recognition increased to 95% while generalization increased to 67%. Obviously increasing the capacity of the network did not inhibit the ability of the network to exhibit better generalization when it was trained with the larger but more informative speech input.

It could be argued that increasing network capacity by adding additional input units does not have the same effect as increasing network capacity by adding additional hidden units. Network capacity is usually measured in terms of the number of connections in a network. It is not clear how one would differentially compare the addition of connections resulting from adding input units with the addition of the same number of connections resulting from adding hidden units. Nor do I know of any research that would support such a contention or such an approach. In the case of Leung and Zue's work, it is certainly not clear how one would explain an increase in both performance and generalization based on some such differential change in the network capacity resulting from the additional input units. After all, one can hardly claim that adding input units causes a decrease in network capacity. Such an argument also fails to take into account the additional information provided to the network by the added units.

Kamm and Singhal report on training feedforward networks using speech input representations containing 35, 65, 125, and 245 milliseconds of spectral data. The change in input data representation was made in each instance by mapping the speech data for the particular temporal input span to 147 input units. Varying amounts of temporal information were provided to the networks even though no change was made in the number of input units in the network. Best performance was exhibited by the network using the 125-ms. data representation.

One final comment on the impact of changes to the quality or information content of a data representation on network generalization is provided by Martin and Pittman. They observe that smoothing of the digit and character input data by convolution with a Gaussian distribution "significantly" improves recognition accuracy. Unfortunately, they do not report numerical results for the recognition and generalization for networks trained with either unsmoothed or smoothed data.

Neither do they indicate that they pursued this comparison across network size and training set size. By neglecting to perform this comparison they may have ignored a factor having potentially far greater impact on network generalization than those they chose to investigate.

Reports of the relationship between data representation and the generalization exhibited by a network which fall into the second category, representation changes in the training patterns, are typified by those of Ahmad and Tesauro [1] and Sietsma and Dow [33]. They investigated the effect of changing the input pattern representation from a binary 0, 1 representation to a binary -1,+1 representation. This simple change in representation increased the generalization significantly (5% to 10% better for a given training set). It also decreased the time required to train the network. Sietsma and Dow report on a follow-up test to their work described above in which the network was trained with random noise added to corrupt the training patterns. The resultant networks exhibited dramatically better generalization than networks trained with clean training patterns. For the networks trained with patterns corrupted by the addition of random noise, they report that pruning the number of connections in the network had only a minimal effect on improving network generalization. Sietsma and Dow offer no explanation as to why the addition of random noise to the training patterns resulted in improved generalization.

Landauer, Kamm, and Singhal [16] report on an investigation of the relationship between the task a network is required to learn and the generalization exhibited by the network. They utilized speech data to train a network which was required both to identify the speech input and to auto-associate the speech input. Their rationale for the combination of a classification task and an auto-association task was that requiring the network to perform two tasks should supply a much greater degree of constraint on the representations formed on the hidden units and that the additional constraints might result in a solution having better generalization for classification. They used a network with 1,125 input units, a variable number of hidden units (usually 20), a number of output units corresponding to the number of phonemes to be classified plus 1,125 units for the autoencoded output. A 150 ms. speech window was stepped across the input in 2 ms. steps. They report preliminary results indicating that generalization for the two task network was better than generalization of the network which performed classification only. No numerical results are presented for either approach. They do not indicate how performance was measured (across the total task, across the classification task or the auto-association task only, or by a vote of both tasks).

The reported results for these investigations of the relationship between other factors (amount of training, data representation, or network task) and the generalization exhibited by a network have been rather consistent. Reports that training a network too much can reduce generalization are provided by Weigend et al and Morgan and Bourlard. Leung and Zue report that a change in the information content of the data representation to include context information resulted in a 13% increase in generalization even though it required in an increase in the size of the input layer and the capacity of the network. Kamm and Singhal report "improved" generalization for a pattern data representation containing more temporal information but the same size input pattern representation. Finally, Martin and Pittman report that a change in the preparation of the data by smoothing with a Gaussian filter results in "improved" generalization. These results regarding the relationship between changes in the quality or information content of the data representation and generalization indicate an area where significant improvements in network generalization can be achieved. The impact of data representation on generalization reported by Ahmad and Tesauro and, also, Sietsma and Dow indicate that minor changes in the training pattern data representations can result in significant increases in generalization of from 5% to 10%. Finally, Landauer et al report that training a network to perform more than one task can constrain the network sufficiently that the internal representations are improved and the network exhibits improved generalization. Additional research could provide added insight into the relationship between network task and network generalization.

1.2.3 Generalization: summary of current research. The research regarding the relationship between network architecture and generalization or training set size and generalization are quite clearly contradictory. Probably the most rigorous empirical studies were those undertaken by Morgan and Bourlard and Martin and Pittman. I have already indicated my concerns regarding the incompleteness of Morgan and Bourlard's work: their investigation was extremely limited in terms of the number of hidden units utilized and apparently terminated at a point at which training recognition was still increasing while generalization showed only a minimal decrease. Martin and Pittman's work was far more thorough in its investigation of the relationship between network architecture and network generalization. Like Morgan and Bourlard's work, however, it neglected other factors which can affect network generalization: network task and network training. The results which have been reported regarding the relationship between data representation and generalization exhibited by a network indicate that this is an area where significant improvements in generalization may be achievable. It should be noted that the latter is an area of investigation where the majority of the results appear to have been a serendipitous result of investigations focused on other factors.

None of these studies attempted a truly systematic investigation of the relationship between network architecture, task, training, training data representations and network generalization. The interactions between these factors were for the most part completely ignored.

1.3 Overview

My research is just such a systematic investigation. How do the architecture of a network, the task which a network must learn, the training of a network, and the data representations used in that training affect the ability of a network to generalize well from previously learned examples to new examples? My research investigates the potential interactions occurring between these factors and their relationships to the generalization exhibited by a network.

Unlike Martin and Pittman, I did not attempt to explore the infinity of potential search spaces that can be created when one tries manually to limit the connectivity of a network or to specify the sharing of weights between units. The research is limited to an investigation of feedforward fully-connected networks. I investigate the relationship between generalization and network architecture not just by varying the number of hidden units but by varying the number of hidden layers and also by varying the tasks of the networks. My guideline was the same overfitting guideline used by Martin and Pittman, Morgan and Bourlard, and others: system models which have too many free parameters develop functions which fit the training data too closely and may not generalize well to new measurements. I intended to manipulate the architecture of the input and output layers as well as that of the hidden layers. Utilizing an architectural viewpoint only, I initially hypothesized that training with physically larger representations, which result in increases in the connectivity and the capacity of the network, reduces recognition performance in a network.

I investigate the relationship between generalization and network task by studying networks which have not only the classification task utilized in much of the research discussed above but which also have several composite tasks similar to that reported by Landauer et al. In my investigation, I formally define network task to mean the set of input/output items that the network is required to learn: $T = (i, o)$, where i is an input pattern and o is an output pattern. A total task may be a composition of two or more subtasks. For example, in the network investigated by Landauer et al, if one group of output nodes o_c is used to classify the input, we define the classification task as $T_c = (i, o_c)$. Another group of output nodes o_a is used to auto-associate the input pattern with itself, defined as $T_a = (i, o_a)$. The total task of the network can then be specified as $T = T_c \circ T_a = (i, o_c \circ o_a)$. Here o_c and o_a are seen to be both semantically and syntactically different output responses.

One hypothesis with regard to compound tasks is: requiring the network to perform more than one appropriately related task should supply a much greater degree of constraint on the representations formed by the hidden units and thus should result in a solution exhibiting better generalization. Adding additional related task requirements forces an increase in the number of input and/or output units and therefore an increase in the number of connections in the network architecture. Such an increase in network capacity, according to the overfitting guideline, should result in a reduction in generalization. Accordingly, I hypothesize that adding additional related tasks should cause a change in generalization. That change may be an increase in generalization, as a result of the added constraints, or it may be a decrease in generalization, as a result of an increase in the network capacity.

Beyond these factors, I investigate the relationship between generalization and network training by using two different training algorithms and several different training schedules. The primary training algorithm utilized in this research is the back-propagation algorithm. For purposes of comparison, however, selected network architectures are trained with both back-propagation and conjugate gradient training. Different training schedules, including incremental training and training by task are also investigated. Different training algorithms may facilitate or hinder the development of generalization.

I investigate the relationship between generalization and data representation by focusing my investigation on the effect of representational changes designed to improve the information content or quality of the data. In this respect, I investigate the effect of such representational changes on generalization within networks having the same architecture and similar tasks. The first investigation compares networks having tasks in which the output data representations are the same but there are different data representations for the same input. The second comparison is for networks having tasks in which the input representations are the same but there are different representations for the same output. I also investigate networks having similar tasks in terms of the output representations but changes in architecture resulting from increases in the information content of the input data. Consideration of previous research results indicate that improvements to the information content or quality of the data is a generally underrated factor which has a significant impact on the generalization exhibited by a neural network.

It is clear that the memorization and generalization exhibited by a system are related to the capacity of the system and the number of training examples used in training the system. What should be equally clear is that one can have an extremely large number of training examples, but if those training examples are poor examples of the concepts which are to be learned, it is unlikely that good generalization will occur. The quality of the data representation is a frequently-unemphasized factor which belongs with system capacity and training set size in the pantheon of factors that affect network generalization.

A system having an extremely large capacity is more than capable of memorizing a very small number of training examples. Whether or not that system can generalize well from those few memorized examples to new examples depends not just upon the number of those examples but also upon whether or not those examples contain sufficient information regarding the similarities underlying the concepts necessary for appropriate generalization to occur. If they do not, it is clear that for generalization to occur the system must be presented with examples which do contain the necessary information. Simply presenting more of the same insufficiently-informative training examples will not guarantee that the system will exhibit good generalization.

On the other hand, a system having a very small capacity cannot be expected to memorize a large number of training examples. Likewise, this same system cannot be expected to generalize well from those examples which it does memorize unless they contain sufficient information regarding the similarities which must be extracted in order for good generalization to occur. It is not clear how one would perform an *a priori* measure of the information content or quality of a data representation. It is nevertheless important to begin research into the relationship between such representational changes in data and their impact on network generalization. What is needed is a means of selecting and preparing data representations having high information content for use in training a network so that it will achieve the best generalization possible given the number of training examples available. To borrow a familiar motto: We need a few good examples!

The ability of connectionist networks to extract the similarities embodied in the training instances permits them to generalize to new or exceptional situations. Appropriate representations contain sufficient evidence of the similarities which must be extracted in order for good generalization to occur. It is possible that an appropriate representation might be physically larger than a less appropriate representation. As a result the capacity of the network would be increased. According to the overfitting guidelines, this should result in a reduction in generalization. Nevertheless, I hypothesize that training with appropriate, even though larger, representations will help the network to to generalize better.

1.3.1 Research focus. A major goal of connectionist learning is to design and train neural networks that generalize well to new examples from a domain after having been trained on a sufficiently large set of training examples selected from that domain. Neither the concepts to be learned nor their representations are stored in a neural network. Rather, what is stored is the connection strengths between the units that allow these concepts to be re-created. Generalization in neural networks is the acquisition of connection strengths which reflect similarities extracted from the representations used in training the network. As yet we do not understand the conditions a network requires in order to do a good job of extracting similarities from the representations used in training the network. We know that, given an appropriate architecture, training algorithm, and sufficient training data, a connectionist network generally can find a function that will memorize the training set. What we do not know is if that function will permit the network to generalize well to data which it has not seen before. My research is an investigation of this complex question: how do the task which the network must learn, the architecture of the network, the training of the network, and the data representation used in that training, both individually and together, affect the ability of a network not just to learn the training data but to generalize well to previously unseen data.

Before discussing the specific experimental methodology used in this investigation it is necessary to consider the problem domain in which this investigation will be carried out.

CHAPTER 2

TASK: SPEECH PERCEPTION

2.1 Experimental Strategy

The four independent variables used in my investigation of network generalization include: network task, network architecture, network training, and data representations. These variables are not all equally easy to manipulate. For example, it is easy to manipulate the network architecture; one need only change the number of hidden units in the network. Likewise it is fairly easy to manipulate the network training. If back-propagation training is being used one can simply change the learning rate being used in order to effect a change in training. Training can also be manipulated by using different training algorithms: back-propagation and conjugate gradient training, for example. Accessing and manipulating network task or manipulating the quality or information content of the data representations used in the training are nowhere near as readily accomplished. One could manipulate the task requirement by requiring a network to accomplish one or the other or both of two or more arbitrarily specified tasks, for example, identify a randomly presented vowel spectrum and, at the same time, complete a randomly presented partial picture of an object. Should a network fail to succeed at the composite task, however, the failure might be ascribed to the unrelatedness of the subtasks or to the composite nature of the tasks. What is needed is a domain of investigation in which two or more related tasks can naturally and logically be required of a network. The more reasonable the relationship between the tasks the more readily analyzable the relationship between the manipulated network task and the generalization exhibited by the network should be. Similar arguments can be made with respect to the relationship between data representation and network generalization. The remainder of this chapter will be devoted to developing the approach to the manipulation of network task that was used in this research. Chapter 3 will be devoted to developing the approach that was used in the manipulation of the quality and information content of the data representations. I defer the discussion of network architecture and network training until Chapter 4.

To rigorously manipulate the task required of the network and observe the resultant effect upon the generalization, I needed to have a domain in which meaningful variations in task requirements could be explored. Certainly it would have been possible to develop an artificial domain in which to carry out my investigation. To do so would require making assumptions about the underlying structures of that domain that might or might not be reasonable when compared to practical applications of interest to researchers such as visual perception or speech perception researchers. Using an artificial domain could result in criticism of my investigation based on several objections: the domain could be considered to be a "toy" domain, objections could be raised concerning the underlying justifications for the task requirements, and it could be claimed that the results of such an investigation simply might not scale-up to larger networks using larger representations of collected experimental data.

Using a real-world domain avoids each of these objections. It also eliminates the time and effort involved in attempting to develop an appropriate artificial domain; such time and effort could be devoted more profitably to investigating the question that forms the focus of my research.

My strategy required that I find a real-world domain in which the ability to generalize from specific instances to a common concept was present. In addition, I needed a domain in which more than one task naturally occurred. Beyond that, the domain should permit more than one quality or level of information content of the data. For such a domain, the tasks carried out in the domain together with the representations appropriate to the domain would assist in the development of alternate network architectures appropriate for use in my investigation.

Speech perception satisfies these requirements. Speech perception is a domain in which the ability of listeners to generalize from specific instances of a spoken word to all instances of that word is an ongoing area of research. Speech perception requires both the ability to listen to and to identify a word. Finally, a variety of different quality or information content representations can be used for speech data. These include both unsmoothed and smoothed spectral representations resulting from Fast Fourier Transforms (FFT) of the speech waveforms and differing temporal spans of the spectral representations. The remainder of this chapter will be devoted to a discussion of generalization in speech perception. I will show how the tasks involved in speech perception can be used to construct neural networks appropriate to my research.

2.2 Generalization: Speech Invariance Problem

How does a listener generalize from the perception of specific instances of a spoken word to all instances of that word? A human listener can extract from the constantly varying speech waveform the phonetic segments which convey the speaker's linguistic message. From those segments the listener is able to make an identification of the linguistic message. This identification coincides with that made for similar phonetic segments uttered by other speakers. If speech were produced in a manner where the acoustic utterance consisted of a simple temporal concatenation of basic acoustic events, the identification problem would be relatively simple: determine the representation for each of the basic acoustic events and use these canonical representations to segment and analyze the utterance. Speech is not produced in this manner [20]. To understand the obstacles which are overcome in identifying a speaker's linguistic message it is necessary to understand something about the relationship between vocal production and the acoustic characteristics of speech. Speech is produced by the vocal organs; the lungs, the trachea, the larynx, the throat, the nose, and articulators such as the tongue and lips which are physically constrained in their movements. Air from the lungs travels up the trachea, through the larynx where the vocal cords are located, and into the vocal tract which consists of the oral cavity and the nasal cavity. If the vocal cords are constricted the air will cause vibrations in the cords and resulting sounds will have a quality that is referred to as **voiced**. Sounds generated without this constriction are called **voiceless**. By manipulating the lips, teeth, tongue, and soft palate various constrictions can be created within the vocal tract and a resulting variety of sounds produced. When the airstream through the vocal tract is obstructed in some fashion so as to produce turbulence, the resulting sound is classified as a consonant. When the airstream is relatively unobstructed and voicing is present the resulting sound is classified as a vowel.

Some differences are occasioned by the movements of the articulators. As the articulators move from one position to another the shape of the vocal tract changes. Since the shape of the vocal tract determines the resonance frequencies, changes in shape occasioned by movements of the articulators result in changes in the resonance frequencies or formant frequencies of a speech sound. Acoustic transitions (coarticulation effects) arise between phonemes as the speaker utters each one in context. It has not been possible to find portions of the speech waveform which uniquely match the perceived phonetic segments or phonemes of the utterance [34]. This is a first obstacle which must be overcome in identifying the linguistic message of a speaker.

In addition to being physically constrained in their movements, articulators differ between individuals. They also differ within an individual as that person grows from infancy to childhood and adulthood. Since the physical characteristics of articulators vary both between speakers and within speaker, the acoustic representation of the same speech percept varies as a function of these physical characteristics. This is a second obstacle which must be overcome in identifying a speaker's linguistic message.

Speech sounds also vary temporally as the speaker changes the rate of speaking or as the surrounding context differs. These three obstacles are among some of the acoustic-phonetic variability problems encountered in the perception of speech. Figures 2.1 and 2.2 depict the spectral representations of the speech waveforms for the word "kill" spoken by two speakers. Note the difference in the time required by each speaker to pronounce the word. The difference in formant frequencies (dark bands) for the male voice and the female voice can be readily seen as can the different formant transitions between the two speakers.

Having briefly considered some of the obstacles which must be overcome in order for generalization to occur in speech perception, I turn now to a discussion of the tasks which are involved in the speech perception process.

2.3 Tasks: A Model of Speech Perception

A number of models of speech perception have been proposed by speech researchers. These models can be broadly classified into two contrasting groups: those which consider the perception of speech to be a separate and distinct process from the production of speech, and those which consider the perception and production of speech to be fundamentally and indivisibly related. Since they are not pertinent to the present investigation, I have confined the discussion of perception-only models of speech perception to Appendix A.2 for the interested reader.

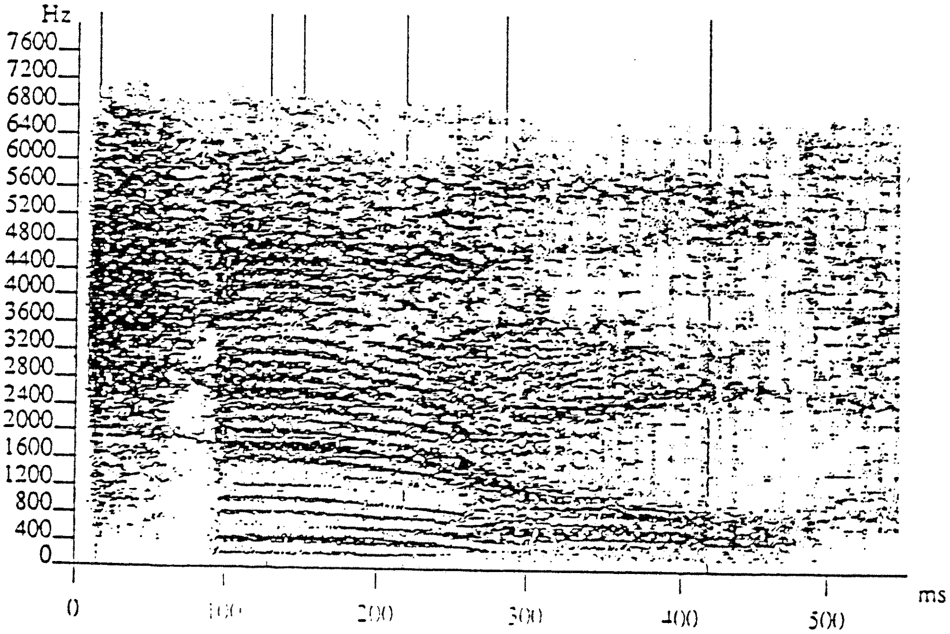


Figure 2.1: Male speaker - "kill"

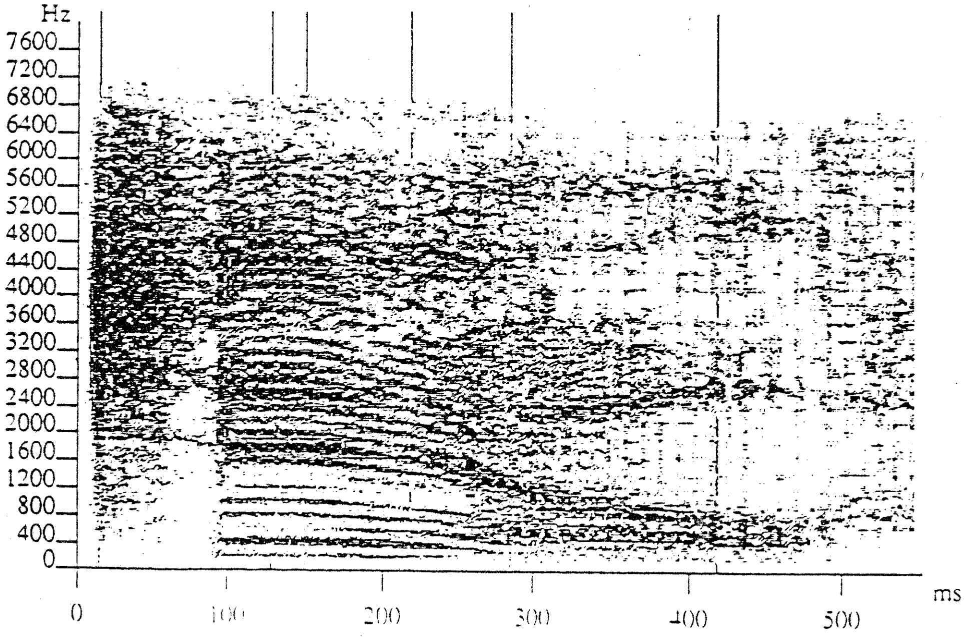


Figure 2.1: Male speaker - "kill"

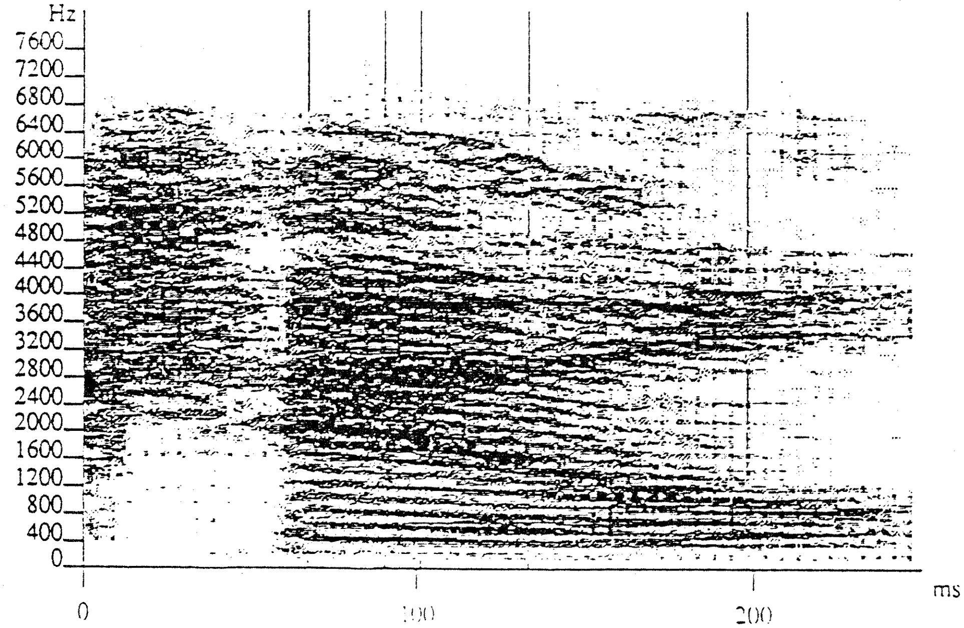


Figure 2.2: Female speaker - "kill"

Several models implicate the production process in the perception of speech. The Motor Theory model is a popular example of this approach. In its earliest form, Liberman and Mattingly's Motor Theory [18], takes as a basic tenet the view that there is a unique speech mode of perception in which the listener attends not to acoustic characteristics common to all auditory perception but to special characteristics specific to speech. These speech-specific characteristics are a result of a unique encoding of speech mediated by neuromotor commands to the production articulators. Neuromotor commands are assumed to exist in the nervous system of the speaker and the listener. According to Motor Theory perceptual generalization of speech occurs through the activation of a neuromotor representation of the phonetic segment.

The speaker's signal must be decoded by the listener. An auditory decoder would process the signal in auditory terms. Such a decoder would require complex processes to deal with the acoustic transitions and variation in the speech signal. To the motor theorist this implies an unparsimonious system in which two entirely separate but equal processes for the encoding and decoding of speech exist side by side. Motor Theory proposes a single system, with appropriate linkages, in which perception is mediated by the neuromotor commands of articulation.

Revised Motor Theory differs from early Motor Theory in that

> ... the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations [20].

These intended gestures are the primitives of both speech production and perception. According to Liberman and Mattingly the relationship between production and perception is not learned, but instead is an innate part of our human heritage. Perception is also said to occur in a specialized speech mode which differs significantly from the standard mode of auditory perception.

Liberman and Mattingly's Motor Theory of speech perception attempts to deal with the problem of acoustic-phonetic variability by hypothesizing that the perception and production of speech are inextricably tied in the human speech perception process. The Motor Theory model of speech perception is a model of a real-world domain which requires that a generalization from a specific auditory input to a common linguistic category be made. In this model we find more than one task naturally occurs; the human listens to a speech sound, identifies that sound, and the identification is mediated by the neuromotor commands. My strategy required that I find a real-world domain in which more than one task naturally occurred. Motor Theory meets that requirement. In the discussion that follows I will show how this model assists in the construction of network models which will be used in my investigation of generalization in neural networks.

2.4 Tasks: Connectionist Implementation

The majority of extant connectionist models for speech recognition are systems whose only task is to classify the speech input. Using the information provided during training, the network is forced to develop its own internal representation for classification. How appropriate this internal representation may be is determined by the generalization exhibited by the network in labeling previously-unseen speech input. Simple classifier networks can provide good recognition performance on the data on which they have been trained, but they sometimes do not generalize at all well to data which they have not seen before. The only task required of the classifier model is that it map the auditory input, i_a, to an output phonemic classification or identification of the sound, o_c. Using the definition of task presented in Section 1.2.2 above, this identification task can be represented as $T_c = (i_a, o_c)$. In my research, a classifier model is used to provide a comparative network against which the performance and generalization of multi-task networks can be measured.

In order for an individual to recognize a spoken word, however, he or she must have heard the word. By analogy, in order for a network to recognize an input speech sound then the network must have heard or listened to that sound. A question arises as to how the ability to listen could be actualized in a network architecture.

Recall that Landauer, Kamm, and Singhal [16] (see Section 1.2.2 above) utilized speech data to train a network which was required both to identify the speech input and to auto-associate the speech input. They reasoned that requiring the network to perform two tasks should supply a much greater degree of constraint on the representations formed on the hidden units and that the additional constraint might result in a solution exhibiting better generalization for classification of speech. Their rationale for requiring an auto-associative task of the network was not based on an analogy with human speech perception. Nevertheless, using an auto-associative task would be one way of actualizing the requirement that the network hear or listen to the input speech sound.

In such an auto-associative network the network is given the task of reproducing the auditory input representation on the output nodes; the auto-associative task would be represented by $T_a = (i_a, i_a)$. In an auto-associative network the hidden layers, which usually have fewer degrees of freedom than the input and output layers, represent a bottleneck for transmission, so the network is forced to develop a compact encoding of the data in the hidden layers. Simple dimensionality reduction does not necessarily lead to a better encoding for the purposes of speech recognition. In addition, people do not necessarily hear the 'exact' sound uttered by a speaker. In my research I extend the auto-associative concept by claiming an *a priori* group-associative relationship between a specific instance of a speech input for a particular word and all other instances of speech input for that same word, whether spoken by the same or other speakers. For example, the input nodes are set to the spectral representation for one speaker pronouncing the vowel "o". Instead of duplicating that input on the output nodes, the output target values are those of the spectral representation of a reference pronunciation of the same vowel. Using this approach, many input speakers pronouncing the vowel "o" are mapped to a single reference pronunciation. This appears to be a small change, yet the network could be led down quite a different path than a simple auto-associative network. 'Group-association' differs from auto-encoding in that the network must search for a coding that identifies between-speaker commonalities. In searching for a new multi-dimensional space, the network must derive one that is insensitive to speaker characteristics. By the group-association process the network can discover the subtle similarities between the speech of one person and the speech of another. The group-association task would be represented by $T_g = (i_a, o_a)$ where o_a may or may not equal i_a.

According to the Motor Theory model of speech perception a human being listens to a speech sound uttered by a speaker and identifies that sound by means of the intended phonetic gestures associated with that sound. These intended phonetic gestures are represented in the brain as invariant motor commands for movements of the articulators appropriate to that sound. The intended phonetic gestures provide the third constraint used in the construction of my network models. While it is impossible to specify exactly what those intended phonetic gestures are for a particular sound, it is possible to assign a unique set of abstract phonetic features to represent phonetic gestures. The articulatory association task would be represented by $T_a = (i_a, o_r)$, where i_a is the auditory input, o_r is the set of abstract articulatory features associated with that input.

Utilizing the phonemic classification task, the auditory group-association task, and the articulatory association task I was able to construct three additional network models for use in my investigation of generalization in neural networks. Combining the phonemic classification task with the auditory group-association task led to the construction of the "Echo" model network. The Echo model has a composite task $T = (i_a, o_c \circ o_a)$, where i_a is the auditory speech input, o_c is the output classifier, and o_a is the auditory reference. Combining the identification portion of the speech perception task with the articulatory association task I constructed the "Mimic" model network. The Mimic model has a composite task $T = (i_a, o_c \circ o_r)$, where i_a is the auditory speech input, o_c is the output classifier, and o_r is the set of abstract articulatory features associated with that input.

By combining all three of these tasks I could have specified a third network model in which the composite task would be specified as $T = (i_a, o_c \circ o_a \circ o_r)$. The Motor Theory model of speech perception requires that a human being listens to and identifies a speech sound uttered by a speaker by means of the intended articulatory gestures associated with that sound. I reasoned that at some point the physical invocation of those articulatory gestures must be associated with that speech sound. Such a process could be that which occurs in an infant "babble" situation. Here the classification task would be described as $T_c = (i_r, o_c)$, the auditory association task becomes $T_{aa} = (i_r, o_a)$, and the articulatory association task is $T_{ra} = (i_r, i_r)$.

Accordingly I incorporated the aspect of articulatory input into the fourth network model which I call the "Full Motor Theory" model. Figure 2.3 shows a diagram which provides a general network representation of the Full Motor Theory model used in my research. At the input layer the network either listens or it speaks (articulates). Both events do not happen at the same time. This network has a composite task which can be described as follows:

i_a = auditory input i_r = articulatory input

o_c = classifier o_a = auditory reference o_r = articulatory features

$T = (i_a \circ i_r, o_c \circ o_a \circ o_r)$

$i_a \Rightarrow \neg i_r$ $i_r \Rightarrow \neg i_a$

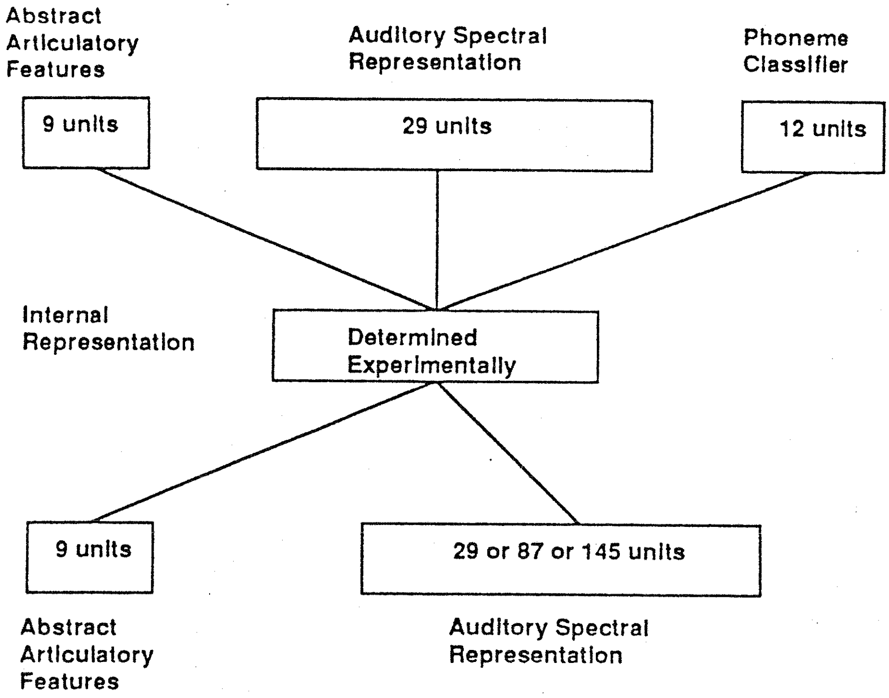


Figure 2.3: Full Motor Theory Model

This total task is a composite of three speech perception tasks: classification of the speech input (auditory or articulatory), association of that speech input with a reference auditory representation for the input, and association of the speech input with a set of abstract articulatory features. In this model, processes for encoding and decoding speech do not exist separately. Rather, this model provides a single system, with appropriate linkages, in which the perception of speech is coordinated with articulation. Note that perception occurs in a specialized speech mode which differs significantly from perception-only models of auditory perception.

Using Liberman and Mattingly's motor theory as my model of speech perception, I created and explored the use of multi-architectural, multi-task neural networks as speech recognition systems. Each of these networks was implemented, trained, and tested using the same speech database with varied data representations.

I have shown that speech perception satisfies the requirements needed for my research: it requires the ability to generalize from specific instances to a common concept; it provides a psychological model, Motor Theory, which naturally embodies more than one task; and, as will be discussed next, it permits more than one kind of data representation to be used in carrying out those tasks.

CHAPTER 3

DATA REPRESENTATION: VOWELS

What is required in training and testing neural networks is data which will provide a good test of the networks' ability to generalize. It is not necessary to investigate the entire set of speech sounds of a language to achieve that goal. Vowels are an important subset of speech sounds which have been the subject of much investigation by both speech researchers and connectionist researchers. What was required in preparing the data was a reasonable means of developing alternate data representations for that purpose. Existing theories of vowel perception can help in determining alternate representations for use in training and testing the networks. Vowels are a good choice for the speech data to be used in my investigation of generalization in neural networks. Because the collection and preparation of the data representations was such a complex portion of this research, I will devote this chapter to a discussion of this topic.

3.1 Theories of Vowel Perception

In the past, vowels have been characterized articulatorily as static vocal tract shapes and acoustically as points in a first and second formant space [15]. This characterization led to a unifying model of vowel perception in which the vowel target was conceived of as a canonical form of the vowel which formed the goal state for a vowel spoken in continuous speech and which is recognized perceptually by the acoustic information provided by the target frequencies of the first two formants of the vowel.

New theories of vowel perception have been developed in the past twenty years as a result of research which characterizes vowels as articulatory, acoustic, and perceptual events [36]. An elaborated target theory of vowel perception represents vowels as target zones in perceptual spaces whose dimensions are specified as formant ratios. This theory of vowel perception represents an attempt to deal with the speaker-normalization problem - the ability to generalize perception across speakers. A dynamic theory of vowel perception differs from the target theory by emphasizing the importance of the formant trajectory patterns, or transitions, in the perception of vowels. This dynamic theory represents an attempt to deal directly with the temporal nature of speech and problems of coarticulation and diphthongization - the ability to generalize perception within a single speaker's speech.

My research utilizes modified forms of both the target theory and the dynamic theory of vowel perception. In the first instance, I utilize one short-term spectrum, ("1-frame" data), selected from the stable portion of the vowel. This short-term spectrum is an exemplar of the central tendency of the vowel. For the modified dynamic transitions, I utilize three short-term spectral frames from three different times in the vowel. These three spectra are exemplars which are selected to represent the temporal sequence in which the vowel unfolds as it is spoken. One is a frame selected from the on-glide formant transition of the vowel, one from the stable portion, and one from the off-glide formant transition. In this "3-frame" representation the middle frame is the same as the single frame used in the static target vowel representation. The dynamic transitions of the vowel are represented in the first and last frames which are samples from the on-glide and off-glide transitions. Thus, representative temporal information for a given target is sampled from the target's surrounding temporal context.

Based upon these two theories of vowel perception, the representations used in my investigation consisted of 1-frame and 3-frame speech input for vowels excised from American English[1]. Network performance was evaluated in three different ways: (1) how well the network recognizes speech data on which it has been trained; (2) how well the network generalizes to new speech data spoken by the training speakers; and (3) how well the network generalizes to speech data spoken by new speakers.

3.2 Speech Database

The speech database used in the training and testing of the networks consisted of a subset of the sounds of English: the twelve vowels which are represented by the single character Arpabet notation /i I e E @ a c o U u A R/ as in heed, hid, hayed, head, had, hod, hawed, hoed, hood, who'd, hud, and heard.

Given the confusion in the field of human speech perception, it is unclear whether the static target-vowel model will be capable of accurate identification, though the dynamic model ought to succeed. To make the test as realistic as possible, it seemed necessary to make the task as strenuous as possible. I wanted to provide a difficult test of the ability of the networks to generalize by emphasizing the variability of the speech data used in training. The problem was to identify appropriate vowel contexts that would emphasize speech variability. I especially wanted to identify contexts that would stress the variability of the dynamic transition representations.

Liberman et al [19] provide an example of two synthesized CV syllables which are perceived as /di/ and /du/. In the /di/ syllable the onset transition shows a rising second formant while in the /du/ syllable the onset transition shows a descending second formant. Contrasting the stops /d/ and /g/ before the vowel /a/, Liberman and Mattingly [21] show that simply changing the onset glide on the third formant from descending, /d/, to ascending, /g/, results in the different perceptions. The differing formant onset transitions occasioned by a word-initial /t/, /k/, or no-consonant provide the rationale for the utilization of these phonemes as specifiers of

[1] In order to avoid much tiresome duplication of the phrase 'American English', all discussions pertaining to speech sounds in this and subsequent sections should be interpreted solely as referring to the speech sounds of American English.

the initial phonemic context of the words collected in the study [39]. Using stops for the final phonemic context provided similar offset transitions, while using voiced phonemes for the final context served to lengthen the vowel.

With these considerations in mind, eight different contexts were used for each vowel. The individual word contexts for each vowel were selected from a list of fifteen potential transition environments. These fifteen possible primary contexts were specified by the phoneme combinations listed in Table 3.1. For example, possible phoneme combinations are tEll, kId, In etc. Using these contexts provided good examples of the variability of the vowels.

Table 3.1: Fifteen Primary Contexts

```
t  \   /  b, d, g
    \ /
k -- V -- n          V = {i I e E @ a c o U u A R}
    / \              - = null
-  /   \  l
```

When it became impossible to find words that conformed to these primary contexts the transition constraints were relaxed to include the optional contexts indicated in Table 3.2. Using these specifications, a table containing 180 possible word contexts was created, with 15 possible contexts for each of the twelve vowels included in the study (Table 3.3).

Table 3.2: Primary and Optional Contexts

```
t[d,s,st] \      / b[p], d[t], g[k]
           \   /
  k[sk,g] - -V -- n               V = {i I e E @ a c o U u A R}
           /   \
    -[h] /      \ l               - = null
```

With the exception of /U/, eight different word contexts were selected from the possible 15 contexts. For /U/ six different words were utilized. Care was taken to

Table 3.3: Speech Tokens for Data Collection

	-i-	-I-	-e-	-E-	-ɑ-	-R-	-A-	-u-	-U-	-o-	-c-	-a-
t-b	teepee	sibling	table*	debit*	tab	turbine*	tub	tuba*		Toby*	daub*	sobbing
t-d	teed*	sitting	stated*	Ted	tadpole*	turd	stud*	Tudor*	stood*	toad	tawdry	toddler*
t-g	fatigue*	stigma*	sake	deck	tag	circle*	tugboat		took*	toga*	talk	toggle*
t-n	teen*	tin*	attain	ten	tan*	turn	tundra*	tuna*		tone*	tawny	Don
t-l	teal	till	entail*	tell*	talent*	sterling	Tull	toolbox*		toll*	tall*	tolerant
k-b	Keebler*	kibble*	cable*	Ceb	cab*	curb	cub*	scoop		Cobe	scaup	cob*
k-d	keyed	kid	skating	Keds	cad	curdle*	cud	cooed*	could*	code*	cawed*	cod*
k-g	geek	giggle*	cake	keg*	gag*	gurgle	McGuckin	cougar*	cook*	cocoa*	gawk*	cog
k-n	keen*	kin*	cane*	kennel*	can*	Kern	cunning	racoon*		cone	gone*	con
k-l	keel*	kill*	kale	Kelly*	Cal	uncurl*	cull*	cool		coal	call*	collar
-b	Hebrew	hip*	Abe	inept	Abner*	herb*	hub*	hubris		oboe*	hobnob*	
-d	heed*	hidden*	aid*	edit*	add	herd*	udder	hooted*	hood*	ode	awed	odd
-g	eke	ignorant	aching*	egg*	Aggies	erg	ugly*	Ugritic	hook*	ogle	auger*	hockey*
-n		in	arcane*	hen*	an	earn*	unknown*			own*	awning*	honor*
-l	eel*	ill	ale*	ell	alley*	earl*	ulcer*	oolong		Olson	all	olive*

The words selected for use in the data collection are starred (*).

ensure that no set of words for a given vowel contained more than two final contexts of the same class, e.g. no more than two -A- words ended in b. Care was also taken to ensure that there was at least one example for each initial context. 94 words in all were selected from this table and used in the collection of the speech data from each speaker. These are marked in Table 3.3 with an asterisk.

3.3 Speaker Population

In order to further test the ability of the networks to generalize, the variability in the speech data was stressed across speakers and data was collected for more than one speaker.

The multi-speaker training and cross-validation set included data from a population of ten speakers, five female and five male. The training and cross-validation databases, therefore, consist of 940 words. Of these, 700 words are used in training the networks and 240 words are used in the cross-validation.

To test the speaker-independent generalization of the networks, speech samples were also collected from three new female and three new male speakers. 564 words were collected from these new speakers, for use in the speaker-independent testing of the network. Four examples of each vowel for each speaker, for a total of 288 words, were randomly selected for use in the speaker-independent testing of the networks' ability to generalize to new data spoken by new speakers.

The majority of speakers were selected from the standard college-age population group, although three of the females and one of the males fell into the middle-age group. All of the speakers spoke English as their native language. One of the speakers was born and spent her youth in Canada. Several of the speakers were non-native speakers of foreign languages. Half of the speakers reported suffering from head colds.

3.4 Speech Data Collection

Speech data was collected in the ICS Speech Lab utilizing the computerized recording facilities of the lab. The 94 vowel context words were permuted and divided into two separate lists so that the data collection process might include a short break halfway through the list for each speaker. In response to computer prompts, speakers spoke into a Shure SM-10A close-talking microphone. Each prompt was displayed twice. At the first prompt the speaker practiced saying the word. At the second prompt, the speakers response was filtered with an elliptical low-pass analog filter with 6.4 kHz cut-off frequency and a 360 dB/octave roll-off and recorded. The sample rate was 16 kHz.

Subsequent to collecting the sampled speech data, it was necessary to process the speech waveform data for use as input patterns to train and test the networks. Each waveform was displayed using the ICS Speech Lab SPEECHBENCH software. Initial and final silence portions of the sample were removed and the reduced waveform was saved. All subsequent processing was applied to the truncated digitized speech sample.

Processing of the truncated data began with a manual verification and segmentation of each of the 1504 tokens used in the three databases. The waveforms were segmented to demarcate the portion of the waveform which corresponded to the vowel. At the recommendation of Dr. Alan Bell of the Department of Linguistics, all tokens were segmented so that the initial vowel segment label occurred at either the release of a preceding stop, when a stop was present, or at the beginning of voicing. The final vowel segment label was positioned at the closure of the vowel. Readers familiar with speech segmentation will recognize that such labeling results in an approximate segmentation of the speech waveform; due to coarticulation effects, an absolute segmentation is difficult to achieve.

Next, it was necessary to extract a spectral frame from the stable portion of the vowel segment for use in the preparation of 1-frame target vowel representations. I also wanted to extract spectral frames from the on-glide and off-glide transitions for use in the 3-frame dynamic transition representations to be used in training and testing the networks. Having demarcated the vowel portion of the speech waveform, a variety of approaches was explored in an attempt to determine an effective technique for placing extraction points within the vowel segment.

I wanted to select a single frame from the stable portion of the vowel. Based upon a suggestion by Dr. Bell regarding average vowel duration for a typical speaker, I hypothesized that a point 120 msec. from the initial segment label should be in a fairly stable portion of the vowel. In order to test the viability of this approach a program was developed to measure the total time between the initial and the final segmentation labels for each vowel segment. Plots of these total time measures showed that all the speakers exhibited consistent production of longer vowels in some tokens than in others. For example, all the speakers took longer to pronounce the vowel /e/ in the token "cane" than they did in the token "table". Within a particular token, for example "cane", some of the speakers exhibited time intervals of less than 300 msec. while others exhibited time intervals of more than 500 msec. or more, see Figure 3.1. For a short vowel such as /I/ the range of times exhibited was from less than 80 msec. to more than 500 msec. Thus, the within-token duration was such

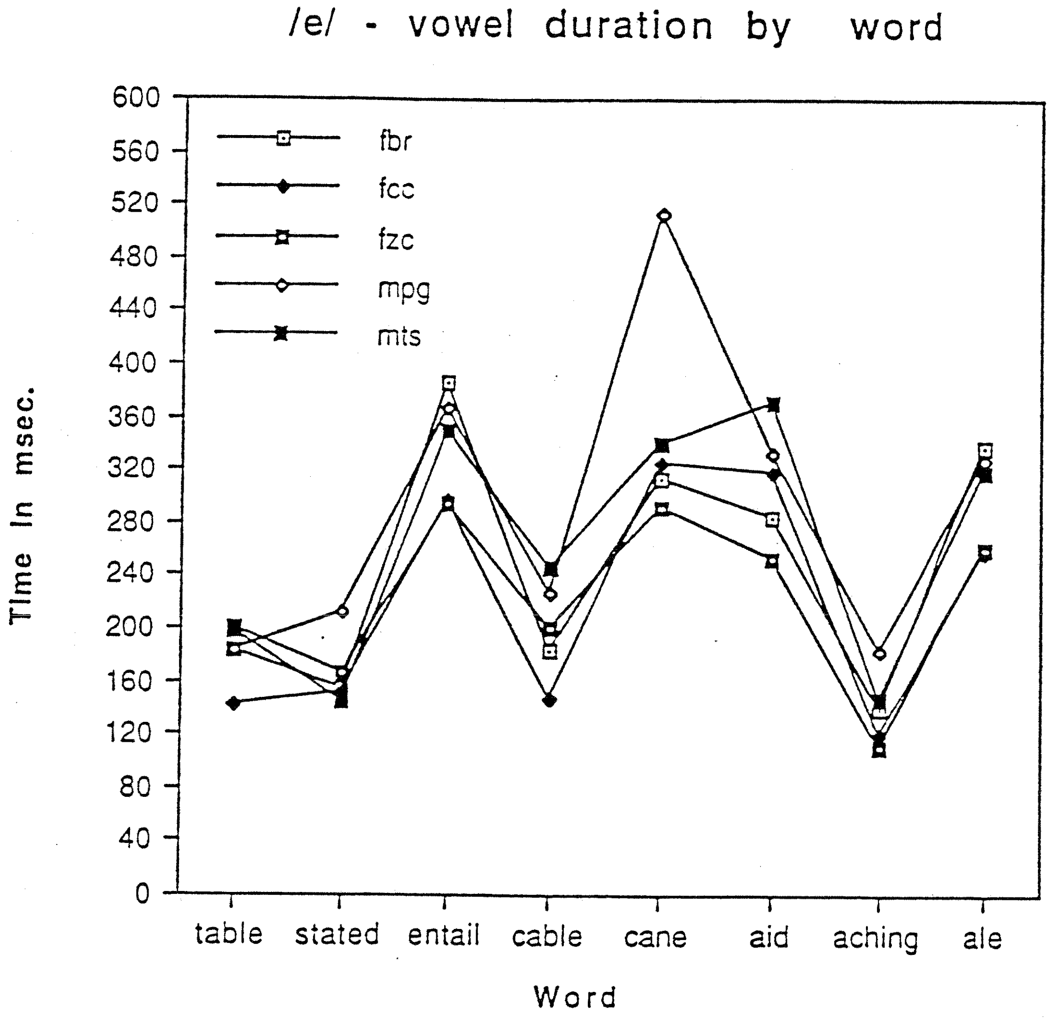


Figure 3.1: Comparative Vowel Durations

that in some instances the 120 msec. point was at the end of, or even beyond, the closure of the vowel segment while it was at the beginning or in the middle of the vowel segment in other instances. Obviously, a fixed duration approach would put the selected extraction point in very different portions of the vowel, depending upon the vowel context and the speaker. This approach to the placing of extraction labels was not deemed viable.

I then tried two automatic approaches to determining the extraction labels. The first approach required a Fast Fourier Transform (FFT) of the speech waveform. The resulting spectral representation was subjected to a minimum Euclidean distance measure between adjacent spectral frames to locate the most stable portion of the spectrum. The Euclidean distances were computed for all words spoken by four of the speakers. Examination of the spectral representation of the tokens showed that selecting a frame using the minimum Euclidean distance approach resulted in many instances in selecting a frame from the final portion of the diphthongs /e/ and /o/. The selected frame was reflecting the second /I/ or /A/ vowel of the diphthong rather than the initial vowel. This approach also was not considered viable.

The second automatic approach computed the point at which the amplitude of the speech waveform was at a maximum. In many instances this procedure proved to be a reasonable approach to selecting an extraction point. In other instances the maximum amplitude occurred in the /I/ or /A/ rather than in the initial vowel of a diphthong. This approach also was not deemed viable.

Visual examination of the speech waveforms indicated that simply extracting a frame at a point halfway between the onset and offset marks might be a viable approach. A third automatic approach was investigated which simply labeled the vowel segment according to this interval. Visual examination of the resulting speech waveforms of all words spoken by four speakers showed that the point midway between the initial and final marks was in the stable portion of the waveform for over

80% of the words. This was a reasonably good extraction point for the target vowel spectral frame representation. The words in which this was not the case were noted. For these words the appropriate fraction of interval location for the stable target vowel extraction point was estimated from a comparison of the waveforms for the four speakers. Using these estimates as a guideline, the exceptional words were then examined for all of the speakers. In the majority of instances the halfway point proved to be a reasonable approximation for the location of the stable target vowel extraction point for all speakers. When necessary, the actual location of the target extraction point was determined by a criteria which specified the point as being between the second and fourth pulse of the 'typical' vowel representation for that speaker.

Using a similar approach resulted in the specification of formant transition selection points at 1/3 and 2/3 of the labeled vowel-interval. These points gave formant transition evidence of the effect of the vowel's contextual environment in all cases.

As a result of the above study, it was decided that the frame selected for use as the 1-frame target vowel representation and for use as the middle frame in the 3-frame dynamic transition representation would be extracted from the labeled vowel segment at a point located halfway between the start and end labels of the segment. The exception was the less than 20% of the vocabulary whose extraction point was manually labeled. For training with dynamic transition data the frames selected for use included this single frame plus those spectral frames at the 1/3 and 2/3 vowel-interval points.

Once the extraction labels were located within the vowel segment the speech waveforms were then processed to provide a representation similar to that developed by the human auditory system.

3.5 Preparing the Data Representation

In order to understand the motivation for the auditory processing used in preparing the spectral representations it is first necessary to consider briefly how the human auditory system processes sounds. While little is understood about how auditory processing of speech sounds is accomplished at the higher levels, there is some knowledge which is pertinent to my research. Research on auditory processing capabilities at the single neuron level has been carried out for more than three decades. Results indicate that the response patterns of neurons in the AI (Auditory cortical field I) region are frequency and intensity sensitive with the characteristic frequency of the neuron related to the place of resonance along the cochlear partition to which it is ultimately connected [5, 9]. There appear to be quantitative differences in the selectivities for human speech sounds in different auditory cortical areas and the cortical auditory system is highly segregated cochleotopically [5]. The interested reader can find a more detailed discussion of how the human auditory system processes sounds in Appendix B.

What is significant for my research is the preservation of frequency and intensity selectivity along the auditory pathway, the existence of selectivities for human speech sounds, and the apparent functional preference for the detection of similarities rather than change in the input. With these constraints in mind, the speech waveform was subjected to FFT processing using a 16 msec. Hamming window and a 3 msec. window shift with a 128-point FFT representation output. Intensity was measured in log dB. The FFT output was then converted from a 128-point linear frequency scale to a 128-point bark scale [43]. The bark scale representation was then compressed to a 48 point scale and smoothed via a two-dimensional Gaussian filter with a standard deviation of 15 msec. in the time dimension and approximately 3 barks in the frequency dimension. The smoothed representation was then again compressed to create a 32 point "brad" [31] scale representation. Using this 32 point brad representation of all of the speech tokens, databases were created containing speaker, context, and vowel information together with the three frames of brad scale FFT data based upon the extraction point labels. These speech databases were used in the preparation of all the smoothed patterns used in training and testing. Figures 3.2 and 3.3 show examples of the smoothed 1-frame auditory representation for each of the twelve vowels spoken by the female speaker (Section 2.2). An alternate version of the database was created by performing the same signal processing with the exception that the bark scale representation was not smoothed via a two-dimensional Gaussian filter.

3.6 Preparing the Training and Testing Patterns

3.6.1 Preparing the auditory patterns. The log dB intensity values for the spectral representations in the database had a potential range which varied from a minimum of 0 to a maximum of 255. The actual range across all speakers for all the speech tokens collected varied from a minimum of 0 to a maximum of 72. The articulatory patterns and the phonemic classification patterns both used a binary 1/0 representation. While one could use the actual intensity values in the auditory patterns, the disparity between those values and the on/off (1/0) nature of the other pattern units used for articulatory and phonemic label representations leads to obvious problems in attempting to train the network using back-propagation techniques. In order to avoid the possibility of activations from the auditory patterns swamping the activations from the articulatory and phonemic patterns and to avoid difficulties in determining the error criterion, the auditory patterns had to be mapped to the range $[0.0, 1.0]$. It is unlikely that the human brain normalizes the intensity of a speech waveform based upon some anticipated maximum intensity value. It did not seem appropriate to so normalize the data in order to map it to the desired range. Other representations, such as normalizing within speaker or within

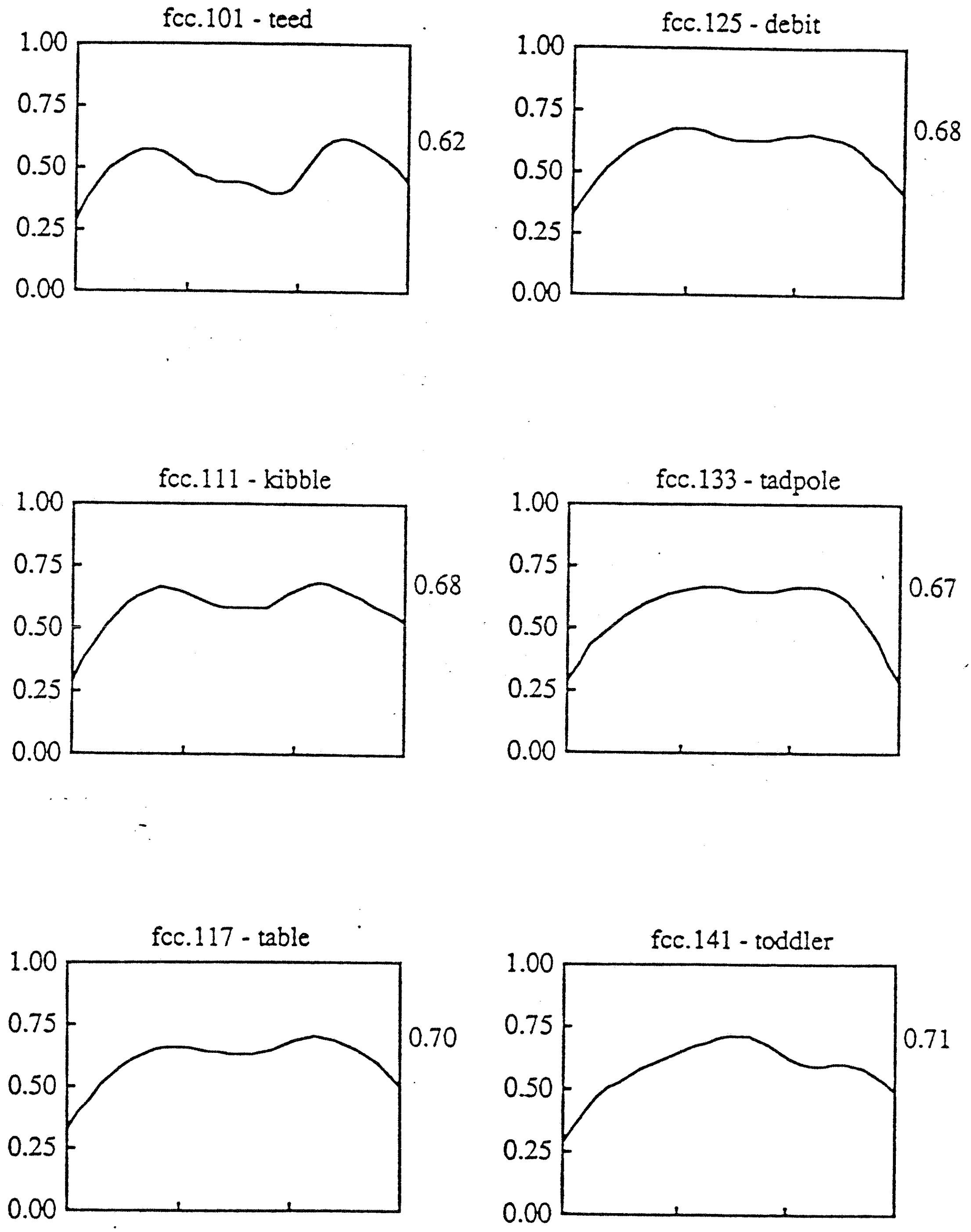


Figure 3.2: 1-frame smooth auditory representations (a - f)

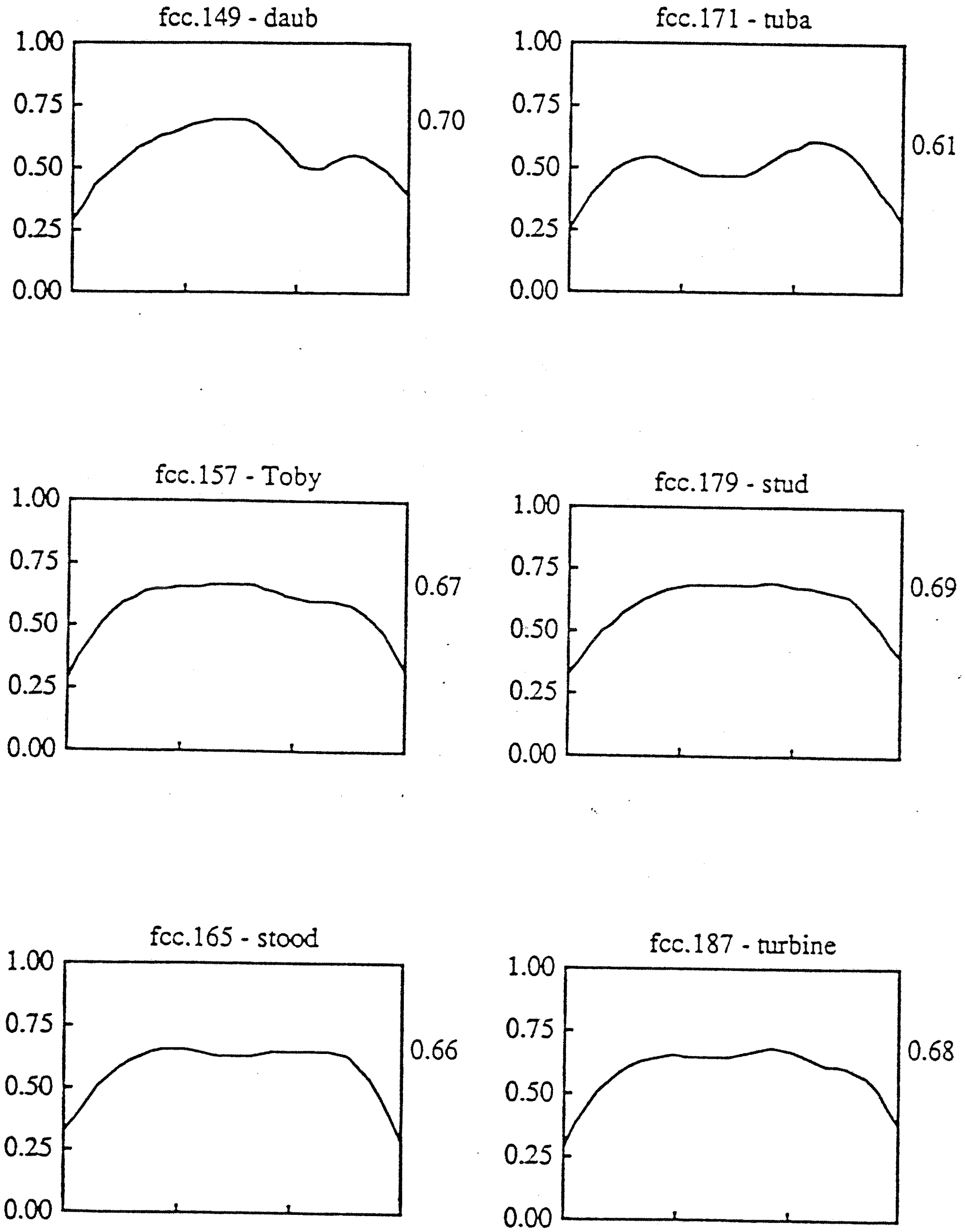


Figure 3.3: 1-frame smooth auditory representations (g - l)

token, would potentially eliminate important speech discrimination information. A representation which maps the potential range, 0 to 255, to the range 0 to 1, while preserving reasonable intensity distinctions within the range actually occuring in the database was used: $a = log(b + 1)/log(256)$.

Auditory training and testing patterns were generated from these spectral auditory representations. All but the three highest-frequency brad points, which represented frequencies above 6 kHz and were beyond the cutoff of the lowpass filter, were used for each frame. The resulting 1-frame patterns have 29 points. The 3-frame patterns were created with 87 points; the lowest 29 points for each of the three selected frames were used to create this pattern. Figure 3.4 shows the discrete activation pattern for the 3-frame dynamic transitions representation for the vowel /I/ in the word "kill" as spoken by the female speaker (see Section 2.2).

3.6.2 Preparing the articulatory patterns.

In addition to preparing the auditory patterns it was necessary to prepare articulatory patterns for use in the training input of the Full Motor Theory model and for the output of the Full Motor Theory model and the Mimic model. Speech sounds have traditionally been classified by linguists in terms of the articulatory features of which they are composed, e.g. voice, place, stop, nasal, lateral, sibilant, height, back, syllabic [15]. Such feature descriptors bear only superficial resemblance to the articulatory parameters used in producing the speech sound and little resemblance to auditory parameters which might be used in perceiving the sound. Other approaches to specifying articulatory features are equally inadequate. The motor theory model of speech perception, however, postulates a relationship between phonemic percept and articulatory and auditory parameters. Two of the nine pattern units allocated to the articulatory patterns were reserved for future use and were arbitrarily set to 'off' and 'on'. The seven unit binary, distributed abstract articulatory representation seen in Table 3.4 was specified for each of the twelve vowels used in this research.

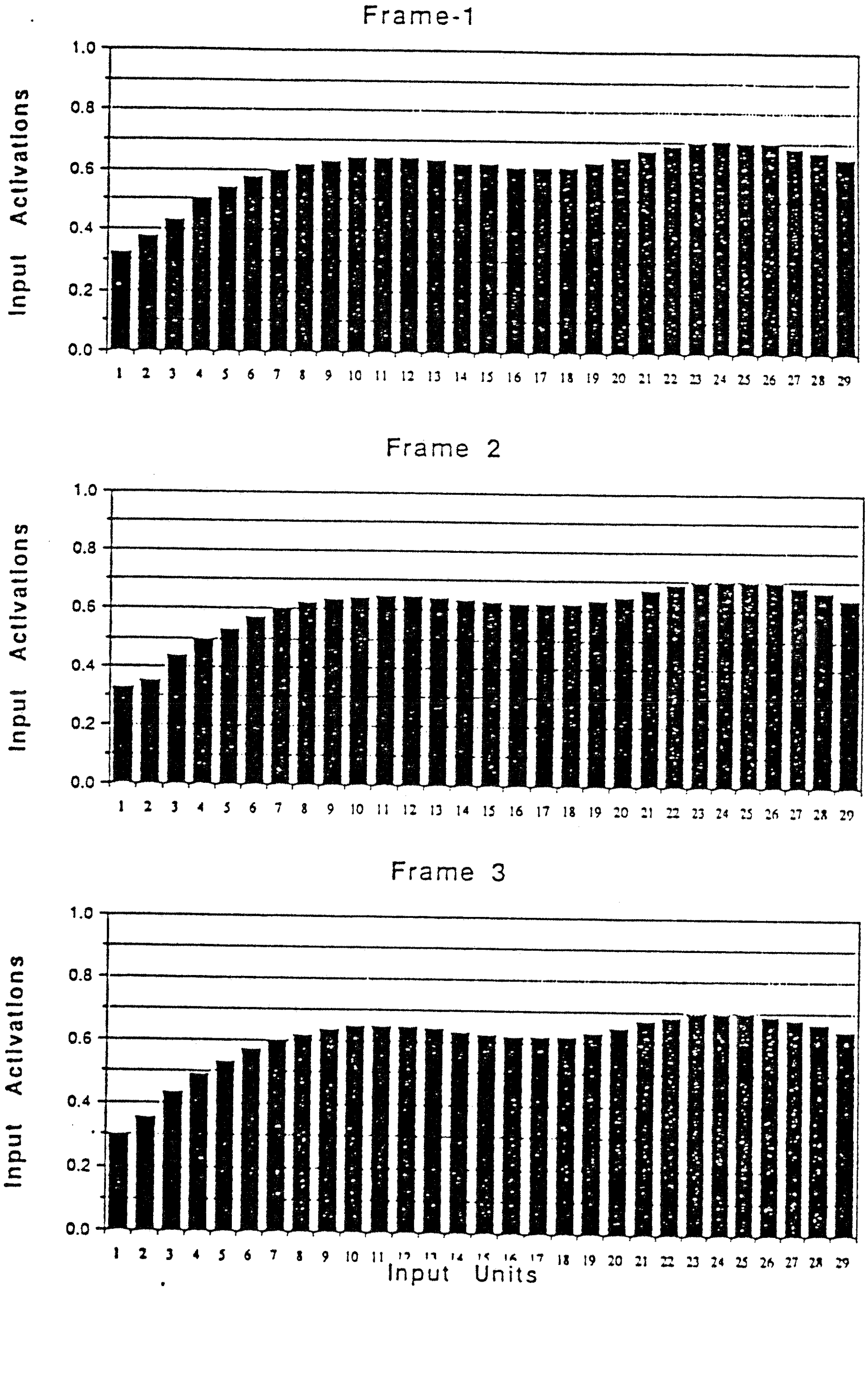


Figure 3.4: 3-frame discrete auditory pattern

Table 3.4: Articulatory Feature Patterns

phoneme	front	back	high	mid	low	rhotacism	rounded
/i/	1	0	1	0	0	0	0
/I/	1	0	1	1	0	0	0
/e/	1	0	0	1	0	0	0
/E/	1	0	0	1	1	0	0
/@/	1	0	0	0	1	0	0
/a/	0	1	0	0	1	0	0
/c/	0	1	0	1	1	0	0
/o/	0	1	0	1	0	0	1
/U/	0	1	1	1	0	0	1
/u/	0	1	1	0	0	0	1
/A/	0	0	0	1	1	0	0
/R/	0	0	0	1	0	1	1

3.7 Allocating the Patterns

As described in Section 3.3, two sets of data were collected for use in creating the patterns used in training and testing. The first, or multi-speaker, set generated patterns from the vowel speech tokens collected from ten speakers, five female and five male. This data was used in the multi-speaker training and cross-validation testing of the networks. The second speaker-independent set generated patterns from the vowel speech tokens collected from six different speakers, three male and three female. This data was used in the speaker-independent generalization testing of the networks.

Recall from Section 2.4 that the network had the task of associating the speech input with a reference auditory representation for that sound. To provide the set of reference auditory representations, a reference speaker was selected at random from among the ten speakers in the multi-speaker data set; this speaker was female. A reference token for each vowel as spoken by that speaker was then selected at random from among the tokens produced by the reference speaker. These reference tokens were used to generate the reference auditory spectral patterns to which the input patterns were to be mapped.

The selection of such a reference speaker at random could result in the selection of a speaker whose vocal characteristics were significantly different from those of the remaining speakers. This could make the search for an appropriate mapping function more difficult and thus limit the ability of the network to learn the training set and generalize to novel data. I wanted to understand what effect this arbitrary selection might have on the training recognition and generalization capabilities of the network. One could train a network several times, each time with a different set of reference patterns provided by a randomly-selected speaker and then compare the results of each training. An alternative way to test this effect would be to train a network twice, once using the randomly selected auditory spectral patterns as target patterns and then, from exactly the same initial conditions, use target auditory patterns which reflect the spectral patterns of all of the speakers in the training set. Such patterns might be called **vowel centroid patterns.** An additional output representation was created using auditory centroid patterns for the vowels.

Input patterns to be used in training the networks were selected randomly from the multi-speaker data with a total of 72 training tokens used for each speaker, including the reference tokens for the reference speaker. The remaining tokens, excluding the reference tokens for the reference speaker, were used in the multi-speaker cross-validation testing of the networks. Of the speaker-independent data, slightly better than half (288) of the available tokens were selected at random for use in the speaker-independent testing.

3.8 Preparing the Human Recognition Baseline Data

I wanted to provide a baseline against which the generalization of the networks could be compared. The obvious choice is to give the data to people, who are currently the best 'experts' available for such tasks. The data used for speaker-independent testing was utilized for this purpose. One half of the tokens for each of the speakers in the speaker-independent test set were allocated for the human baseline experiment. The problem was to provide an auditory stimulus that would make the human identification task as easy as possible without providing information about the surrounding consonantal context. Vowel extraction was specified in the following manner: (1) the portion to be extracted should include, at a minimum, all of the waveform between the 1/3 and 2/3 of vowel-interval extraction points inclusive; (2) to provide transitions into and out of this portion, where possible an additional minimum of 2 to 3 pulses of the waveform prior to and subsequent to the 1/3 to 2/3 vowel interval should also be extracted; (3) the extracted vowel segment should be 'played' and checked for vowel identifiability and duration; (4) if necessary, the extracted area of the vowel waveform was extended to increase the vowel duration without crossing into consonantal contextual regions of the waveform. The information provided to the human listener in the baseline experiment was significantly greater than that provided the networks trained with either the 1-frame or the 3-frame auditory data.

Having examined in detail the collection and preparation of the speech data and the data representations in training the networks and in implementing the human recognition comparison experiment, I now turn to an examination of the architecture of the neural networks and the training and testing methodology and guidelines used in my research.

CHAPTER 4

ARCHITECTURE, TRAINING AND MEASURES

I now describe the the architectural variations tested during the investigation. As appropriate, I indicate the task and data representation manipulations which provided the rationale for those variations. I then present the detailed specification of the various network architectures. Following this architectural specification, I will describe how the networks were trained and how performance and generalization were measured.

4.1 Architectural Variations

The relationship between generalization and network architecture was investigated both by manipulating the hidden structure for each of the network types and by comparing networks having the same hidden structure but differing input and output architectures. Recall that the input and output architecture of a network reflects the task which a network is required to perform. In Section 2.4 I presented a general description of each of the four network models that are used in my research: the full Motor Theory model, the Echo model, the Mimic model, and the Classifier model. In my research I compare the training performance and generalization exhibited within a network for each of the subtasks ascribed to that network and across networks for each of the different network types.

The time limitations involved in collecting and processing the speech data required that the amount of training and testing data available be limited to approximately 700 training patterns, 240 multi-speaker testing patterns, with 574 patterns available for speaker-independent testing. I wanted to determine the architectural design which would exhibit the best generalization for each of the network models.

The two basic guidelines for the design of neural networks which are used by many connectionist researchers are:

(1) The overfitting guideline: If a network can learn a problem, then the fewer the number of free parameters in the network, the better the network is likely to generalize.

(2) The dataset-size guideline: The larger the number of free parameters in a network, the more data needed to train it.

Keeping these guidelines and the fixed size of my training set in mind, it seemed clear that I should allocate the smallest number of hidden units possible that still enabled the network to learn the training set. According to the guidelines, that network also should exhibit the best generalization.

In accord with the target vowel theory of vowel perception and the dynamic theory of vowel perception, I wanted to study the effect of using both 1-frame and 3-frame spectral representations upon the ability of a network to recognize vowels. I explored the relationship between network generalization and data representation by training the networks with at least four different input representations, the 1-frame smoothed and unsmooted data representations and the 3-frame smoothed and unsmoothed data representations. Limited exploration of generalization after training with 5-frame smoothed input data representations was also undertaken. The full Motor Theory network and the Echo network were, in all instances, limited to a 1-frame auditory output.

The architectural specification for both the input and the output layers of the network reflect the representations used in training the network. Using 3-frame spectral input data requires using three times as many auditory input units as is required for 1-frame speech input data. According to the design guidelines this implies that

- More pattern training data might be required in order to achieve the same level of performance in the 3-frame input networks as in the 1-frame input networks.
- The 3-frame networks, having a larger input layer, might require more hidden units and, therefore, might not generalize as well as the smaller 1-frame networks for a given set of training instances.

4.2 Architecture Specification

Figure 4.1 shows the general description for the full Motor Theory Model of speech perception as presented in Section 2.4. I will present the architectural specification for this model. The variations for each of the remaining three models are indicated in terms of this specification.

The model is implemented as a feedforward fully-connected multi-layer network with a sigmoid activation function on the hidden and output units. At the input layer the network either receives a 'motor' pattern of articulatory activation or an acoustic signal, but not both. The network has three tasks that it is required to accomplish: classify the input phonemically, associate the input with an auditory spectral representation, and associate the input with a set of articulatory features. These task requirements together with the input and output speech representations serve to specify the architecture of the input and output layers of the network.

I begin with a discussion of the output architecture. Since the speech data consists of twelve English vowels, twelve binary localist output units were allocated to code the phonemic labels. As noted previously (Section 3.6) the spectral representations use 29 real-value points to represent each short-term segment. The auditory output nodes, therefore, consist of 29 real-value units. The distributed articulatory features listed in Table 3.4 were represented by seven binary units allocated for this purpose. Two additional binary units were allocated for future extensions of

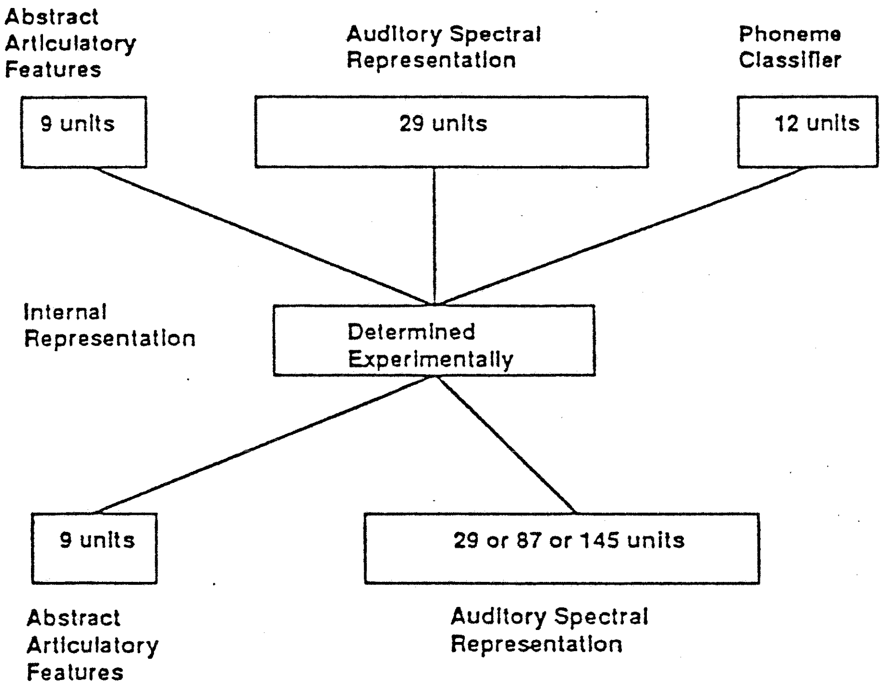


Figure 4.1: Motor Theory Model

this work. One was always 'on', while the other was always 'off'. The output layer consists of a total of 50 units, 21 of which are binary and 29 of which are real-valued.

The input layer for the full Motor Theory model includes an auditory group and an articulatory group. As in the output layer, nine binary units were allocated for the representation of the articulatory features. For experiments utilizing the 1-frame target vowel spectral input, a set of 29 real-value units was allocated to the auditory representation. Experiments utilizing the 3-frame dynamic input allocated 87 real-valued units to the auditory representation. Recall that at the input layer the network either receives a 'motor' pattern of articulatory activation or an acoustic signal, but not both. To facilitate the network's ability to distinguish between the two types of input a single binary unit was allocated to indicate whether articulatory activation was on or off.

The three remaining models, the Echo, Mimic, and Classifier models, only received acoustic input. Consequently, the input architecture for each of these models was determined by the input representation used in training the network: a 1-frame representation had 29 real-value inputs while a 3-frame representation had 87 real-value inputs. In the Echo model the network classifies this auditory input in addition to associating it with a 1-frame auditory spectral representation. On the output layer the Echo model had 12 binary units allocated for phonemic classification and 29 real-value units allocated for auditory representation. The Mimic model classifies the input in addition to associating it with the reference articulatory representation. The output layer of the Mimic model includes 12 binary units allocated for phonemic classification and 9 binary units allocated for articulatory features. In the Classifier model the network simply classifies the input. 12 binary units are allocated for the phonemic classification on the output layer.

As the preceeding discussion makes clear, much of the input/output architecture of the networks was determined by the Motor Theory model. The remainder was determined by decisions regarding the spectral and articulatory data used in training the networks. Nevertheless, there was still much room available for research into the relationship between the hidden structure of the network and generalization exhibited by the network. The hidden structure is not subject to the same design constraints as the input and output layers and so had to be experimentally determined. For each network model, the relationship between the number of hidden units and hidden layers used and the generalization exhibited by the network was open for exploration.

I now discuss how the networks were trained and how their recognition performance and generalization were measured.

4.3 Training and Measures

The networks were implemented and trained with Yoshiro Miyata's StarNet neural network simulator. The primary training technique is back-propagation. [1] With the exception of the experiments described in Section 5.8.2, training was accomplished with random presentation of the training patterns and limited manipulation of the learning rate. Three types of patterns and pattern units were used in the output units: real-value, binary-localist, and, binary-distributed. Establishing a single error criterion that would provide an appropriate measure of error across all of these three different types of units quickly proved to be a difficult task. Since I was interested in the ability of the networks to exhibit good generalization, I decided to track the training performance of the networks and the multi-speaker and speaker-independent generalization that the networks exhibited rather than tracking training error. Training recognition and multi-speaker and speaker-independent generalization were measured at regular intervals during the training process. Training was

[1] A limited comparison with conjugate gradient training was performed for the 1-frame Classifier networks.

discontinued after 40,000 epochs. The epoch limitation was an empirical determination resulting from the extensive training time required to reach that point and the limited increase in generalization that appeared to occur for most of the networks beyond that point.

The next question to be answered was: how can one measure the generalization which is accomplished by a connectionist network? Testing a network with a cross-validation testing pattern set is generally considered to be a global measure of generalization. In my research there are two measures of generalization. The first measure utilizes multi-speaker data and measures the ability of the network to generalize to new speech input for previously-encountered speakers. This is the kind of cross-validation testing that is usually used to measure the ability of a network to exhibit good generalization to novel inputs from the same domain. A second, more difficult, generalization measure utilizes speaker-independent data and measures the ability of the network to generalize to new input for never-before-encountered speakers. This is a test of the network's ability to generalize to novel inputs from a similar, but not the same, domain.

The usual approach to determining the correctness of a network output pattern requires comparing the actual output pattern with the desired target pattern for all output nodes. For two of the four network models the output layer contains both binary-valued units and real-valued units. A moment's reflection on the distinction between binary-valued units and real-valued units should reveal the difficulties that such an approach would create for determining the appropriate recognition of a speech input. An output pattern which contains an exact replica of the desired target spectral representation could be misclassified as a result of otherwise minimal errors between the binary output units. This potential for misclassification militates against using the total task for measuring the performance of a network.

As an alternative, recognition measures were computed for each task in the network. Another measure was computed by task vote of all the tasks in the network. Thus four recognition measures were computed for the Full Motor Theory networks, three recognition measures were computed for the Mimic and Echo networks, and one recognition measure was computed for the Classifier network. The performance based on each of these measures will be discussed in Chapter 5. In the course of my research there were many different networks to be trained and evaluated. It was decided that recognition results would be more comparative across networks if phonemic classification was determined by a simple maximum activation test across the twelve phonemic classifier units rather than by using signal detection techniques to determine optimum thresholds for each classifier node in each network. The continuous articulatory feature values were converted to binary values by a simple 0.5 threshold test on the activation values. The output articulatory pattern was compared to the set of reference articulatory patterns associated with each vowel using a city-block measure to determine recognition of the closest vowel. The real-value auditory spectral association performance was measured using a Nearest Neighbor algorithm with a Euclidean distance measure against the set of twelve target auditory spectral patterns associated with each vowel. For the voting measure, each task was allocated one vote. For the phonemic classification task, there were no cases of two or more phonemes exhibiting the same maximum value; the phonemic vote was always given to the winner for that task. For the spectral representation, if there were two or more patterns at an equal distance from the target pattern then the vote was split between them. Otherwise the vote went to the phoneme having the spectral pattern closest to the target. For the articulatory representation, if the actual output pattern matched the target pattern the vote went to the target phoneme otherwise the vote was divided among the phonemes having equally close representations to the desired target pattern.

Performance measures do not specifically reflect the relationship between the generalization exhibited by a network and the individual factors which affect that generalization: architecture, task, training data representations, and training of that network. I utilized statistical multiple regression techniques to isolate those experimental factors which are the primary determiners of the networks' ability to exhibit good generalization.

One additional measure was used in comparing the performance of the networks: learning efficiency. Learning efficiency is a measure which can be used to provide a comparison of networks based upon the number of connection weights in the networks and the learning rates at which the networks are trained.

In general, ignoring external factors such as computer speed, user load, etc., if two networks can learn a data set under exactly the same conditions and at the same learning rate then one can expect that the network having the larger number of connections will take longer to learn the training set. Similarly, if two networks are architecturally alike and can learn the training data under the same conditions but at differing learning rates then one can expect that the network learning at the higher learning rate will learn the data faster than the network learning at a lower learning rate. How "efficiently" the network connections are learned is, therefore, directly related to the learning rate and inversely related to the number of connections in the network. The "connection learning efficiency" can be defined as $c = \mathcal{F}(\eta, n)$, where n is the number of connections in the network and η is the learning rate used in training the network. If we approximate $\mathcal{F}$ with a simple exponential function, ie. $c = a(\eta/n)$ then, letting $a = 1.0$, we can obtain an approximate measure of the network's connection learning efficiency. For example, consider the case of two networks, both having the same input and output layers. Each network contains the same number of hidden units. One of the networks has a single hidden layer while the other has two hidden layers. The single hidden layer network has slightly more connections than the two hidden layer network, for example, 2870 versus 2690, but the rate at which it learns each connection is an order of magnitude faster than the rate at which the two hidden layer network learns each connection - 0.1 versus 0.01. The connection learning efficiency for the single hidden layer network is $34.8 * 10^{-6}$ while the connection learning efficiency for the two hidden layer network is $3.7 * 10^{-6}$. The single hidden layer network exhibits a connection learning efficiency approximately 10 times greater than that exhibited by the two hidden layer network.

This concludes the discussion of the experimental framework and methodology used in my investigation. I turn now to a discussion of the results of that investigation.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Research Focus

To reprise, my research investigates the question: how do the task which the network must learn, the architecture of the network, the training of the network, and the data representation used in that training, both individually and collectively, affect the ability of a network not just to learn the training data but to generalize well to previously unseen data. The research is embedded in an experimental framework which utilizes three multi-task connectionist models derived from the Motor Theory of speech perception and a standard classifier model for speech recognition. The speech data used in training and testing the four networks consists of 12 American English vowels. 3-frame vowel pattern representations are based on the dynamic-transitions model of vowel perception, while 1-frame vowel representations are based on the target model of vowel perception. Both smoothed and unsmoothed versions of the input data were used for training and testing.

In Chapter 1, I presented the guidelines and hypotheses used in my research. As a courtesy to the reader, I briefly reiterate them here. I then present baseline measures of the difficulty of the learning task required of the networks. These measures provide a lower bound against which the performance of the networks can be compared. This is followed by a discussion of the performance of the best networks for each of the four network models and for each type of data representation. The best networks show the actual upper bound on the performance exhibited by the networks. With these upper and lower bounds in mind, I then discuss the relationship between the generalization exhibited by the networks and each of the four factors involved in the investigation: architecture, task, training, and data representation.

5.2 Guidelines and hypotheses

The guidelines available to assist in this research were the following:

- The overfitting guideline: If a network can learn a problem, then the fewer the number of free parameters in the network the better the network is likely to generalize.
- The dataset-size guideline: The larger the number of free parameters in a network the more data needed to train it.

The following hypotheses were also investigated:

- Adding additional tasks should cause a change in generalization. That change may be an increase in generalization as a result of the added constraints, or it may be a decrease in generalization as a result of an increase in the network capacity,
- Training with appropriate, even though larger, representations will help the network to generalize better.

5.3 Baseline measures of performance

In order to judge the difficulty of the learning task it is important to have baseline measures of recognition performance. Two baseline measures were collected to indicate the difficulty involved in recognizing the vowel data. The first is a measure of how difficult it is for humans to recognize the isolated vowel tokens. This involved presenting selected examples from the speaker-independent database to human subjects for recognition. The second measure was obtained by determining the recognition performance of an accepted speech recognition system.

Using 144 speech tokens extracted from the speaker-independent testing database, four subjects attempted to recognize the selected vowels under the following experimental conditions:

- Vowel excerpts were presented by computer program.
- At each presentation a numbered list of twelve words containing the twelve vowels to be identified were displayed on the computer terminal, e.g. heed, hid, etc.
- Listeners had the option of being able to listen to each vowel sound from one to three times before entering their identification of the vowel.
- Listeners identified the presented vowel by typing in the identifier (0 ... 9, a, b) corresponding to the word containing the vowel sound they had identified in response to a prompt for identification.
- The selected vowel identification was echoed back to the listener and the listener was able to either continue to the next vowel presentation or to correct their identification if they felt they had made a mistake.

A training block in which five vowel sounds were presented was used to familiarize the subjects with the task. Subjects were permitted to repeat the training program as many times as desired until they felt comfortable with the procedure. The full identification program was then presented to the subject. Recall that the 144 speech tokens were deliberately extracted so as to make identification of the vowel as easy as possible for the listener. Nevertheless, average vowel recognition performance by the four listeners was only 54%.

Tajchman [37] describes an experiment in which human listeners were asked to listen to and identify vowel segments extracted from the TIMIT database. The procedure used for extracting the vowel segments is not specified. The average human recognition performance for 5 individuals for this task was 60%, closely paralleling the performance of my subjects.

The second baseline measure was the K-nearest Neighbor algorithm. This is a widely accepted process for recognizing speech. A set of baseline measurements using a K-nearest Neighbor algorithm were made on the multi-speaker and speaker-independent testing databases. The Nearest Neighbor (k = 1) and the best K-nearest Neighbor recognition performances were as follows:

```
Nearest Neighbor Performance (k = 1):
          multi-speaker = 46.3%
    speaker-independent = 41.0%


Best K-nearest Neighbor Performance:
          multi-speaker = 47.5% at k = 43,
    speaker-independent = 42.7% at k = 15
```

Note that the best K-nearest Neighbor performance for multi-speaker data occurs at k = 43 while the best performance for speaker-independent data occurs at k = 15. The K-nearest Neighbor algorithm resulted in an automated speech recognition rate of less than 50% for both multi-speaker and speaker-independent data.

The lower bound for the recognition task established by the human recognition baseline is 54%. The lower bound for recognition established by the K-nearest Neighbor algorithm is 42.7%. Based on these results it seems reasonable to expect that network speaker-independent generalization results should be between 42.7% and 54%.

5.4 Performance of the Best Networks

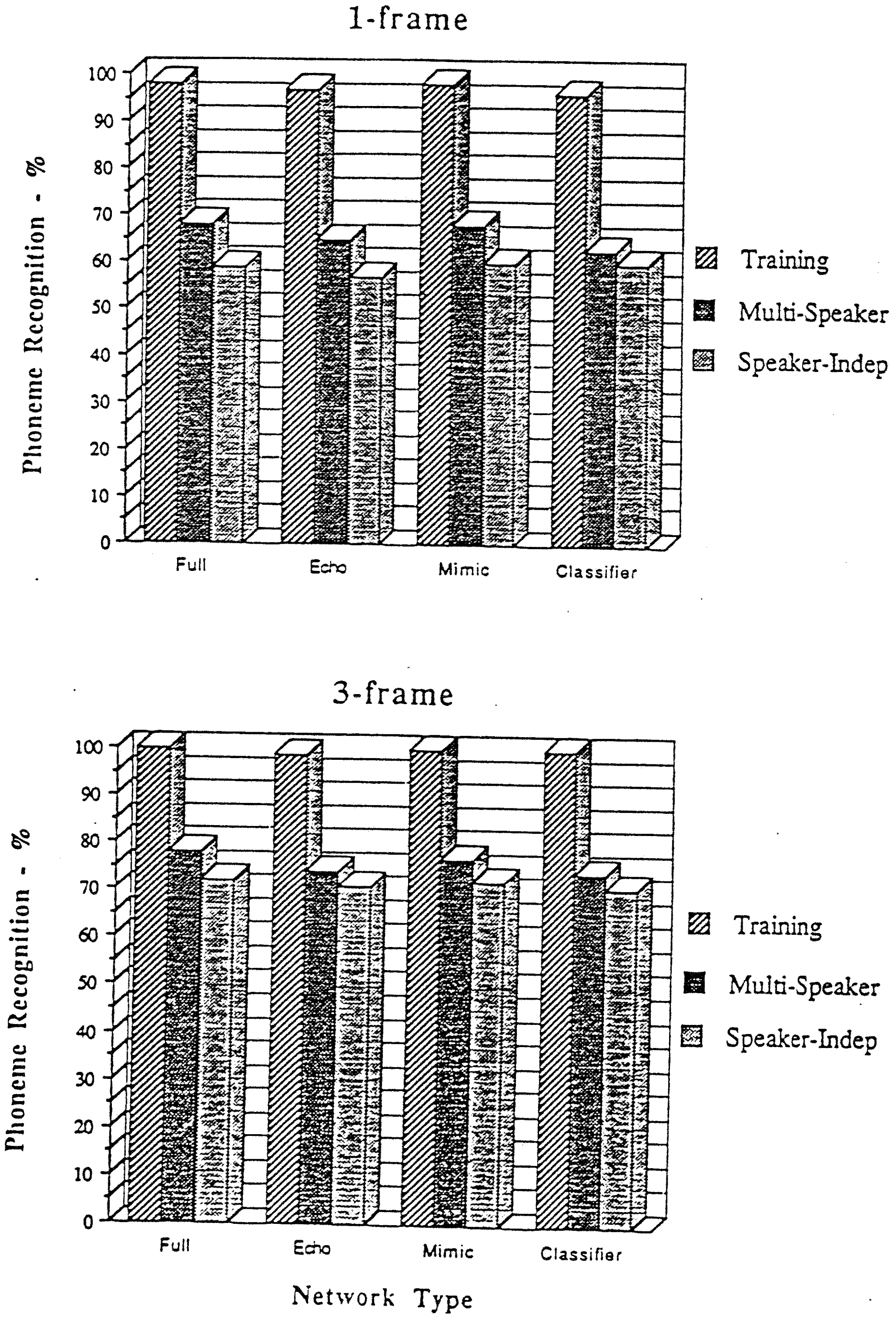
In discussing the results of the investigation attention will be focused upon the performance and generalization exhibited both within and across network models. In order to place those results in the proper perspective it is important to have some understanding of the best performance exhibited by a given network model for a given data representation. Table 5.1 presents comparative measures of the best performance and generalization exhibited by each of the network models investigated in my research. These are presented for each of the four models using 1-frame smoothed and 3-frame smoothed training data and for the Full Motor Theory model and the Classifier model using 1-frame unsmoothed and 3-frame unsmoothed training data.

Table 5.1: Best Performance Networks

Data Type	Network Type	Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
1-Smooth	Full	80	98	68	59
	Echo	70	97	65	57
	Mimic	80	98	68	60
	Classifier	90	96	63	60
3-Smooth	Full	70	100	78	72
	Echo	85	100	75	72
	Mimic	95	100	77	72
	Classifier	80	100	70	74
1-Unsmooth	Full	90	100	75	70
	Classifier	70	100	68	69
3-Unsmooth	Full	70	100	71	65
	Classifier	90	100	80	71

The number of hidden units used in the best network is as specified in the table. Each of these networks has a single hidden layer. Figures 5.1 and 5.2 provide graphic comparisons of the performance of these networks.

For the networks trained with 1-frame smoothed data, the Full Motor Theory and Mimic networks exhibit the best multi-speaker and speaker-independent generalization: 68%, and 59%/60% respectively. For the two network models trained with 1-frame unsmoothed data, best generalization results are exhibited by the Full Motor Theory model: multi-speaker generalization of 75%, and speaker-independent generalization of 70%. Comparing the 1-frame smoothed with the 1-frame unsmoothed results indicates that the networks trained with the unsmoothed data



1-frame

3-frame

Figure 5.1: Best Performance Networks - Smoothed Data

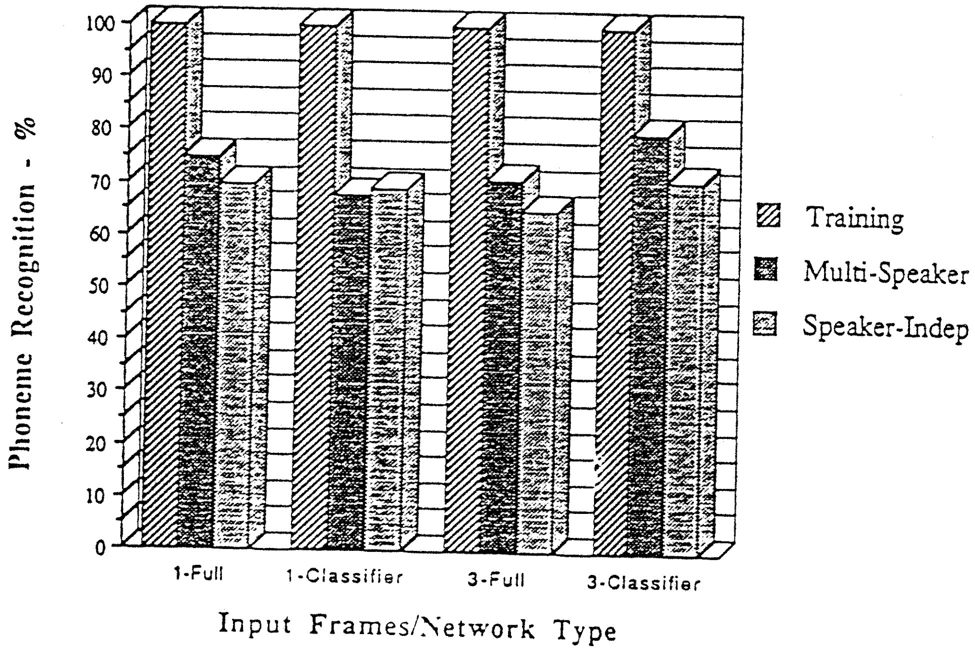


Figure 5.2: Best Performance Networks - Unsmoothed Data

performed significantly better.

All of these 1-frame results exceed the human recognition baseline (54%) by more than 5% and the K-nearest neighbor baseline (42.7%) by more than 16%. These 1-frame results are also extremely good when compared to performance reported by other researchers. For example, Muthusamy and Cole [28] reported speaker-independent generalization results of 51.40% on vowels after training a simple classifier network using a conjugate gradient algorithm with a single frame of data from the TIMIT database. These are not strictly comparable with my results since their network was trained and tested using data from a much larger group of speakers.

Networks trained with 3-frame smoothed data exhibited significantly increased generalization capabilities over the 1-frame smoothed data networks: approximately a 10% increase for multi-speaker generalization, from 68% to 78%, and approximately a 13% increase for speaker-independent generalization, 59% versus 72%. Again, the Full Motor Theory and Mimic networks are virtually identical in exhibiting the best multi-speaker generalization and speaker-independent generalization for the 3-frame smoothed data networks: 77%/78%, and 72%, respectively. For the two network models trained with 3-frame unsmoothed data, best results are obtained with the Classifier network which exhibits multi-speaker generalization of 80%, and speaker-independent generalization of 71%.

Across all of these networks the best overall generalization is exhibited by the 3-frame unsmoothed Classifier network (80%, 71%) while the 3-frame smoothed Full Motor Theory network exhibits almost the same performance (78%, 72%). Muthusamy and Cole reported speaker-independent generalization results of 55.66% after training a network with three frame vowel data from the TIMIT database.

These results were all based upon a twelve vowel phonemic classification. Although presented with prompts designed to elicit differing responses for the vowels /a/ and /c/, for example - 'hod' and 'hawed', none of the 16 subjects used in the data collection process actually distinguished between these two vowels. Using an 11-vowel performance measure in which these two vowels are grouped together the best training recognition, multi-speaker generalization, and speaker-independent generalization for the networks are shown in Table 5.2.

Table 5.2: Best Performance Networks - 11 Vowels

Data Type	Network Type	Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
1-Smooth	Full	80	98	73	64
	Echo	70	97	72	63
	Mimic	80	98	73	65
	Classifier	90	96	68	64
3-Smooth	Full	70	100	81	76
	Echo	85	100	79	76
	Mimic	95	100	82	76
	Classifier	80	100	74	78
1-Unsmooth	Full	90	100	79	73
	Classifier	70	100	73	74
3-Unsmooth	Full	70	100	78	68
	Classifier	90	100	84	73

The best multi-speaker generalization and speaker-independent generalization for the 1-frame smoothed data networks again occur with the Full Motor Theory and Mimic networks: 73%, and 64%/65% respectively. For the two network models trained with 1-frame unsmoothed data, best results are obtained with the Full Motor Theory model which exhibits multi-speaker generalization of 79%, and speaker-independent generalization of 73%.

For the 3-frame smoothed data networks, the Full Motor Theory and Mimic networks are again virtually identical in exhibiting the best multi-speaker generalization and speaker-independent generalization: 81%/82%, and 76%, respectively. For the two network models trained with 3-frame unsmoothed data, best results are again obtained with the Classifier network which exhibits multi-speaker generalization of 84% and speaker-independent generalization of 73%. Across all of these networks the 3-frame smoothed Full Motor Theory and Mimic networks exhibit the best overall generalization (81%/82%, 76%).

The 11-vowel class recognition measure, which is a more accurate reflection of the collected speech data, results in an increased multi-speaker generalization of 5% or better and an increased speaker-independent generalization of 4% or better for the networks trained with 1-frame smoothed data. For the networks trained with 1-frame unsmoothed data there is an increase in generalization of 3% or more in both multi-speaker and speaker-independent testing. A similar increase can be seen in the case of the networks trained with 3-frame smoothed data. For the networks trained with 3-frame unsmoothed data, an increase of 4% or better can be seen in multi-speaker generalization and 2% or better occurred in speaker-independent generalization.

Using a still broader classification based upon the abstract articulatory features specified for each of the twelve vowels in the speech data, it is possible to determine the performance characteristics of the best networks in terms of the following eight phonetic characteristics: front-high, front-mid, front-low, center-mid, center-low, back-high, back-mid, and back-low. The corresponding phonetic performance measures for the networks are shown in Table 5.3. The best multi-speaker and speaker-independent generalization for the 1-frame smoothed data networks again occur with the Full Motor Theory and Mimic networks: 80%/81%, and 76%/77%, respectively. For the two network models trained with 1-frame unsmoothed data, the best results are obtained with the Full Motor Theory model which exhibits multi-speaker generalization of 84%, and speaker-independent generalization of 81%.

For the 3-frame smoothed data networks the Mimic network alone shines in exhibiting the best multi-speaker generalization (89%) and speaker-independent

Table 5.3: Best Performance Networks - 8 Phonetic Classes

Data Type	Network Type	Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
1-Smooth	Full	80	99	80	76
	Echo	70	98	79	77
	Mimic	80	99	81	77
	Classifier	90	98	77	74
3-Smooth	Full	70	100	86	82
	Echo	85	100	85	82
	Mimic	95	100	89	83
	Classifier	80	100	80	83
1-Unsmooth	Full	90	100	84	81
	Classifier	70	100	80	79
3-Unsmooth	Full	70	100	82	78
	Classifier	90	100	88	81

generalization (83%). For the two network models trained with 3-frame unsmoothed data, the best results are again obtained with the Classifier network which exhibits multi-speaker generalization of 88%, and speaker-independent generalization of 81%. The network exhibiting the best overall performance is the 3-frame smoothed Mimic network (89%, 83%).

Comparing the 8-phonetic-classes measure with those of the 11-vowel and the 12-vowel measures, reveals even greater improvements in generalization. For the networks trained with 1-frame smoothed data there is a minimum increase in multi-speaker generalization of 7% and 12%, respectively, and a minimum increase in speaker-independent generalization of 10% and 14% respectively. In the case of the 1-frame unsmoothed data networks the minimum improvements in multi-speaker generalization are 5% and 9%, respectively, with corresponding minimum improvements in speaker-independent generalization of 5% and 10%. The 3-frame smoothed data networks show minimum improvements of 5% and 8% for both generalization measures, while 3-frame unsmoothed data networks show minimum improvements of 5% and 9%.

Across all three measures - 12-vowel, 11-vowel, and 8-phonetic-classes - the 3-frame smoothed data representation consistently leads to the highest levels of generalization: 78%/72%, 82%/76%, and 89%/83%. From these results it would appear that the 3-frame smoothed data representation is the most appropriate representation used in training and testing these networks.

5.5 Exposition Approach

I will now discuss the results of my investigation into the relationship between the generalization exhibited by a network and the four independent factors: architecture, task, training, and data representation. Unfortunately, attempting to disentangle the effects of each of these four factors is almost as difficult as the problem faced by speech researchers as they seek to understand the variability problem of speech perception (Section 2.2). For example, it is possible to ascribe changes in network performance to differences in network architecture when all three of the other factors can be held fixed and only the hidden structure of the network is changed. However, it is much more difficult to untangle the effects of each factor when discussing the differences in network performance exhibited when a network is trained with differing input representations which require differing input architectures (for example, the 1-frame data representation versus the 3-frame data representation). Given the definition of task provided in Section 1.2.2, $T = (i, o)$ it can also be argued that such a change in input data also results in a change in task, since the i of the network trained with the 1-frame representation differs from the i of the network trained with the 3-frame representation. In so far as possible, I will try to untangle the experimental results with regard to each factor. Where there is an interaction of factors, such as the data representation change which results in subsidiary changes in the architecture and task of the network, I will discuss the results in terms of what I believe is the primary factor effecting changes in performance and generalization while noting the possibility of effects from subsidiary factors. The reader may, of course, disagree as to which is the primary factor in a particular situation. I can only beg the reader's indulgence in attempting to follow the pathway which I will wend through this rather amazing thicket.

5.6 Architecture

I begin with the experimental results of the investigation into the relationship between the architecture of the network and its ability to generalize well to novel data. Here it is possible to isolate the factor of architecture by keeping the inputs and outputs constant. Although the training parameters are held constant, it is obvious that the initial state of the network connections vary as the architecture of the hidden structure varies. In a fully-connected strictly feedforward network there are two basic approaches to varying the hidden structure of the network. First, one can vary the number of hidden units used in that structure and, second, one can vary the number of hidden layers over which those units are distributed. I first discuss the results exhibited by networks having a single hidden layer and a varying number of hidden units. These results are then compared to those obtained from networks having the same number of hidden units but with those hidden units distributed over more than one hidden layer.

5.6.1 Architecture: one hidden layer.

Using the 1-frame smoothed data representation, I first describe in detail the results for the Full Motor Theory model network as the architecture of the hidden layer is varied. I then discuss the results for each of the remaining three network models as the architectures of their hidden layers are varied. Similar results are discussed for each of the three remaining data representation types: 3-frame smoothed, 1-frame unsmoothed, and 3-frame unsmoothed. Wherever possible I will limit repetitious discussion by describing performance in terms of similarities to previously described behavior. Consideration will be given to the relationship between architecture and generalization capability of the networks as it is revealed within each data representation type. With respect to the overfitting guideline, in order to compare these results with those reported by others, the multi-speaker generalization is of greatest interest. Speaker-independent generalization presents results for vowel identification for tokens selected from a different speaker domain than that used in training the networks. This is an even broader test of generalization than that reported by most researchers.

Recall that the training set consisted of 700 training patterns selected from the multi-speaker database. 240 patterns were selected from the same database for multi-speaker testing. 574 patterns were used for testing speaker-independent generalization. According to the dataset-size guideline, I needed to develop an architecture which could be trained by 700 training patterns. It was obvious from the beginning that Widrow's suggested training sample size (10 times the number of connections in a network) probably could not be used in this investigation. 700 training patterns would limit the 1-frame Full Motor Theory network to a maximum of 8 hidden units and the 3-frame Full Motor Theory network to a maximum of 5 hidden units!

A series of experiments were run to determine the number of hidden units required for a single-hidden-layer Full Motor Theory network to learn the training set. All of the networks were trained for 40,000 epochs. To minimize the effect of minor variations in performance, mean performance was computed over the last 20,000 training epochs. I wanted to determine experimentally the minimum number of hidden units and hidden layers that the networks would require in order to learn the training set and generalize well to new speech data. Initially, memorizing the training set was defined as exhibiting at least a 90% recognition rate on training data.

Figure 5.3 depicts the mean training recognition, multi-speaker generalization, and speaker-independent generalization results versus number of hidden units for the experiments in which Full Motor Theory networks were trained with 1-frame smoothed speech input. Performance and generalization results are also presented in Table 5.4.

All of the single-hidden-layer 1-frame networks having 40 or more hidden units exhibit a 90% or better recognition rate on the training data. Interestingly, while the overfitting guideline would lead us to believe that the best generalization would occur in such a network which has the smallest number of hidden units (40), it is apparent that as good or better multi-speaker generalization occurs when there are more than that number of hidden units in the hidden layer. This is true in the

Table 5.4: 1-frame Smoothed Data - Full Motor Theory Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
30	85	54	50
40	90	59	49
50	94	59	52
60	96	61	56
70	96	62	54
80	97	64	57
90	97	60	56
150	97	62	58

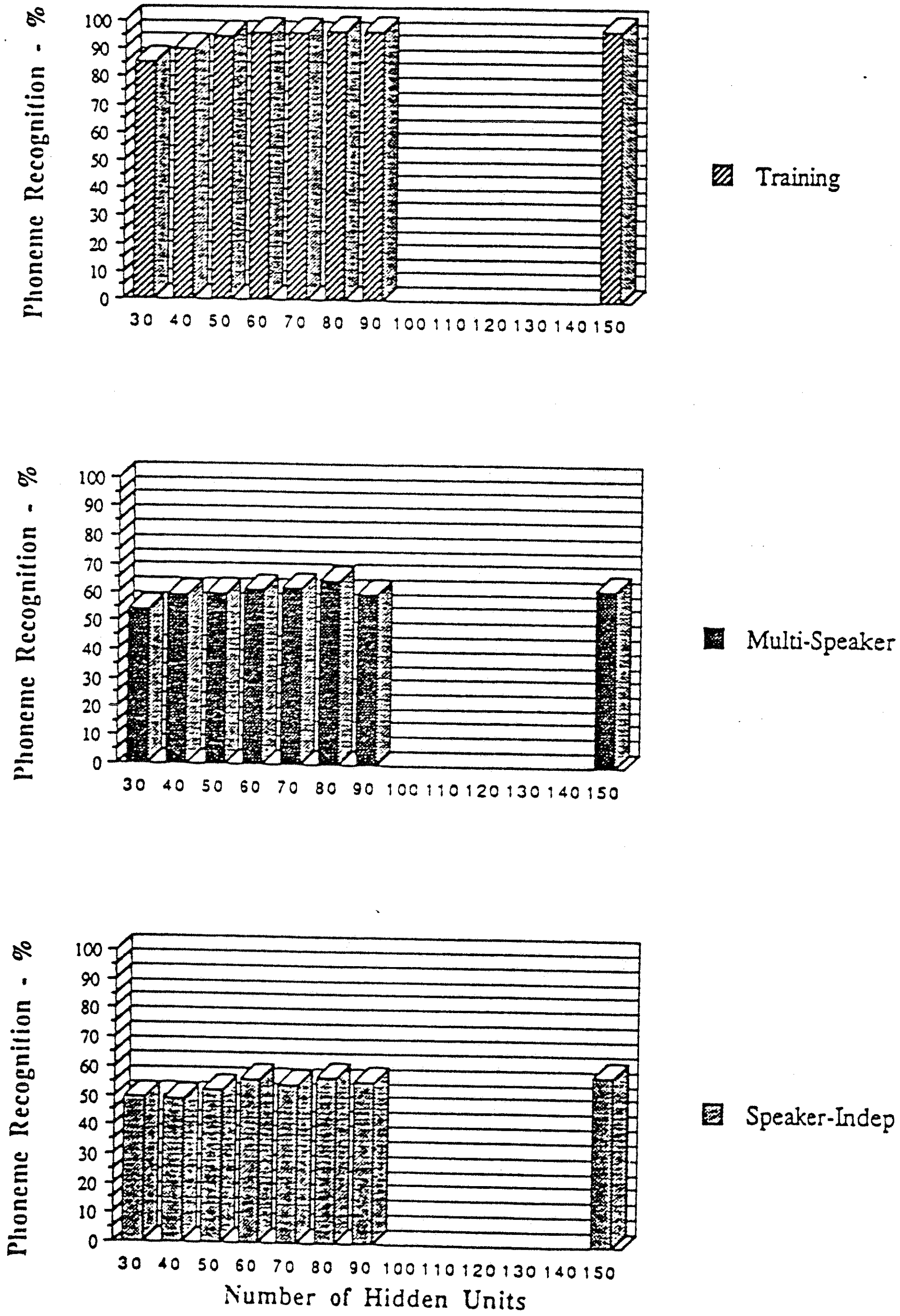


Figure 5.3: 1-frame Smoothed Data - Full Motor Theory Model

case of speaker-independent generalization, as well. For example, the best mean multi-speaker generalization, 64%, occurs with 80 hidden units while the best mean speaker-independent generalization, 58%, occurs with 150 hidden units.

It can be argued that the 40 hidden unit network really did not memorize the training set, since the networks having more hidden units correctly recognize 97% of the training items. The mean training recognition rate across all of the networks which satisfy the initial 90% or better training recognition rate learning criterion is 95%. This can be used to establish a new recognition criterion of 95%. Using this more stringent criterion does not significantly change the outcome. Full Motor Theory networks having from 60 to 150 hidden units in the hidden layer each exhibit better than a 95% training recognition rate. Multi-speaker generalization results for each of these networks range from 60% to 64%, respectively, with a mean multi-speaker generalization for these five networks of 62%. Similar results hold for the speaker-independent generalization: a range of 54% to 58% with a mean speaker-independent generalization of 56%.

What is most striking about these networks is the comparative equivalency of the generalization results. Each network was trained from different initial conditions. The initial connection weights differed for each network. The number of hidden units for each network differed greatly (60 - 150 hidden units). The number of connections to be learned varied from 5,340 to 13,350 and yet each of these networks was able to satisfy the new 95% training recognition rate criterion using exactly the same amount of training data. In all five of these networks the training recognition deviates by less than 1% from the average, 96.6%. In all five of the networks the multi-speaker generalization ranges from 60% to 64%. In all five of the networks the speaker-independent generalization ranges from 54% to 58%. The capacity of the largest network is more than twice that of the smallest. Nevertheless, the smallest network did not generalize any better; this contradicts the overfitting guideline.

Finally, it should be noted that the networks containing 80, 90, and 150 hidden units, respectively, each exhibit the same mean training recognition rate, 97%. The largest network does exhibit a slightly lower mean multi-speaker generalization, 62% versus a high of 64%, but a slightly higher mean speaker-independent generalization, 58% versus a low of 56%. With exactly the same training recognition rate it would appear to be difficult to ascribe the slightly lower multi-speaker generalization to the network's having over-learned the training set, especially in light of the slightly higher level of speaker-independent generalization.

Similar results were obtained with the Echo, Mimic, and Classifier networks trained with 1-frame smoothed speech input. Figures 5.4, 5.5, and 5.6 show the mean training recognition, multi-speaker generalization, and speaker-independent performance versus number of hidden units for each of these network models. Tables 5.5, 5.6, and 5.7 [1] present the same data in tabular form.

Table 5.5: 1-frame Smoothed Data - Echo Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
30	90	57	55
40	93	59	47
50	93	63	53
60	95	59	50
70	95	62	54
80	95	61	53
90	95	63	52
150	95	60	52

For the Echo model, networks having from 60 to 150 hidden units all exhibit 95% training recognition. However, networks having from 70 to 150 hidden units

[1] Table 5.7 does not show performance values for all of the Classifier network values seen in Figure 5.6.

exhibit better multi-speaker and speaker-independent generalization than does the network having 60 hidden units.

Table 5.6: 1-frame Smoothed Data - Mimic Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
30	88	59	50
40	93	67	51
50	94	58	49
60	95	63	51
70	96	62	53
80	97	66	57
90	97	63	55
150	97	63	58

For the Mimic model, networks having from 60 to 150 hidden units all exhibit 95% or better training recognition. A network having 80 hidden units exhibits much better overall generalization than does the network having 60 hidden units while a network having 150 hidden units exhibits at least as good multi-speaker generalization. Both the 80 hidden units network and the 150 hidden units network exhibit better speaker-independent generalization: 57%/58% versus 51%.

Table 5.7: 1-frame Smoothed Data - Classifier Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
30	91	61	48
40	93	59	54
50	95	59	57
60	95	61	54
70	95	62	55
80	96	61	53
90	95	61	56
150	95	62	55

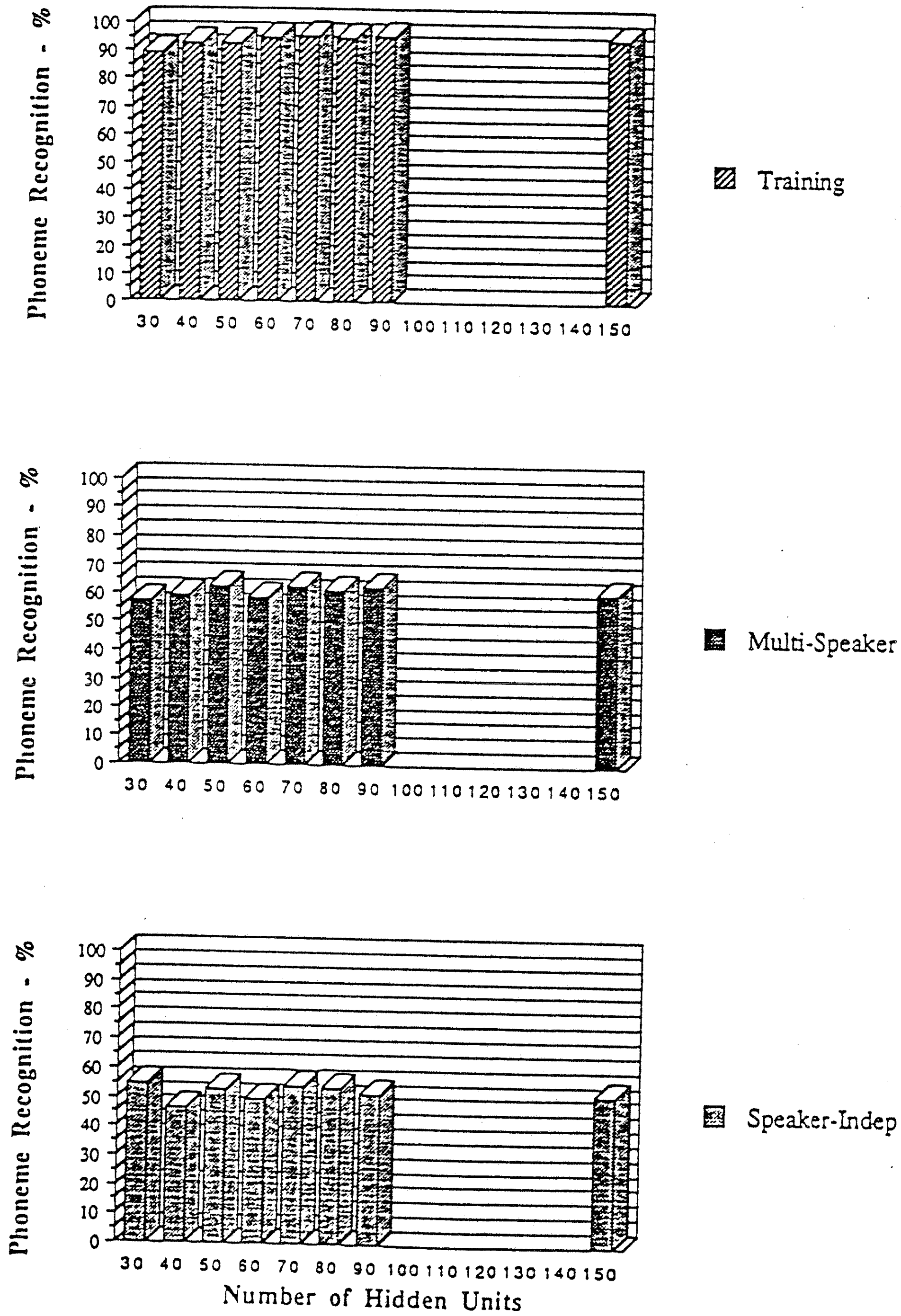


Figure 5.4: 1-frame Smoothed Data - Echo Model

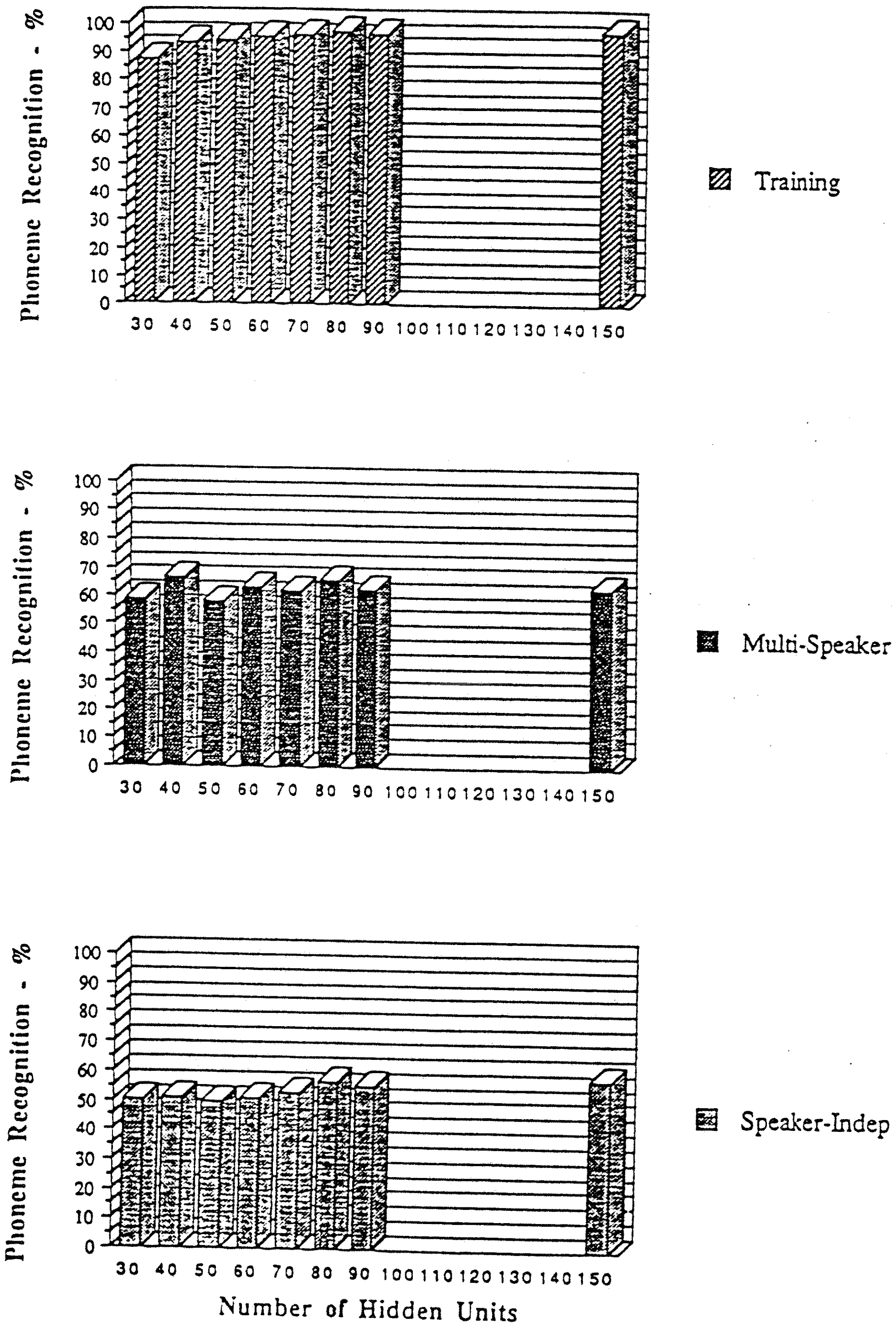


Figure 5.5: 1-frame Smoothed Data - Mimic Model

For the Classifier model, networks having from 50 to 150 hidden units all exhibit 95% or better training recognition. Networks having 70, 90, and 150 hidden units all exhibit the best overall generalization: 62%/55%, 61%/56%, and 62%/55% versus 59%/57% for the 50 hidden unit network.

Neither the overfitting guideline nor the dataset-size guideline appear to apply to the four network models trained with 1-frame smoothed data. The experimental results indicate that across a broad range of hidden units, the initial state of the network did not significantly affect the ultimate results of training. These networks exhibit comparable levels of generalization across a broad range of hidden units. The insensitivity of network performance to variations in the number of hidden units is one of the more striking results of these experiments.

A series of similar experiments were run using 3-frame smoothed input. Again, all networks were trained for 40,000 epochs and mean performance was computed over the last 20,000 training epochs.

Figure 5.7, shows the mean training recognition, multi-speaker generalization, and speaker-independent generalization results for each of the Full Motor Theory networks trained with 3-frame smoothed speech data input. The actual performance and generalization values are listed in Table 5.8. Like the 1-frame networks, the Full Motor Theory networks trained with 3-frame smoothed speech data input also exhibit a rather striking comparability in training recognition and generalization, even though each network was trained from different initial weight states.

Networks having from 60 to 150 hidden units all exhibit 99% training recognition. The number of connections to be learned varied from 8,820 to 22,050. Multi-speaker generalization results for each of these five networks ranges from 70% to 74%, respectively with an average generalization of 72.6%. Each of these networks differs from that average by less than 3%. Best speaker-independent generalization,

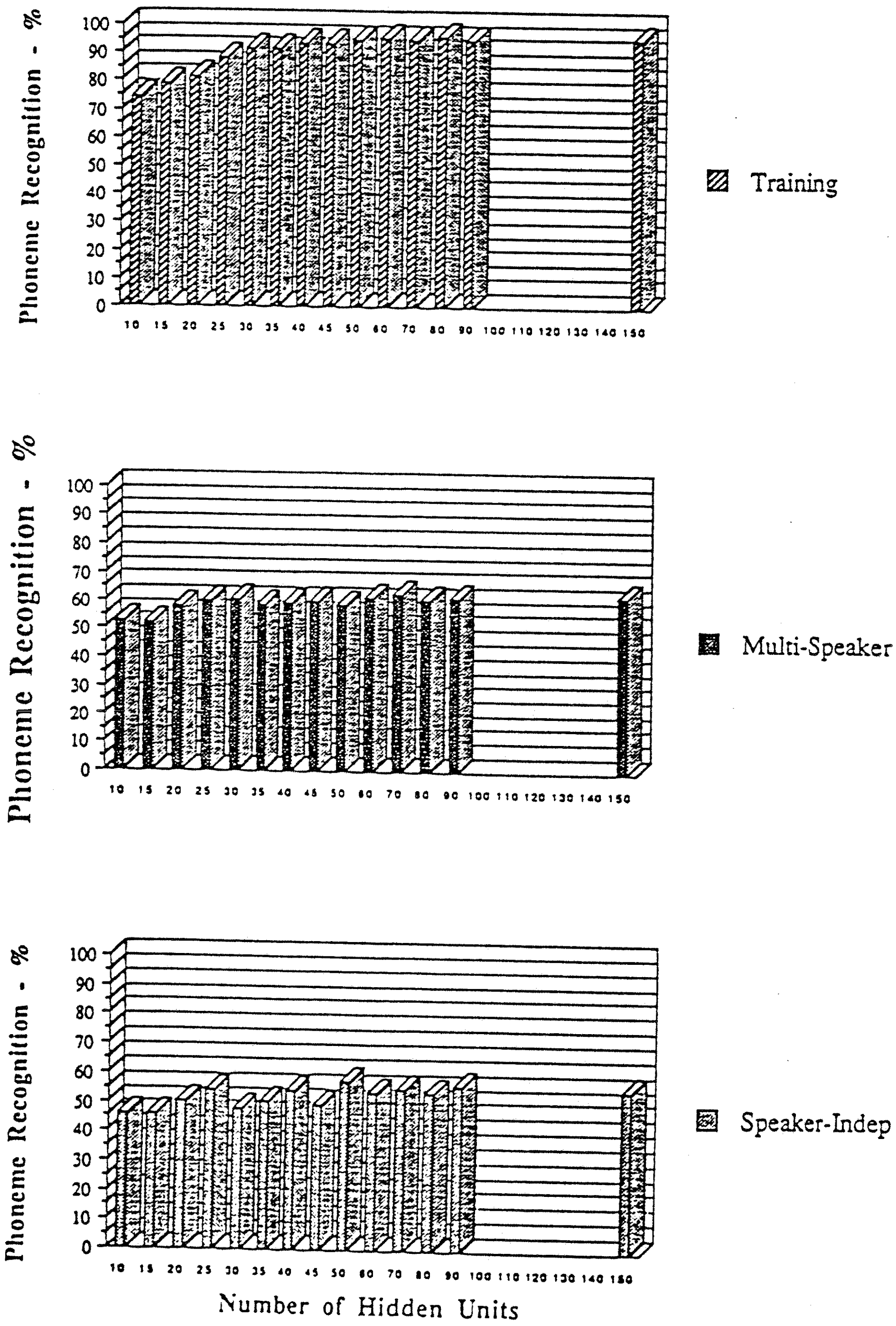


Figure 5.6: 1-frame Smoothed Data - Classifier Model

Table 5.8: 3-frame Smoothed Data - Full Motor Theory Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
30	97	69	67
40	98	68	63
50	98	69	64
60	99	70	67
70	99	74	71
80	99	73	68
90	99	73	71
150	99	73	71

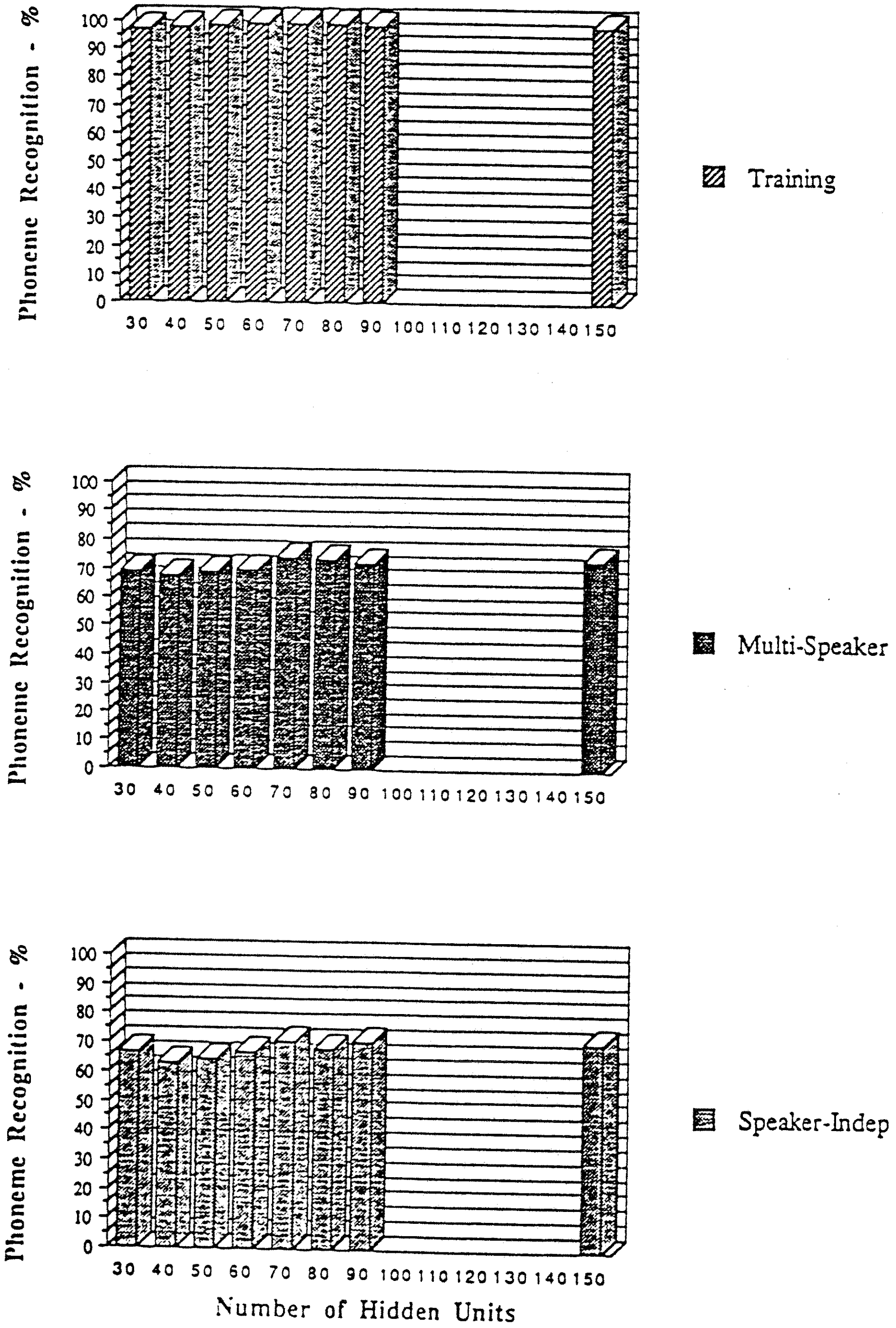


Figure 5.7: 3-frame Smoothed Data - Full Motor Theory Model

71%, occurs for networks having 70, 90, and 150 hidden units. Contradicting the overfitting guideline, networks having from 70 to 150 hidden units exhibit quite similar multi-speaker and speaker-independent generalization, all of which is better than that exhibited by the 60 hidden unit network.

Similar arguments can be made with respect to the Echo, Mimic, and Classifier networks which have been trained with 3-frame smoothed input. Figures 5.8, 5.9, and 5.10 show the mean training recognition, multi-speaker generalization, and speaker-independent generalization results by number of hidden units. Corresponding values for these networks are listed in Tables 5.9, 5.10, and 5.11 [2]. For the Echo model, networks having from 35 to 150 hidden units all exhibit 98% or better training recognition. The best multi-speaker generalization occurs in the network containing 75 hidden units, 75%, while the best speaker-independent generalization occurs in the networks containing either 95 or 150 hidden units, 70%.

Table 5.9: 3-frame Smoothed Data - Echo Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
35	98	69	61
45	99	70	65
55	99	71	67
65	99	68	68
75	99	75	65
85	99	73	69
95	99	71	70
150	98	72	70

For the Mimic model, networks having from 35 to 150 hidden units all exhibit 99% or better training recognition. The best multi-speaker generalization occurs in the network containing 85 hidden units, 75%, while the best speaker-independent generalization occurs in the networks containing either 95 or 150 hidden

[2] Values reported here are for selected Classifier networks only.

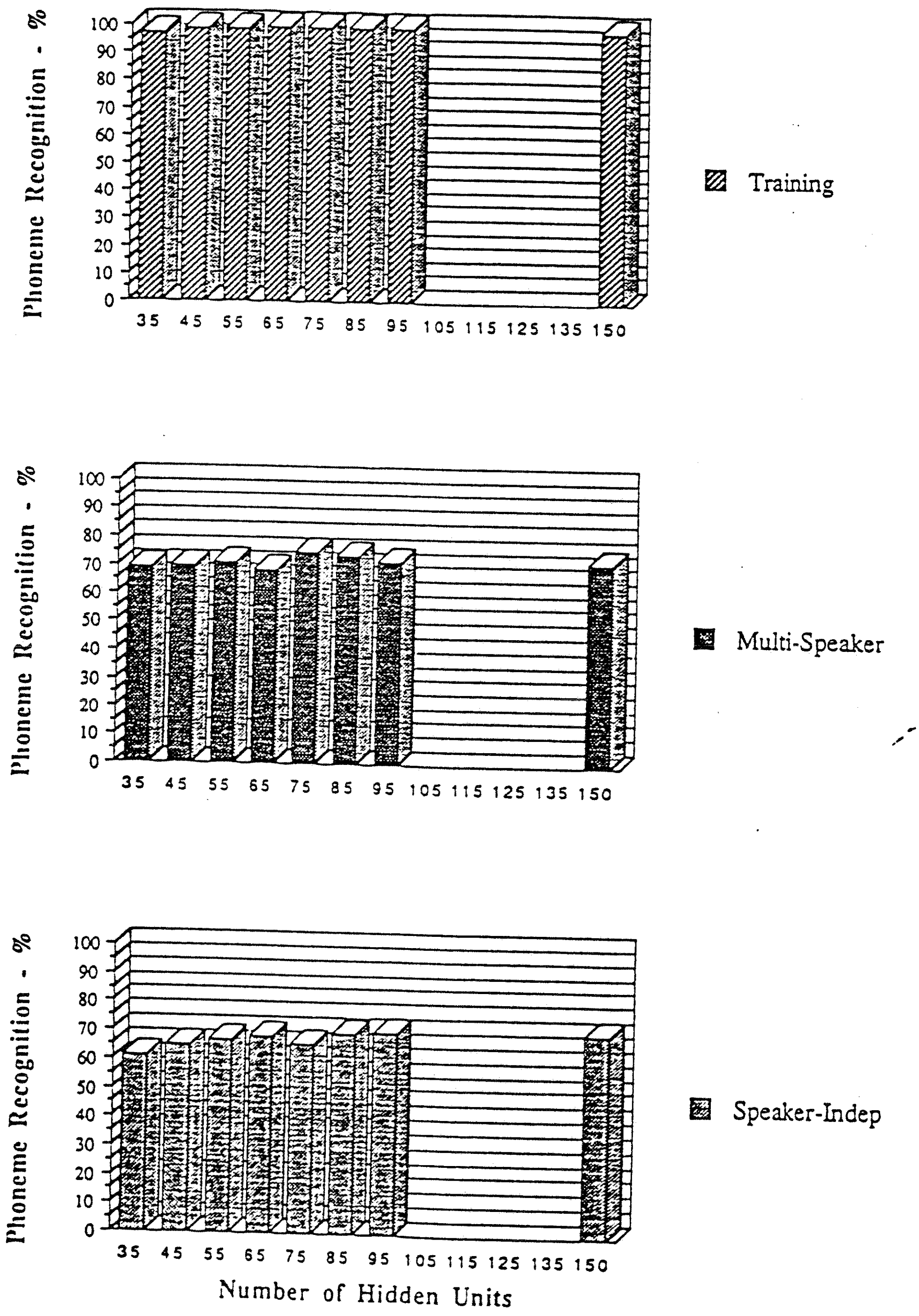


Figure 5.8: 3-frame Smoothed Data - Echo Model

Table 5.10: 3-frame Smoothed Data - Mimic Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
35	99	69	62
45	99	73	65
55	99	74	69
65	99	72	69
75	99	73	69
85	100	75	67
95	99	74	70
150	99	72	70

units, 70%.

Table 5.11: 3-frame Smoothed Data - Classifier Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
30	98	68	61
40	99	73	64
50	99	72	66
60	99	69	66
70	99	71	65
80	99	69	69
150	99	71	69

For the Classifier model, networks having from 30 to 150 hidden units all exhibit 98% or better training recognition. The best multi-speaker generalization occurs in the network containing 40 hidden units, 73%. The best speaker-independent generalization occurs in the networks containing either 80 or 150 hidden units, 69%.

Neither the overfitting guideline nor the dataset-size guideline appear to apply to the four network models trained with 3-frame smoothed data. The results indicate that across a broad range of hidden units, the configuration of the network did not significantly affect the ultimate results. These networks exhibit comparable levels of generalization across a broad range of hidden units.

A series of similar experiments were run using 1-frame unsmoothed speech data input. These experiments were limited to a subset of the models: the Full Motor Theory model and the Classifier model. Like the smoothed data sets, the unsmoothed data training set contained 700 training examples. Again, all of the networks were trained for 40,000 epochs and mean performance was computed over the last 20,000 training epochs.

Figure 5.11 shows the mean training recognition, multi-speaker generalization, and speaker-independent generalization results versus number of hidden

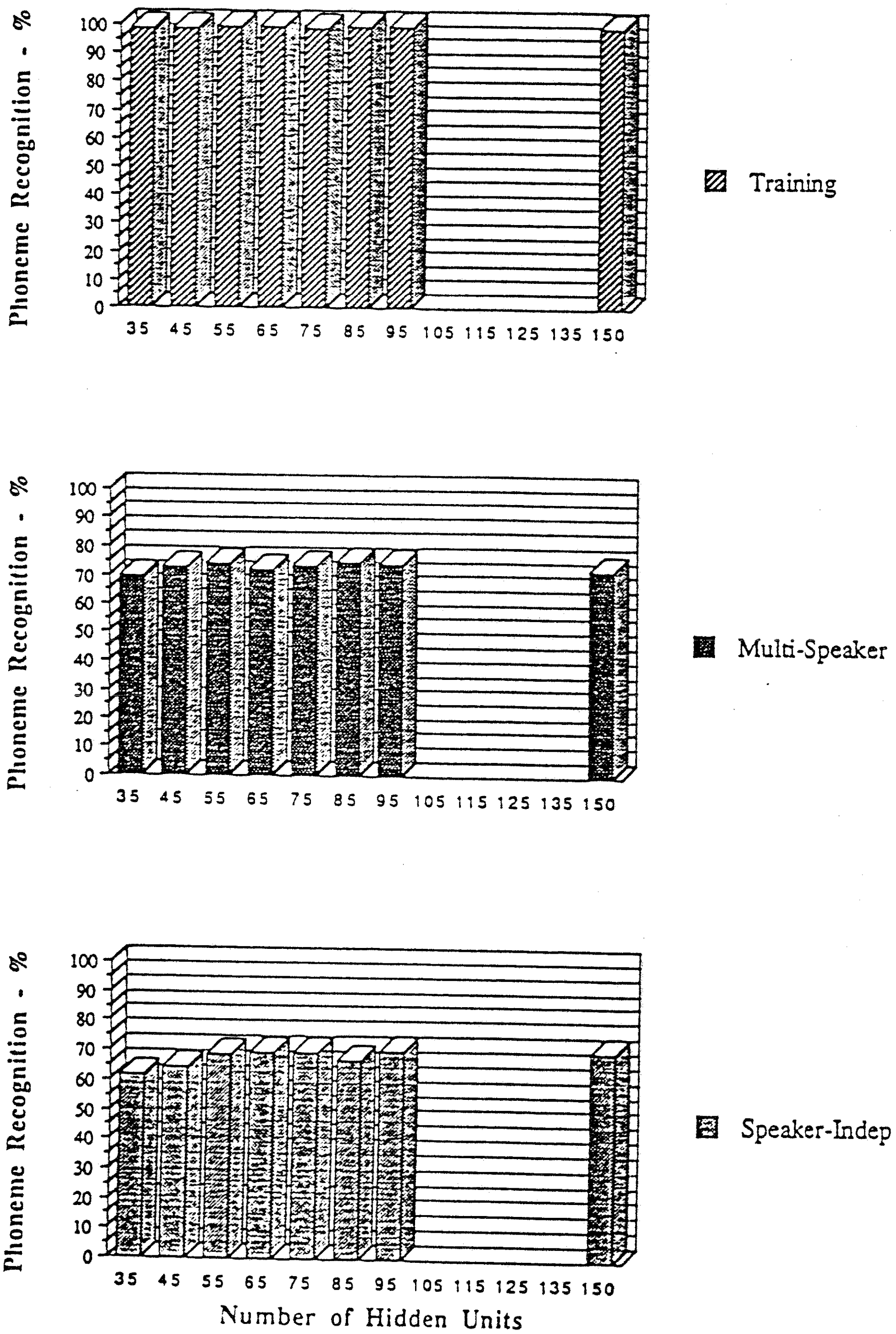


Figure 5.9: 3-frame Smoothed Data - Mimic Model

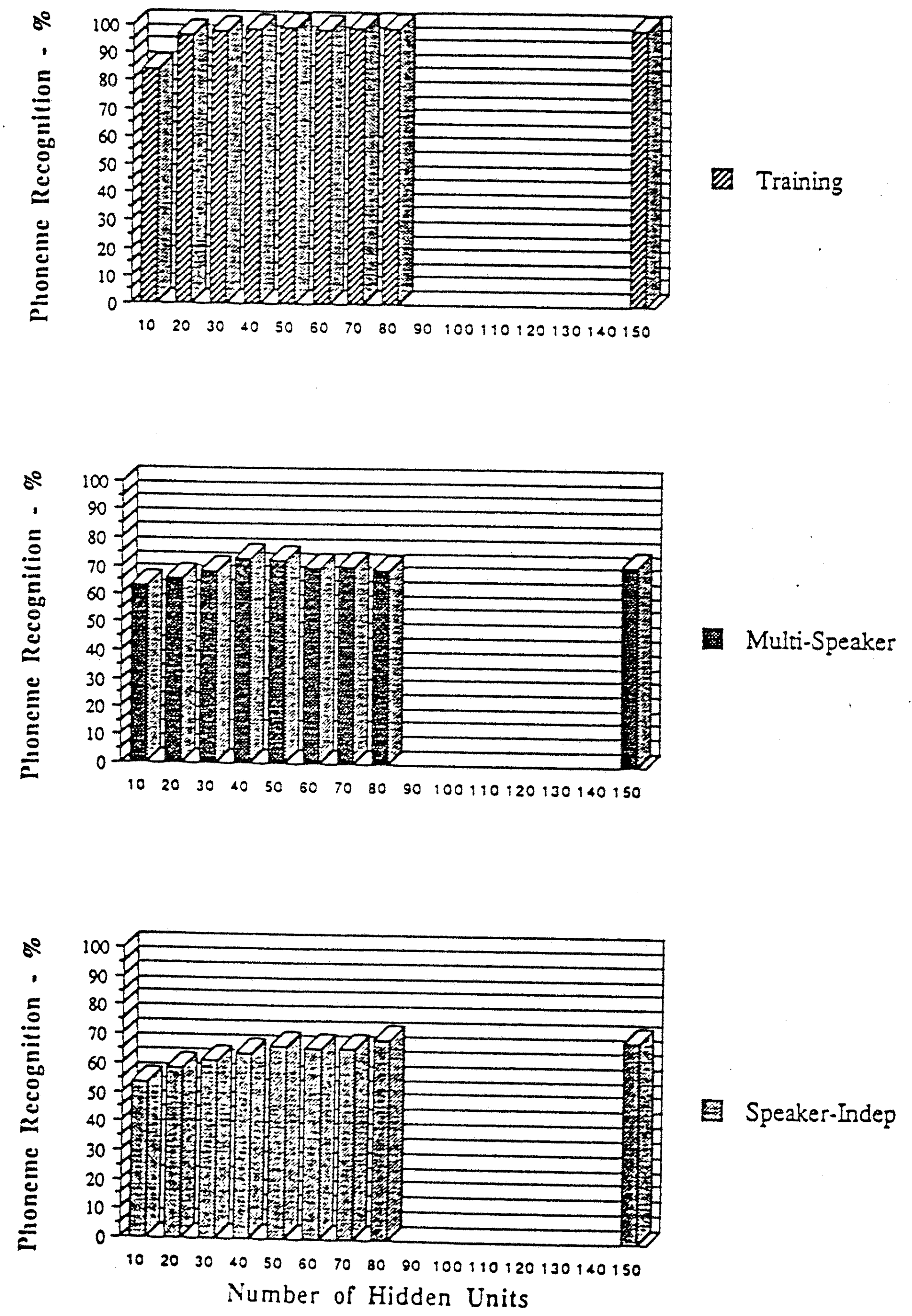


Figure 5.10: 3-frame Smoothed Data - Classifier Model(eta = 0.1)

units for the Full Motor Theory networks trained with 1-frame unsmoothed input. The corresponding numbers may be seen in Table 5.12. Full Motor Theory networks having from 50 to 150 hidden units each exhibit 100% training recognition. Multi-speaker generalization results for these networks range from 66% to 71%. For speaker-independent generalization, results range from 62% to 67%. While the overfitting guideline would lead us to believe that the best generalization would occur in such a network which has the smallest number of hidden units, 50 units, better overall generalization occurs when there are 90 hidden units in the hidden layer.

Table 5.12: 1-frame Unsmoothed Data - Full Motor Theory Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
15	91	59	49
30	98	60	55
50	100	69	63
80	100	66	62
90	100	71	67
150	100	70	65

Each of these networks was trained from the same initial conditions as the corresponding 1-frame smoothed input networks. These networks also exhibit comparable levels of generalization across a broad range of hidden units. Also evident is a broader range of variation in multi-speaker and speaker-independent generalization than was seen in the networks trained with 1-frame smoothed speech input data.

A similar pattern of results was found in tests of the Classifier networks trained with 1-frame unsmoothed input. Figure 5.12 shows the mean training recognition, multi-speaker generalization, and speaker-independent generalization results versus number of hidden units for these networks. Corresponding values appear in Table 5.13. Classifier networks having from 50 to 150 hidden units each exhibit 100% training recognition. Multi-speaker generalization ranges from 65% to 68%

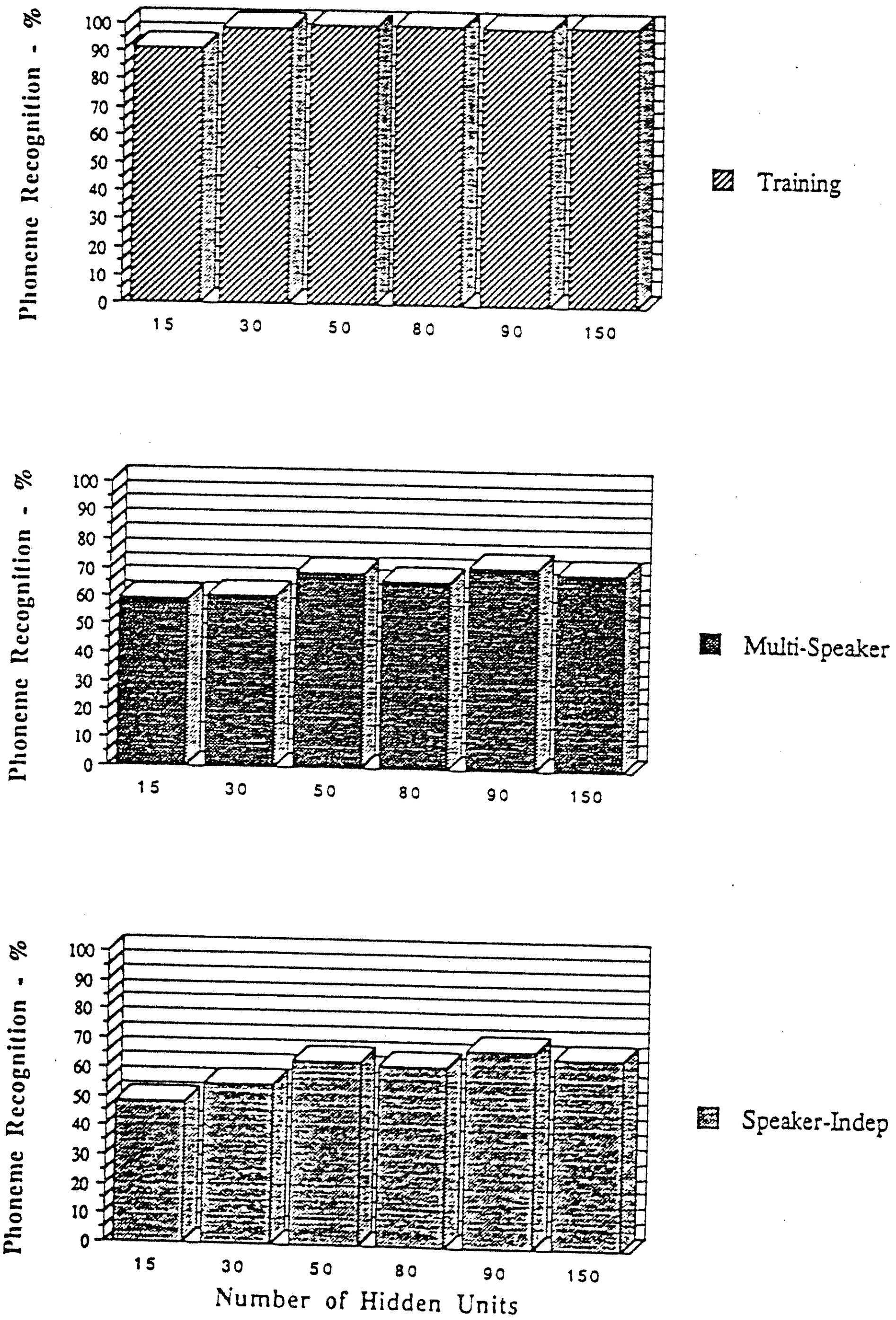


Figure 5.11: 1-frame Unsmoothed Data - Full Motor Theory Model

Table 5.13: 1-frame Unsmoothed Data - Classifier Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
15	95	57	50
30	99	63	56
50	100	65	63
70	100	68	63
90	100	65	63
150	100	68	60

across these four networks. Speaker-independent generalization ranges from 60% to 63%. Contradicting the overfitting guideline, the network exhibiting the best overall generalization is the one containing 70 hidden units.

For these two models, neither the overfitting guideline nor the dataset-size guideline appear to apply to the networks trained with 1-frame unsmoothed data. The insensitivity of network performance to variations in number of hidden units, can be seen in these networks also.

Yet another series of experiments were run using 3-frame unsmoothed input. Figure 5.13 shows the mean training recognition, multi-speaker generalization, and speaker-independent generalization results for Full Motor Theory networks trained with different numbers of hidden units. Corresponding values appear in Table 5.14. Networks containing 50, 70 and 90 hidden units each exhibit training recognition of 100%. Contradicting the overfitting guideline, it is clear that the network having 90 hidden units exhibits the best overall generalization.

A similar pattern of results was found with the Classifier networks trained with 3-frame unsmoothed input data. Figure 5.14 depicts the mean training recognition, multi-speaker generalization, and speaker-independent generalization exhibited by these networks. Corresponding values are listed in Table 5.15. Networks containing from 30 to 90 hidden units each exhibit 100% training recognition. Contradicting

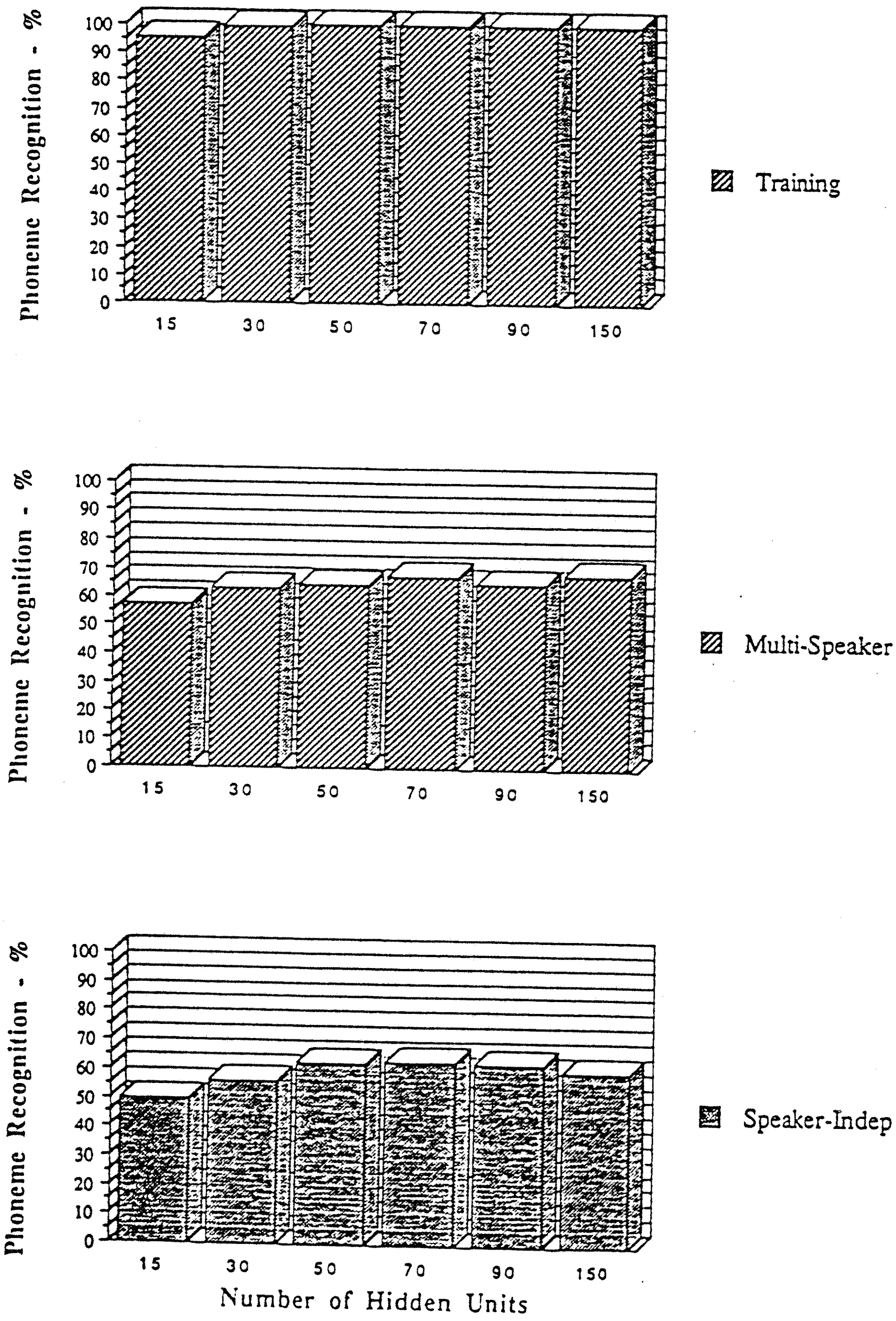


Figure 5.12: 1-frame Unsmoothed Data - Classifier Model

Table 5.14: 3-frame Unsmoothed Data - Full Motor Theory Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
15	98	61	50
30	99	68	63
50	100	67	58
70	100	68	63
90	100	76	66

Table 5.15: 3-frame Unsmoothed Data - Classifier Model

Mean Phoneme Recognition

Hidden Units	Train %	Multi-Spkr %	Spkr-Indep %
15	98	60	52
30	100	70	63
50	100	75	64
70	100	71	66
90	100	76	68

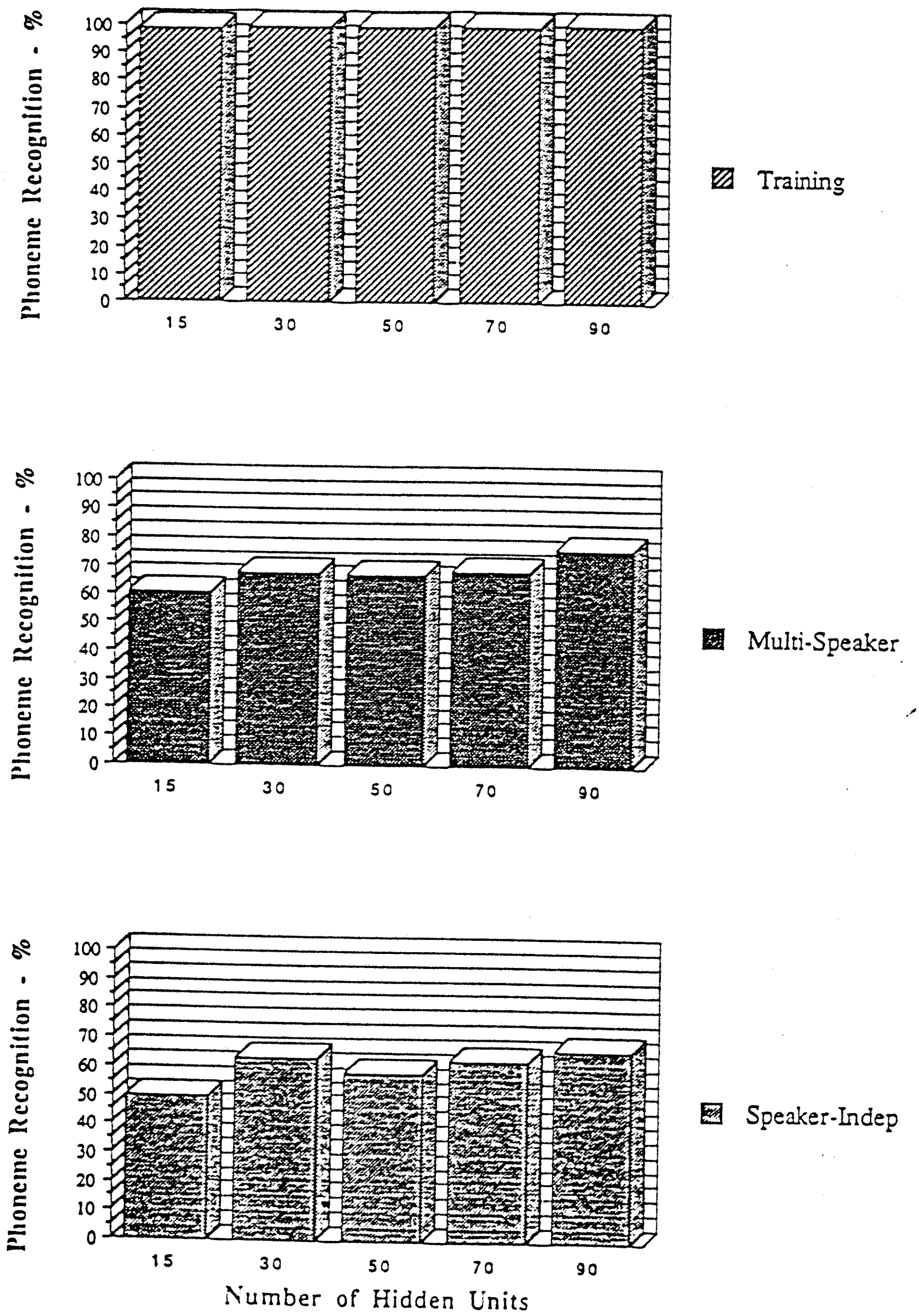


Figure 5.13: 3-frame Unsmoothed Data - Motor Theory Model

the overfitting guideline, the network exhibiting the best overall generalization is the one containing 90 hidden units.

Neither the overfitting guideline nor the dataset-size guideline appear to apply to the two models trained with 3-frame unsmoothed data. These networks exhibit comparable levels of training recognition and generalization across a broad range of hidden units.

5.6.2 Architecture: multi-hidden-layers. All of the preceeding results were for single-hidden-layer networks trained with back-propagation using a learning rate of 0.1. It has been suggested that networks having more than one hidden layer might be capable of exhibiting better training and generalization performance. The rationale underlying such a suggestion is that the added layers provide the network with additional dimensions in which to reorganize the material to be learned.

Little is known about the relationship between network generalization and architecture for multi-hidden-layer networks. This is doubtless due to the multitude of potential architectures that can be specified once one opens the door to the possibility of having more than one hidden layer. One approach is to fix the total number of hidden units to be used in the network. The hidden units can then be distributed over the multiple hidden layers and a comparison made between the resulting performance and the performance exhibited by a single-hidden-layer network having the same total number of hidden units.

In addition to investigating single-hidden-layer networks, a limited investigation of both two hidden layer and three hidden layer networks was carried out. The multi-hidden-layer networks proved to be much more sensitive to learning rate than were single-hidden-layer networks. This sensitivity to learning rate of the multi-hidden-layer networks will be described in greater detail in Section 5.8.1. Suffice it to say at this point that it was necessary to explore the effect of learning rate on

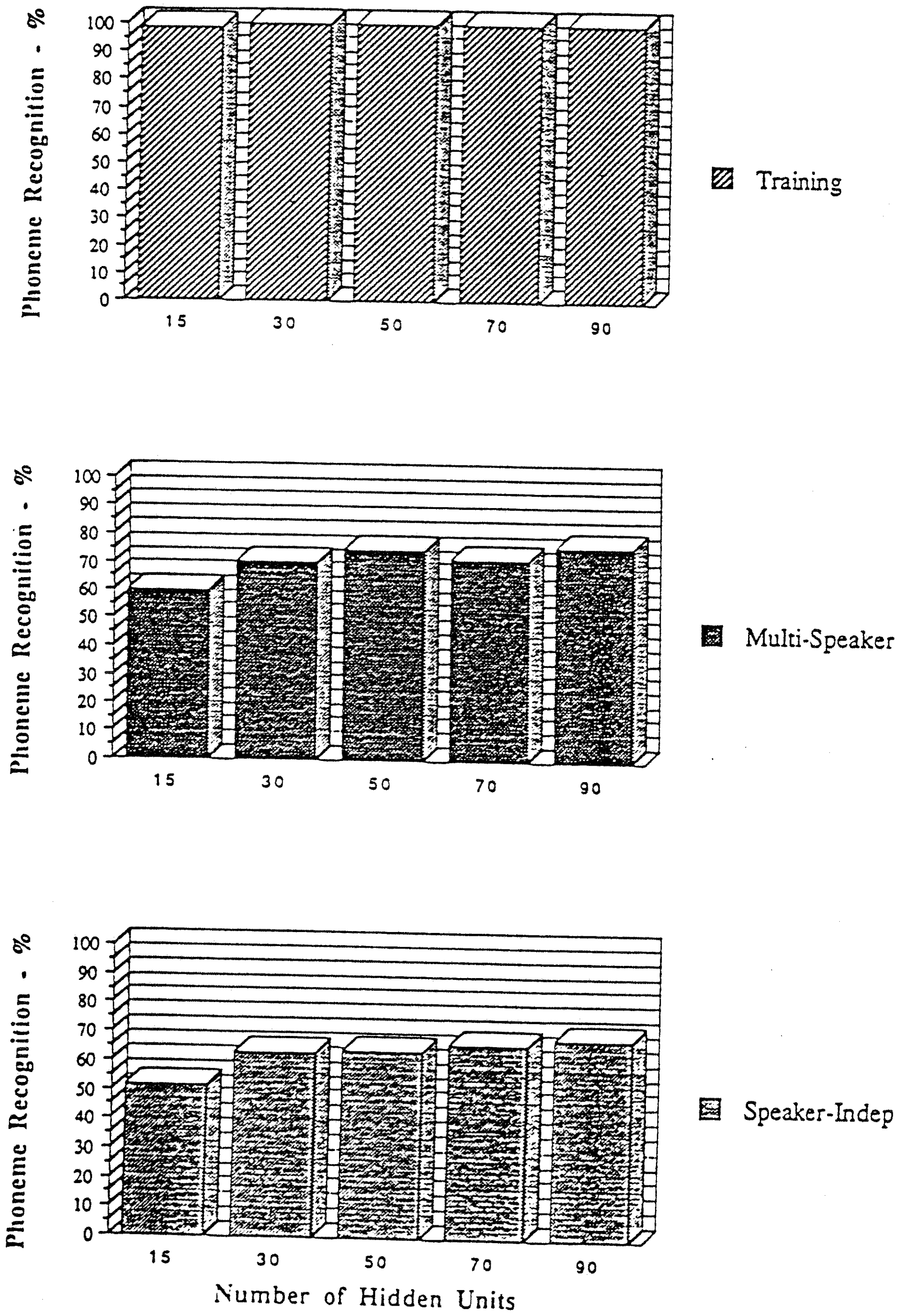


Figure 5.14: 3-frame Unsmoothed Data - Classifier Model

the training of the multi-hidden-layer networks in order to establish an appropriate rate at which to train the networks. While all of the single-hidden-layer networks were trained with a learning rate of 0.1, it was necessary to train the two and three hidden layer networks with a learning rate of 0.01.

Figure 5.15 shows the mean performance of a two-hidden-layer network having 70 hidden units distributed over the two hidden layers compared with the mean performance of a single-hidden-layer network containing the same number of hidden units. The training recognition rate and the multi-speaker generalization exhibited by the two-hidden layer network trained with a learning rate of 0.01 are essentially equivalent to that exhibited by the single-hidden-layer network containing 70 hidden units and trained with a learning rate of 0.1. The single-hidden-layer network has slightly more connections than the two hidden layer network - 2870 versus 2690 - but the rate at which it learns each connection is an order of magnitude faster than the rate at which the two hidden layer network learns each connection - 0.1 versus 0.01.

Using the learning efficiency measure described in Section 4.3, the learning efficiency for the single-hidden-layer network is, approximately, $34.8 * 10^{-6}$ while the learning efficiency for the two hidden layer network is $3.7 * 10^{-6}$. The single-hidden-layer network exhibits a learning efficiency approximately 10 times greater than that exhibited by the two-hidden-layer network.

For a three-hidden-layer Full Motor Theory network trained with the same input data, training recognition and generalization results similar to those exhibited by single-hidden-layer networks were observed. The learning efficiency for the 90 hidden unit single-hidden-layer network was an order of magnitude greater than that for the three-hidden-layer network, $10.7 * 10^{-6}$ versus $1.8 * 10^{-6}$.

Both two-hidden-layer and three-hidden-layer networks were trained with 3-frame smoothed speech data. The results of these experiments were essentially the

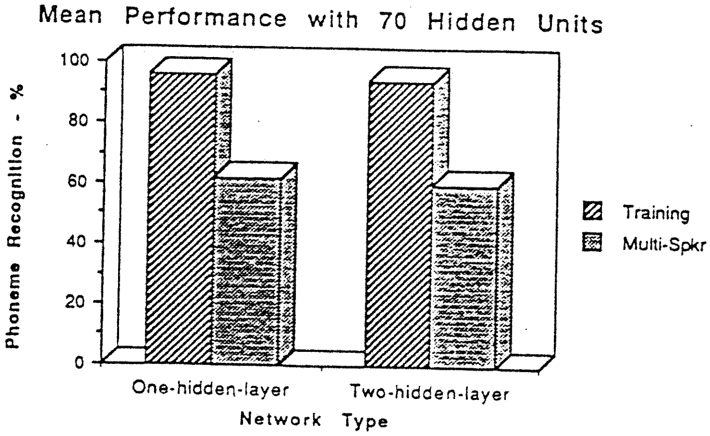


Figure 5.15: Performance as a Function of Hidden Layers

same as those reported above in the case of the 1-frame smoothed data training. The multi-hidden-layer networks proved to be extremely sensitive to learning rate. The training recognition and the multi-speaker generalization were no better than that exhibited by the single hidden layer networks. Considering the difference in learning efficiency, all remaining experiments were carried out using single-hidden-layer networks.

These empirical results are similar to the theoretical results reported by Judd [13]: for a given number of hidden units, it is better to contain those units in a single-hidden-layer than to distribute them over two or more hidden layers. There appears to be no difference in the training recognition and generalization results as a result of this restriction. In addition, the learning efficiency of the single-hidden-layer networks was an order of magnitude greater than that of the multi-hidden-layer networks.

5.6.3 Architecture: conclusions. The experimental results indicate that interestingly comparable levels of training recognition, multi-speaker generalization, and speaker-independent generalization are exhibited across all of these networks trained with four different data representations. The overfitting guideline which states that if a network can learn a problem, then the fewer the number of free parameters in the network the better the network is likely to generalize simply does not apply in the case of these networks. For the networks that satisfied not just the initial training recognition criterion but an even higher empirically determined training recognition criterion, those having more free parameters appeared to train and generalize as well or even better than those having fewer free parameters.

Likewise, as the experimental results exhibit, the dataset size guideline which holds that the larger the number of free parameters in a network the more data needed to train it also does not apply in the case of these networks. It is quite possible that additional data might have permitted the networks to learn a better mapping function and, thus, to generalize better. That is not the point being made here. What is pertinent to the current discussion is the fact that, for the given training data, networks having more free parameters trained and generalized as well, if not better, than networks having fewer free parameters.

It is worthwhile to note that networks which exhibit either the same or only minimally different training recognition can exhibit markedly different generalization. For example, the 1-frame smoothed Echo networks having 60 and 70 hidden units each exhibit a training recognition of 95%. The 60 hidden unit network exhibits generalization of 59% and 50% versus 62% and 54% for the 70 hidden unit network. Similarly, the 3-frame smoothed Mimic networks having 35 and 95 hidden units exhibit a training recognition of 99%; the generalization for the former is 69% and 62% versus 74% and 70% for the latter. It is important that researchers track generalization directly rather than merely tracking error rate or training recognition. This point will be discussed further in Section 5.8.4.

One additional result, which parallels the results reported by Judd [13], is that for a given number of hidden units, it appears to be better to contain those units in a single-hidden-layer than to distribute them over two or more hidden layers. There appears to be no difference in the training recognition and generalization results as a result of this restriction and, in the case of the networks investigated here, the learning efficiency of the single-hidden-layer networks was an order of magnitude greater than that of the multi-hidden-layer networks.

5.7 Task

I turn now to a discussion of the results pertaining to the relationship between the task of the network and the ability of the network to generalize well to novel data. The composite tasks are described in Section 2.4.

As described in Section 4.3 above, a procedure external to the network simulator was used to compute a classification measure for each subtask. In addition, a classification measure was computed by a majority vote of all subtasks. Here we find a single task, classification, having multiple realizations across various architectural/data representations. The correlations between each of these subtask recognition measures and the total vote recognition measure were extremely high, $r > .99$. In addition, the phoneme classification measure, which occurred in each of the four network models, was virtually always within less than 0.5% of the total vote recognition measure. This provided a convenient means of directly reading the phoneme classification accomplished by each of the networks. The phoneme classification measure was used in developing a multiple regression analysis.

Recall that I had hypothesized that adding additional tasks should cause a change in generalization. That change, as a result of the added constraints, may be an increase in generalization, or, as a result of an increase in the network capacity, it may decrease generalization. As the hypothesis indicates, variations in task necessitate variations in network architecture. In Section 4.2, the composite task was used, in combination with the appropriate data representation, to specify the architectural details of each network. It can be argued that the architecture and data representation factors can also be implicated in the results described. I contend that the primary factor in the current discussion is the network task. At least in the case of architectural manipulation, my contention received solid support from the lack of effect of architectural manipulation discussed in Section 5.6.

5.7.1 Task: analysis of the experiments. In Section 5.6.1 I presented the experimental results for the training recognition, multi-speaker generalization, and speaker-independent generalization exhibited for each of the four network models trained with 1-frame and 3-frame smoothed speech input and 1-frame and 3-frame unsmoothed speech input. I have already discussed the striking similarity of these results for each of the network models. In comparing models having the same hidden structure trained with the same auditory input an interesting similarity of training recognition and generalization results can be seen. This is true even for the Full Motor Theory networks which have a slightly larger input architecture as a result of the added articulatory feature input. For example, Table 5.16 shows the comparative results for all the networks trained with 1-frame smoothed auditory input data representations.

Notice that as the number of hidden units increases from 30 to 60 there is an increase in training recognition until an average training recognition of 96% is achieved across all four network types. A similar increase can be seen in multi-speaker generalization, with an average generalization of 62%. An increase in speaker-independent generalization to an average of 53% occurs across the same range of hidden units. From 70 to 150 hidden units the networks exhibit virtually no change in either training recognition or multi-speaker generalization. There is a gradual increase in speaker-independent generalization from an average 53% to 56% across this range. There are markedly comparable levels of performance across differing network types having the same number of hidden units once the networks contain 70 or more hidden units.

The comparable levels of performance and the similarity of results is made even more striking by examination of a multiple regression analysis of the networks' training recognition and generalization. First, all of the networks trained with 1-frame smoothed speech data input were analyzed using a multiple regression analysis where the input architecture, number of hidden units, and output architecture were used as independent variables. The 1-frame representation is reflected in the input architecture, which is the same for all but the Full Motor Theory model. The tasks assigned to the networks differ primarily in the differing outputs for each of the four models. The architectures of the networks differ primarily in the number of hidden units in each network and, secondarily, in the aspects of the input and output

Table 5.16: 1-frame Smoothed Data - All Models

Mean Phoneme Recognition

Hidden Units	Network Type	Train %	Multi-Spkr %	Spkr-Indep %
30	Full	85	54	50
	Echo	90	57	55
	Mimic	88	59	50
	Classifier	91	61	48
40	Full	90	59	49
	Echo	93	59	47
	Mimic	93	67	51
	Classifier	93	59	54
50	Full	94	59	52
	Echo	93	63	53
	Mimic	94	58	49
	Classifier	95	59	57
60	Full	96	61	56
	Echo	95	59	50
	Mimic	95	63	51
	Classifier	95	61	54
70	Full	96	62	54
	Echo	95	62	54
	Mimic	96	62	53
	Classifier	95	62	55
80	Full	97	64	57
	Echo	95	61	53
	Mimic	97	66	57
	Classifier	96	61	53
90	Full	97	60	56
	Echo	95	63	52
	Mimic	97	63	55
	Classifier	95	61	56
150	Full	97	62	58
	Echo	95	60	52
	Mimic	97	63	58
	Classifier	95	62	55

architecture which reflect the tasks of the network. The training recognition, multi-speaker generalization, and speaker-independent generalization are the dependent variables. The t-values and associated probabilities for this analysis are listed in Table 5.17.

Table 5.17: 1-frame Smoothed Data - Multiple Regression Analysis

Measure	X	t-value	Probability
Training	Input	0.18	.8550
	Hidden	7.17	.0001*
	Output	0.42	.6761
Multi-spkr	Input	0.72	.4748
	Hidden	4.66	.0001*
	Output	0.15	.8845
Spkr-indep	Input	1.32	.1978
	Hidden	4.85	.0001*
	Output	1.11	.2772

* significant at $p < .02$

The criterion of $p < .05$, which is commonly accepted as the threshold to reject the null hypothesis in the analysis of human data may not be appropriate for the analysis of data resulting from computational models. Models do not exhibit the same random variation. In order to assure that the analysis provides results which are truly significant I have established a threshold of $p < .02$ for rejection of the null hypothesis.

The results of this analysis indicate that for all four network models the differing tasks required of the networks, reflected in the input and output architectures, are not significant predictors of variance in network performance. The primary predictor of training recognition, multi-speaker generalization, and speaker-independent generalization variance is the number of hidden units in the networks. An examination of Table 5.16 indicates that initial increases in the number of hidden units result in increasing training and generalization performance for all models.

These four different network types differ in the required tasks. They are similar in that each network is required to perform the phoneme classification task. They are also similar in the input data with which they are trained. The networks can be grouped together based upon these similarities, a phoneme classification subtask and input data representation, and upon a common indifference to the total task required of each. Such a grouping of networks can be labeled a "family" of networks. The 1-frame smoothed input networks form a family of networks.

A similar multiple regression analysis with the input architecture, number of hidden units, and output architecture used as the independent variables and the training recognition, multi-speaker generalization, and speaker-independent generalization as dependent variables was performed for all of the networks trained with 3-frame smoothed input. The t-values and associated probabilities for this analysis are detailed in Table 5.18.

Table 5.18: 3-frame Smoothed Data - Multiple Regression Analysis

Measure	X	t-value	Probability
Training	Input	0.24	.8125
	Hidden	2.08	.0064*
	Output	0.49	.6285
Multi-spkr	Input	0.25	.8044
	Hidden	4.75	.0001*
	Output	0.05	.9572
Spkr-indep	Input	0.73	.4719
	Hidden	7.74	.0001*
	Output	0.47	.6324

* significant at $p < .02$

The results of this analysis indicate that for all four models the primary predictor of variance not just in training recognition, but also in multi-speaker generalization and speaker-independent generalization, is the number of hidden units in the hidden structure of the networks. Like the 1-frame input networks, initial increases in the number of hidden units in these 3-frame input networks result in increasing training and generalization performance for each model.

Once again, there was no difference in training recognition, multi-speaker generalization, or speaker-independent generalization based upon a difference in the task of the model. Like the 1-frame networks, the 3-frame networks have a common phoneme classification subtask and can be said to form a family of networks whose primary predictor of variance in performance is the number of hidden units.

As the above analyses show the 1-frame smoothed input networks and the 3-frame smoothed input networks form families in which the members exhibit remarkably similar training recognition and generalization. One explanation for this somewhat surprising grouping is that it could be a result of the smoothed nature of the input data. Martin and Pittman [22] attributed similarity in performance for their networks to such an explanation. A statistical analysis of the networks trained with unsmoothed data may shed light on this question.

A multiple regression analysis was performed for the two network models, the Full Motor Theory model and the Classifier model, trained with 1-frame and 3-frame unsmoothed data. The results were similar to those described above with the exception that the results regarding hidden units were somewhat unclear. Bearing in mind the limited number of data points available, I performed a linear regression analysis using number of hidden units as an independent variable. The results of this analysis for the networks trained with 1-frame unsmoothed data can be seen in Table in Table 5.19.

For the two models trained with 1-frame unsmoothed data the results are somewhat unclear. The significance of variation in the number of hidden units on

Table 5.19: 1-frame Unsmoothed Data - Linear Regression Analysis

Measure	X	t-value	Probability
Training	Hidden	2.21	.0544
Multi-spkr	Hidden	3.32	.0089*
Spkr-indep	Hidden	24.09	.0392

* significant at $p < .02$

training recognition and speaker-independent generalization do not meet the stringent criterion of $p < .02$. They are significant at $p < .05$, the commonly accepted threshold for rejection of the null hypothesis. As was discussed in Section 5.6.1, the 1-frame unsmoothed data results in greater experimental variation in performance than does the 1-frame smoothed data. In Section 3.1 it was noted that the 1-frame target vowel data reflects a static theory of vowel perception that may not be appropriate for the recognition of vowels. If this 1-frame data representation is less appropriate than the 3-frame representation then the utilization of unsmoothed data, while resulting in an increase in performance, (see discussion below in Section 5.9.2 for additional comments regarding this phenomenon) may also tend to mask the underlying commonalities between the models for the limited set of networks that have been investigated here. I contend that the analysis does not argue against the conclusion that these networks form a family based upon a similarity in input data and phoneme classification subtask.

The results of a similar analysis for the two network models trained with 3-frame unsmoothed data can be seen in Table 5.20. Although not as clear as the results for the corresponding networks trained with 3-frame smoothed data, they indicate that for these network models also the primary predictor of training recognition, multi-speaker generalization, and speaker-independent generalization variance, is the number of hidden units in the networks. For these networks there was no difference

Table 5.20: 3-frame Unsmoothed Data - Linear Regression Analysis

Measure	X	t-value	Probability
Training	Hidden	3.406	.0114*
Multi-spkr	Hidden	3.292	.0133*
Spkr-indep	Hidden	3.464	.0105*

* significant at $p < .02$

in training recognition, multi-speaker generalization, or speaker-independent generalization that depended upon a difference in the task of the network model. These networks form a family based upon a similarity in input data, 3-frame unsmoothed input, and a common phoneme classification subtask.

5.7.2 Task: conclusions.

As a result of these analyses I conclude that, with respect to the networks investigated in this research at least, networks using the same data input and having a common phoneme classification task form families of networks. These families exhibit a common indifference to to the total task required of each model. The difference in training recognition and generalization performance of the members of a particular family can only be statistically ascribed to differences in the number of hidden units used in the network. Based upon the analysis of networks trained with four different types of data input it does not appear reasonable to ascribe the common behavior of network family members to the input representation with which they are trained.

On the basis of these results then, the first hypothesis that adding additional tasks should cause a change in generalization which may be an increase in generalization (as a result of the added constraints) or it may be a decrease in generalization (as a result of an increase in the network capacity) must be rejected. Apparently adding constraints by requiring additional tasks of a network did not cause an increase in network generalization. The increase in network capacity which resulted from the increase in output architecture required by the additional tasks also did not cause a decrease in network generalization. Generalization simply did not change as a result of adding additional tasks to a network.

5.8 Training

There is an extensive body of literature reporting on a variety of different approaches to the training of neural networks. Consequently only a limited investigation was carried out exploring the relationship between training factors and the ability of the network to generalize well to new data. This investigation includes a limited manipulation of the learning rate, the training schedule, an alternative training algorithm, and quantity of training data. I turn now to the question of the relationship between factors affecting the training of the network and the ability of the network to generalize well to novel data.

5.8.1 Training: learning rate. As mentioned in Section 5.6.1, multi-hidden-layer networks proved to be sensitive to learning rate. Figure 5.16 presents the dynamic behavior of three 1-frame Classifier networks where 50, 70, or 90 hidden units were distributed among two hidden layers. The networks were trained with a learning rate of 0.1, the learning rate used in training single-hidden-layer networks. Note that the network having 50 hidden units was incapable of even beginning to learn the training data. The training recognition of the networks having 70 and 90 hidden units first increases to a high of approximately 87%. The training recognition then begins to exhibit an oscillatory behavior with recognition alternately decreasing then increasing. The overall trend in each case indicates a decrease in training recognition from the original high. The networks obviously are incapable of learning the speech training data.

Figure 5.17 shows the performance of the 70 hidden unit two-hidden-layer network when it is trained with a learning rate of 0.01. Reducing the learning rate

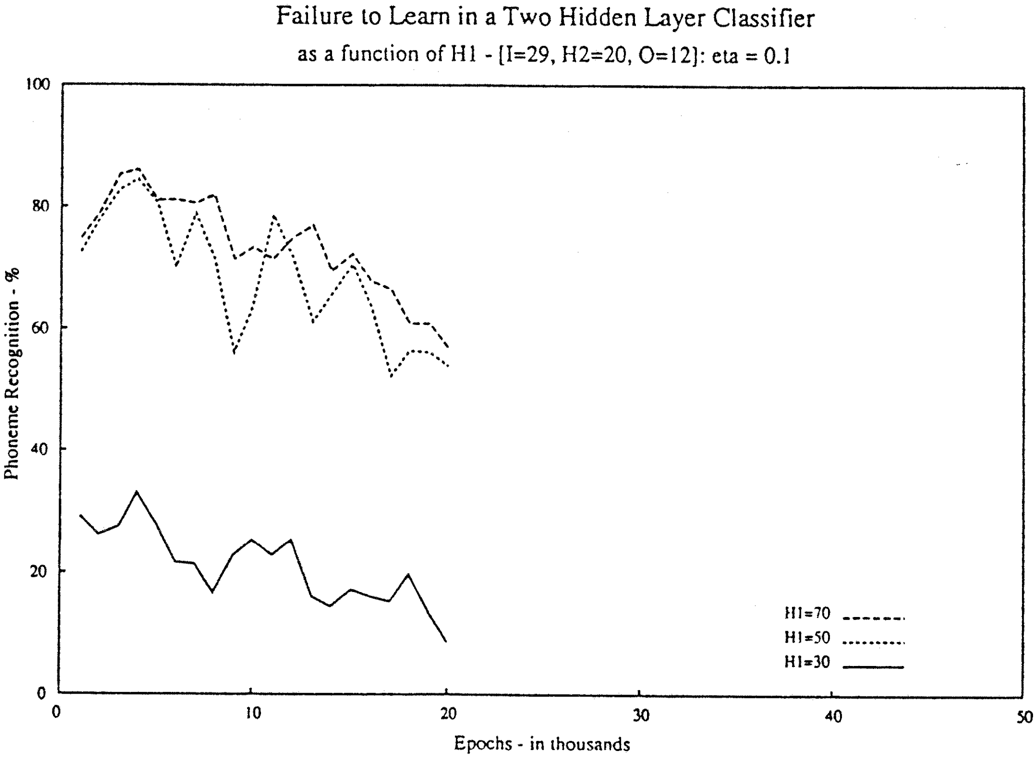


Figure 5.16: Failure to Learn in Two-Hidden-Layer Networks (eta = 0.1)

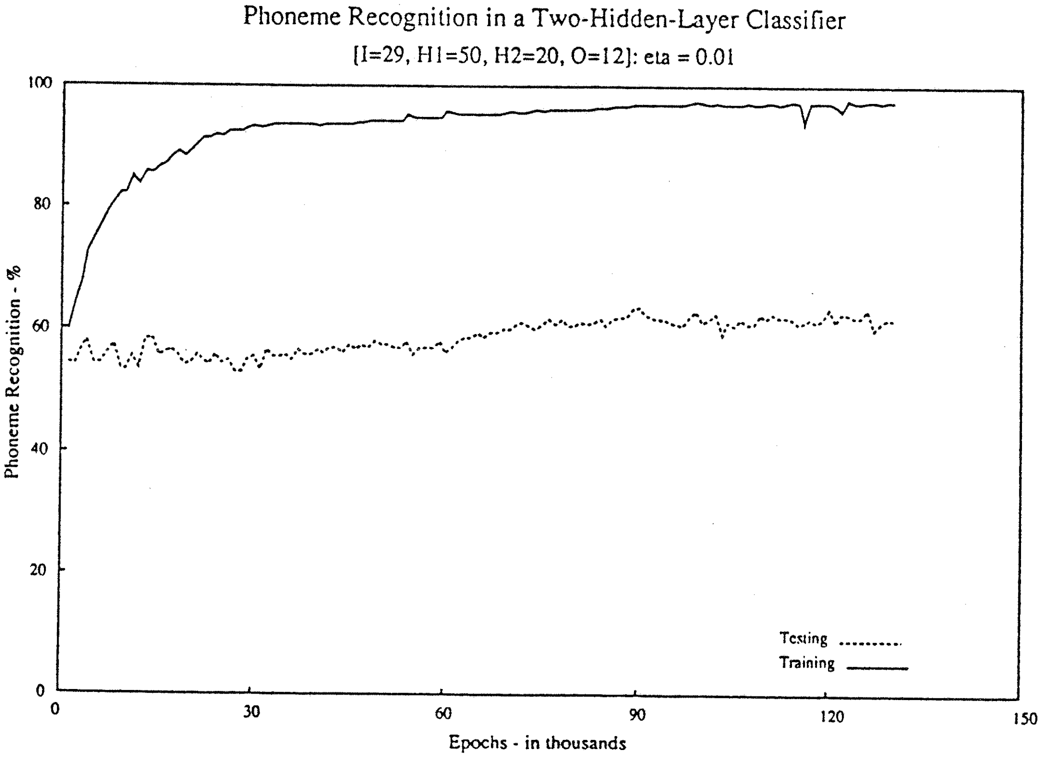


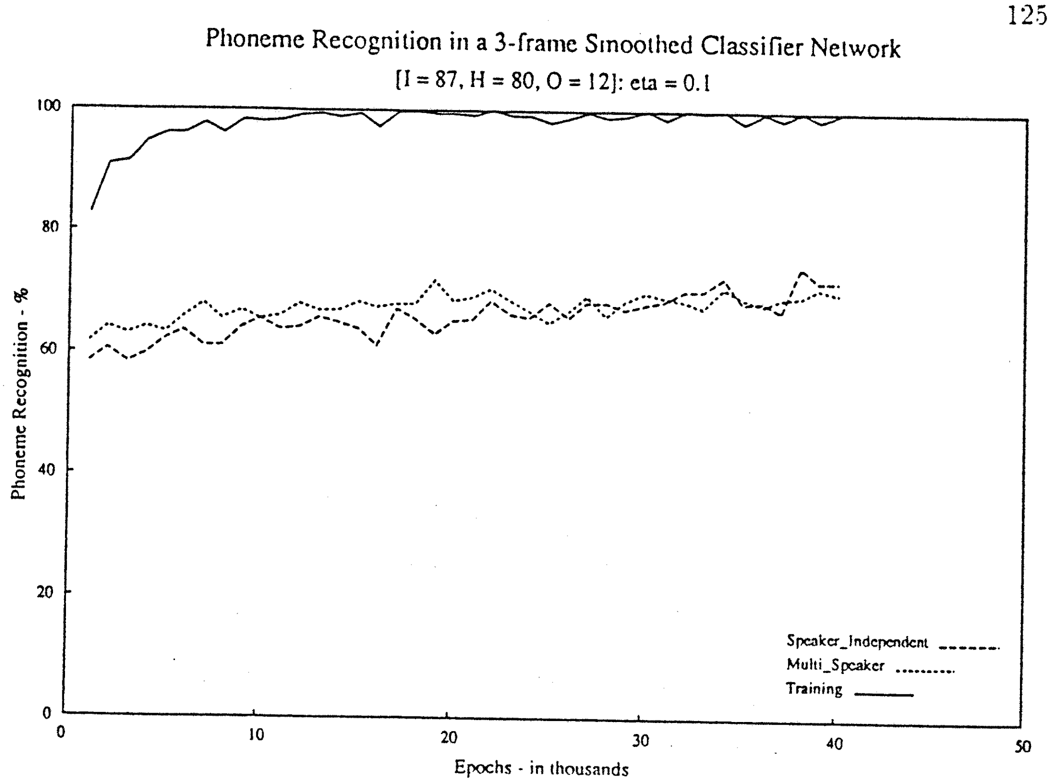
Figure 5.17: Performance in a Two-Hidden-Layer Network

by an order of magnitude permitted the network to learn the training data. These results indicate that multi-hidden-layer networks are more sensitive to learning rate than single-hidden-layer networks. It is necessary to find an appropriate learning rate when training these networks.

A study was undertaken to investigate the effect of varying the learning rate on the training of single-hidden-layer networks. The networks used for this study were Classifier networks trained with 3-frame smoothed input. For a given number of hidden units, the initial state of the network was the same in each case. The networks were trained using two learning rates which differed by an order of magnitude, 0.1 and 0.01.

Figure 5.18 depicts the performance exhibited for a network having 80 hidden units when trained with a learning rate of 0.1 (top) and when trained with a learning rate of 0.01 (bottom). Training performance in the case of a .1 learning rate is essentially flat across the final 25,000 epochs. Generalization, on the other hand, exhibits a slight upward trend throughout the training period. Both training performance and generalization in the case of a .01 learning rate are essentially flat across the last 30,000 epochs. Similar results are exhibited for each of the other networks when comparing training with a learning rate of .1 to training with a learning rate of.01.

Figure 5.19 compares the mean training recognition, multi-speaker generalization, and speaker-independent generalization exhibited by the two sets of networks. It is apparent from the figure that networks trained with the lower learning rate exhibit lower performance in all three categories. A multiple regression analysis of the networks was performed using number of hidden units and learning rate as the independent variables. Training recognition, multi-speaker generalization, and speaker-independent generalization were used as the dependent variables. The results are somewhat mixed as to the impact of each of the independent variables upon



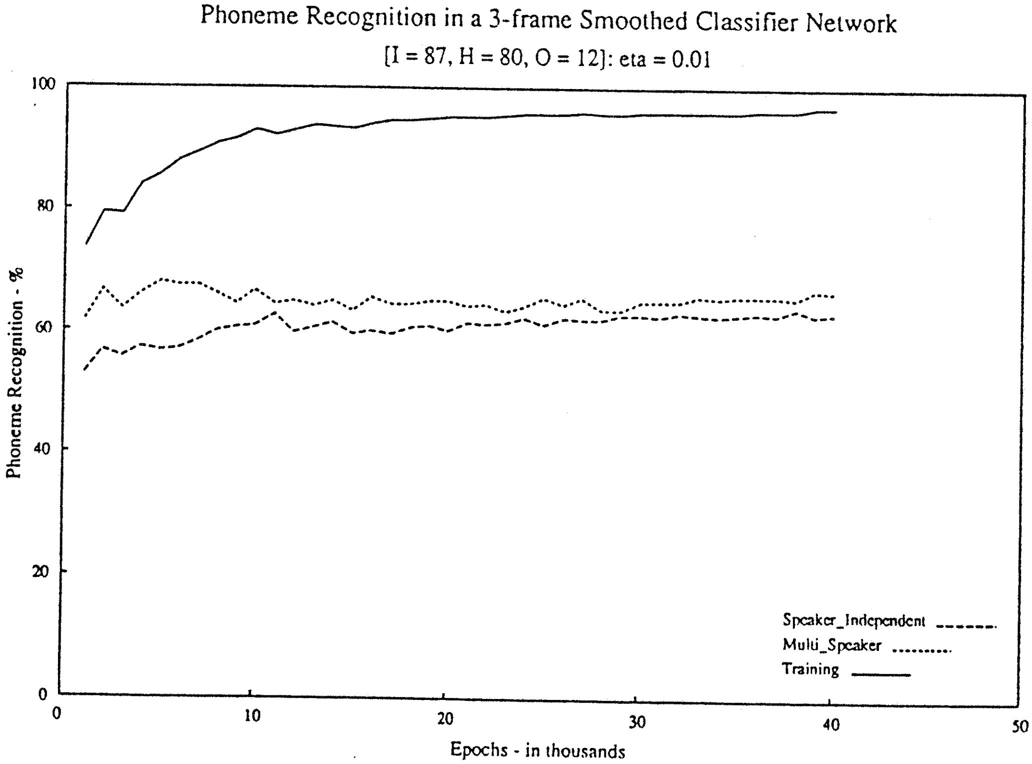


Figure 5.18: Comparative Performance for Two Learning Rates

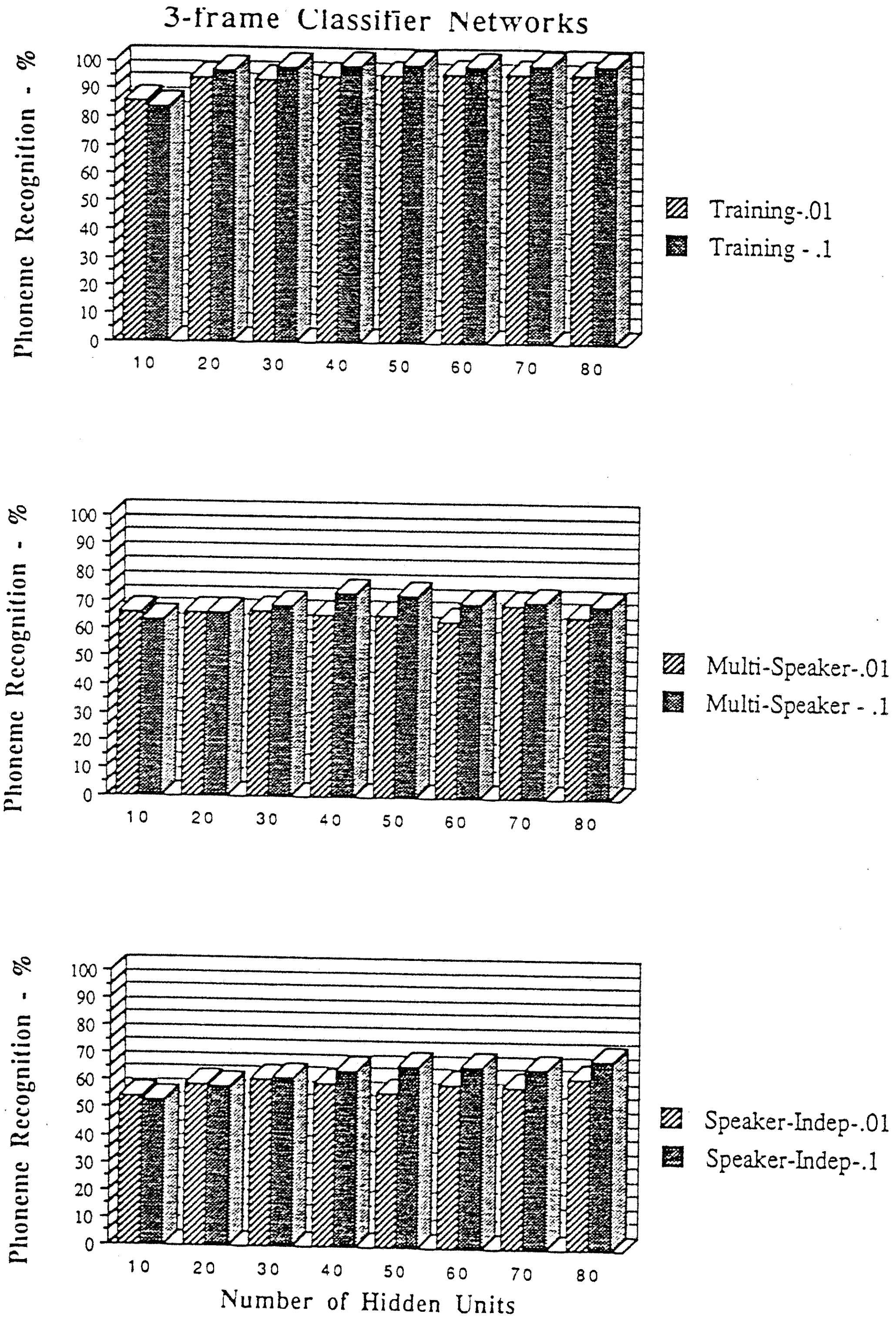


Figure 5.19: Classifier Performance as a Function of Learning Rate

the dependent variables, as can be seen in Table 5.21. Clearly, as far as training recognition is concerned, the number of hidden units is the primary predictor of variance in the performance results. This reflects the increase in training recognition performance resulting from the increase in number of hidden units. In terms of multi-speaker and speaker-independent generalization the training rate assumes a predictive role. I hesitate to draw strong conclusions from such a limited investigation and such mixed results. It appears that consideration should be given to the effect of both the design of the network's hidden structure and also to the effect of the learning rate upon the network's ability to generalize.

Table 5.21: Effect of Learning Rate on Classifier Networks

3-frame Smoothed Data Classifier Networks

Measure	X	t-value	Probability
Training	Hidden	2.69	.0196*
	Learning Rate	1.25	.2357
Multi-spkr	Hidden	1.68	.1193
	Learning Rate	2.72	.0186*
Spkr-indep	Hidden	3.69	.0031*
	Learning Rate	2.78	.0165*

* significant at $p < .02$

5.8.2 Training: schedule of training.

As indicated previously the networks described in Section 5.6.1 were trained with random presentation of the training patterns. Two alternative schedules of training were also explored: incremental training by number of vowel patterns and training by task.

By incremental training I mean the following: first, the network is trained to learn one example of each vowel spoken by each speaker; then the number of examples is increased by an additional example of each vowel spoken by each speaker and the network is retrained using the expanded training set; this last step is repeated until the network is trained with the full training set. While there is no theoretical motivation for training a network by the incremental increase in training set size, previous experience indicated that such training can sometimes lead to somewhat better or faster training. Here I tried an incremental approach to training a Full Motor Theory network. The results of this investigation quickly indicated that the network only achieved recognition rates equivalent to those achieved with random presentation of the entire pattern set. Training also took longer using the incremental approach. As a result, this approach was pursued no further.

An approach with somewhat greater theoretical motivation is that of training the network in a manner akin to that in which a child appears to learn language. First, children hear speech sounds, then they "babble", and finally, understand and recognize spoken language. Using this approach I first tried training a Full Motor Theory network to perform the auditory association task until recognition reached an asymptote. The task of articulatory association was added until there was no further improvement in articulatory recognition. Finally, the task of phonemic labeling was included in the training regime. This task-oriented approach to training was tried both without and with freezing the appropriate connection weights for the tasks being trained. Suffice it to say that neither approach resulted in any increase in the ultimate training recognition of the network as measured across any of the three tasks and both approaches took longer to effect final training than did the simple all-at-once random presentation approach.

One final aspect should be mentioned with regard to training schedule. As noted previously, the majority of networks investigated were trained for 40,000 epochs. Figure 5.17 showed the results of training a two-hidden-layer with a learning rate of .01 for up to 128,000 epochs. Note that, in contrast to the results reported by Morgan and Bourlard [26] (Section 1.2.1), this network does not appear to exhibit reduced generalization as a result of extensive training. This result is true for all of the networks in this research which were extensively trained.

5.8.3 Training: alternative algorithm. For purposes of comparison, Classifier networks fed with 1-frame smoothed input data were trained using the conjugate gradient algorithm in addition to standard back-propagation. The random seed used to generate the connection weights was the same for both algorithms. The software used for this training was provided to us by Mark Fanty at the Oregon Graduate Institute. The conjugate gradient training was significantly faster than the back-propagation training algorithm. However, the results of this training are somewhat questionable. The training recognition and multi-speaker generalization results are summarized in Table 5.22. While the results for multi-speaker generalization are comparable across the two algorithms, the training recognition results simply are not. I was never able to resolve the differences between the two training algorithms in this respect. David Shaw, a doctoral student in Psychology, who was also using this software and experiencing similar problems, passed along the results of these experiments to Mark Fanty at OGI. The response was as follows:

Table 5.22: Conjugate Gradient Training versus Back-Propagation

	Conjugate Gradient		Back-Propagation	
Hidden Units	Train %	Multi-Spkr %	Train %	Multi-Spkr %
10	61.6	57.9	74.6	53.0
15	58.7	56.3	78.9	52.6
20	60.9	56.7	81.6	58.1
25	59.6	53.8	88.2	60.3
30	60.3	55.0	91.3	60.7
35	61.1	57.9	91.6	58.7
40	59.1	52.9	93.5	59.4
45	61.6	55.8	93.2	59.8

> All I can suggest is that the conjugate gradient descent pushes the network to a local minimum for some reason....One other person reported a similar problem.

Different random seeds did not help.[3]

5.8.4 Training: data quantity.

Seventy out of 94 tokens for each of ten speakers, for a total of 700 tokens, were allocated to the training set. If one can reduce the size of the training set and still achieve equivalent training recognition and generalization, it is possible to reduce the amount of time required to train a network. Networks were trained with 1-frame and 3-frame smoothed data representations using either a full training set or a reduced training set. In the reduced training set two tokens/vowel/speaker had been removed resulting in a training set containing 460 tokens.

As Figure 5.20 shows, training recognition for a network trained with a full training set and a network trained with a reduced training set are essentially the same in the case of 1-frame smoothed auditory input data. Multi-speaker generalization using the same multi-speaker testing set is the same: 64%. In the case of speaker-independent generalization, however, there is a slight degradation in generalization: 54% versus 57%. It would appear that the full training set is not providing the

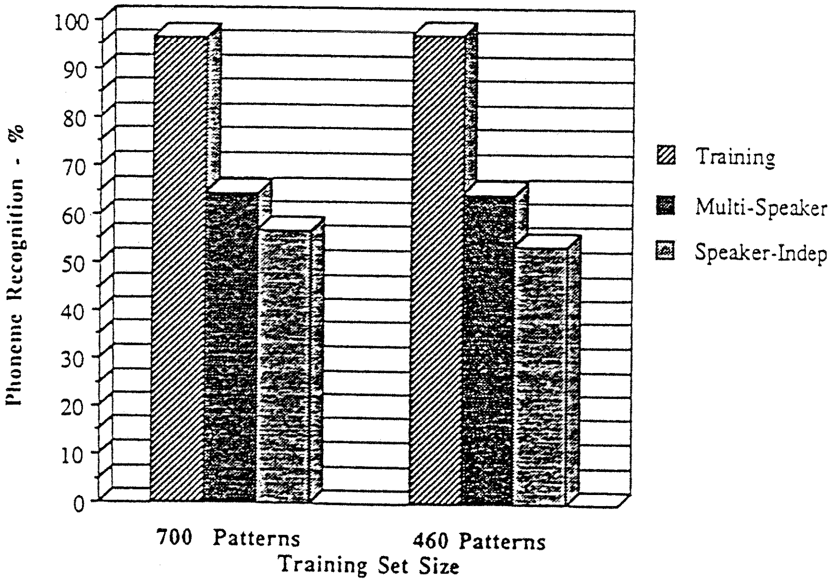


Figure 5.20: 1-frame Performance as a Function of Training Set Size

[3] Personal communication from Mark Fanty.

network with that much more evidence of underlying similarities than is the reduced training set. These results do not indicate that providing additional 1-frame training data would have resulted in an improvement in network performance.

In the case of a network trained with 3-frame smoothed auditory input data the situation shown in Figure 5.21 is somewhat different. Once again the training recognition results are essentially the same. The generalization for both multi-speaker and speaker-independent data are significantly degraded for the network trained with the reduced training set: 68% versus 74% for the multi-speaker data and 63% versus 71% for the speaker-independent data. Obviously, the underlying similarities being identified by the network given the reduced 3-frame training set were not as universal as those identified for the full training set. As a result of these experiments it did not appear fruitful to pursue the possibility of training the networks by further reducing the training set. These results indicate that providing additional 3-frame training data might have resulted in an improvement in network performance.

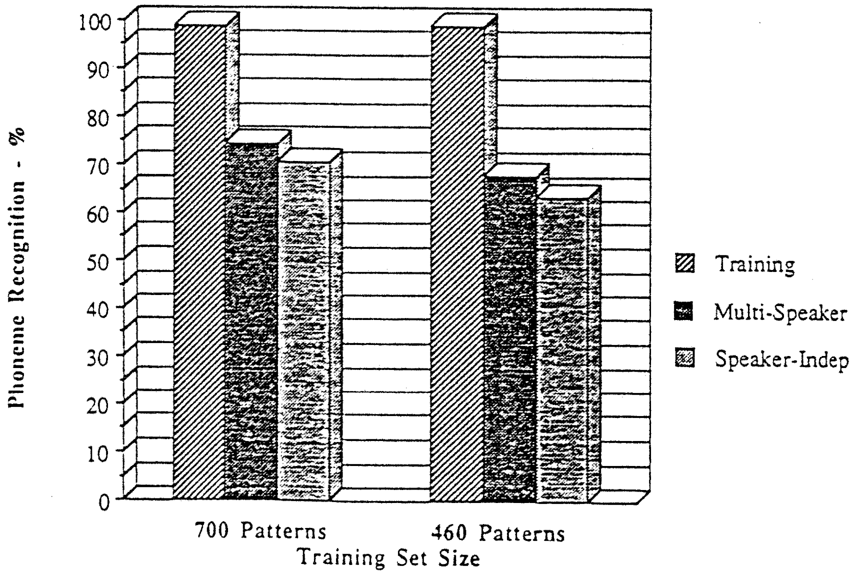


Figure 5.21: 3-frame Performance as a Function of Training Set Size

Recall from Section 1.2.1 that Cheung et al reported on an approach to improve network generalization by dynamically enlarging the training set through a process that includes additional copies of training patterns which were more poorly trained than others. I tried a modified version of this approach in which a subset containing the most poorly trained patterns was selected every 100 cycles instead of every cycle. The resulting network exhibited little better than a 1% performance increase on the training set, no improvement in generalization, and a perceptibly longer training time. Based on this experience, I did not find this a promising line of investigation.

5.8.5 Training: conclusions.

Of the three aspects of my investigation touched upon thus far: architecture, task, and training, the results regarding the relationship between the training of the network and the ability of the network to generalize well to novel data were the least productive and promising. I was able to conclude that one should give consideration not just to the design of the hidden structure of the network but also to a limited investigation of the effect of the training learning rate upon the network's ability to generalize.

In contrast to the results reported by Morgan and Bourlard, the networks investigated in this research did not exhibit a decrease in generalization as a result of overtraining. A network trained up to 128,000 epochs showed no evidence of poor generalization.

The investigations into different training schedules and an alternative training algorithm were essentially unproductive. With respect to the effect of varying the number of training exemplars for a given training class, the negative results of these experiments forced me to conclude that it did not appear fruitful to pursue further the possibility of training the networks with either a reduced or a dynamically enlarged training set.

One result occurred repeatedly throughout the investigation. Figure 5.22 shows the training recognition and generalization exhibited by a Full Motor Theory network having 150 hidden units and trained with 1-frame unsmoothed data. The figure also shows the mean square training error rate and the mean square multi-speaker and speaker-independent error rates. Several items of interest are evident in this figure. First, although the training error rate decreases to a minimum and then increases slightly, the training recognition increases to a high level and remains relatively stable thereafter. Second, the mean square multi-speaker and speaker-independent error rates decrease to a minimum and remain relatively stable; yet the multi-speaker and speaker-independent generalization continue to increase throughout the training of the network! The increase in generalization contrasts strongly with the stable error rates. If the network were attempting to fit noise in the data the network would exhibit a decrease in generalization with increased training. Obviously the network is not exhibiting effects resulting from overtraining. Apparently the network is able to reorganize the error on the individual units in a manner that results in an improvement in generalization while the mean square error across all units remains stable. Similar results were observed in other networks as well. These results suggest that researchers should take the precaution of tracking both training performance and generalization during the network training process rather than simply tracking the training error rate.

5.9 Data Representation

I complete my report of the experimental results of the investigation by describing the results pertinent to the question of the relationship between the data representations used in training the network and the ability of the network to generalize well to previously unseen data. Comparisons are made both within and between the smoothed data and the unsmoothed data.

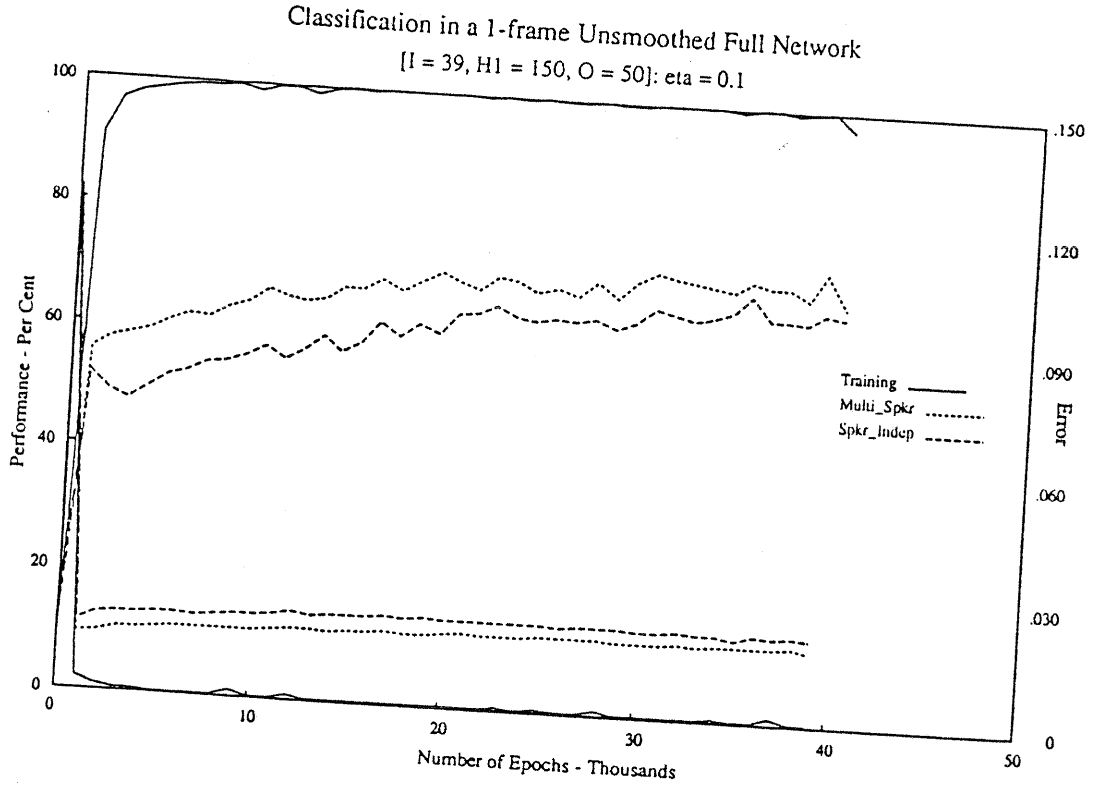


Figure 5.22: Performance, Generalization and Mean Error Rate

5.9.1 Data representation: 1-frame versus 3-frame. I hypothesized that training with appropriate representations will help the network to generalize better. Experiments were run to determine the effect of different representations on generalization. In the case of 1-frame versus 3-frame representations the differing representations required a different input architecture. Shifting from a 1-frame to a 3-frame architecture also had the effect of increasing network capacity. According to the dataset-size guideline, training an increased capacity network with a fixed size training set should have resulted in a decrease in generalization. My hypothesis contradicts this guideline by asserting that the appropriateness of the data representation is of greater significance than the physical size of that representation. This was suggested by the results of Cheung et al [6].

I first consider the performance exhibited by networks trained with 1-frame smoothed data and 3-frame smoothed data. The results show that for a given model the training recognition, multi-speaker generalization, and speaker-independent generalization vary as a function of the 1-frame versus 3-frame input representations.

In Sections 5.4 and 5.6.1, I discussed the fact that the 3-frame input representation required no more hidden units than the smaller 1-frame input representation in order for a network to exhibit good training recognition and generalization results. For example, in the case of the best Full Motor Theory networks the 3-frame representation actually results in a slightly higher training recognition, 100% versus 98%, significantly better multi-speaker generalization, 78% versus 68%, and significantly better speaker-independent generalization, 72% versus 59%. Figure 5.1 compares the training recognition, multi-speaker generalization, and speaker-independent generalization results for the best networks trained with smoothed data. It is worthwhile to note that the best 1-frame Full Motor Theory network has a total of 7,120 connections while the best 3-frame Full Motor Theory network has a total of 10,290 connections - 3,170 more connections to be learned during the training process. Nevertheless, utilizing the same number of training patterns, the 3-frame network is able to learn the training set better than the 1-frame network. It is also able to generalize that knowledge to previously unseen multi-speaker data significantly better than the 1-frame network, 78% versus 68%, a difference of 10%. This difference in generalization ability to previously unseen data is even more striking when one considers the speaker-independent data. The 3-frame network exhibits 72% recognition of new speaker-independent data while the 1-frame network only exhibits 59% recognition of the same data, a difference of 13% in generalization.

Comparison of Figures 5.3 and 5.7 show that similar results for the mean training recognition, the mean multi-speaker generalization, and the mean speaker-independent generalization were obtained for all of the smoothed data Full Motor Theory Networks. Similar results can be seen in the other network models across the 1-frame and 3-frame input representations. It can be seen that, given the same number of hidden units and the same network task, the network trained using the 3-frame input representation, in general, exhibits a higher training recognition and better generalization than the 1-frame representation.

An analysis of all of the network models trained with both the 1-frame smoothed data and the 3-frame smoothed data was performed using a multiple regression procedure in which the input architecture/input data representation, the output architecture, and the hidden structure of the networks serve as independent variables and training recognition, multi-speaker generalization, and speaker-independent generalization are dependent variables. The results of this analysis are summarized in Table 5.23.

The analysis shows that across all four of the network models trained with smoothed data the primary predictors of variance are the number of hidden units and the input representation used in training the network. I have previously noted that

Table 5.23: All Smoothed Data Networks - Multiple Regression Analysis

Measure	X	t-value	Probability
Training	Input	6.73	.0001*
	Hidden	7.25	.0001*
	Output	0.06	.9505
Multi-spkr	Input	17.25	.0001*
	Hidden	6.88	.0001*
	Output	2.31	.0245
Spkr-indep	Input	21.83	.0001*
	Hidden	8.56	.0001*
	Output	1.56	.1241

* significant at $p < .02$

initial increases in the number of hidden units result in increasing training and generalization results for each model. A change from a 1-frame to a 3-frame smoothed input representation also makes a difference. The 1-frame representation provides information only about the vowel target (Section 3.1), while the 3-frame representation provides additional information regarding the dynamic vowel transitions and is considered to be a potentially more appropriate representation for vowel perception by students of vowel perception theory.

Experiments were conducted using an extended 5-frame smoothed input consisting of five spectral frames in which two additional frames are selected at the starting and ending vowel segmentation points. The results indicate that this additional 5-frame information provided no further improvement in mean training recognition (99%) and, in fact, results in a 3% degradation in mean generalization when compared with the results for the 3-frame network (Figure 5.23). In contrast to the results exhibited by the 1-frame network, the 5-frame network does exhibit significantly better training recognition and generalization results. With respect to the smoothed 1-frame and 3-frame data representations, I conclude that, as the above discussion indicates, the hypothesis which states that training with appropriate representations will help the networks to generalize better is valid. With respect to the smoothed 1-frame and 5-frame data representations, this hypothesis is equally valid. With respect to the 3-frame and 5-frame data representations, it would appear that the two representations are equally appropriate. Appropriateness of representation cannot be measured in terms of the physical size of the representation; it can only be estimated by an *a posteriori* comparison of network performance.

Let us now consider the results for the 1-frame versus 3-frame representations in the case of unsmoothed data. Comparing the performance between the Classifier networks trained with 1-frame and 3-frame unsmoothed data shows similar variation to that seen with the 1-frame and 3-frame smoothed networks. For the full

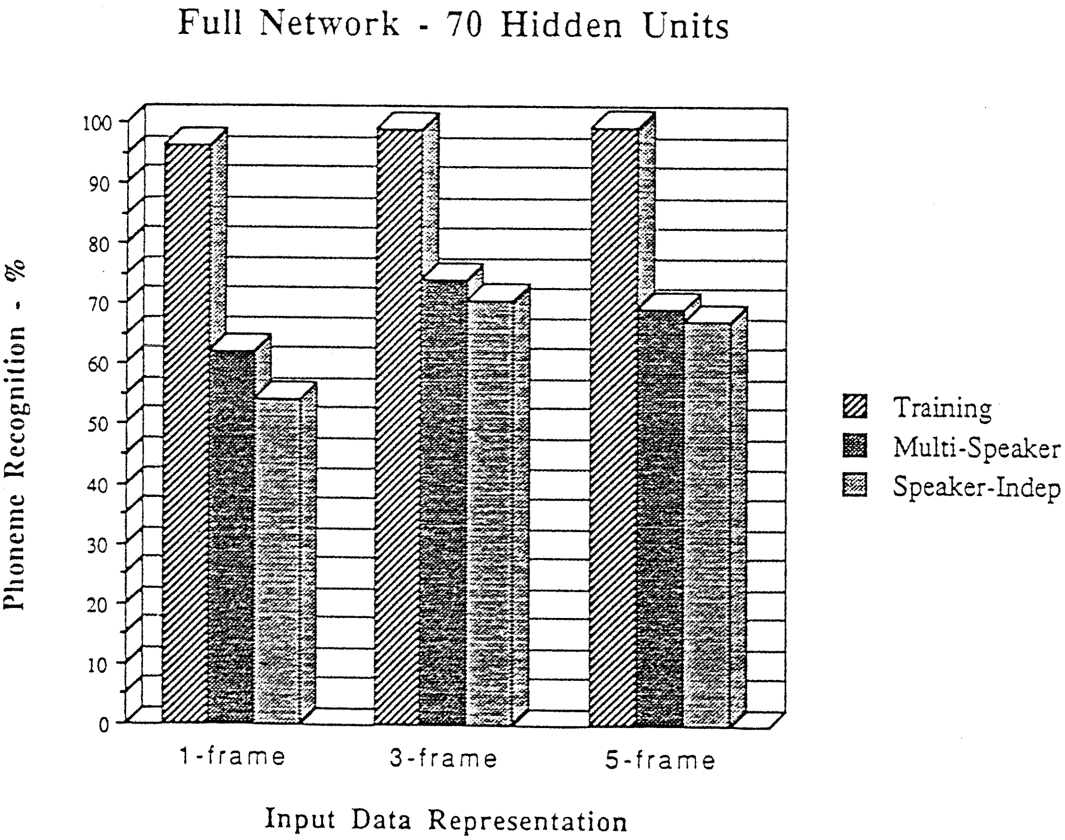


Figure 5.23: Network Performance as a Function of Input Representation

Motor Theory networks trained with unsmoothed data, a slightly different pattern was observed. Both the 1-frame and 3-frame networks exhibit reasonably similar behavior. It is not clear why this should be so. One potential explanation is based upon the concept of task. On this view the additional tasks required of the Full Motor Theory network so constrained the internal representations developed by the network that the network was unable to take advantage of any additional information that might have been provided by the 3-frame unsmoothed data. An alternative explanation based upon the concept of appropriate representation combined with the task of a network will be developed in the next section.

5.9.2 Data representation: smoothed versus unsmoothed. Figure 5.24 compares the effect of training a Full Motor Theory network with 1-frame smoothed versus 1-frame unsmoothed data. Using unsmoothed data leads to better training recognition, multi-speaker generalization, and speaker-independent generalization than does the smoothed representation. Figure 5.25 depicts similar results for the 1-frame Classifier network. In Figure 5.26, however, the situation is reversed for the 3-frame Full Motor Theory network. Using the smoothed training data leads to equivalent training recognition and somewhat better multi-speaker and speaker-independent generalization than does the unsmoothed data representation. In Figure 5.27, the 3-frame Classifier networks given smoothed and unsmoothed training data are equivalent in their training recognition and generalization.

A multiple regression analysis of training recognition and generalization results for the 1-frame Full Motor Theory and Classifier networks trained by both smoothed and unsmoothed data is shown in Table 5.24. Obviously both number of hidden units and the smoothed/unsmoothed data representation are significant predictors of variance. As Figure 5.24 and Figure 5.25 indicate, for the 1-frame networks, training with the unsmoothed data results in better training recognition and generalization. The analysis supports that conclusion.

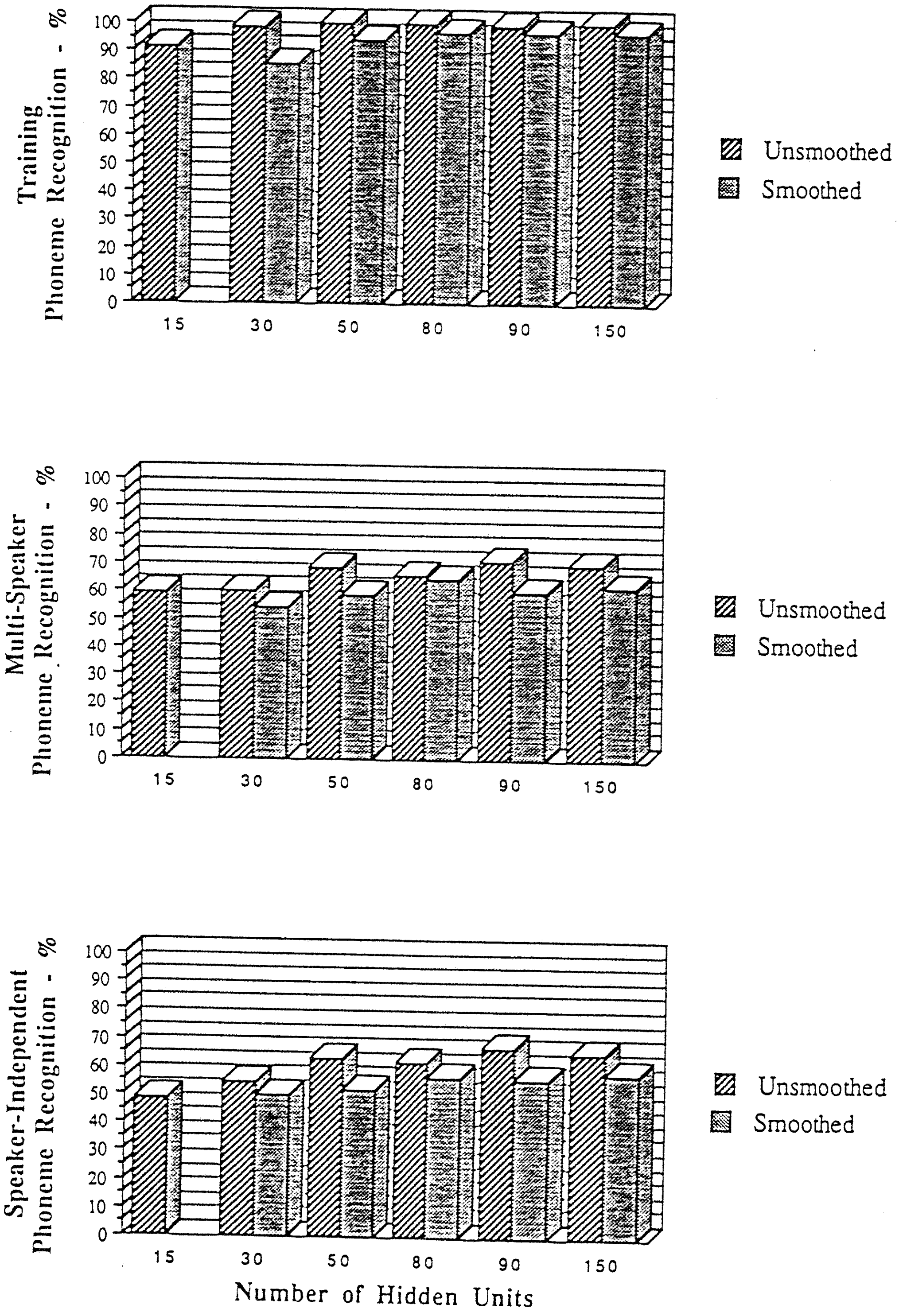


Figure 5.24: Smoothed vs. Unsmoothed Data - 1-frame Motor Theory

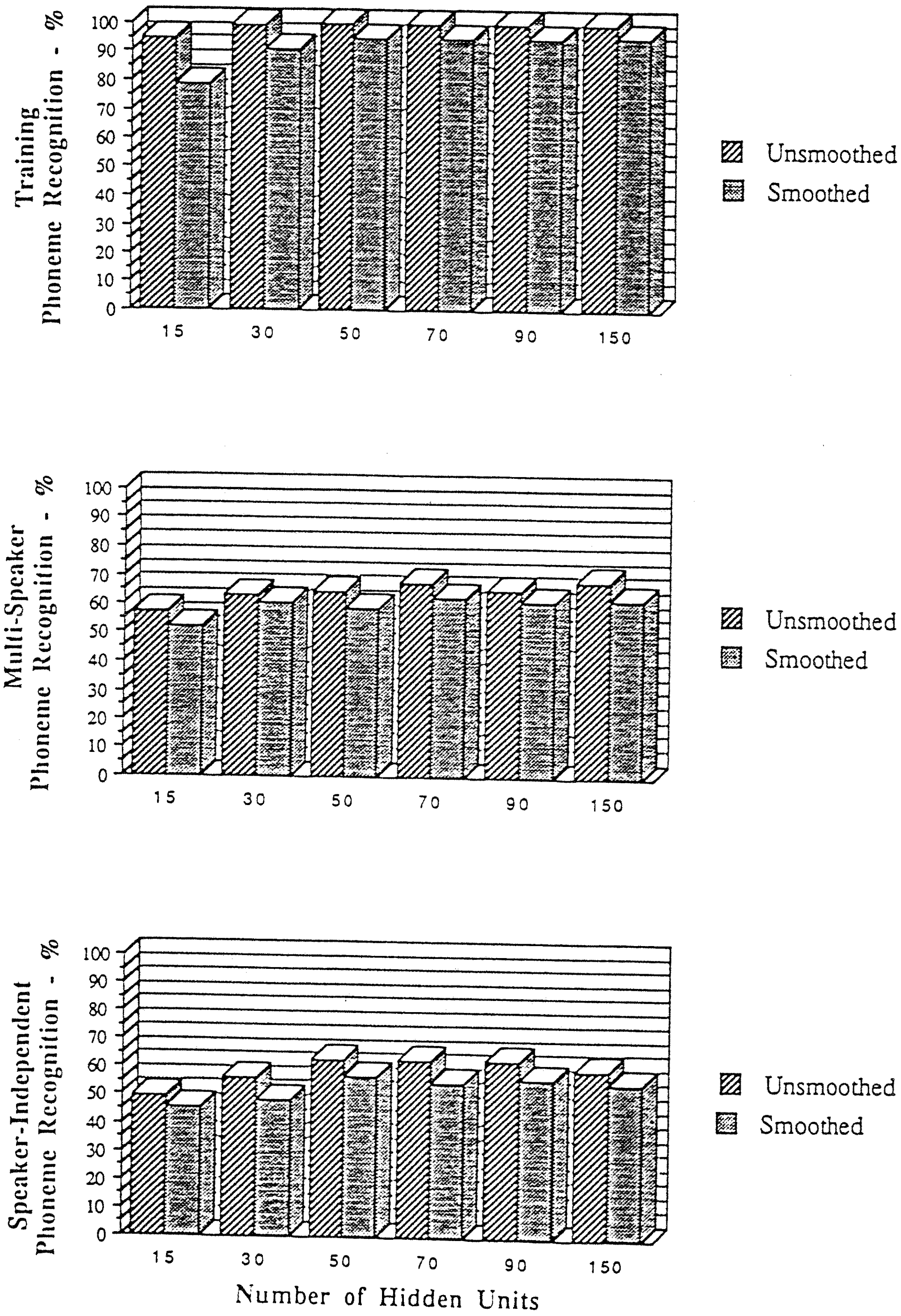


Figure 5.25: Smoothed vs. Unsmoothed Data - 1-frame Classifier

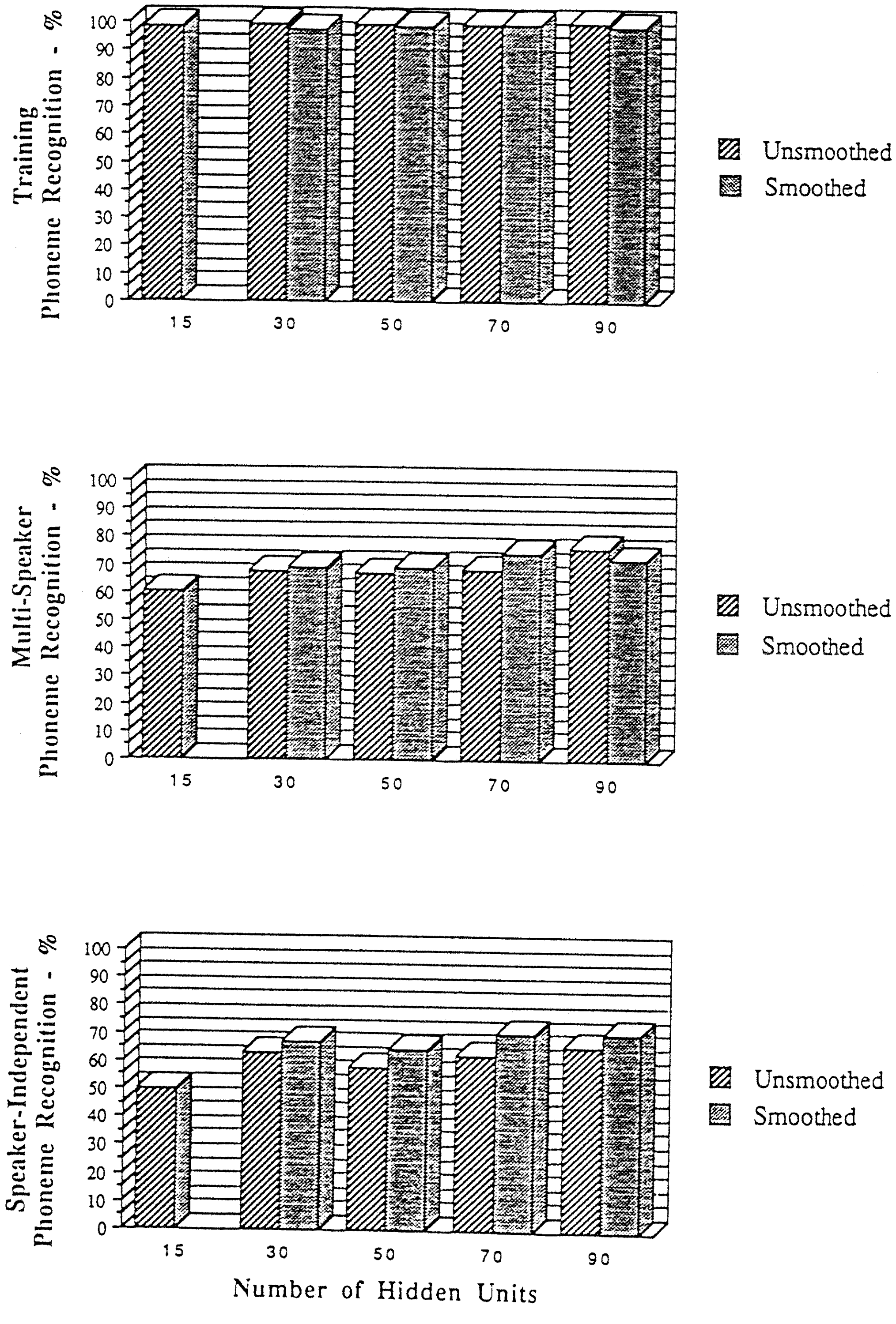


Figure 5.26: Smoothed vs. Unsmoothed Data - 3-frame Motor Theory

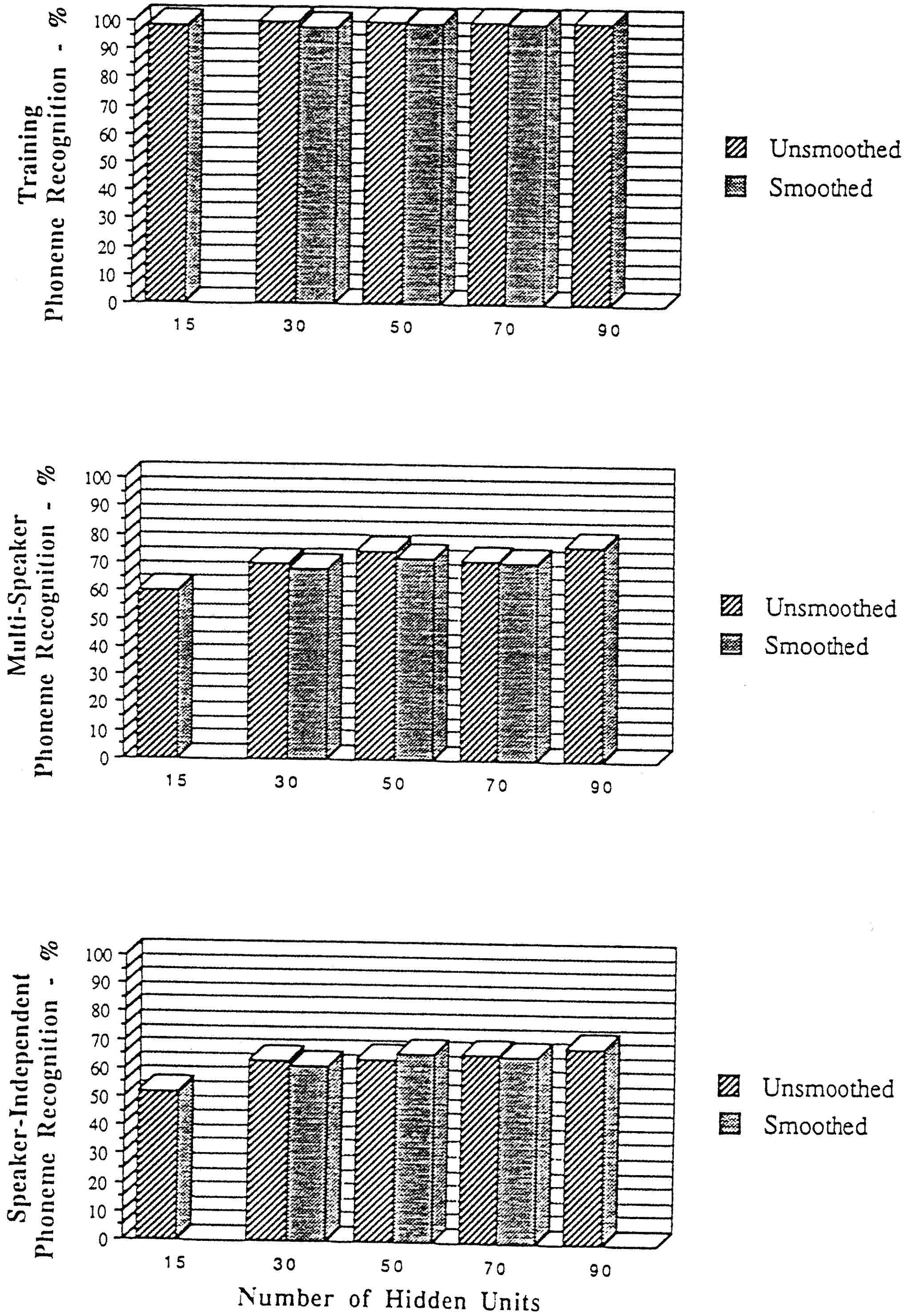


Figure 5.27: Smoothed vs. Unsmoothed Data - 3-frame Classifier

Table 5.24: All 1-frame Data Networks - Multiple Regression Analysis

Measure	X	t-value	Probability
Training	Hidden	4.51	.0001*
	Unsmooth/Smooth	4.22	.0002*
Multi-spkr	Hidden	4.22	.0001*
	Unsmooth/Smooth	5.25	.0001*
Spkr-indep	Hidden	4.72	.0001*
	Unsmooth/Smooth	4.85	.0001*

* significant at $p < .02$

A multiple regression analysis of recognition and generalization results for the 3-frame full Motor Theory and Classifier networks as trained by both smoothed and unsmoothed data is shown in Table 5.25.

Table 5.25: All 3-frame Data Networks - Multiple Regression Analysis

Measure	X	t-value	Probability
Training	Hidden	2.90	.0081*
	Unsmooth/Smooth	1.197	.2467
Multi-spkr	Hidden	3.88	.0007*
	Unsmooth/Smooth	1.09	.2876
Spkr-indep	Hidden	5.35	.0001*
	Unsmooth/Smooth	2.60	.0157*

* significant at $p < .02$

It is clear from this analysis that in the case of 3-frame data there is no advantage to the network by training with either smoothed or unsmoothed data except for a slight effect in the case of speaker-independent generalization. This was depicted in Figure 5.26 where the unsmoothed data resulted in poorer generalization and in Figure 5.27 where the generalization is approximately equal for both data types. The results of a multiple regression analysis of all of the networks trained with both smoothed and unsmoothed data can be seen in Table 5.26. The analysis shows once again the previously discussed effect of the hidden units on training recognition and generalization. The significant effect of the input representation for network training performance and generalization can be clearly seen. What is of interest here is the fact that the difference between the smoothed and unsmoothed input data representations is only statistically significant for training recognition. It is not significant for either multi-speaker or speaker-independent generalization. One cannot help but wonder if similar results would be exhibited by an analysis of other research purporting to exhibit the beneficial effect of training networks with

Table 5.26: All Data Networks - Multiple Regression Analysis

Measure	X	t-value	Probability
Training	Input	6.77	.0001*
	Hidden	6.32	.0001*
	Output	0.51	.6085
	Unsmooth/Smooth	4.63	.0001*
Multi-spkr	Input	13.43	.0001*
	Hidden	6.42	.0001*
	Output	1.62	.1085
	Unsmooth/Smooth	2.03	.0454
Spkr-indep	Input	13.11	.0001*
	Hidden	7.29	.0001*
	Output	0.96	.3416
	Unsmooth/Smooth	1.58	.1175

* significant at $p < .02$

so-called noisy data.

One explanation for the contradictory results exhibited in the 1-frame and 3-frame unsmoothed network performance was offered in Section 5.9.1. That explanation was based upon task and the constraints that tasks can place upon generalization exhibited by a network. An alternative explanation can be based upon the concept of an appropriate representation. On this view, a 3-frame dynamic representation is considered to be a more appropriate representation for vowel perception than is a 1-frame representation. For smoothed data, the 3-frame representation results in better training recognition and better generalization than does the 1-frame representation. Similar results have been exhibited in the case of the 1-frame to 3-frame unsmoothed data representation comparison. Comparing the smoothed to the unsmoothed data representation, however, the 1-frame unsmoothed representation results in better performance than does the 1-frame smoothed representation while the 3-frame unsmoothed representation results in somewhat poorer performance than the 3-frame smoothed representation in the case of the Full Motor Theory network and, at best, equivalent performance in the case of the Classifier network.

Sietsma and Dow [33] indicated that the addition of pattern noise to training patterns resulted in better performance in networks trained with such patterns. In the smoothed data representations, the noise present in the original input has been blurred. In the unsmoothed data representations this is not the case. In contrast to being trained with smoothed representations, networks trained with 1-frame unsmoothed data may be responding in a fashion similar to networks which have been trained with patterns having added pattern noise. If the 1-frame representation is an inappropriate representation then the retention of this unblurred noise in the data may have an effect similar to the addition of pattern noise to a training pattern. This does not imply that the 1-frame unsmoothed data is a more appropriate representation than the 1-frame smoothed data representation. Rather, it would seem to have implications with respect to how networks are currently trained. After all, why should training a network with the addition of pattern noise help the network to learn better? This is certainly a question worthy of future research. Until it is answered, it is inappropriate to conclude that the 1-frame unsmoothed data representation is a more appropriate representation than the 1-frame smoothed data representation. In light of existing theories of vowel perception and our knowledge of how the the human auditory system seeks similarities rather than dissimilarities in data an unsmoothed representation seems inappropriate.

For the 3-frame representation, the retention of the original signal noise in the unsmoothed representation may, in fact, serve to make it more difficult for the network to discern transition information. Again, in light of our knowledge of how the human auditory system seeks similarities rather than dissimilarities in data, this would seem to be true. From the experimental results, it is clear that the 3-frame unsmoothed representation hinders the network in learning to generalize. In the Full Motor Theory network, where additional constraints are placed upon the network by the multiple tasks, the combination of additional constraints plus the retained noise in the data representation might make it more difficult for the network to learn the required tasks and to generalize well.

5.9.3 Data representation: target representation.

Recall that the reference auditory spectral target patterns to which input patterns were mapped were selected at random from tokens spoken by a randomly selected reference speaker (Section 3.7). Random selection could potentially stress the mapping function the network is required to learn because the speaker might be unrepresentative of the population. This would hinder the ability of the network to learn the training set and to generalize to previously unseen data. I wanted to know how this selection of auditory reference patterns might affect the generalization capabilities of the network. While I could simply train a network to a second reference speaker and compare the results, a more appropriate way to test this effect would be to train a network twice, once using the randomly selected spectral patterns and a second time using target patterns which reflect the spectral properties of all of the speakers in the training set. This was done by computing vowel centroid patterns. These are patterns in which the intensity repesentation at each frequency point is the average across all of the intensity values at that point for all examples of that vowel spoken by all speakers in the training set.

Figure 5.28 compares the mean training performance and generalization for a 1-frame network trained with reference speaker target patterns with the mean training performance and generalization for a 1-frame network trained with centroid target patterns. Training performance is virtually the same in both instances. The multi-speaker generalization of the network trained using centroid patterns (65%) is obviously superior to that of the network trained using randomly selected reference patterns (59%). The speaker independent generalization is also improved (56% versus 52%).

The impact of training to a centroid target disappears in the case of a 3-frame network. Figure 5.29 compares the training performance and generalization for a 3-frame network trained with vowel centroid target patterns with a 3-frame network trained with reference target patterns. Here, both training recognition and generalization results are essentially the same.

Obviously, shifting from a randomly selected reference speaker to a centroid target has an effect on the ability of the network to learn a training set and to generalize in networks trained with 1-frame input data. In the case of networks trained with 3-frame input data the potential advantage appears to have already accrued to the network through the use of the larger input representation. No further advantage is gained by using the centroid representation. It is clear that some, but not all, of the information provided the network by the 3-frame input

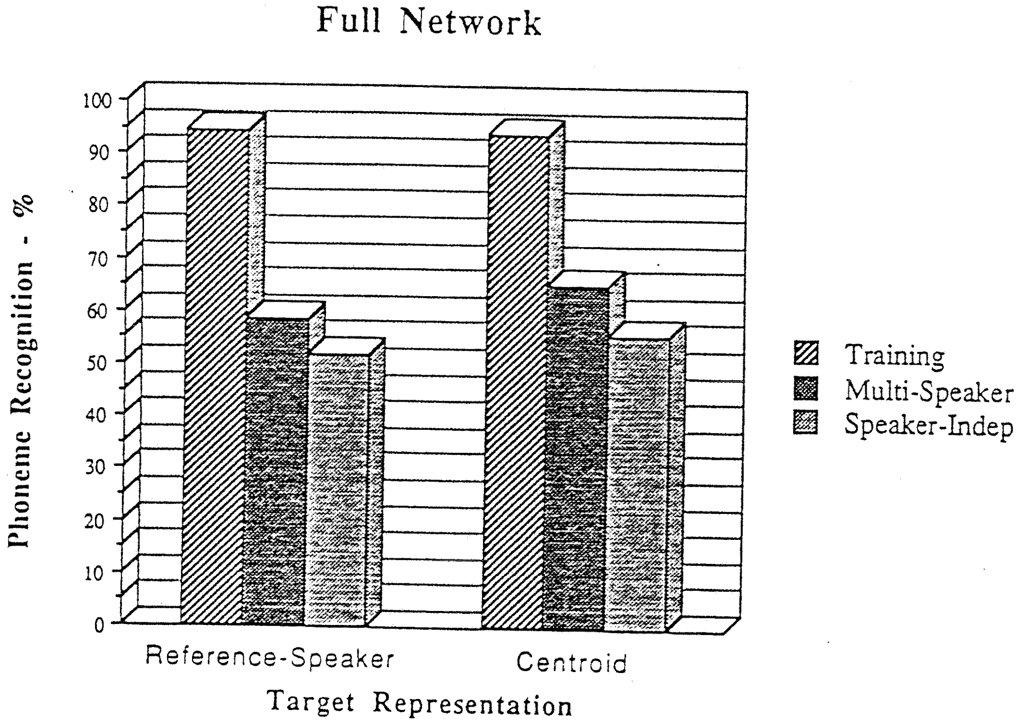


Figure 5.28: Performance in a 1-frame Network using a Centroid Target

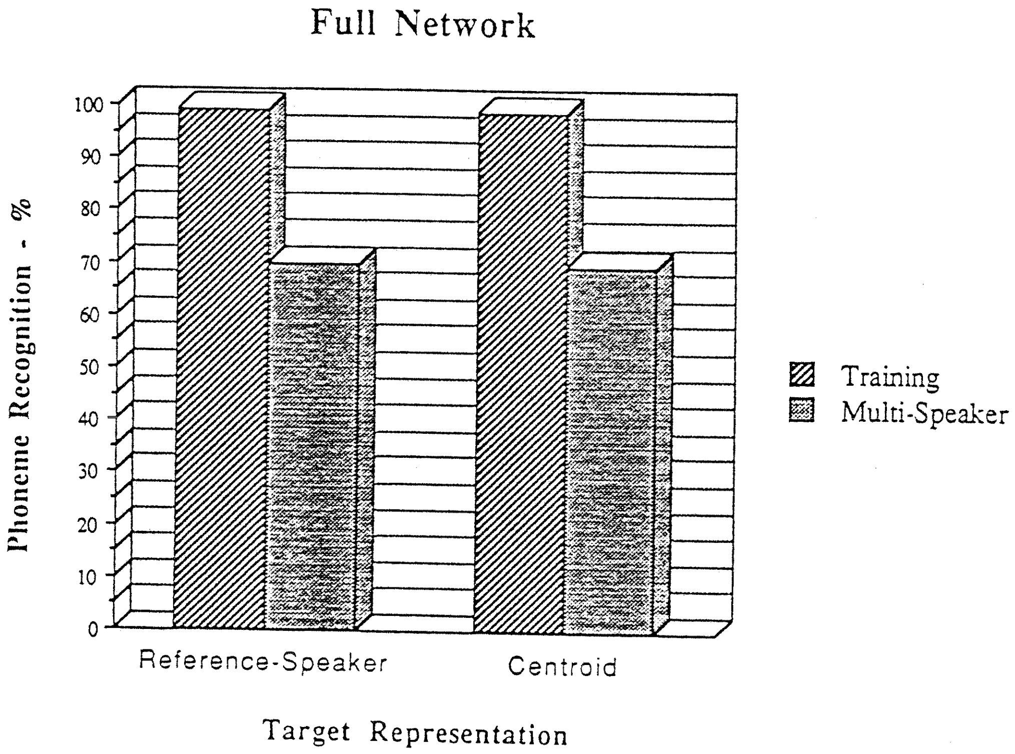


Figure 5.29: Performance in a 3-frame Network using a Centroid Target

information can also be provided to the network by means of an appropriately chosen target representation, in this case the centroid vowel patterns. Choice of target representation can make a difference!

5.9.4 Data representation: conclusions. Comparing the performance between the networks trained with 1-frame and 3-frame smoothed data indicates that the 3-frame data representation, felt to be more appropriate by students of vowel perception theory, is also more appropriate for the training of the four network models used in this research. Using the 3-frame smoothed data representation for training the networks clearly results in networks exhibiting better generalization than does the 1-frame smoothed data representation. In the case of the unsmoothed data representations, similar results are seen for the Classifier networks. The apparently nonconforming behavior evidenced in the case of Full Motor Theory networks trained with the unsmoothed data has already been discussed above. In a network such as the Full Motor Theory network, where there are additional constraints placed upon the network by the added tasks, the combination of additional constraints plus the retained noise in the representation may, in fact, make it more difficult for the network to discern the transitional information and learn the required tasks and, thus, generalize well.

While using a centroid target rather than a randomly selected reference speaker target makes no difference in performance for 3-frame representations, it does result in better performance for the less appropriate 1-frame input representation. The choice of target representation can make a difference

Finally, with the exception of the somewhat unclear evidence surrounding the interaction between training, pattern noise, and network task, in the case of unsmoothed vowel data representations, the hypothesis, which states that training with appropriate, even though larger, representations will help the network to generalize better must be deemed valid.

CHAPTER 6

CONCLUSIONS

My research has been an investigation of this complex question: how do the architecture of the network, the task which the network must learn, the training of the network, and the data representation used in that training, both individually and collectively, affect the ability of a network to learn the training data and to generalize well to previously-unseen data. Utilizing a standard classifier model for speech recognition and three multi-task connectionist models whose tasks are derived from the Motor Theory of speech perception to classify 12 American English vowels, I explored this question in depth.

The experimental results regarding the relationship between the architecture of the network and the generalization exhibited by a network were discussed in Section 5.6. Each of the network models, when trained with a fixed size training set, exhibited comparative levels of performance and generalization across a broad range of hidden units. Networks having fewer hidden units did not generalize as well as networks having more hidden units, even though the increase in number of hidden units implied a concomitant increase in total network capacity. For networks having exactly the same training recognition, better generalization was generally exhibited by those networks which had more hidden units. This was true for networks trained with each of the four representations used in this research (1-frame, 3-frame, both smoothed and unsmoothed). These results concerning network architecture contradict the overfitting guideline, which states if a network can learn a problem, then the fewer the number of free parameters in the network the better the network is likely to generalize. The dataset-size guideline which states the larger the number of free parameters in a network the more data needed to train it also does not apply.

With respect to the relationship between the task required of a network and the generalization exhibited by that network I showed in Section 5.7 that all four network models, having from one to three tasks and trained with the same input representation, proved to be a family of networks whose primary predictor of performance variance was the number of hidden units in the networks. Increasing the number of hidden units resulted in a significant increase in training recognition and generalization. This family relationship held true across all four of the representations used in this investigation. The tasks required of the networks, and the differing input and output architectures and concomitant changes in network capacity resulting from those tasks, did not affect the performance and generalization exhibited by the networks. The hypothesis that adding additional tasks should cause a change in generalization was specifically rejected by these results.

A possible objection can be made to this conclusion in the case of the 1-frame networks trained with vowel centroid data. It should be noted that this is not an instance where additional tasks were required of the network. There is no change in network capacity; rather there is a change in the output representation of a particular task. The improved performance in this case is more appropriately ascribed to a change in the information content of the data representation than to a change in the network task.

The question remains as to why these additional task constraints did not help the network to develop better internal representations and thus to generalize better. One possible answer is that the Motor Theory of speech perception is incorrect. This explanation would certainly serve to explain why the additional task constraints did not help the full Motor Theory model develop better internal representations and thus to generalize better than the other network models. What it does not explain is why all of the network models were able to develop internal representations which resulted in essentially similar training performance and generalization. As a consequence, it is my firm opinion that a rejection of the Motor Theory of speech perception on the basis of this research would be inappropriate.

An alternative explanation might be that the Motor Theory model was implemented incorrectly. Some support for this explanation can be gained by noting that both the auditory association task and the articulatory association task can be described as classification tasks. In fact, classification measures were developed and used for these tasks as well as for the specific phonemic classification task included in the model. The phonemic classification task mapped a real-value distributed input representation to a binary localist phonemic output. The auditory association task mapped this same input to a similar output. The articulatory association task mapped this same input to a binary, distributed articulatory output. All the network knows is what it can derive from the presented information. The output is trained with twelve sets of information in which the same phonemic feature occurs with the same auditory and articulatory representation. I sought to develop better internal representations by forcing a network to identify commonalities from the auditory signals. This was done by mapping all instances of a particular vowel to a reference instance for that vowel. It may have been better to auto-associate the auditory input. Such an approach would have mapped the 700 training inputs to 700 training outputs in which there were commonalities across the phonemic and articulatory subtasks but differences in the auditory subtask. I will take up this question regarding task constraints and generalization once again, after I present my conclusions regarding the relationship between data representation and network generalization, below.

The experimental results regarding the relationship between the training of the network and the generalization exhibited by a network were discussed in Section 5.8. I described the impact of learning rate in back-propagation training. I also compared back-propagation training with conjugate gradient training. Reducing the number of training examples for a less appropriate 1-frame training set does not have a significant effect on the performance and generalization exhibited by the networks. On the other hand, reducing the number of training examples for a more appropriate 3-frame training set significant reduces the performance and generalization exhibited by the networks. I also showed a divergence in performance measures: even though the training error rate reaches a minimum and remains relatively stable does not mean that the network has reached the maximum level of generalization. Improvement in network generalization can continue as the network reorganizes the mean square error across the output units. Most significantly, I verified the necessity for tracking generalization directly rather than observing training error rate or training recognition.

In Section 5.9 I showed that training networks with the 3-frame smoothed representation resulted in better performance and generalization than exhibited by either the 1-frame smoothed data representation or the unsmoothed data representations. The use of speech noise and of a centroid target representation to enhance the performance and generalization exhibited by the less appropriate 1-frame data representation was also discussed. From this I conclude that choice of target pattern representation can make a difference in the generalization exhibited by a network.

The answer to why additional task constraints did not help the networks to generalize better can be found in the area of data representation. The Motor Theory model of speech perception proposes that speech perception is mediated by the neuromotor commands of articulation. As was pointed out in Section 3.6.2 it is extremely difficult to determine the articulatory parameters used in the production of a particular instance of a given speech sound. Accordingly, the approach used to represent those parameters in the current research consisted of specifying a unique abstract representation for each vowel as it might have been spoken. Perhaps a better approach, at least for purposes of multi-speaker training and generalization testing, would have been to specify a particular representation for each vowel as it was spoken in all contexts by each speaker. It is not clear what implications such an approach might have had for speaker-independent generalization since there is no way that the networks could have guessed the unique abstract articulatory representations that would have been assigned to speakers that it had not previously encountered. An alternative might have been to provide a set of articulatory features in which some aspects unique to each speaker and some aspects which reflected commonalities appropriate to each spoken vowel were combined. Such an approach would have provided both the full Motor Theory network model and the Mimic network model with additional information which might have resulted in improved performance and generalization exhibited by each of these network models.

In research, as well as in life, hindsight is often better than forsight. At the beginning of this project all previous experience and research seemed to indicate that the most significant effect on network generalization would be that provided by constraints resulting from the specification of the network tasks. At the end of this project, after having explored the relationship between network task and generalization using networks trained with four different data representations, it is clear that this is simply not the case. Decisions regarding data representations obtain an even greater significance than they were originally given. In preparing the data representations, considerable emphasis was placed upon developing as good an auditory representation as possible. As was discussed, the impossibility of obtaining a set of accurate articulatory features led to a decision to represent them with a set of unique abstract features. At the conclusion of this project, however, it is my considered opinion that the articulatory data representations used in training the full Motor Theory network provided insufficient information to the network. The network was unable to take advantage of the added articulatory task constraint. With the use of the auto-associative auditory task and the further enhanced articulatory data representations perhaps the results of future research will be different; networks trained with such enhanced representations might exhibit clear differences based upon the tasks required of the networks.

It is clear from this investigation of the relationship between network architecture, network task, network training, training data representation and the generalization exhibited by a network that given an appropriate architecture, training algorithm, and sufficient training data, the data representation itself is the primary determiner of a network's ability to generalize well to new data.

BIBLIOGRAPHY

[1] Subutai Ahmad and Gerald Tesauro. Scaling and generalization in neural networks: A case study. In **Advances in Neural Information Processing Systems**, pages 160–168, 1988.

[2] Eric B. Baum and David Haussler. What size net gives valid generalization. **Neural Computation**, 1(1):151–160, 1989.

[3] Sheila E. Blumstein and Kenneth N. Stevens. Phonetic features and acoustic invariance in speech. **Cognition**, (10):25–32, 1981.

[4] Gary L. Bradshaw. **Learning to understand speech sounds: A theory and model.** PhD thesis, Carnegie-Mellon University, 1984.

[5] John F. Brugge and Richard A. Reale. Auditory cortex. In Alan Peters and Edward G. Jones, editors, **Cerebral Cortex**, pages 229–271. Plenum Press, New York, 1984.

[6] Raymond K. M. Cheung, Irving Lustig, and Alain L. Kornhauser. Relative effectiveness of training set patterns for back propagation. In **International Joint Conference on Neural Networks**, volume 1, pages 673–678, 1990.

[7] Randy L. Diehl. Feature detectors for speech: A critical reappraisal. **Psychological Review**, 89(1):1–18, 1981.

[8] Peter D. Eimas and John D. Corbit. Selective adaptation of linguistic feature detectors. **Cognitive Psychology**, 4:90–109, 1973.

[9] Moise H. Goldstein and Moshe Abeles. Single unit activity of the auditory cortex. In Wolf Dieter Keidel and William D. Neff, editors, **Handbook of Sensory Physiology - Auditory System**, volume 5, pages 199–218. Springer-Verlag, 1975.

[10] R. Paul Gorman and Terrence Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. **Neural Networks**, 1(1):75–89, 1988.

[11] Steven Greenberg. Acoustic transduction in the auditory periphery. **Journal of Phonetics**, 16:3–17, 1988.

[12] Geoffrey E. Hinton. Connectionist learning procedures. **Artificial Intelligence**, 1988.

[13] J. Stephen Judd. **Neural Network Design and the Complexity of Learning**. The MIT Press, Cambridge, Massachusetts, 1990.

[14] Candace A. Kamm, Lynn A. Streeter, Yana Kane-Esrig, and David J. Burr. Comparing performance of spectral distance measures and neural network methods for speech recognition. 1989.

[15] Peter Ladefoged. **A Course in Phonetics**. Harcourt, Brace, Jovanovich, second edition, 1982.

[16] T. K. Landauer, C. A. Kamm, and S. Singhal. Teaching a minimally structured back-propogation network to recognize speech sounds. In **Proceedings of the Ninth Annual Conference of the Cognitive Science Society**, pages 531–536, 1987.

[17] Hong. C. Leung and Victor W. Zue. Some phonetic recognition experiments using artificial neural nets. In **International Conference on Acoustics, Speech, and Signal Processing**, pages 422–425, 1988.

[18] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. **Psychological Review**, 74(6):431–461, 1967.

[19] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. **Psychological Review**, 74:431–461, 1967.

[20] Alvin M. Liberman and Ignatius G. Mattingly. The motor theory of speech perception revised. **Cognition**, 21:1–36, 1985.

[21] Alvin M. Liberman and Ignatius G. Mattingly. A specialization for speech perception. **Science**, 243:489–494, 1989.

[22] Gale L. Martin and James A. Pittman. Recognizing hand-printed letters and digits. In **Advances in Neural Information Processing Systems 2**, pages 405–414, 1990.

[23] James L. McClelland and David E. Rumelhart, editors. **Parallel Distributed Processing Explorations in the Microstructure of Cognition: Foundations**, volume 1. The MIT Press, Cambridge, Massachusetts, 1986.

[24] Brian C. J. Moore. **An Introduction to the Psychology of Hearing**. Academic Press, New York, second edition, 1982.

[25] Brian C. J. Moore. **Frequency Selectivity in Hearing**. Academic Press, New York, 1986.

[26] N. Morgan and H. Bourlard. Generalization and parameter estimation in feed-forward nets: Some experiments. In **Advances in Neural Information Processing Systems**, 1989.

[27] Michael C. Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In **Advances in Neural Information Processing Systems**, pages 107–115, 1989.

[28] Yeshwant K. Muthusamy and Ronald A. Cole. Speaker-independent vowel recognition: Spectograms versus cochleograms. In **International Conference on Acoustics, Speech, and Signal Processing**, 1990.

[29] Gregg C. Oden and Dominic W. Massaro. Integration of featural information in speech perception. **Psychological Review**, 85(3):172–191, 1978.

[30] R. Plomp. The ear as a frequency analyzer. **The Journal of the Acoustical Society of America**, 36(9):1628–1636, September 1964.

[31] Elizabeth Lake Richards and Gary Lee Bradshaw. Optical filtering of speech waveforms. In **Proceedings of the Rocky Mountain Conference on Artificial Intelligence**, page forthcoming, 1989.

[32] Nelson Yuan sheng Kiang and William T. Peake. Physics and physiology of hearing. In S. S. Stevens, editor, **New Handbook of Psychology**, volume 1, pages 277–326. Wiley, 1989.

[33] Jocelyn Sietsma and Robert J. F. Dow. Creating artificial neural networks that generalize. **Neural Networks**, 4(1):67–79, 1991.

[34] Kenneth N. Stevens and Sheila E. Blumstein. The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. L. Miller, editors, **Perspectives on the Study of Speech**, pages 1–38. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1981.

[35] Kenneth N. Stevens and Morris Halle. Remarks on analysis by synthesis and distinctive features. In W. Wather-Dunn, editor, **Models for the Perception of Speech and Visual Form**, pages 88–102. 1967.

[36] Winifred Strange. Evolving theories of vowel perception. **The Journal of the Acoustical Society of America**, 85(5):2081–2087, 1989.

[37] Gary Tajchman and Marcia Bush. Contextual effects in the perception of naturally produced vowels. Talk at 120th ASA Meeting, San Diego, 1990.

[38] Georg von Bekesy and Walter A. Rosenblith. The mechanical properties of the ear. In S. S. Stevens, editor, **Handbook of Experimental Psychology**, pages 1075–1115. Wiley, New York, 1951.

[39] Richard M. Warren. **Auditory Perception**. Pergamon Press, New York, 1982.

[40] A. S. Weigend, B. A. Huberman, and D. E. Rumelhart. Predicting the future: A connectionist approach. In **Advances in Neural Information Processing Systems**, 1990.

[41] I. C. Whitfield. The role of auditory cortex in behavior. In Alan Peters and Edward G. Jones, editors, **Cerebral Cortex**, pages 329–349. Plenum Press, New York, 1984.

[42] Bernard Widrow. Adaline and madaline. In **First International Conference on Neural Networks**, pages 143–158, 1987.

[43] E. Zwicker. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). **The Journal of the Acoustical Society of America**, 33(2):248, February 1961.

APPENDIX A

OTHER MODELS OF SPEECH PERCEPTION

A.1 Perception-only models

Perception-only models assume that ordinary auditory processes are sufficient to explain the perception of speech sounds. In feature detector theory, the first of these models, the search for auditory feature detectors was motivated by the existence of specialized visual feature detectors [8]; the existence of analogous specialized auditory feature detectors was assumed. It is further assumed that the feature detection process is neither learned nor modifiable; rather, it is innate. The feature detection process utilizes specialized auditory receptors to analyze the acoustic signal into a set of independent acoustic features. These acoustic features are then passed to a higher level processor which compares and evaluates them with respect to known features that define perceptual units in long-term memory. The results of this process, the input acoustic features and potential matching prototypes, are then passed on to a higher level processor for further evaluation [29].

Feature detector theory finesses the noninvariance problem by assigning its resolution to higher level processors[8]. Feature detector theory explains categorical perception by permitting detector outputs to vary according to differences in within-category stimuli and assigning higher order processing elements the role of quantizing, and thus categorizing, those feature detector outputs [7, 8]. Feature detector theory does not attempt to explain how the higher level processes obtain their knowledge of known perceptual features nor how these perceptual features are then related to higher level symbolic concepts. From a systems viewpoint, feature detector theory provides only a partial explanation of the speech perception process.

Blumstein and Stevens propose a perception-only model in which an invariant acoustic property can be found for each feature. The properties correspond to phonetic dimensions of the speech sounds used in language. On their view, the inventory of possible speech sounds is constrained by both the articulatory mechanism and the perceptual mechanism [34]. The implication is that these invariant acoustic properties can be derived directly from the input signal by means of unique property detectors with which the auditory system is equipped [3].

In contrast to feature detector theory, Blumstein and Stevens' approach has the advantage that it acknowledges the existence of constraints imposed by the physical articulatory and auditory mechanisms involved in the perception and production of speech. Their model does not utilize these constraints in the perception process nor does it provide an explanation of the process by which knowledge of acoustic features is acquired or the process by which they are related to conscious language percepts.

Property integration theory [4] rejects the hierarchical processing approach used in feature detector theory. Instead it posits a central decision unit which utilizes information about selected acoustic events, passed to it via the peripheral property detectors, to make a decision regarding the identification of the input signal. Property integration theory is based upon a learning theory in which language is acquired through experience with words.

Property integration theory has a distinct advantage over feature detector theory in that it provides an explanation as to how knowledge of the pertinent properties is acquired and utilized in the speech perception process. Unlike Blumstein and Stevens, however, it does not seek to incorporate constraints on speech sounds imposed by the physical articulatory and auditory mechanisms into the decision process.

I now turn to an examination of one of the models in which the perception of speech sounds is assumed to be fundamentally and inextricably tied together with the production of speech.

A.2 Perception-production models

In analysis by synthesis, the production and perception of speech are conceived of as a single system. Here the perception of speech involves an internal synthesis of patterns generated by the same phonological rules used by the production system. The input pattern is iteratively matched against the synthesized patterns until an appropriate match is found [35]. On this view, the noninvariance problem is accomodated by means of an iterative matching process.

Analysis by synthesis is obviously a computationally-intensive model of speech perception. The time required to find and synthesize an appropriate match for a speech utterance should be virtually immediate if the analysis-by-synthesis model is to correspond to human speech perception capabilities (approximately 50 phonemes/second [4]). We know, however, that the human brain is only capable of sequentially executing approximately 100 instructions/second [23]. Assuming that more than two instructions would be involved in iteratively synthesizing and matching patterns, the human brain does not appear to be fast enough to meet the computational demands of the analysis-by-synthesis model.

The other main perception-production model of speech perception is Liberman and Mattingly's Motor Theory of perception. This model is discussed in detail in Chapter 2.

APPENDIX B

HUMAN AUDITORY SYSTEM

An extensive body of literature is available describing the neurophysiology and psychoacoustic characteristics of the human auditory system, [24, 32, 38, 39]. In the discussion that follows I will present a brief review of the auditory periphery and cortex and what is known about its functionality. I then turn to the more specific problem of how human speech sounds are represented in the neural output of the auditory periphery. The auditory nerve is the main pathway by which speech input reaches the higher processing centers of the brain.

Sound is funneled into the ear via the pinna and the external auditory meatus or ear canal (Figure B.1). The sound wave strikes the tympanic membrane and its movement in turn causes the three small ossicles of the middle ear, the malleus, incus, and stapes to move. The stapes is attached to the oval window of the cochlea and movement of the stapes causes displacement of the fluid within the upper chamber, the scala vestibuli. There is an opening, the helicotrema, at the apical end of the cochlear spiral between the scala vestibuli and the lower chamber, the scala tympani, which serves to equalize the fluid displacement between the two chambers. Since the net volume of fluid within the cochlea must remain constant, there is an equal amount of fluid displaced at the round window, the membrane opening between the scala tympani and the middle ear. The two canals are otherwise separated by two membranes, Reissner's membrane and the basilar membrane, which together form the cochlear duct. This duct is closed at the helicotrema. Lying on the basilar membrane is the organ of Corti within which the auditory receptors, the hair cells, are found.

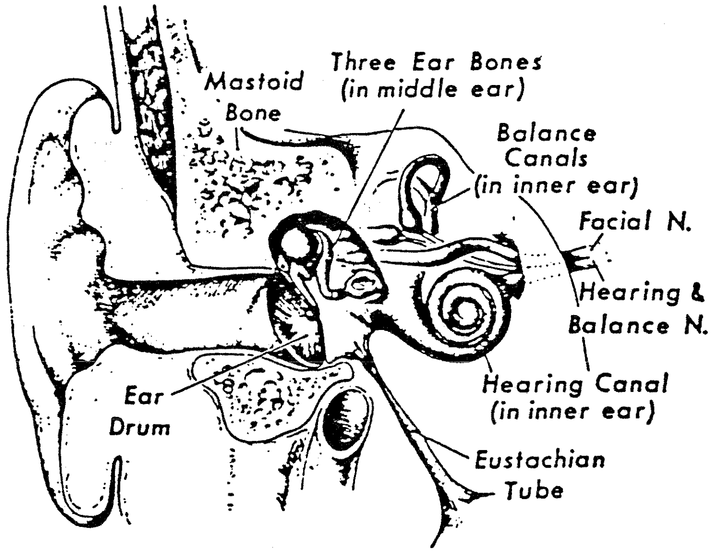


Figure B.1: The Human Ear

There are three rows of outer hair cells which lie close to the cochlear wall and one row of inner hair cells. The wave which results from the displacement of the fluid within the cochlea moves along the cochlear spiral and the basilar membrane moves in response to that wave. The tips of the stereocilia projecting from the hair cells on the basilar membrane interact with the tectorial membrane, which lies directly above them and between the basilar membrane and Reissner's membrane. Present understanding of the reception process indicates that the shearing deflection of the stereocilia by the tectorial membrane causes electrochemical changes in the receptor cells and resulting stimulation of the associated auditory nerve fibers. The tapered shape of the basilar membrane together with its variant stiffness causes it to respond differentially and in a nonlinear fashion to the fluid displacement for different frequencies of sound, with maximal sensitivity to high frequencies at the base and maximal sensitivity to low frequencies at the apex [25, 30]. The response functions of the basilar membrane also vary with sound pressure level with broader responses being seen at high-amplitude levels [11]. For the basilar membrane nonlinearity occurs at sound pressure levels in the normal speech range.

From the above discussion, it can be seen that the sound transduction process is both mechanical and electrical in nature. What is of interest to my research from the above discussion are the following observations: the auditory parameters operate within a system whose task is the perception of sound; this auditory unit, which is physically constrained (e.g. by fluid displacement requirements and by basilar membrane shape and variant stiffness) exhibits tonotopic organization of the basilar membrane and a nonlinear response to different sound frequencies and intensities.

While little is understood about how auditory processing of speech sounds is accomplished at the higher auditory levels there is some knowledge which is pertinent to my research. For more than three decades research on the auditory processing capabilities of a listener at the single neuron level has been carried out. Results indicate that the response patterns of neurons in the AI (Auditory cortical field I) region are frequency and intensity sensitive with the characteristic frequency of the neuron related to the place of resonance along the cochlear partition to which it is ultimately connected [5, 9]. There appear to be quantitative differences in the selectivities for human speech sounds in different auditory cortical areas and the cortical auditory system is highly segregated cochleotopically [5]. It is suggested that the function of sensory cortex is the detection of similarities among stimuli, not the detection of differences [41].