Spring 1992

Formant Estimation from DCTC's Using a Feedforward Neural Network

Shubhangi U. Kelkar

Old Dominion University

FORMANT ESTIMATION FROM DCTC's USING A FEEDFORWARD NEURAL NETWORK

by

Shubhangi U. Kelkar
B.E.(Instrumentation), June 1986,
College of Engineering, Pune, INDIA.

A Thesis submitted to the Faculty of
Old Dominion University in Partial Fulfillment of
the Requirement for the Degree of

MASTER OF SCIENCE
in
ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
May, 1992

Approved by:

Dr. David Livingston (Director)

Dr. Stephen Zahorian (Director)

Dr. Jack Stoughton

FORMANT ESTIMATION FROM DCTC's USING A FEEDFORWARD NEURAL NETWORK

by

Shubhangi U. Kelkar
B.E.(Instrumentation), June 1986,
College of Engineering, Pune, INDIA.

A Thesis submitted to the Faculty of
Old Dominion University in Partial Fulfillment of
the Requirement for the Degree of

MASTER OF SCIENCE
in
ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
May, 1992

ABSTRACT

FORMANT ESTIMATION FROM DCTC's USING A FEEDFORWARD NEURAL NETWORK

Shubhangi U. Kelkar
Old Dominion University, 1992
Directors: Dr. David Livingston
Dr. Stephen Zahorian

Formants are the natural frequencies of the human vocal tract. Existing methods for estimating formants from speech signals are computationally complex and subject to errors for certain type of speech sounds. This thesis describes a method for estimating vowel formant frequencies from Discrete Cosine Transform Coefficients (DCTC's), a form of cepstral coefficients, using a feedforward neural network with back-propagation training. Experimental results are based on a large multispeaker data base. The results are obtained for both a linear transformation and a feedforward neural network with a nonlinear hidden layer. In general, the neural network transformation is superior to the linear transformation for formant estimation. Thus our experiments indicate the nonlinear nature of the relationship between DCTC's and formants. However, since the results are always much better for training data as compared to test data, a large data base is necessary for adequate neural network training. Vowel classification experiments show that estimated formants can discriminate vowels nearly as well as 14 DCTC's.

ACKNOWLEDGEMENTS

I wish to express a special thanks and sincere gratitude to Dr. Livingston, for his patience and helpful advise throughout the research work. I am sincerely grateful to Dr. Zahorian for his invaluable guidance and endless assistance. I would also like to thank Dr. Stoughton for serving on my thesis committee.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 FORMANTS

Man's primary means of communication is speech. His ability to convey information with his voice is unique among all the animals. The operation of the vocal system as a whole can be divided into two functions, excitation and modulation. Excitation occurs primarily at the glottis and modulation is caused by various organs of the vocal tract. As the acoustical wave is transmitted through the vocal tract, its frequency content is modified by the cavity resonances in the tract. These natural resonances are called **formants** and are a function of the shapes of the various regions of the vocal tract. In continuous speech, the formant resonances vary in time as the vocal tract changes shape. Formants are identified by number in order of increasing frequency: F1, F2, F3, etc. Each vowel has a different pattern of resonances than the others. Further, not all speakers have the same formant frequencies for the same vowels. The formant frequencies are higher for smaller vocal tracts. The ratio of different cavity sizes to one another also determines the formants. In particular, the voices of women have higher frequency formants than for males, and those of small children still higher frequencies.

1.2 IMPORTANCE OF FORMANT FREQUENCIES

Formants are of interest for a number of reasons. They are the most important source of articulatory information. They have been used as primary features in speech recognition, speech synthesis and also as the major information to be transmitted in encoded speech. The first three formants contain sufficient information to classify the vowels. If a coordinate system is set up using F1 and F2 as a basis, vowels cluster in specific regions [18]. Formants are efficient and important acoustic cues because they constrain global shape to a large degree. Thus, a dependable and automatic algorithm for computing these frequencies would be useful for many aspects of speech research.

The vocal tract dimensions, which determine the formant frequencies, are rather difficult to obtain. Since the information about formants is contained in the spectrum envelope, all formant estimators either implicitly or explicitly make use of the spectrum envelope. The problem proves to be challenging for a number of reasons:

1. There is a high degree of overlap between frequency regions in which formants may be located. Thus formant estimation is not as simple as locating the peaks of the spectrum in non-overlapping frequency bands.

2. Normally, the maxima in the spectral envelope are due only to formants. However, when a single formant splits, spurious peaks occur which are difficult to distinguish from the formant peaks.

3. Sometimes, the frequencies of adjacent formants are too close to resolve. Here, the practical difficulties are in recognizing a blend.

4. The speech signal is affected both by the properties of the source and by the vocal tract. If the source spectrum has a zero close to one of the formant frequencies, it is extremely difficult to determine the frequency of that particular formant. A side branch element such as the nasal cavity creates a similar problem.

1.3 FORMANT ESTIMATION VIA NEURAL NETWORKS

Formant estimation from the speech signal has been of interest for many years. Several methods have been developed over the years and are successful in varying degrees [13][16][23]. The present study describes a method for estimating vowel formant frequencies from Discrete Cosine Transform Coefficients, a form of cepstral coefficients, using a feedforward neural network with back-propagation training. Two approaches have been studied and compared: a one layer linear transformation and a feedforward network with nonlinear hidden layers. The objective of this study was to determine if formants can be reliably estimated from DCTC's with a neural network, and to compare the performance of the neural network for this task to a linear transformation.

1.4 THESIS STRUCTURE

Chapter two reviews some of the major previous efforts in the area of formant estimation. The different existing approaches for formant estimation are described which include analysis-by-synthesis, peak-picking and root-solving. An alternative approach, formant estimation from linear transformation of the LPC cepstrum, is also presented. The basic back-propagation algorithm is described and DCTC's are explained.

The basis for the current research is discussed in chapter three. The learning and generalizing abilities of neural networks are considered when trained with back-propagation. The description of the database, experiments and results achieved for various test runs is contained in chapter four. A summary of the current work in the context of previous literature and an examination of possible future research is presented in chapter five.

CHAPTER II

BACKGROUND

2.1 PREVIOUS RESEARCH ON FORMANT ESTIMATION

A number of previous efforts have been directed towards formant estimation and are briefly reviewed here. There are several approaches including: analysis-by-synthesis, peak-picking from smoothed spectra, and computation of pole frequencies through root-solving in an all pole model of speech spectra.

1. ANALYSIS-BY-SYNTHESIS: In this method [1], a spectrum generated from estimated formants is compared with the actual spectrum and the formants in the estimated spectrum are varied until the difference becomes minimal. Since the analysis-by-synthesis method takes into account the entire spectral shape rather than simply relying upon the spectral peaks, one spurious peak would not change the results drastically. Its disadvantages include the need for complex processing and its dependency upon the accuracy of the speech model.

Bell et al. [1] are some of the early researchers who estimated formants using analysis-by-synthesis. The vocal tract transfer function over the ranges of first three formants was duplicated using a model consisting of a set of three tunable filters. The frequencies and bandwidths of these filters were varied until they matched the observed spectral envelope. They matched the model to filter bank outputs. Olive[16] also used the analysis-by-synthesis approach to estimate formants. He devised a Newton-Raphson technique to find a least-squares fit and matched all three filters simultaneously instead of sequentially.

2. PEAK-PICKING: This approach makes use of the fact that the maxima in the spectral envelope are primarily due to formants. Hence the appropriate peaks from a smoothed spectrum are selected to be the formants. The skill of the algorithm lies in recognizing spurious peaks and/or formant blends.

McCandless [13] used the peak-picking technique to estimate formants from linear prediction spectra. The algorithm first fills formant slots with the available peaks in a frame, based on frequency position relative to an educated guess. However, in the case of formant merging, there is a deficiency of peaks and slots remain unfilled. Special routines are then called to recompute the enhanced spectrum to deal with these cases.

3. ROOT-SOLVING: For this approach, a transfer function model of the is first used to model the vocal tract. Typically this model is a linear prediction (LP) all-pole model. Polynomial root solving techniques are used to find the roots of the LP denominator polynomial, that is, the poles of the transfer function. These poles include the formant frequencies plus additional extraneous poles. Additional processing is required to identify the formants and eliminate the extraneous poles. Formant estimation using these approaches have been developed by various researchers.

Talkin [23] introduced a dynamic programming approach for formant tracking. First root-solving of the LP polynomial is used to identify up to 5 formant candidates per frame. Dynamic programming then finds the lowest cost path through a lattice of formant candidates over a segment of speech. The cost of each path is the sum of the "local" costs and "transition" costs. The deviation of formant candidates from expected formant values and the candidate bandwidths determine the local cost. The transition costs are proportional to the deviation of formant values from frame to frame, since formants generally vary gradually over time. The formants are also constrained such that F3 > F2 > F1. The path obtained by dynamic programming is then the best set of formants, according to the criteria as described. This technique has recently been refined by Zahorian and Jagharghi [26].

All of the above-mentioned approaches have their individual drawbacks. They all require complex processing. They work well for most speech segments but suffer from occasional large errors. Even the dynamic programming approach, which appears to be the most robust formant tracking algorithm currently available, is not completely error free and requires very time-consuming calculations.

Another approach for estimating formants was suggested by Pols et al. [20] and was implemented by Broad and Clermont [2]. They proposed that a linear relationship exists between the cepstral coefficients and formants. Regression analysis was used to compute this linear relationship. The method is not highly accurate, but is robust in the sense that large errors are rare. It gives better results if tuned to individual speakers, but its performance degrades considerably for more than one speaker.

This thesis describes a method for estimating vowel formants from Discrete Cosine Transform Coefficients (DCTC's), a form of cepstral coefficients. We propose that the relation between the DCTC's and formants is nonlinear and use a feedforward network trained with the back-propagation algorithm to compute the nonlinear relationship. It will be shown that the neural net provides a more accurate transformation of DCTC's to formants than does a linear transformation, for both training and test data.

2.2 BACK-PROPAGATION ALGORITHM

The back-propagation algorithm [22] [17] [25] is one of the most popular and widely used techniques for training neural networks. The algorithm is based on gradient descent where the weights in a feedforward network are changed to learn a training set of input-output pairs. Figure 2.1 shows a feedforward network with one hidden layer.

The governing equations for a back-propagation algorithm are briefly reviewed here. A feedforward network consists of an input layer, a number of hidden layers (usually one or two) and an output layer. The neurons in the input layer simply store the input values. For the neurons in the hidden and output

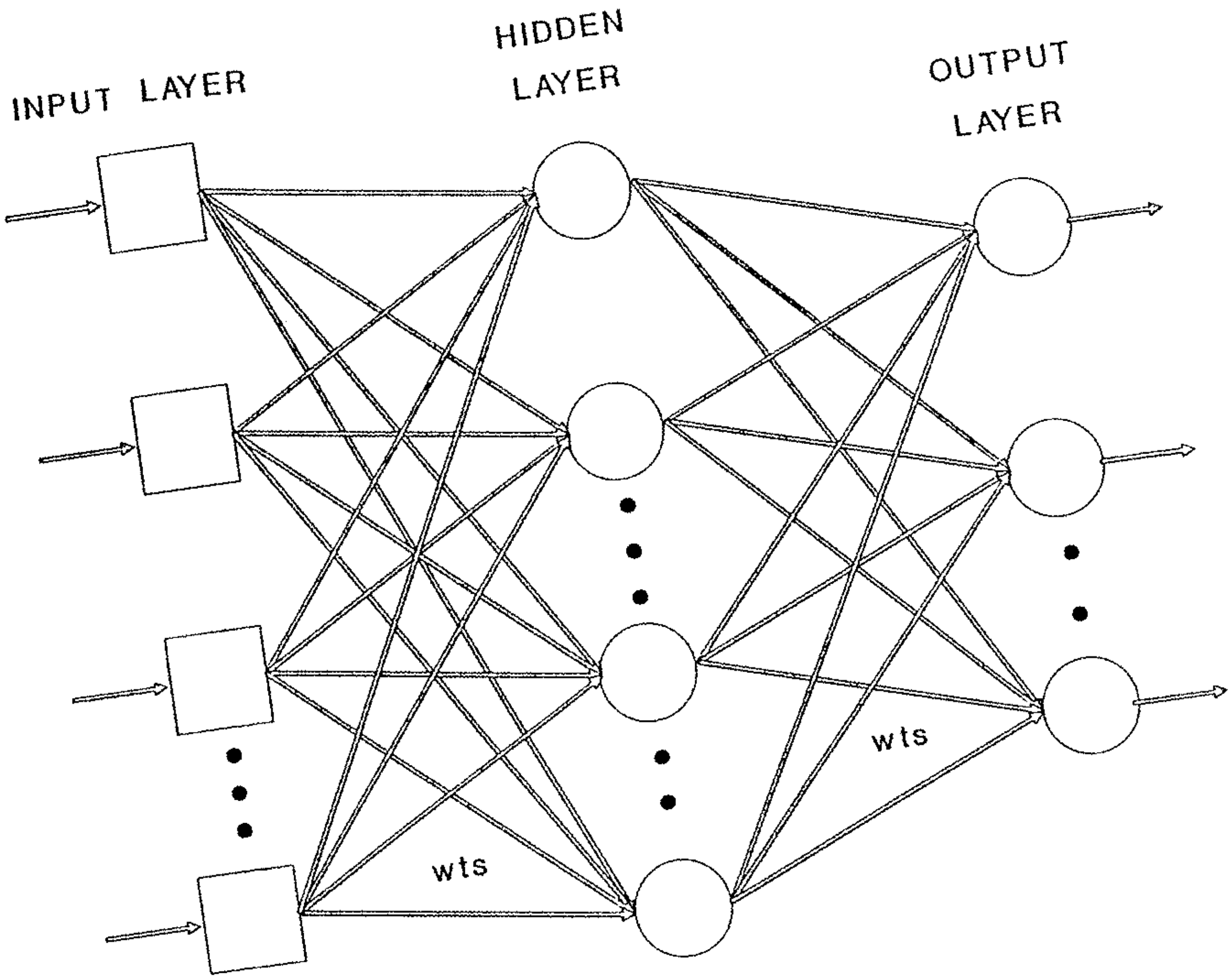


Figure 2.1. Feedforward network with one hidden layer.

layers, two calculations are carried out. First, the net input for each neuron is calculated as

$$net_j = \sum W_{ij}\, a_i \tag{2.1}$$

where the w_{ij}'s are the weights connecting to the previous layer.

Second, the activation of the neuron, a_j, is calculated as a nonlinear function f of net_j as

$$a_j = f\,(net_j) \tag{2.2}$$

The network learns by making changes in its weights in a direction to minimize the sum of squared errors between its output signals and a training data set. The minimization is done using gradient descent.

$$W_{ij}(t+1) = W_{ij}(t) + \eta\ \delta_j\ a_i \tag{2.3}$$

where

$$\delta_j = f'\,(net_j)\ .\ (T_j - O_j)$$

for output units, and

$$\delta_j = f'\,(net_j) \sum \delta_k\ W_{jk}$$

for hidden units. The term η is the learning rate and is used to control how fast and to what degree of accuracy the squared error is minimized.

The application of back propagation thus includes two phases. During the first phase the input is applied and a forward pass is made through the network to compute the output value for each unit. This output is then compared with the target, and an error term for each output unit is obtained. In the second phase, a backward pass calculating delta terms for each unit in the network is made. Once these two passes are complete, the weights are changed according to equation 2.3. The weights can be changed on a pattern-by-pattern basis, or the weight changes may be accumulated over the ensemble patterns.

2.3 DCTC's

DCTC's are the coefficients in a Discrete Cosine Transform of a selected frequency range of the magnitude spectra. They encode the smoothed overall shape of the spectrum and are defined as follows: Let H(f) be the magnitude spectrum of a speech frame, H'(f) be a nonlinearly amplitude-scaled version of H(f), H'(f') be a nonlinearly frequency warped version of H'(f) and [H'(f')] be a portion of H'(f') over a selected frequency range. Then the DCTC's are the a_n's in the equation

$$[H'(f')] = \sum a_n \cos[(n-1)\pi f'] \qquad (2.5)$$

Thus DCTC's are a form of cepstral coefficients (Rabiner and Schafer [21]) with different frequency and amplitude scaling and range selection. Note that DCTC1, the coefficient of a constant term, is a measure of the average level of the spectrum; DCTC2, the coefficient of a half-cycle of a cosine, is a measure of the spectral tilt; and so on. A smoothed spectrum can be obtained from the DCTC's, where the degree of smoothing depends on the number of DCTC's used.

The theoretical development for estimating formants from DCTC's is discussed in chapter three.

CHAPTER III

THEORETICAL DEVELOPMENT

3.1 GLOBAL SPECTRAL SHAPE

Extensive research has been performed by Zahorian and Jagharghi [26] and Nossair and Zahorian [15] to compare the overall global spectral shape with formants in classifying vowels and stop consonants. In their experiments, the overall spectral shape was represented by DCTC's. The DCTC's proved to be superior to formants for classifying both vowels and stops. Since the DCTC's encode the overall global spectral shape and the formants can be viewed as acoustic cues, a subsequent hypothesis is that formants can be computed from DCTC's. That is, since a smoothed spectrum can be computed as a linear transformation of the DCTC's, and since formants can be computed using peak peaking of the spectrum, it should be possible to compute the formants directly from the DCTC's with a single complex transformation.

3.2 CEPSTRAL COEFFICIENTS, DCTC'S AND FORMANTS:

Broad and Clermont [2] observed the correlations between the formant frequencies and linear combinations of cepstral coefficients as given by Plomp et al. [19] and Pols et al. [20]. They hypothesized that the formants can be obtained from linear transformations of cepstral coefficients and successfully demonstrated this with data from one speaker. However in tests with multiple speakers (four), the accuracy of the transformation was significantly degraded and they concluded that the linear relation does not hold for the multiple speaker case.

DCTC's are the traditional cepstral coefficients with an added flexibility in selecting frequency range, frequency scaling and amplitude scaling. We believe that the DCTC's and formants are functionally related, although not through a linear transformation. We hypothesized a nonlinear relationship and decided to use neural networks to investigate this unknown relationship.

3.3 NEURAL NETWORKS FOR LEARNING AND GENERALIZING

Neural networks (NN's) have caught the attention of a number of people in a wide variety of areas. Tasks that are intractable using a conventional digital computer, but are typically performed by the human brain are ones which are the goals for neural net applications. These networks can learn and remember. NN's are generalized models for representing various input-output relationships. One of the attractive feature of NN's is their ability to correctly handle noisy or erroneous input signals. Speech recognition has been one of the fields of application of neural networks [12].

The back-propagation algorithm is a popular algorithm used for finding optimum weights in a multilayer network in many pattern recognition applications and many people have investigated its capabilities. It is an example of a mapping network that learns to approximate a nonlinear function, y = f(x), from sample x, y pairs. A feedforward network with a nonlinear hidden layer can be trained using back-propagation algorithm to approximate many nonlinear functions. To approximate a particular set of functions $F_i\{x_k\}$ to a given accuracy, at most two hidden layers are needed [6]. Hornik et al. [8] established that multilayer feedforward networks with as few as one hidden layer (with a sufficient number of hidden units with logistic signal functions) are capable of approximating any measurable function from one dimensional space to another. Cybenko [5] proved that only one hidden layer is enough to approximate any continuous function. This justifies the use of neural networks to approximate a nonlinear mapping. Since the back-propagation algorithm has proven its power in an enormous number of applications, it was selected to train the feedforward network in our experiments.

3.4 MAPPING DCTC's to FORMANTS

The objectives of the present study were to investigate whether any relation exists between DCTC's and formants; and to test whether this relation is a linear or nonlinear one. To determine the nature of this hypothesized complex mapping between DCTC's and formants, a feedforward network with back-propagation algorithm was used. That is, the network would learn to map the DCTC's to formants if they are functionally related. In our experiments, both the linear and nonlinear transformations were performed to estimate formants from DCTC's and the results were compared.

3.5 LEARNING LINEAR AND NONLINEAR RELATIONS

In our experiments, for learning the nonlinear relations between inputs and outputs, a feedforward network with a nonlinear hidden layer was to be trained with the back-propagation algorithm. A network without any hidden layers was used to learn a linear relation. Such a network has an input layer and a linear output layer and learns by adapting the weights in between the two layers. This type of network represents the linear transformation case where the inputs and outputs are assumed to have a linear relationship. The adaptation of weights can be viewed as solving for the linear regression coefficients, as used by Broad and Clermont. A feedforward network with a nonlinear hidden layer, by means of weights in the hidden layers, tries to determine the functional relation between a set of input and target values, as supplied to the network as training data. Given sufficient training data, this relation should also generalize to other data outside the training set.

The hypothesis was tested experimentally, i.e; the network was trained and tested on a large database and the results for the linear and nonlinear transformation were compared. Chapter four describes the various experiments performed and discusses the results.

CHAPTER IV

EXPERIMENTS AND RESULTS

4.1 DATABASE

The database for these experiments was that used in previous work [15] [26]. It consisted of 99 CVC (consonant-vowel-consonant) syllables spoken by each of 30 speakers. Appendix A lists these CVCs. Ten of the speakers were adult males (M), 10 were adult females (F), and 10 were children (C) between the ages of 7 and 11 (5 male, 5 female). These speakers were all native speakers of English and approximately half of the speakers were natives of Virginia. The other half were selected from other regions, mostly from the northeast United States. The CVC syllable list contained ten vowels /aa, iy, uw, ae, er, ih, eh, ao, ah, uh/. The initial consonant was one of /b, d, g, p, t, k, h, l, w/. The final consonant was one of /b, d, g, p, t, k, v, s/.

For generating the target formant values, a 10th-order linear predictor (LP) model of the signal was computed. The roots of the LP polynomial were calculated to obtain five formant candidates per frame. The steady state portion of each vowel was equally divided into nine such frames. Finally, the dynamic programming formant tracking routine described in chapter two was used to select the first three formants for each frame. This algorithm was not fully automatic in the sense that the speaker category (M, F, or C) and expected formant values for each vowel were specified to the program. The operation of this tracking program was verified through visual inspection of several hundred formant tracks. The formant values obtained by this procedure were found to be very accurate.

The DCTC's were computed as the d_j's in equation 2.5 for j varying from 1 to 15. The number of DCTC's determine the degree of smoothing of the original spectrum. A 14th order model was chosen for the initial experiments. The effect of the number of DCTC's on formant estimation was tested in an additional experiment. The first coefficient is a measure of average level of spectrum and hence was not used in these formant tracking experiments. More detailed information about the database can be found in references 15 and 26.

4.2 PROGRAM

The experiments were performed with a feedforward network trained with back-propagation. An important feature of the network program was the flexibility in choosing various parameters; i.e., the number of hidden layers, the number of nodes, the learning rate and momentum, data files, optional test files, the number of DCTC's, scaling options, etc. The network was set up to have linear input and output layers and nonlinear hidden layers. The sigmoid (0 to 1) nonlinearity function was chosen (equation 2.2) for the hidden layers. The input and output values were analog values and were scaled to have zero mean and a standard deviation of one. The program was set up to train the network with specified training data files and to periodically examine the performance on training data as well as on optional test data. Note that for the case of separate training and test data, the test data was obtained from different speakers. Thus test results are a measure of speaker-independent performance. In these experiments, the performance was measured in terms of the average absolute error and correlation coefficient between the actual and estimated formants.

The correlation coefficient r is a measure of the linear association between two variables x (actual formants) and y (estimated formants). It is defined as

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \tag{4.1}$$

where

$$\sigma_x = \sqrt{\Sigma(x_i - \overline{x})^2}$$

is the variance of x,

$$\sigma_y = \sqrt{\Sigma(y_i - \overline{y})^2}$$

is the variance of y, and

$$\sigma_{xy} = \sqrt{\Sigma(x_i - \bar{x}).(y_i - \bar{y})}$$

is the covariance between x and y. Note that the correlation coefficient r lies between -1 to +1; when r is close to +1, the values of the actual and estimated features are approximately equal.

The same program could be used to function as a neural network classifier. For the classifier mode, one hidden layer and one output node were used for each vowel category. All nodes were logistic sigmoids. As described in more detail later, a classification experiment was conducted both for original features and estimated formants.

4.3 EXPERIMENTS AND RESULTS

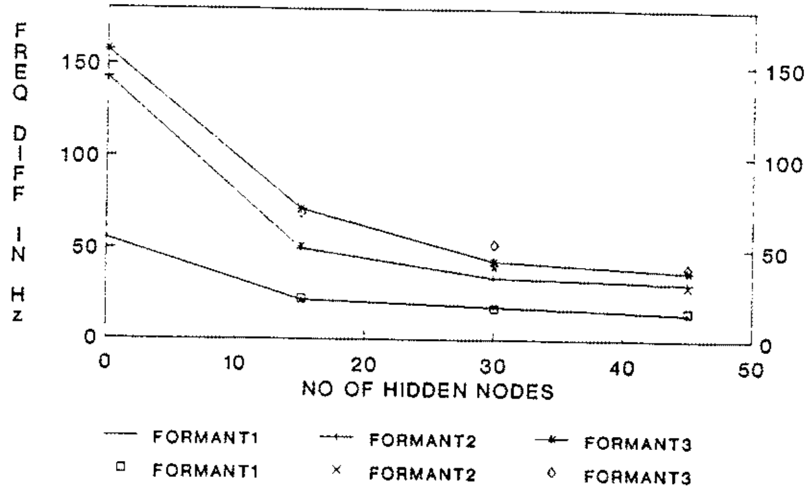
Several experiments were conducted on the database for four cases: males (M), females (F), children (C), and all speakers (A). The network was provided with 14 DCTC's (coefficients 2 to 15) as inputs and calculated corresponding formants as target values. DCTC's and formant targets for the middle (fifth) frame in the steady state portion of the vowel constituted the data set for that vowel. For all the experiments, one training iteration consisted of the presentation of one sample from each category (vowel). The correlation coefficients and absolute error differences between actual and estimated formants were collected periodically after a specified number of iterations.

4.3.1 Tests Various tests were performed which can be summarized as follows:

A. **TRAINING ONLY** These experiments were conducted to investigate the potential of this approach using all data for training. Several runs were made with different numbers of hidden layers and hidden nodes for all four cases (i.e. M, F, C, A). Figures 4.1 through 4.4 display the effect of the number of hidden nodes on the error and correlation coefficient for all the cases: M, F, C and A respectively.

The results shown are after 100,000 training iterations. As the number of hidden nodes increases, the performance improves. However, after about 30 hidden nodes, the error decreases very little as the number of hidden nodes increases. For the two hidden layer case, the total number of weights in the network was chosen to be approximately the same as for the corresponding one hidden layer case. For example, the two hidden layer network with 9 and 8 hidden nodes is equivalent to the 1 hidden layer network with 15 hidden nodes, the two hidden layers with 14 and 14 nodes respectively correspond to 1 hidden layer with 30 nodes; and the two hidden layer network with 18 and 18 hidden nodes correspond to 1 hidden layer network with 45 hidden nodes.

The increase in the number of hidden layers does not seem to have a substantial effect on error.



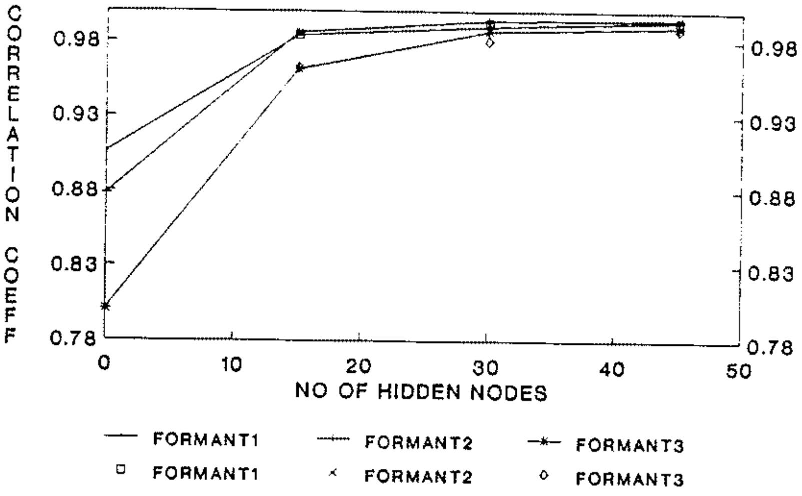
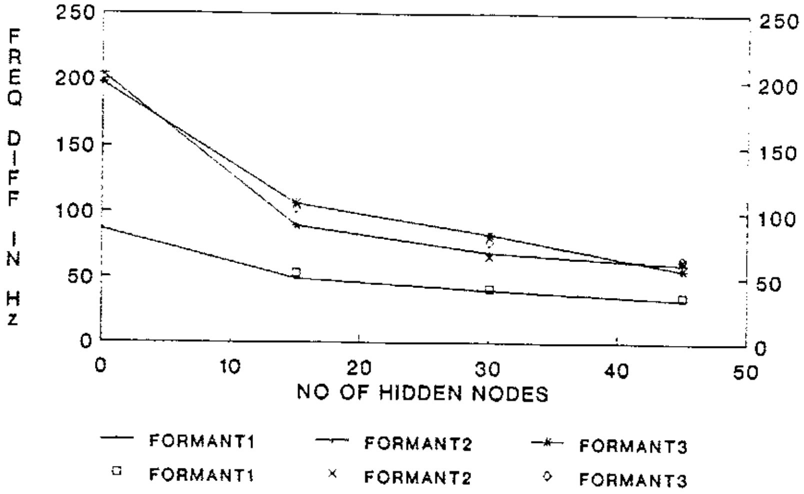


Figure 4.1. Training results for male speakers.



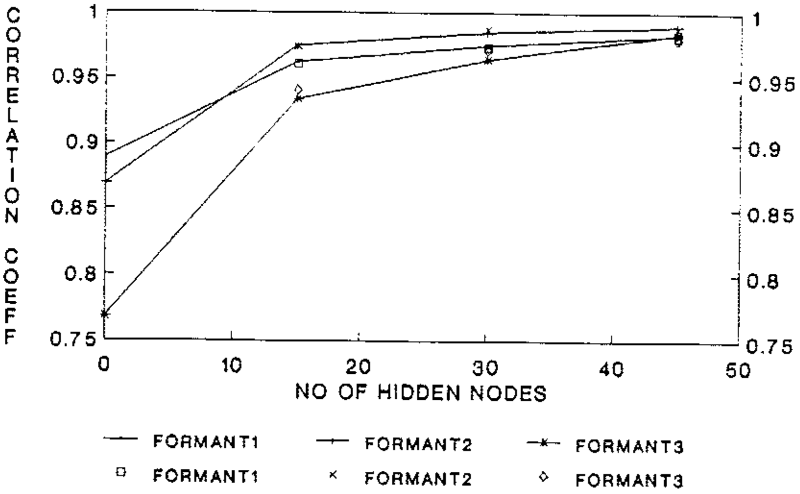
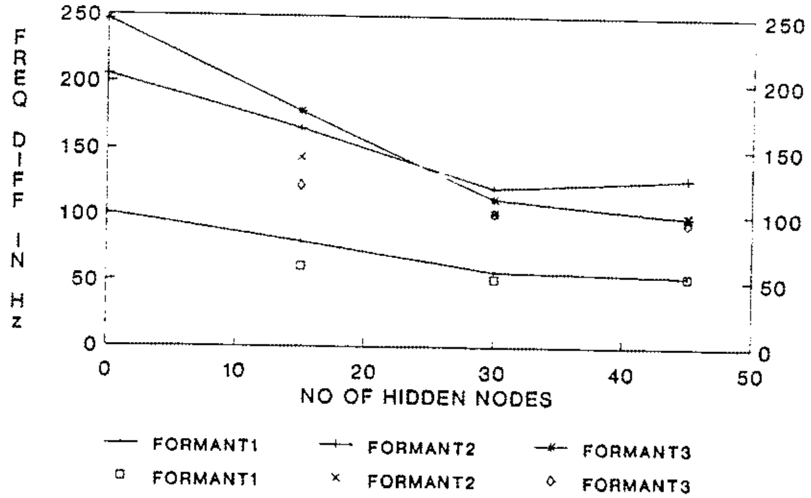


Figure 4.2. Training results for female speakers.



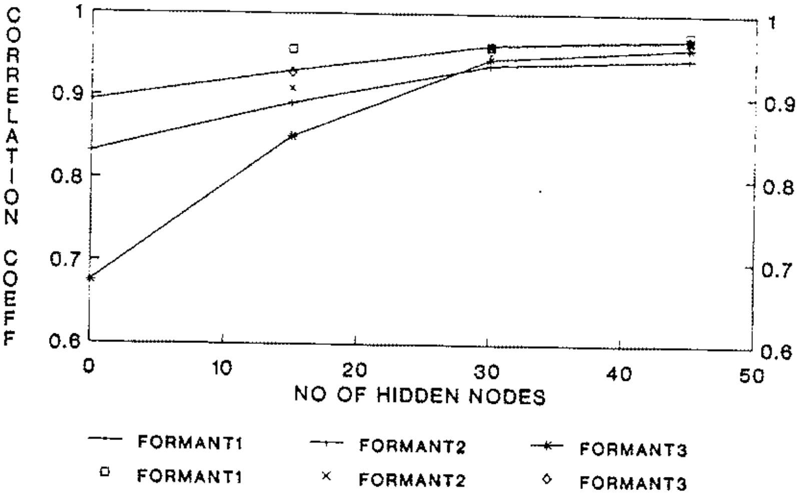


Figure 4.3. Training results for child speakers.

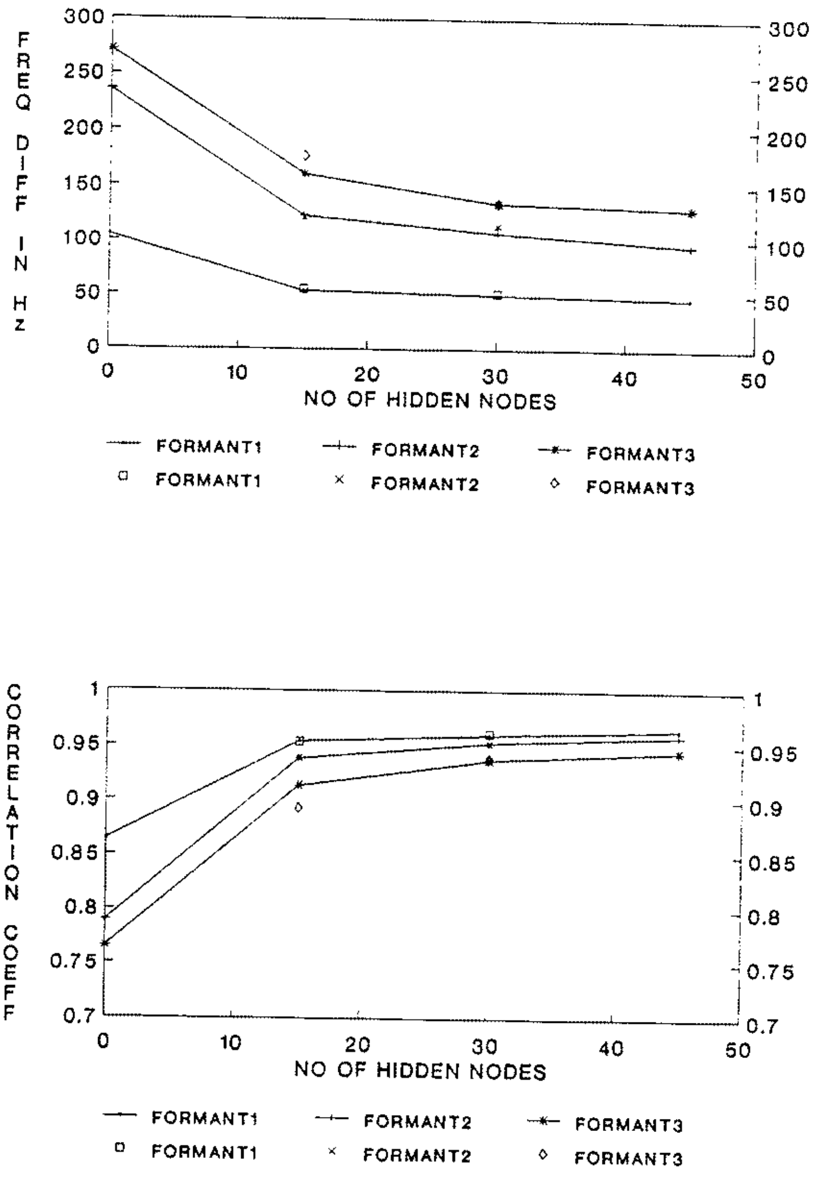
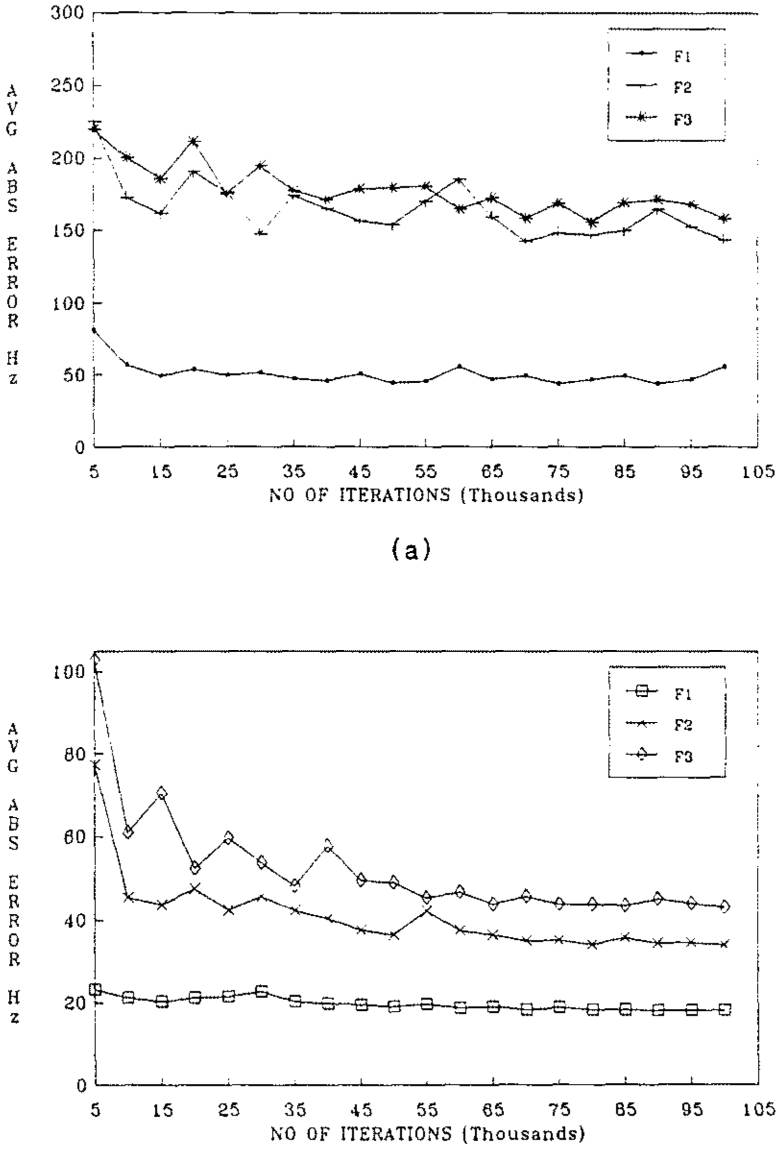


Figure 4.4. Training results for all speakers.

Correlation coefficients of 0.95 to 0.99 were achieved with a one hidden layer network (with 30 hidden nodes) as compared to 0.68 to 0.90 from the linear network. For the all speaker cases, the average absolute errors for F1, F2, F3 were 49 Hz, 106 Hz, and 134 Hz respectively versus 104 Hz, 236 Hz, and 277 Hz for the linear network. These results clearly show that the nonlinear network results in a much better estimate of the formants as compared to the linear network. Figure 4.5 shows the effect of the number of iterations on average absolute error for both the linear and nonlinear networks. The results are collected after each set of 5000 iterations for the male data case on training data. For the nonlinear network, the training results improve as the number of iterations are increased. The linear network, on the other hand, shows no obvious improvement in learning with a greater number of iterations.

From the above test runs, a one hidden layer network with 30 hidden nodes was selected as the optimal nonlinear case and compared with the linear case in the tests that follow.



(a)

(b)

Figure 4.5. Differences between target and estimated frequencies (Hz) collected at various number of iterations for male training data
(a) linear (0 hidden layer) case
(b) nonlinear (1 hidden layer) case.

B. TRAINING AND TESTING For each of the four cases, approximately half the database was used to train the network and the other half was used to test the network performance on unseen data. The training and test data were then interchanged and the average of the two sets of test results was computed for increased statistical reliability. A linear (zero hidden layer) network and a network with one hidden layer with 30 hidden nodes were implemented for each test. The test results are important since they represent the ability of the network to generalize. Figures 4.6 to 4.9 show bar graphs for comparing the results, again as average formant errors and correlation coefficients for both training and test data, for cases M, F, C and A respectively.

As mentioned above, the network was trained on half the database and tested independently on both the training and testing data. Using the linear transformation, correlation coefficients, after 3000 training iterations, range from 0.50 to 0.93 on training data and from 0.47 to 0.85 on test data. The corresponding correlation coefficients for the nonlinear transformation are 0.91 to 0.99 on training data and 0.79 to 0.94 on test data. Thus the higher correlation coefficients for the nonlinear mapping, especially on the test data, indicate that on the average the nonlinear mapping is a better formant estimator. The correlation results after 20000 iterations (not shown) are also higher for the nonlinear network versus the linear network. The correlation coefficients for the

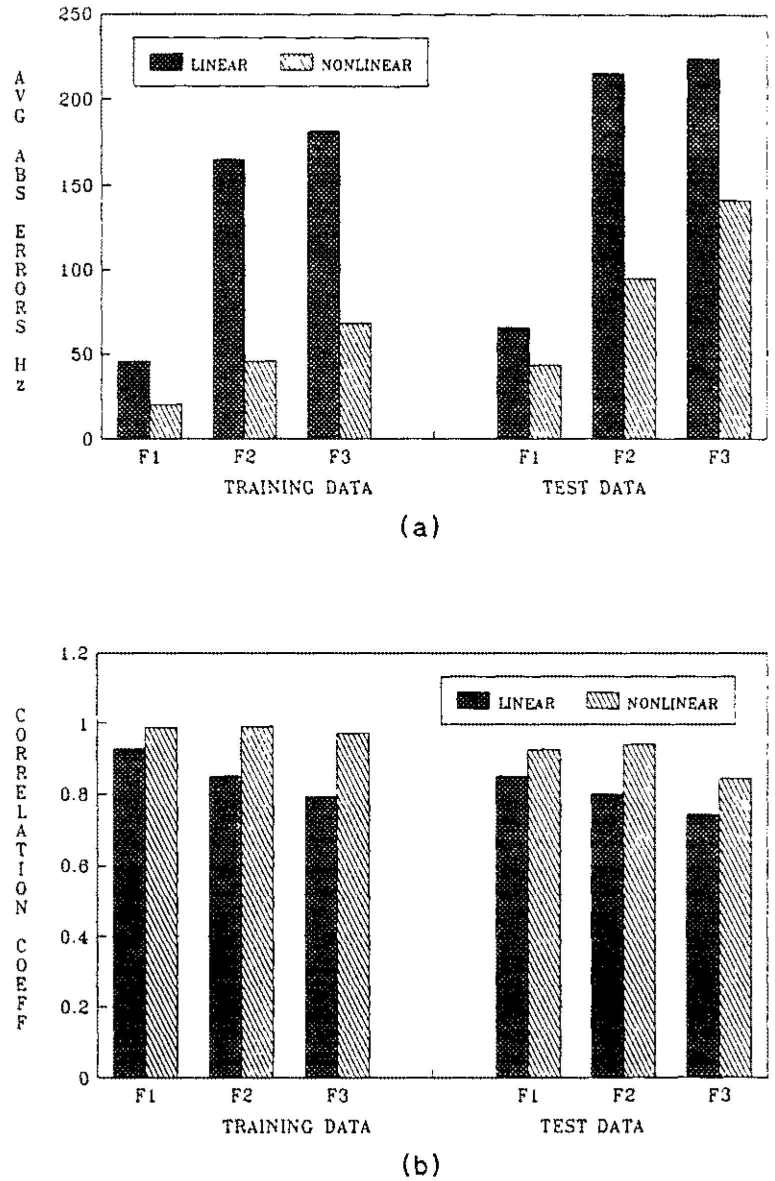


Figure 4.6. Comparison of results for linear(0 hidden layer) and nonlinear (1 hidden layer) case after 3000 iterations for male data.

(a) Differences between target and estimated frequencies (Hz)
(b) Correlation coefficients.

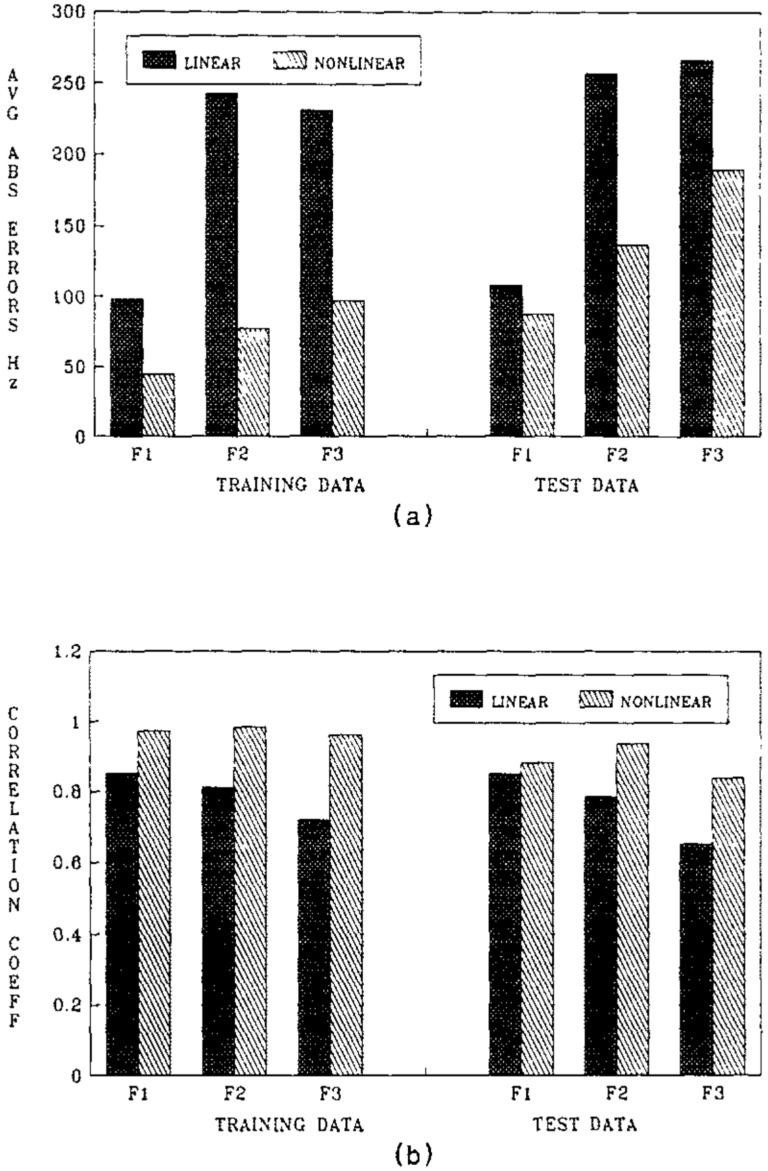


Figure 4.7. Comparison of results for linear(0 hidden layer) and nonlinear (1 hidden layer) case after 3000 iterations for female data.

(a) Differences between target and estimated frequencies (Hz)
(b) Correlation coefficients.

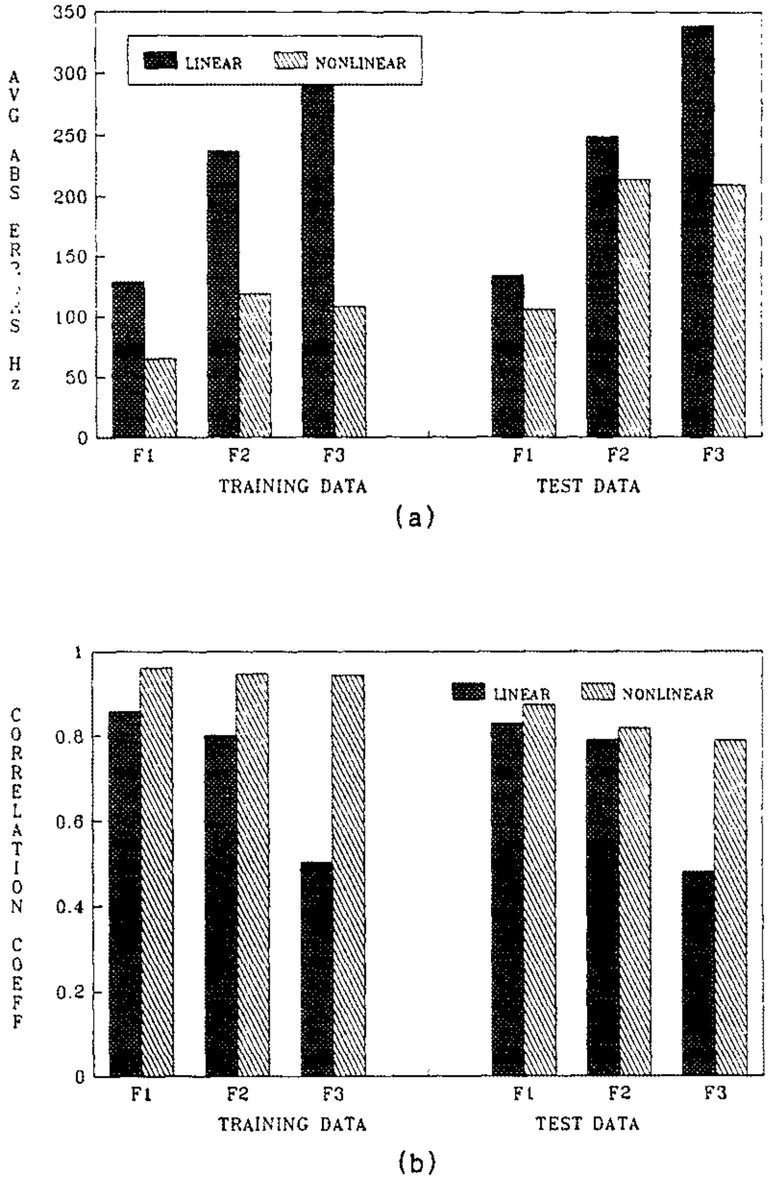


Figure 4.8. Comparison of results for linear(0 hidden layer) and nonlinear (1 hidden layer) case after 3000 iterations for child data.

(a) Differences between target and estimated frequencies (Hz)
(b) Correlation coefficients.

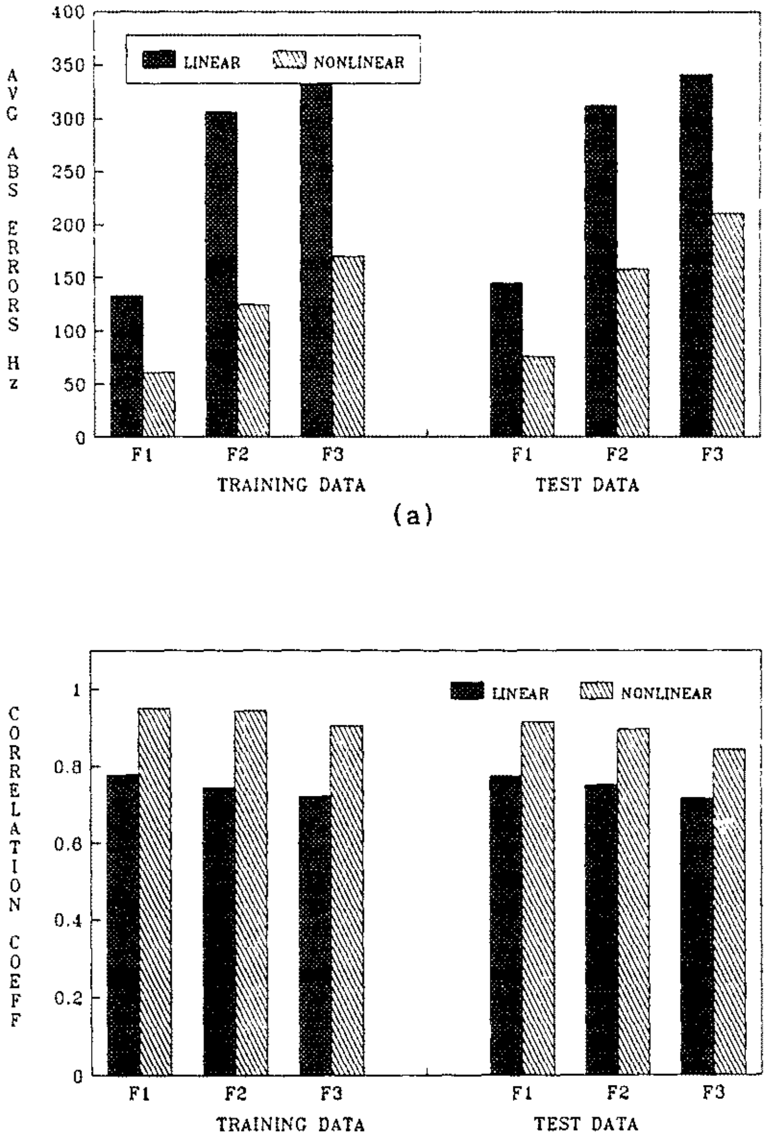
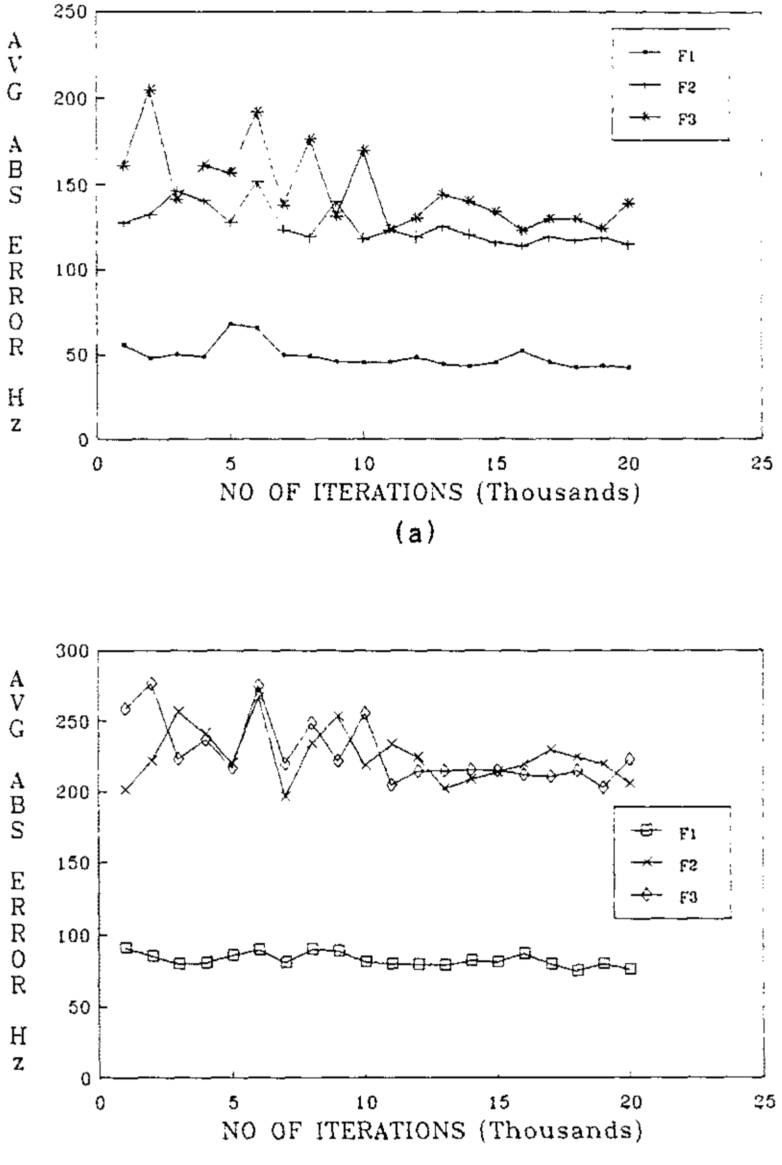


Figure 4.9. Comparison of results for linear(0 hidden layer) and nonlinear (1 hidden layer) case after 3000 iterations for all data.

(a) Differences between target and estimated frequencies (Hz)
(b) Correlation coefficients.

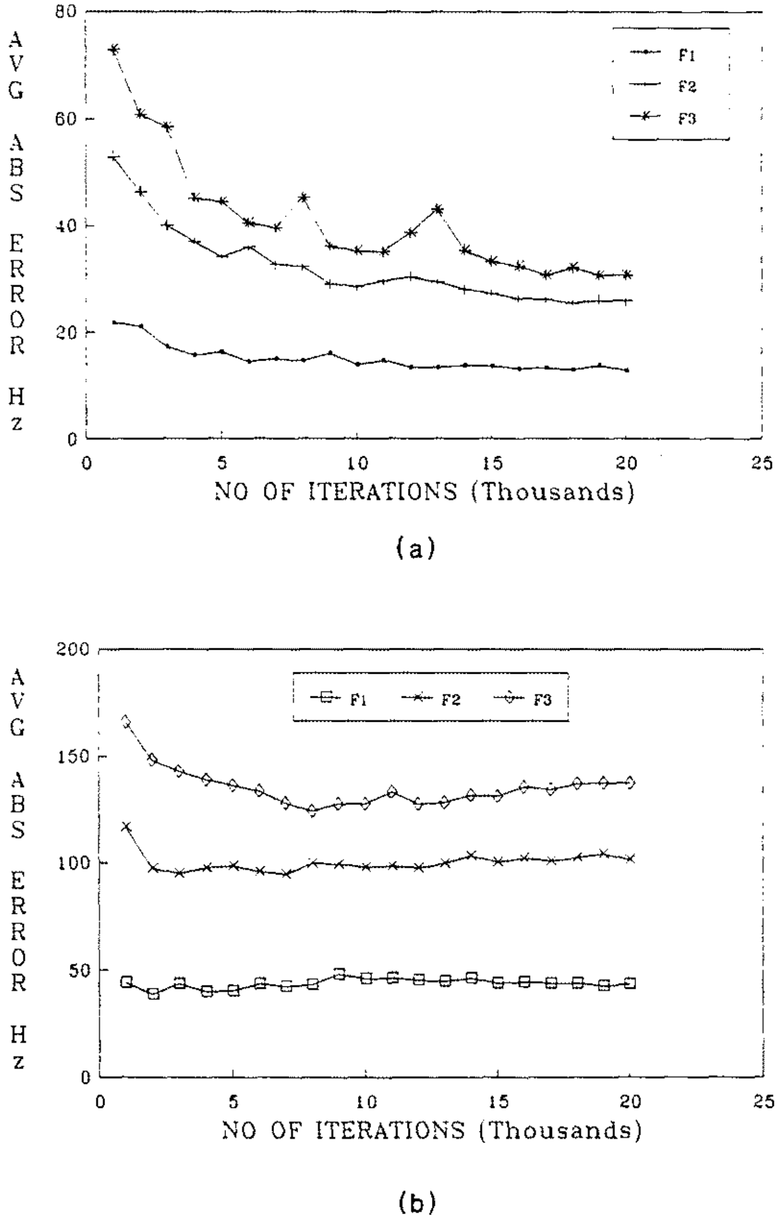
linear case after 20000 iterations (not shown) are 0.675 to 0.94 on training data and 0.662 to 0.886 on test data. The corresponding values for the nonlinear case are 0.971 to 0.995 on training data and 0.675 to 0.94 on test data. The careful reader will note that all values have increased for training data. However on the test data, values increase for the linear transformation but decrease slightly for the nonlinear transformation. This implies that the nonlinear network is "overtrained." Additional testing, with more iterations, yielded no improvement in performance for the linear transformation. This can be seen from figure 4.11 which shows the absolute error differences for the linear transformation up to 20000 iterations. The corresponding absolute error differences for the one hidden layer case are also given in figure 4.11. Hence the nonlinear network is more successful in learning as well as in generalizing as compared with the linear transformation. However, the possibility for overtraining appears to be specific to the nonlinear network.



(a)

(b)

Figure 4.10. Differences between target and estimated frequencies (Hz) collected at various number of iterations for male data for 0 hidden layer (linear) case. a. training results b. testing results.



(a)

(b)

Figure 4.11. Differences between target and estimated frequencies (Hz) collected at various number of iterations for Male data for 1 hidden layer (nonlinear) case. a. training results b. testing results

C. TIME-DELAY BACK-PROPAGATION A time-delay back-propagation network takes into account the previous and present input values to compute present outputs. A feedforward network with one time delay was trained for different numbers of hidden nodes. In particular the network was set up to estimate formants, given the DCTC's of both the current and previous frame. The experiments were performed for training only as well as with training and testing. The data set used for these experiments was slightly different than that used for other runs[2]. Nine DCTC's were used as input parameters and the data set contained three vowels spoken by the same 30 speakers mentioned previously.

The average formant errors and correlation coefficients for the Male case are given in figure 4.12. The results are collected at 20000 iterations since as the size of the network increases, it takes a longer time to learn. The time delay improves the performance by a small amount. However, since the effect was very small, the time-delay approach was not pursued with more detailed testing.

4.3.2 TESTS FOR COMPARING RESULTS:

Using the weights generated by the network during training experiments, input DCTC's were transformed to obtain the estimated output formants. For a more detailed comparison (as compared to the correlation coefficients and average absolute errors already mentioned) with the target formants, two approaches were implemented.

[2] For the other data, DCTC's had been computed for one frame only.

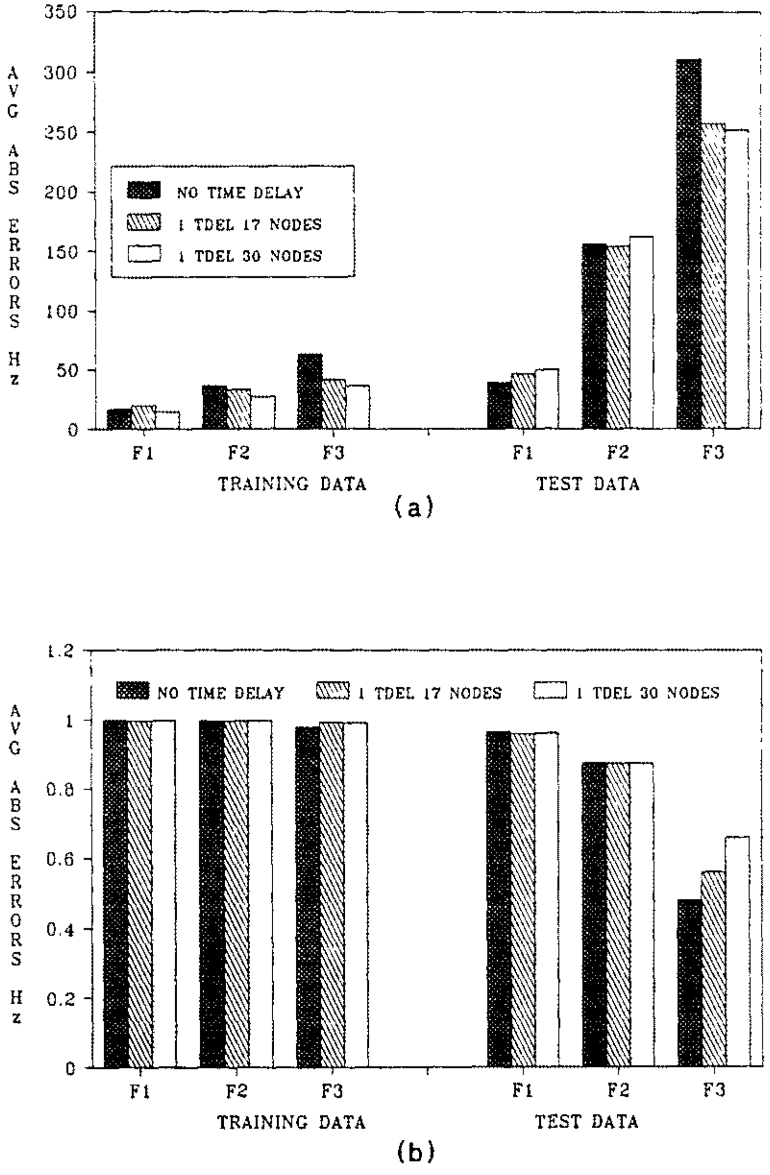


Figure 4.12. Comparison of results for networks with and without time delay. Results are shown after 20000 iterations for male data case.

(a) Differences between target and estimated frequencies (Hz)
(b) Correlation coefficients.

A. HISTOGRAMS

The results given previously only indicate average performance. This is not a very complete comparison. That is, for example, although the nonlinear network has better average performance than the linear transformation, there might be a relatively large number of very large errors. Such an effect would be detrimental in some applications. To investigate performance in more detail, error histograms were created. The histograms for the male and all data cases are shown in figures 4.13 to 4.16. The percentage error is the difference between target and estimated formant frequencies relative to the target formant. The histograms show the percentage of tokens for which the percentage error is the indicated amount.

The error histograms are plotted for both the nonlinear and linear case. The figures show that the error distribution of the nonlinear case is excellent as compared to the linear case. For example, for the male training data (Formant 1) about 82% (367 tokens out of 444) fall within 5% error for the nonlinear case shown in figure 4.6; whereas 40% (179 tokens out of 440) show up in that region for the linear case. There are also relatively much fewer "large" errors for the nonlinear transform versus the linear transform. Similar differences hold for other histograms. The histograms plotted for male test data also show that the majority (80%) of the estimation errors are less than 10% for the nonlinear case. Histograms obtained from training data naturally indicate better performance as compared to the test histograms. Performance based on training data can be

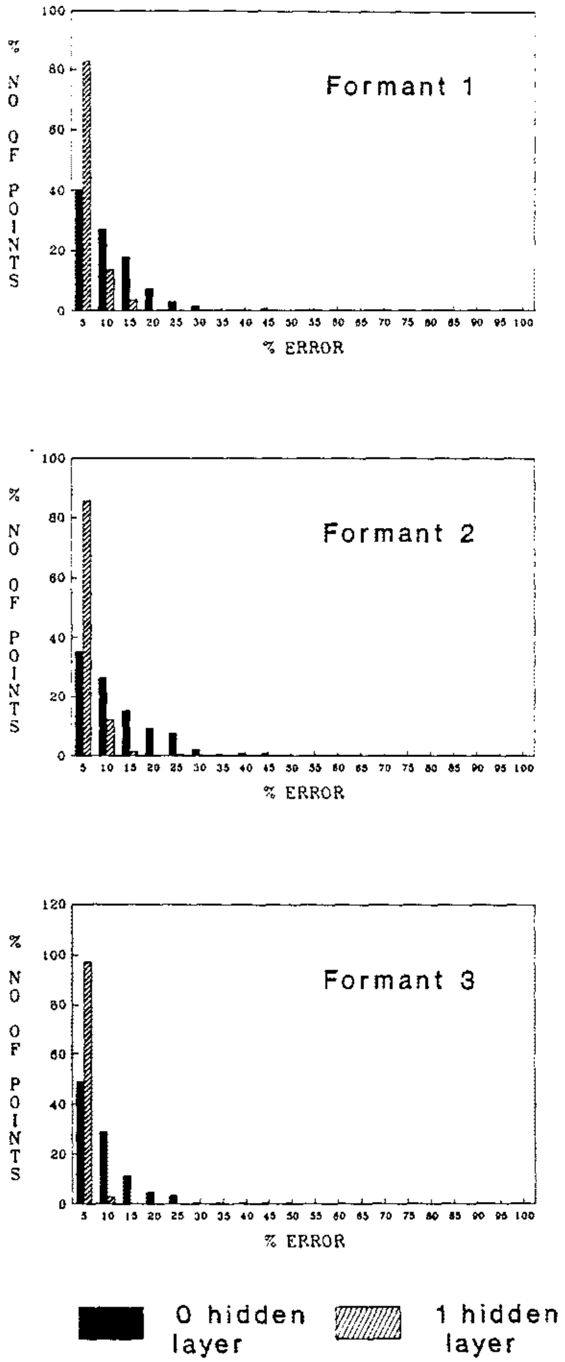


Figure 4.13. Error histograms for training data, male speakers.

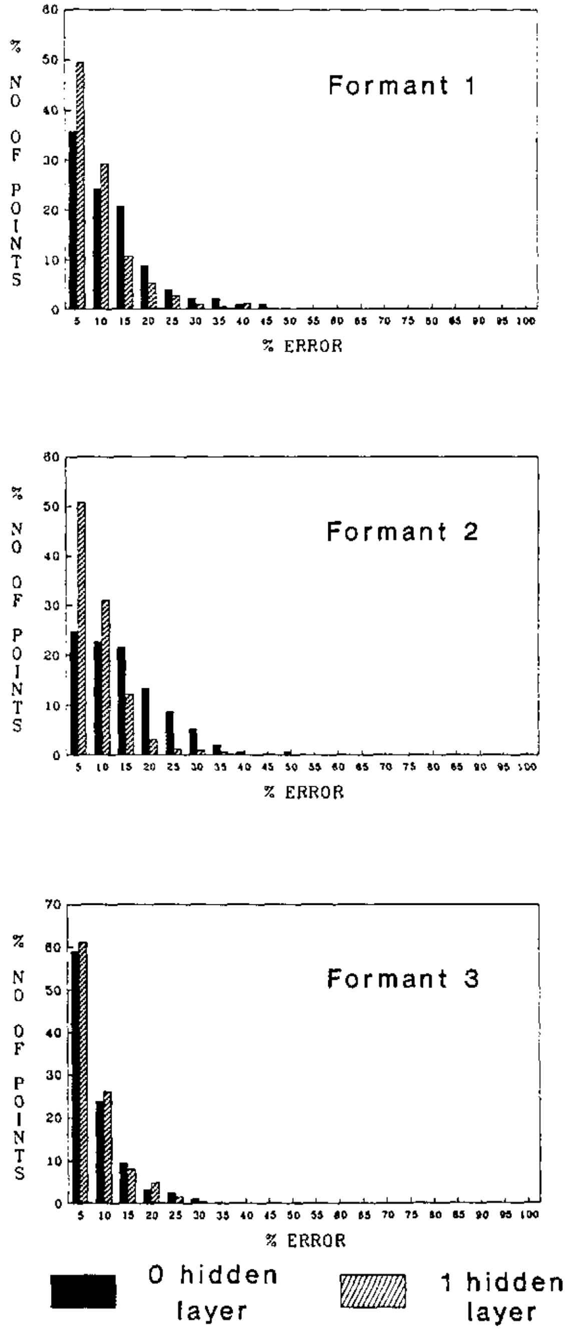


Figure 4.14. Error histograms for test data, male speakers.

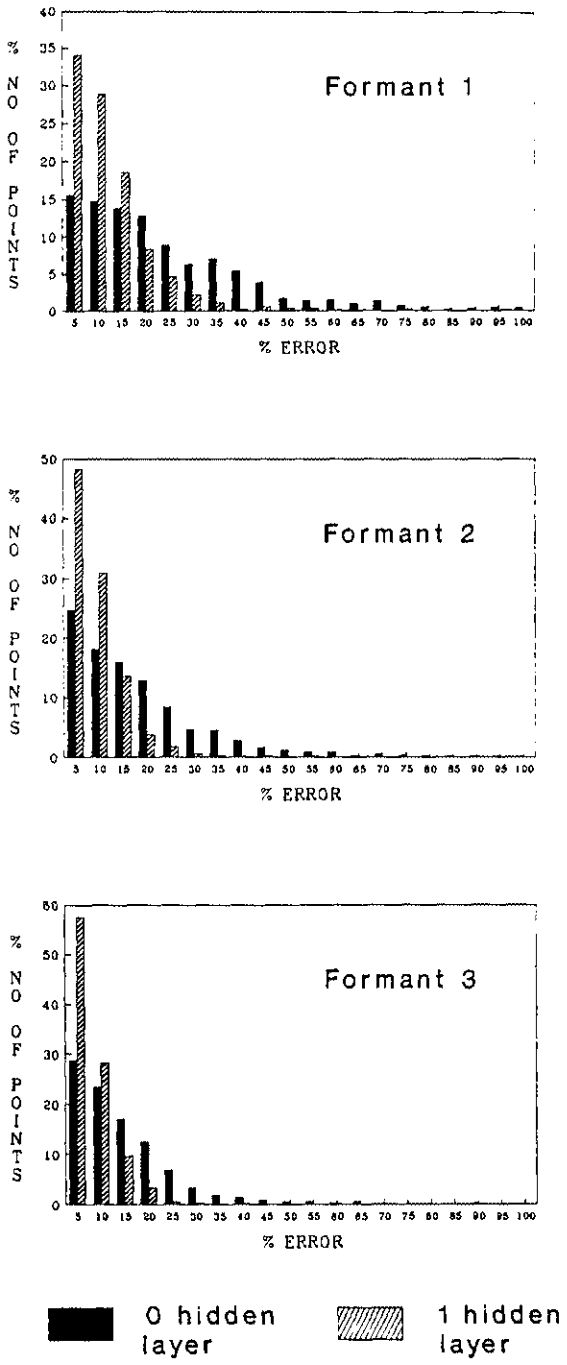


Figure 4.15. Error histograms for training data, all speakers.

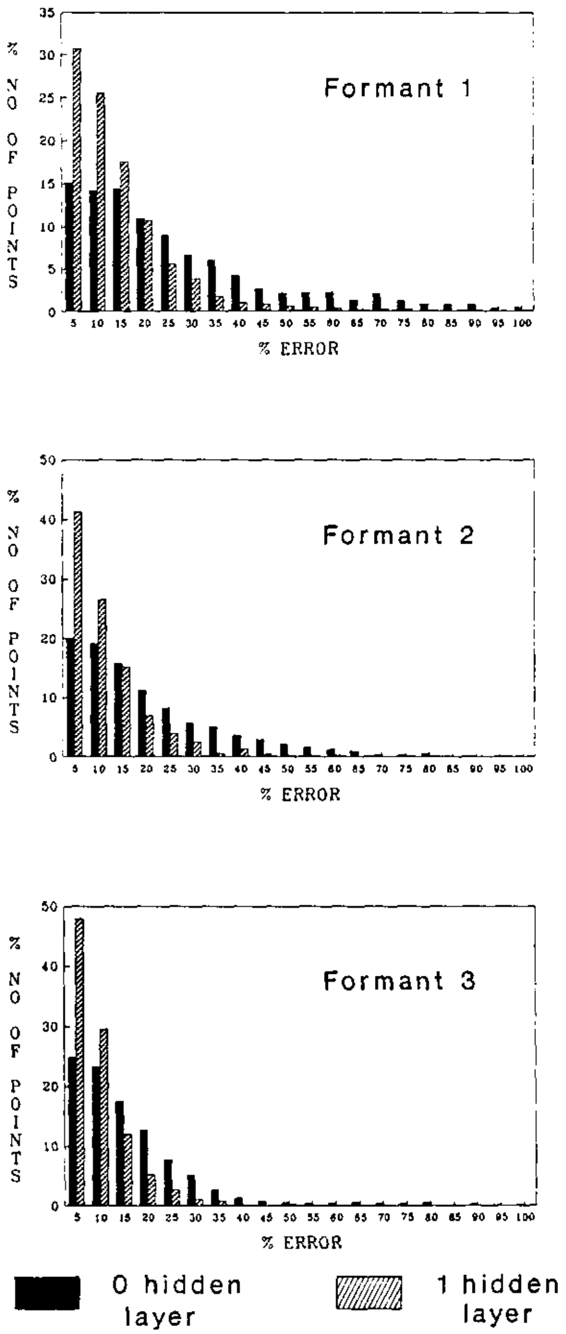


Figure 4.16. Error histograms for test data, all speakers.

considered as an upper limit on potential performance. Since, even for the training data, there are still a few cases where the error is above 30%, we cannot conclude that no large errors would occur with this procedure.

B. CLASSIFICATION

The results already discussed clearly show that the nonlinear transformation results in better formant estimation than for the linear transformation, but that there still are significant errors. Nevertheless, the mapping of 14 DCTC's to a 3-dimensional space could result in a 3-dimensional feature space which is as suitable for vowel recognition as are formants. To examine this possibility, vowel classification experiments were run on formants, DCTC's, and estimated formants.

From the estimated formant frequencies, vowel classification was performed using a neural network classifier and the results obtained were compared with the classifier output when run with the target formants and the input DCTC's. The recognition rates for the linear and nonlinear cases are given in tables 4-1 and 4-2. The overall high performance for the nonlinear case supports its superiority. On the training data, the nonlinear case recognition rates are approximately the same as the recognition rates with actual (target) formants. However, for the test data case, the nonlinear case recognition rates are lower than that of actual (target) formants. For all data types, the linear case recognition rates are lower than or at the most equal to the nonlinear case.

Training data	DCTC's	TARGET	LINEAR	NON LIN
MALE	98.874	86.937	79.279	86.486
FEMALE	98.621	88.276	81.379	87.126
CHILD	98.624	79.817	70.413	80.275
ALL	80.913	72.319	52.624	73.080

Test Data	DCTC's	TARGET	LINEAR	NON LIN
MALE	76.591	82.717	64.318	70.682
FEMALE	72.562	75.283	59.637	59.410
CHILD	80.930	80.465	63.023	64.419
ALL	72.311	72.998	50.191	65.980

Table 4-1. Comparison of vowel classification results after 3000 iterations. Results are expressed as percentage recognition rates for each case.

Training data	DCTC's	TARGET	LINEAR	NON LIN
MALE	99.324	88.514	83.333	88.964
FEMALE	99.540	91.264	87.126	92.874
CHILD	99.771	84.174	79.817	86.486
ALL	87.681	79.001	59.316	79.087

Test data	DCTC's	TARGET	LINEAR	NON LIN
MALE	73.864	80.227	62.727	72.045
FEMALE	71.882	73.696	59.184	57.370
CHILD	78.140	75.814	63.488	63.953
ALL	71.243	75.133	52.937	71.548

Table 4-2. Comparison of vowel classification results after 20000 iterations. Results are expressed as percentage recognition rates for each case.

The vowel classification results based on the DCTC's, as given in tables 4-1 and 4-2, illustrate the potential of DCTC's for vowel discrimination. Note that the rates for training data are very high, as compared to formants, but the test rates are generally lower. This effect indicates insufficient training data for the high dimensionality features. In general test results for the best estimated formants on the DCTC's are comparable to test results, indicating that the estimated formants do contain most of the information in the DCTC's for vowel discrimination.

4.4.3 THE EFFECT OF THE NUMBER OF DCTC's ON FORMANT ESTIMATION

The number of DCTC's used determines the degree of smoothing of the original spectrum. Hence a large number of DCTC's imply a more detailed spectral description. To investigate the effect of the number of DCTC's on formant estimation, an experiment was conducted. Figure 4.17 is a graph of the average correlation coefficient (over all three formants) vs. the number of DCTC's for test data. This experiment was performed for male speakers only. The results shown contain the maximum average correlation coefficients and the average correlation coefficients obtained after 3000 and 20000 iterations respectively. The results are best when 10 to 14 DCTC's are used. The correlation coefficients degrade otherwise due to the fact that the size of the network increases.

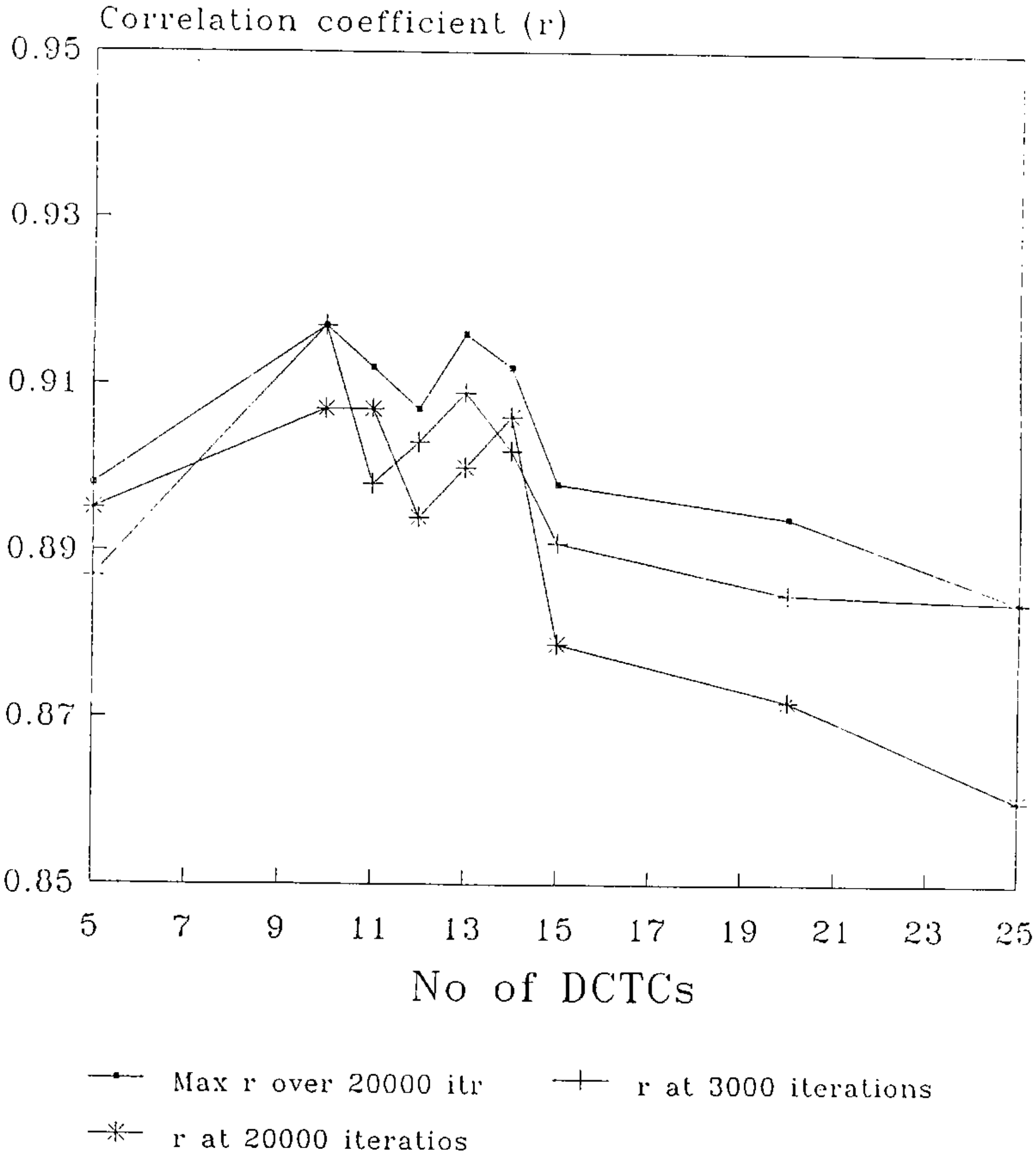


Figure 4.17. Correlation coefficient vs. number of DCTC's.

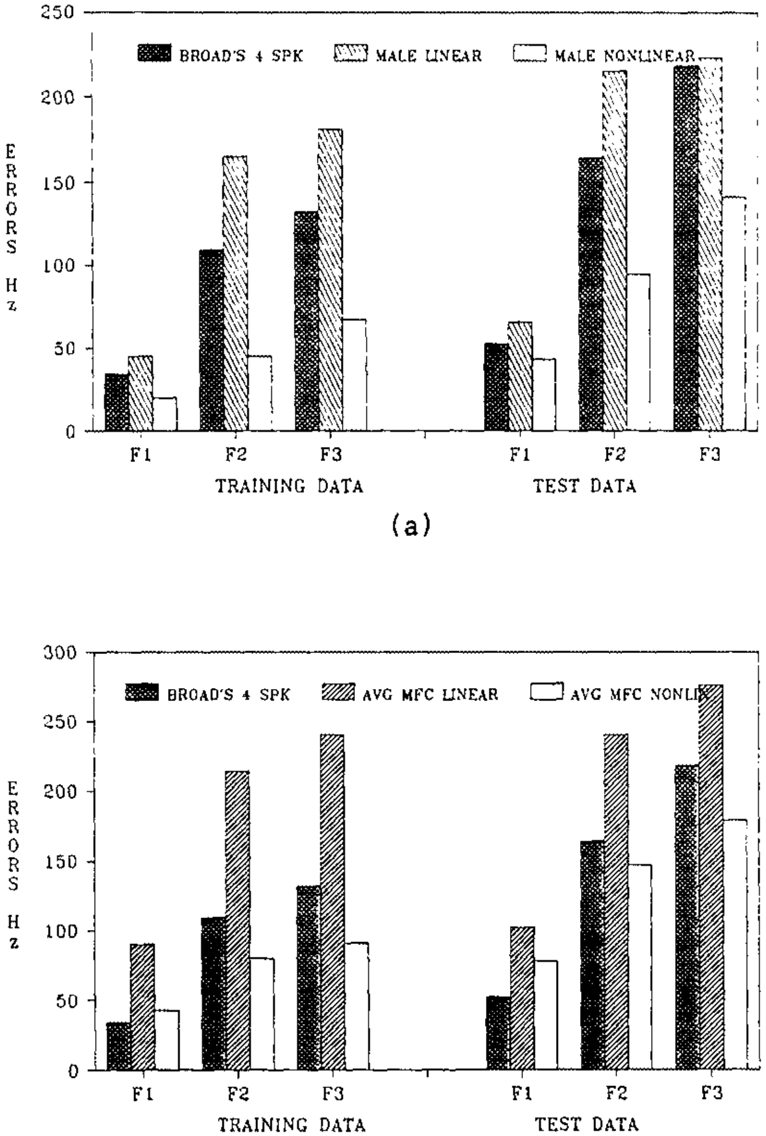
4.4 GENERAL DISCUSSION

From all the experiments, it is clear that the network is fairly successful in learning to estimate formants from DCTC's. For most tokens, the errors are quite small. Maximum errors are typically on the order of 30%, depending on the details of the data used, the formant (Fl, F2, F3), etc.

In comparison with the Broad and Clermont [2] data base, our database was much larger and diverse. It came from 30 speakers as opposed to four speakers for the Broad data. Even the M, F, C cases each used data from ten speakers. Ten vowels were used with each vowel paired with at least one instance of /b, d, g, p, t, k/ in the initial and final positions, whereas Broad's data contained nine vowels in CV (consonent vowel) context with differing consonants only in the starting position.

In their experiments on four speakers, they trained the network on the data from three speakers and tested it on data from the fourth using 75% of the total data on training. As mentioned above, we have used 50% of the data for training and 50% of the data as the test data. For example, in the male (M) case, data from 5 speakers is used for training the network and the data from the other 5 is used as the test data.

Broad's most general (4 speakers: 3 training, 1 test) case can be compared with our male data(10 speakers: 5 training, 5 test) as shown in figure 4.18. For the linear transformation, both the training and test errors are higher than that of



(a)

(b)

Fig 4.18: Comparison of results with Broad's general case (4 speakers);

(a) with results of male data (10 speakers)
for 0 hidden (linear) and 1 hidden layer (nonlinear) case;
(b) with averages of results on male, female and child data cases
for 0 hidden (linear) and 1 hidden layer (nonlinear) case.

Broad's general case data base. The nonlinear case errors are substantially smaller than those resulting from the linear transformation as well as Broad's case for both training and test data. Figure 4.18 also shows the errors for our linear transformation and nonlinear case, averaged over M, F and C speakers cases. Again, the linear transformation errors are higher than Broad's errors and also than linear transformation errors for male data, due to the fact that more diverse data were used. The nonlinear case errors (avg. M, F and C) are substantially lower than for the linear transformation. Even for this case, despite the larger differences in data base, the nonlinear errors are lower than those from Broad's data for F2 and F3.

Thus the results show that the nonlinear network performed considerably better than the linear transformation (which can be considered to be equivalent to Broad's results, but with a larger data set). However, the experimental results differ from Broad's results in two aspects. Broad's results are shown in terms of rms errors while our results are shown in terms of average absolute errors, although this is probably a minor factor for comparing the performance. Also, Broad et al. have used the LPC cepstrum to compute the cepstral coefficients, while we have used the smoothed spectrum for obtaining DCTC's.

CHAPTER V

CONCLUSIONS

Two major issues regarding formant estimation were investigated. First, whether the formants can be computed from the overall global spectral shape. DCTC's were used as the global spectral shape parameters. Secondly, whether the relation between the DCTC's and formants is linear or a nonlinear one. A feedforward network with the back-propagation algorithm was successful in learning to estimate formants from the DCTC's. The performance of a nonlinear hidden layer network was found to be superior to that obtained with a linear transformation network, supporting the hypothesis that the relation between DCTC's and formants is nonlinear. The potency of neural networks for establishing unknown relations is worthwhile in handling these types of problems.

A number of suggestions may be given for possible future research in this respect. The exact relation between DCTC's and formants is yet unknown and may be investigated. In these experiments, DCTC's were computed previously. Future research may involve real time DCTC's for estimating formants. The use of LPC cepstrum may prove better since it emphasizes the spectral peaks.

In these experiments, the networks should not be overtrained. This is important since although network performance continues to improve on training data, the test data results degrade after a number of iterations. In our experiments, the number of iterations, at which the network training should be stopped, was obtained manually by observing the network performance at various points. A heuristic may be used to determine the point at which the test data results are at maximum.

Clearly, the neural network did represent the fundamental relationship between DCTC's and formants. Since the network performed much better on training data than on test data, it also seems likely that the neural net formant estimation procedure would perform much better with a larger data base for training.

APPENDIX A

List of 99 CVC syllables recorded for vowel database. The DARPABET phonetic spelling for each syllable is also given.

pot/paat/	bah/baa/	tog/taag/	dot/daat/
got/gaat/	top/taap/	cob/kaab/	pod/paat/
pock/paak/	hod/hhaad/	peep/piyp/	beet/biyt/
teak/tiyk/	deep/diyp/	keep/kiyp/	geese/giys/
peeb/piyb/	keyed/kiyd/	league/liyg/	heed/hhiyd/
boot/buwt/	poop/puwp/	toot/tuwt/	dupe/duwp/
coop/kuwp/	gook/guwk/	tube/tuwb/	sued/suwd/
moog/muwg/	who'd/hhuwd/	pat/paet/	bat/baet/
tack/taek/	dad/daed/	cap/kaep/	gap/gaep/
tab/taeb/	tag/taeg/	had/haed/	bird/berd/
dirt/dert/	curb/kerb/	perk/perk/	turk/terk/
gerp/gerp/	durg/derg/	heard/hherd/	dip/dihp/
tick/tihk/	kit/kiht/	give/gihv/	bib/bihb/
bid/bihd/	pig/pihg/	hid/hihd/	pep/pehp/
bet/beht/	ted/tehd/	debt/deht/	keg/kehg/

get/geht/	web/wehb/	peck/pehk/	head/hhehd/
bought/baot/	caught/kaot/	daub/daob/	gawk/gaok/
talk/taok/	paup/paop/	baud/baod/	cawg/kaog/
gawp/gaop/	hawd/hhaod/	but/baht/	tuck/taht/
putt/paht/	dug/dahg/	cup/kahp/	gut/gaht/
cub/kahb/	bud/bahd/	hud/hhahd/	book/buhk/
took/tuhk/	put/puht/	could/kuhd/	good/guhd/
hood/hhuhd/	boat/bowt/	dope/dowp/	goad/gowd/
code/kowd/	pope/powp/	toad/towd/	coke/kowk/
goag/gowg/	coab/kowb/	hoed/hhowd/	

REFERENCES

[1] Bell, C. G., et al. (1961). "Reduction of speech spectra by analysis-by-synthesis technique," J. Acoust. Soc. Amer., vol 50, 637-655.

[2] Broad, D. J. and Clermont, F. (1989). "Formant estimation by linear transformation of the LPC cepstrum," J. Acoust. Soc. Amer., 86(5), 2013-2017.

[3] Cole, R., et. al. (1989). "Speech recognition, VLSI and neural networks," Proc. Northcon/89 Electron Show and Convention, Portland, OR.

[4] Chen, J. R. and Mars, P. (1990). "Stepwize variation methods for accelerating back-propagation algorithm," Proc. IJCNN-90-WASH-DC.

[5] Cybenko, G. (1989). "Approximation by superpositions of a sigmoidal function," Mathematics of Control, Signals, and Systems, 2, 303-314.

[6] Cybenko, G. (1989). "Continuous value neural networks with two hidden layers are sufficient," Math Contr. Signal and sys vol 2, 303-314.

[7] Jacobs, R. A. (1988). "Increased rates of convergence through learning rate adaptation," Neural networks, Vol 1, 295-307.

[8] Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer Feedforware Networks are Universal Approximators," Neural Networks 2, 359-366.

[9] Kinoshita, J. and Palevsky, N. (1987). "Computing with neural networks," High technology, May 87, 21-31.

[10] Kolmogorov, A. N. (1957) "On the representation of continuous function of many variables by superposition of continuous function of one variable and addition," Dokl. Akad. Nauk USSR, 114, 953-956.

[11] Lippmann, R. P. (1987). "An introduction to computing with neural nets," IEEE ASSP magazine, 4-22.

[12] Lippaman, R. P. (1989). "Review of Neural Networks for Speech Recognition," Neural Computation 1, 1-38.

[13] McCandless, S. S. (1974). "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. ASSP-22, 135-141.

[14] Movellan, J. (1990). "Error functions to improve noise resistance and generalization in backpropagation networks," Proc. IJCNN-90-WASH_DC.

[15] Nossair, Z. B. and Zahorian, S. A. (1991). "Dynamic spectral shape features as acoustic correlates for initial stop consonants," J. Acoust. Soc. Amer., 89(6), 2978-2991.

[16] Olive, J. P. (1971). "Automatic formant tracking by a Newton-Raphson technique," J. Acoust. Soc. Amer, vol 50, 661-670.

[17] Parker, D. (1982). "Learning Logic," Invention report S81-64, Office of Technology Licensing, Stanford University.

[18] Peterson, G. E. and Barney, H. L. (1952). "Control methods in a study of the vowels," J. Acoust. Soc. Amer. 24, 175-184.

[19] Plomp, R., Pols, L. C. W., and Geer, J. P. (1967). "Dimentional analysis of vowel spectra," J. Acoust. Soc. Amer. 41, 707-712.

[20] Pols, L. C. W. and Plomp, R. (1973). "Frequency analysis of Dutch vowels from male speakers," J. Acoust Soc. Amer. 53, 1093-1101.

[21] Rabiner, L. R. and Schafer, R. W. (1978). Digital Processing of Speech Signals (Prentice-Hall, Inc., Englewood Cliffs, New Jersey).

[22] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). "Learning internal representations by error propagation," Parellel Distributed Processing," (D. E. Rumelhart and . L. McClelland, the MIT Press, Cambridge Massachusetts), Volume I, Chapter 8, 318-364.

[23] Talkin, D. (1987). "Formant trajectory using Dynamic programming with modulated transition costs," J. Acoust. Soc. Amer. 82, 555.

[24] Tollenaere, T. (1990). "SuperSAB: fast adaptive back-propagation with good scaling properties," Neural Networks, vol 3, 561-573.

[25] Werbos, P. J. (1974). "Beyond Regression: New Tools for Prediction and Analysis in the behavioral Sciences", Ph.d. Thesis, Harward University, Cambridge.

[26] Zahorian, S. A. and Jagharghi, A. J. (1991). "Speaker normalization of static and dynamic vowel spectral features," J. Acoust. Soc. Amer. 90, 67-75.