Order Number 9325063

Phoneme recognition with neural network preprocessing filters and backpropagation

Alankar, Sudhir, Ph.D.

City University of New York, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Phoneme Recognition with Neural Network Preprocessing Filters and Backpropagation

by

SUDHIR ALANKAR

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1993

This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

2/22/93
Date

Michael Anshel
Chair of Examining Committee

2/22/93
Date

Stanley Hahn
Executive Officer

Professor John Antrobus

Professor Stefan Burr

Dr. Jeffrey Fookson
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Phoneme Recognition with Neural Network Preprocessing Filters and Backpropagation

by

Sudhir Alankar

Advisor: Professor John Antrobus

Hearnet is an artificial neural network designed to simulate speaker independent speech recognition. The design of the model exploits neurophysiological findings on the afferent pathway from the outer ear to the auditory cortex. HearNet consists of a feedforward network, FilterNet, and a recurrent backpropagation network, LearnNet. FilterNet transforms the speech signal into phonemic information through cascaded, spatially distributed, neural network filters. The output of the FilterNet is fed into LearnNet which learns the successive invariances in the speech signal inorder to recognize phonemes.

The purpose of this research was to investigate the advantage of a speech recognition model based on the architecture of the human auditory system. FilterNet filters are modelled after the neurons in the cochlear nucleus and other

neural bodies, that are presumed to extract salient features of speech. Preprocessing of the speech signal by FilterNet simplifies the computational load of LearnNet. Each of the parallel filters of FilterNet passes relevant acoustic feature information of speech, in compacted form, to LearnNet. The filters of the HearNet model distinguish it from known speech recognition systems which require an enormous amount of computing power and are confounded by speaker variability.

HearNet recognizes a small number of speaker independent phonemes of monosyllabic words, in isolation with a recognition rate of 100%.

Acknowledgements

It is with great pleasure to acknowledge a number of people who contributed to the completion of this dissertation.

Professor John Antrobus served as my mentor and advisor. His knowledge has greatly influenced me. He introduced me to the concept of designing a speaker independent speech recognition system by exploiting the neurophysiological properties of the auditory system. His seminars in speech perception, neural networks, neurocognition, psycholinguistic theory gave me the opportunity to explore the field in depth. He deserves special credit for making the lab environment an exciting and stimulating place. I thank him for his support, advise and encouragement through all phases of my research.

I thank Professor Michael Anshel for being the chairman of my dissertation committee, for his support and patience in monitoring the progress of my work.

I thank Professor Stefan Burr and Dr. Jeffrey Fookson for being in my committee.

I owe a particular debt of gratitude to Professor Daniel McCracken, Nick Papavassiliou and George Kleiner who understood the need for computing

resources and helped provide the necessary computer equipment.

Dr. Jeffrey Fookson deserves credit for helping me in important aspects of programming. I am grateful for his contributions.

I thank Maura Pilotti, Chien-Ming Yang and Sophie Molholm in helping me record the speech data. Chien-Ming Yang's assistance with the graphics and figures has been invaluable and I thank him for his help.

I thank Sophie Molholm for her valuable feedback and for proofreading the thesis. I thank Chaim Tarshish for helping me in modifying the diagrams. Ibrahim Adali deserves a special thanks for his help in some of the aspects of programming.

Finally I thank my family and friends for their support and encouragement throughout the three years of this dissertation.

To my Mother and Father

Contents

List of tables

List of figures

1 Introduction to Speech Recognition

1.1 Speech Recognition: A Difficult Task

The brain has been subject of research almost since the time human beings started "thinking".Computer scientists, psychologists, philosophers, physiologists, and scores of other scientists have been applying varied techniques to explain the complex mechanisms of the brain, some of them looking at anatomical circuit diagrams, others at functional control loops and still others searching for basic organizing principles (building blocks). The prospect of therapeutic manipulation has been a longstanding driving factor for much of the research on speech recognition. Understanding the brain as an "intelligent information processing machine" has brought computer scientists, psychologists, applied mathematicians, physicists and engineers into the fold of the interdisciplinary field of "neural networks" for delving into the intricacies of brain modeling, in particular speech recognition, with renewed vigor.

One of the primary objectives of computer science is to achieve a natural interface between people and machines. Despite other advances in computer science, it has been extremely difficult to build machines which recognize continuous speech. Although humans understand speech spoken in different environments by

a variety of speakers, we are not aware of the amount of processing involved in the neurophysiology of the human auditory system. The process of speech recognition might be considered at a number of stages of coding.

The temporal nature of speech makes the recognition a very difficult task. Machine interpretation of the meaning of complete sentences is a very difficult task. This has been accomplished only when the syntax has been limited. Despite decades of intensive research, no machine has yet been able to recognize general, continuous speech produced by an arbitrary speaker, when no speech samples have been supplied. The variability problem in the human acoustic speech signal, large inter-speaker variability (in loudness, frequency band, the articulation rate due to age, gender, accent and the distance from the listener), intra-speaker variability due to emotional state, and intra-word variability due to coarticulation, is one of the greatest challenges for theories of speech perception and for artificial continuous speech recognition.

Two classes of variability, namely coarticulation (the production of a stop consonant is conditional upon, or influenced by, the production of the adjacent phonemes) and speaker variability, and their interactions have been particularly troublesome for artificial speech recognition systems. Because the coarticulation and speaker variability problems have not yet been satisfactorily solved, they provide

critical criteria against which theories and systems of speech recognition may be evaluated (Tartter, 1986; Martin et al., 1981).

1.2 Neurophysiological Findings of the Auditory System

The human auditory system carries out a sequence of transformations en route from cochlea to the auditory cortex so that the computations required by the auditory cortex are substantially simplified relative to that required by the auditory nervous system modelled by current continuous speech recognizer systems. The auditory nerve sends the output of the cochlea to the segmented cochlear nucleus, each segment of which is capable of carrying out parallel, independent transformations of the auditory nerve signal. Subsequent less well studied nuclei (inferior colliculus, the medial geniculate body), and a variety of smaller nuclei then collectively transform the signal before it reaches the auditory cortex.

Some of the neurons whose characteristics have been simulated with FilterNet (Antrobus, Alankar, Deacon, & Ritter, 1992; Antrobus, Alankar, & Fookson, 1993; Antrobus, & Alankar, 1991; Antrobus, Bushell, Alankar, & Tartter, 1990), listed by location in the auditory pathway, are:

a. Cochlea and Auditory Nerve (Kiange et al., 1988; Rhode, 1991):

i. Scaling by F_0 harmonics, log frequency (pitch) and loudness.

ii. Narrow and broad (for fricatives and aspirations) frequency band widths.

iii. Although the firing rate of most units increases with loudness, many units response selectively to a particular loudness band.

b. Cochlear Nucleus (Kiange et al., 1988; Rhode, 1991):

i. Ramp frequency (or slope ("chirp") detectors (for formant transitions).

ii. Pauser neurons (emit a onset spike and then continue to accumulate input from other neurons to some threshold. Because the off period varies across pauser neurons, they are ideal for timing voice onset and vowel duration).

iii. Onset neurons (has characteristics by a transient increase in rate at the onset of tone burst, with aspiration to spontaneous level after the tone is maintained a while).

iv. Off neurons (for signalling event onsets that are context dependent on the termination of other events).

c. Primary Auditory Cortex:

Layers 4a and 4b: Delta Loudness Neurons. Some 4b cells of the monkey

fire in response to voice onset - an increment in loudness preceded by a smaller increment in loudness of a critical minimum duration (Steinschneider, 1992).

d. All Locations

i. Off Detectors

ii. Product neurons, neurons whose output is a function of the product of some set of its weighted inputs, rather than the sum.

The segments of the cochlear nucleus are defined by different concentrations of the neuron types described above. The segmental organization suggests that each segment is dedicated to a different class of acoustic feature analysis. If so, the cochlear nucleus is the first of a series of nuclei en route from the cochlea to the auditory cortex, by which salient information in the acoustic signal is extracted so that the computations of the auditory cortex can be carried out on a spatially represented, but greatly compacted signal.

2 Theory of Neural Networks

2.1 Introduction

The basic processing element in the human brain is the neuron. The neuron has several dendrites, usually branched, which receive information from other neurons and single axon which outputs the processed information usually by the propagation of a "spike" or an "action potential". The axon ramifies into various branches that make synapses onto the dendrites and cell bodies of other neurons (Kandel & Schwartz 1992).

Artificial neural networks or connectionist models have risen to some prominence over the last few years. They are modelled on the gross structure of the human brain, a collection of nerve cells, or neurons, each of which is connected to as many as 10,000 others, from which it receives stimuli: inputs and feedback, and others to which it sends stimuli. Some of these connections are strong, others are weak. Whatever the name, all these models attempt to achieve good performance via dense interconnection of simple computational elements. The neural networks used by engineers are only loosely based upon the structure of the brain. Since we have only fragmentary information about how the brain works, it will be a long time before we can recreate in a machine all the capabilities of the human brain.

Neural network models have greatest potential in areas such as speech recognition. Instead of performing a program of instructions sequentially as in a conventional computer, neural network models explore many competing problems simultaneously using massively parallel networks composed of many computational elements connected by links with variable weights.

Several artificial neural network models have been studied in the past to achieve automatic human like speech recognition. These models are composed of many nonlinear computational elements operating in parallel and arranged in certain way (Waibel, A. 1989; Kohonen, T. 1988). The earliest speech recognition schemes were heavily influenced by naive ideas about phonetics and a fascination with analog electronics. Some of these systems worked surprisingly well on carefully pronounced speech. By the late 1970s small vocabulary recognition algorithms were in commercial use. These were most reliable for a single speaker using isolated words. The advent of hidden Markov modelling (HMM), in the early 1980s, has since enabled recognition system to become speaker independent, even over telephone lines. The more difficult task of recognizing continuous fluent speech has been tackled in the 1980s by returning to the concept of phoneme recognition, this time via acoustic elements called, 'sub-word units'.

Hidden Markov models use a stochastic model of individual word production.

In this case, the similarity measure is the probability that a word model produced the observed speech. However, this method does not account for the specific discriminations that are required in a recognition system that employs a particular vocabulary, since the standard training algorithms for these techniques only employ within class information. Hence, it might be expected that these systems would not perform particularly well on highly confusable vocabularies in conditions that produce a large range of differences between different utterances of the same word (Makhoul, Jelinek, Rabiner, Weinstein, & Zue, 1989).

2.2 Neural Networks - The Basic Structure

A neural network consists of a collection of processing elements. Each processing element has many input signals, but only a single output signal. The output signal fans out along many pathways to provide input signals to other processing elements. The processing that each element does is determined by a transfer function, which is a mathematical formula that defines the elements output signal as a function of whatever input signals have just arrived and the adaptive coefficients present in the local memory. A neural network is divided into layers. Depending on the design of the neural network, the processing elements either operate continuously or are updated episodically. A scheduling function determines

which way and how often each processing element is to apply its transfer function.

Computational elements or nodes are connected via weights that are typically adapted during use to improve performance. Every connection entering a processing element has an adaptive coefficient called a weight assigned to it. This weight, which is stored in the local memory of the processing element, is generally used to amplify, attenuate, and possibly change the sign of the signal in the incoming connection. The transfer function sums this and other weighted input signals to determine the value of the processing elements next output signal. Thus the weights express the strength of the connection from neighboring processing elements.

The weights are modified in a backpropagation network. Most transfer functions include a learning law which modifies all or some of the weights in response to the input signals and the values supplied by the transfer function. In effect, the learning law allows the processing elements response to change with time, depending on the nature of the input signals.

In supervised training, the network is supplied with both input and desired output data. After each trial, the network compares its own output with the correct answers, corrects any differences, iterating until the output error reaches an acceptable level. After training, the network is ready to process new inputs.

3 Network Architecture of HearNet

3.1 Speaker Independent Speech Recognition System: HearNet

The network architecture of HearNet (Figure 1) consists of an artificial auditory filter system, FilterNet, and a learning neural network, LearnNet for the perception of speech.

The FilterNet is a series of parallel interconnected auditory artificial neural filters designed after the human auditory system outlined in Chapter 2, and LearnNet is a recurrent backpropagation network that learns phonemic features of speech.

FilterNet is designed to approximate characteristics of the auditory filter system so that the speech perception performance of HearNet will approximate that of the human listener - more so than do those of systems built with general pattern recognition algorithms.

The design of FilterNet is guided by the performance of the entire HearNet system, of which it is the major part. The process of construction is a continual cycle of testing, modification, testing and so on.

LearnNet learns both to select the FilterNet information that is useful for recognizing the features of a particular language and to compute the linear and

nonlinear combinations of this filter information so that each phoneme is accurately identified.

Although all normal humans learn to recognize speech, the heterogeneous ways in which this learning is accomplished, provides a basis for preferring one model of the process over another. LearnNet, the learning model used here, which is a modified back propagation model, is employed as a way of evaluating FilterNet, and does not claim to be a definitive model of all of the interacting ways in which speech perception is learned.

Recurrent loops within many of these filters allow them to sustain their activity until it is overridden by new information. These recurrent loops take advantage of speaker's identity and capture the relevant features that carry over from prior speech. As the accuracy of this output increases, LearnNet is provided with an increasingly distinct basis for discriminating among ambiguous and noise embedded linguistic units. In the feedforward or recognition mode, this increased latency mimics that of human listeners to noisy or ambiguous speech.

In general, sensory neurons have higher maximum firing rates and they habituate more rapidly than cortical neurons (Tucker, & Williamson, 1984). As the fine grain temporal changes in acoustic information are selectively filtered by nuclei in the auditory pathway the compacted information may be passed on, without

information loss, at a slower rate. FilterNet simulates this slowing by collapsing information over adjacent time frames. Collapsing adjacent pairs of frames cuts the input cycles to LearnNet in half so that a monosyllabic word may occupy 90 rather than 180 frames. In addition to reducing computing time, this accelerated learning by decreasing interference from adjacent words.

3.2 An Artificial Auditory Filter System: FilterNet

FilterNet carries out parallel sequences of transforms of the power spectra of the acoustic signal (.05 - 4.5 Khz, in 22 ms. Hamming windows in successive 1/2 window steps; energy is expressed to the power of 0.6 and, within each band, is squashed between 0 & 1, except where noted; Note, window size determines sensitivity to formant slopes). Filters are constructed from units and networks of units whose links and weights are set by the designer and are not modified by training. Each filter selectively passes different forms of context-dependent information. For example, the Auditory Nerve Filter includes 1) broad 2 octave bands in 1 octave steps to 4.5 kHz. (Figure 6), 2) narrow frequency band 1/3 octave units in 1/6 octave steps to 4.5 kHz. (Figure 5), and 3) sets of harmonically ordered intervals within the 0.1 - 2 kHz range (Figure 7). (Note, harmonically ordered values will be called frequencies; octave or log ordered values (Miller, 1989), pitch). Units

representing 3 levels of loudness are nested within each frequency band unit. FilterNet consists of 2226 units, of which 145 are Auditory Nerve units and 229 output to LearnNet.

The steps involved in FilterNet are: 1) the selection of the next phoneme that HearNet is to be trained to discriminate. The phoneme should be a neighbor to one or more of the previously studied phonemes. For example, all of the bilabials should be modelled before starting on the alveolars or nasals so that the necessary degree of discriminability is achieved for the most confusable phoneme neighbors. The following vowel: /æ/, the consonants: /b/, /p/ and /d/, the fricative: /s/ have been evaluated, and only through step 7. 2) Update the literature review and consult on all of the linguistic, psycholinguistic, auditory neurophysiologic, and speech recognition research related to the phoneme and its phonemic neighbors, 3) on the basis of this review, identify all of the subphonemic information that LearnNet must have to learn to discriminate this phoneme from its neighbors, 4) describe the neurons and networks in the auditory system that may function as filters for this information, 5) modify existing networks, or add new artificial neurons and filter networks to simulate these natural filters, 6) using monosyllabic words, both isolated and in continuous speech, that contain the relevant phonemes recorded by a small heterogeneous set of speakers, examine the output of individual

nets within FilterNet to determine whether the critical information successfully passes each of the appropriate filters. If unsuccessful, return to step 5. If successful, 7) link FilterNet to LearnNet and train the network with 20 speakers to discriminate this phoneme from each previously learned phoneme. If recognition of the new phoneme is poor or if response to previously discriminated phonemes is degraded, evaluate each part of the system starting at step 5. If successful, 8) test for generalization to 20 independent speakers. If unsuccessful, carry out additional generalization tests, described above, to determine the type of generalization failure and redesign the system accordingly. If successful, return to 1), and repeat until all of the phonemes are successfully recognized.

Coarse Frequency Coding Filters

Much of the variance in the speech signal may be reduced by taking account of the multiple contexts of the acoustic components. For example, the **b - p** difference in voice onset time (VOT) is the difference in the duration of the time between 2 successive increments in loudness: the first an increment above the ambient noise, the second an increment above the last increment. The loudness increment, from the stop to onset of voicing is also conditional on the loudness of the voicing, and on the prior level of voicing, and the pitch of the stop. To create

a refractory period for **b - p** discrimination based on the amplitude, a delta-db unit which has a slow decay unit with a feedback weight of +0.9 is used, so that the delta unit cannot fire again until the slow decay dissipates. The array of ON units compute small increments in loudness, and the OFF units indicates the end of burst (Figure 6). The other difference between /bad/ and /pad/ is the overall shape of the spectrum. For example, just prior to the onset of voicing, a weak burst of turbulence noise is generated at the constriction of the syllable /bad/, resulting in a burst spectrum. This increment in energy is computed by Figure 6-2. The spread of energy for /spad/ spectra is more diffused and the overall shape of the spectra is quite different. The frication which is present in /spad/ is detected by Figure 6-3.

Starting with the three classes of Auditory Nerve input, a series of filters compute different context-dependent functions. Some phonetic features such as fricatives and affricatives consist of broad frequency bands of white noise. On the assumption that the precise frequency boundaries and center frequencies of these bands are not critical for phoneme perception and that the contour of the loudness envelop is simpler to compute when the frequencies are coarse coded, a variety of loudness contour mini filters can be constructed. Loudness comparisons across different time frames is accomplished with delay lines (Hampshire & Waibel, 1990). The slopes of the contours is computed by adding Pauser units. One such

CoarseFreqLoudPosDelta network (Figure 6) was designed to produce the voice onset response observed by Steinschneider (1992) in the monkey cortex 4a layer. When linked to LearnNet, it provided sufficient information to discriminated /pa/ from /ba/.

Harmonic Filters

The harmonically ordered units of the Auditory Nerve Filter are arranged to form a series of harmonic sieve filters (Cohen, Grossberg & Wyse, 1992, Figure 7) that locate values of F_0 in the 0.1 - 0.25 kHz range. Each F_0 unit is linked to 6 harmonic intervals, and is inhibited by the inharmonic background. Higher harmonics are not discriminative. These F_0 values provide the contexts for several later filer nets. A Pauser F_0 Filter uses Pauser units to time F_0 duration, which in turn, is used in a cortical Formant Interval Filter net to compute and represent the pitch interval distances between F_0 and the speech formants, whether moving or stationary. LearnNet relies heavily on this matrix for vowel identification. A set of Echoic F_0 filters, Pauser Units with slow decay times, compute a running average of both F_0 pitch and duration, both of which contribute to a stable speech normalization process - even when the speaker is not voicing. Each of the above values is represented by 3 units of graded loudness sensitivity.

Formant Transitions and Formants

Filters must be able to pass all of the format pattern information that human listeners use to perceive vowels and voiced consonants, and they must deliver it in the form that listeners appear to use. The design of the current set of filters follows the dynamic specification approach of Strange (1989). It also incorporates the log frequency scale, and the sensory reference (current, fast and slow changing values of F_0) and auditory-perceptual space concepts proposed by Miller (1989), but without the constraints imposed by his search for a linear model of vowel perception.

It is proposed that the human peripheral auditory system analyzes auditory stimuli with approximately a 1/3-octave frequency resolution (Searle, Jacobson, & Kimberley, 1980). Because bandwidth is proportional to frequency, 1/3-octave bandwidth allows spectral resolution of low frequencies as well as temporal resolution at high frequencies. Spectral resolution of low frequencies enables separation of the first and second formants, while temporal resolution of high frequencies provides accurate timing information for the rapid onset of bursts.

Because formants are rarely stationary, even in the sustained portion of a vowel, filters must respond selectively to the moving frequencies of individual formants and changing distances between adjacent formants. The key to these filters

is the Ramp Pitch unit, (Figure 5) modelled after ramp frequency neurons in the Cochlear Nucleus. Using delay lines of up to 4 ms., they respond (as Sigma Pi units) selectively to a pitch change in the 1/3 octave scale. Although the ramp filter can discriminate a wide range or slopes, the current model groups all slopes into positive, negative or stationary. Although slopes may be computed from the edges or other loudness gradients, the current Ramp Pitch filter uses lateral inhibition to locate the formant mean pitch and computes the formant ramp from these running means. Pauser units measure their duration.

This local first derivative is passed on to a larger Relative Pitch filter net - putatively a cortical net - that represents this information as pitch distances between dominant pitch values, particularly between F_0 and higher formant pitches. Representing stationary and ramp pitches as distances from F_0 minimizes interspeaker variance in inter formant distance due to age, gender and other vocal tract variation, as well as differences associated with dipthongization and affect.

3.3 Speech Perception Learning System: LearnNet

LearnNet is a back propagation network with one hidden layer, and one recurrent "Echoic" state layer (Figure 10). The Echoic layer is a copy of the Out layer from the previous time window. LearnNet accepts input from FilterNet in 11 ms. cycles. This information is transformed at the Hidden layer and passed to the Out layer where it is joined by the output of the Echoic layer. Learning or correction of unit weights occurs only at the Hidden and Out layers. The Echoic layer holds a 11 ms. old memory of the Out layer. But the Out information can be recycled indefinitely depending on the learning demands of the Teacher layer. Therefore the duration of the Echoic memory is not fixed.

Although no claim is made that the Echoic layer represents what is often called echoic memory (Crowder, 1982), the Echoic layer may represent the ability of units in auditory cortex columns to pass information back and forth among each other. It may be noted that many units in FilterNet have recurrent loops with fixed weights that provide auditory memories that vary with the characteristics of each filter. The recurrent loop is biased to capture the relevant features that carry over from prior speech. This is very critical in distinguishing the speaker's identity.

LearnNet's success to date suggests that it provides a satisfactory method for evaluating and comparing different models of FilterNet. The use of a

backpropagation model, with a recurrent loop that includes a state layer, to learn serially ordered events was first proposed by Jordan (1986). In his model, the state layer receives input from the output layer and returns it to the hidden layer. Norris (1990) suggested a modification in which both the input and output of the State layer are connected to the hidden layer. In an unpublished study, using binary coded artificial speech as input, Antrobus and Alankar (1992) compared three recurrent models (Figure 9) and found that speech recognition was most accurate with the architecture described in Figure 10. The superiority of this model was greatest when there were no gaps between successive words.

In order to learn that a particular temporal interval in the power spectra of a particular speaker contains a particular phoneme or feature of speech, the feature and its location is first identified and coded by an expert user of the language. The code is the criteria by which the back prop Teacher guides the learning - the changes in the weight structures - that identify a particular phoneme or feature. The processes of back propagation of error by which the network adjusts its weights over a series of exposures to multiple exemplars so that its output approximates the Teacher code has been described by Rumelhart, Hinton, and Williams (1986).

The Alignment of Acoustic Information with Teacher Code

The multiple sequential dependencies in the motor articulation of speech, called coarticulation, create corresponding overlapping dependencies of the subphonemic features conveyed by the auditory signal. Learning to identify subfeatures that overlap and modify one another is a straight-forward task for a backpropagation network because mutual modification is a form of nonlinear interaction that back prop models are designed to solve. But coarticulation blurs the temporal boundaries between adjacent linguistic features so that it may be difficult, even for an expert judge to determine the point at which the information relevant to a particular feature begins or ends.

The recurrent loop in LearnNet eliminates the need for exact alignment of Input and Teacher. We have demonstrated that LearnNet can identify phonemes in a serially presented monosyllabic whole word format. Back propagation network should be able to learn to recognize a particular feature or phoneme in any coarticulated word context if the Teacher has been supplied with a sufficient number of coded examples. It is unnecessary to align the teacher code with the acoustic interval within the word. On the other hand, if the task is to identify several or all of the phonemes in each word, the Teacher must have a separate slot for each possible phoneme in each possible ordered position. Again, alignment with the

coarticulated and independent portion of each phoneme is unnecessary. Once HearNet has learned the phonemes, recognition proceeds serially and all of the relevant phonemes remain active until presentation of the word is complete. Because redundancy exists in the order of most phonemes, HearNet shows, after the first phoneme, partial activation of the likely subsequent phoneme candidates even before they are presented, and the correct phoneme may become fully active when the minimal discriminative filter information arrives. In this respect, HearNet demonstrates some part of the response latency behavior that human listeners exhibit.

4 Experimental Results

4.1 Stimuli Recording Procedure

The database used in this research contains monosyllabic CVC words of 25 male speakers. It consists of words which were difficult to discriminate. It is made up of the stop consonants, /b/, /p/, /d/; the fricative /s/, and the vowels /æ/ and /a/. The words used for this research were, /bad/, /pad/ and /spad/. The recording was done in a relatively quite room. Speech was recorded digitally at 11 kHz. directly on a 80386 DOS machine using a ProAudio Spectrum Thunderboard with additional software developed in house. The digitized speech was transferred through the local Sun 4 network back to the, SUN 3/60 in our lab. The FFTs were performed on the speech data using HIPS2 software (Landy, M. 1990). The networks of FilterNet and LearnNet were built within the Parallel Processing Simulator developed by the Rochester computer science department (Goddard, Lynne, Mintz, & Bukys, 1989).

The consonant-vowel-consonant (CVC) were pronounced in continuous speech by native American speakers (25 male and 40 female), some of whom had regional accents. Speakers were asked to read the words in sets of 8 words, each set beginning with 'dog' and ending with 'dig'.Loudness and pitch are highest at the

beginning of an utterance and then gradually decline (Tartter, C. V. 1986). Therefore, the first and last words were not part of the experimental learning set. The remaining 8 words in each set were arranged in random order.

Energy values in the acoustic signal are extracted in descending order using a cascade of bandpass filters. Energy is extracted in three different ways. 1) broad 2-octave band-widths in 1-octave steps to 4.5 kHz., 2) narrow frequency band-widths of 1/3-octave set at 1/6-octave steps to 4.5 kHz., and 3) sets of harmonically ordered intervals within the 0.1 - 2 kHz. range. Each word was analyzed using a Hamming window, and then frequencies were extracted using a 256 point FFT. The resulting DC spectrum was logarithmically spaced on a frequency scale. Energy was then raised to the power of 0.6 (Green, 1988).

Energy in each band-width was represented by the activation of the corresponding input neurons of FilterNet. The data was fed into HearNet every 11 mss. in the following manner: Words of one set consisting of 4 different vowels and 4 different consonants each set preceded by the word 'dog' of one speaker, followed by a similar set of another speaker and so on. The initial word, 'dog' was used to help the network acquire a normalized framework for each speaker.

After each set of 8 words (plus 'dog'), a set from a new speaker was introduced. Long words lists with a single speaker were avoided in order to prevent

catastrophic forgetting, a weakness of back propagation where old learned material is lost when new material is learned.

4.2 Experiment Technique

The LearnNet has three layers: an input layer with 229 units, corresponding to the compacted output of FilterNet; a hidden layer with 30 units; an output layer with 9 units, whose number of units corresponds to the number of coding units for the Teacher; and a state layer with 9 units. Throughout a learning rate of 0.1, and a momentum term of 0.1 was used. The LearnNet consists of the following configuration:

Input Units

The number of input units to the HearNet is 229 which is the output of the FilterNet. This input is the combination of the outputs from the broad-band 2-octave net, the sieve net and the narrow-band 1/3rd octave net is number of input units is already in a compacted form and may still be reduced, which is left for further research.

Hidden Units

Hidden Units

The number of hidden units is 30 and the number of layers is one. Not much guidance from any literature is available for the exact number of hidden units. The number 30 was derived from several experiments.

Output Units

The number of output units is 9 which is the used to discriminate the phonemes. The code for output units is in binary form as shown in table 2.

State Units

The number of state units is 9 which is used as a memory buffer to hold the speech information.

Teacher Code

The code for teacher units for each word follows the principle with all initial consonants /b/, /p/ and /sp/ in the beginning, followed by the vowels /æ/, and /a/, followed by the last consonant /d/ (see Table 2).

e.g. /pad/ = 01001001

Connections of the HearNet

The initial range of weights of the network is -1 to +1. All the 229 input units are connected to all hidden units with initial random weights ranging from 0.001 to 0.01. The hidden units are then connected to all the output units with initial random weights of 0.005 to 0.1. The output units are connected to a state layer with a fixed weight of 1, and with a delay of 2. The state units are then connected back to the output units in a one-to-one fashion with random weights of 0.01 to 0.13. The delay of 2 for the link from the output units to the state units was chosen so that information at time 1 being held at the state units is available when the information at time 2 from the hidden units arrives.

A momentum term is added to the backpropagation network, which involves adding a term to the weight adjustment that is proportional to the amount of the previous weight change. Once an adjustment is made it is remembered and serves to modify all subsequent weight adjustments. Using the momentum method, the network tends to follow the bottom of narrow gullies in the error surface, if they exist, rather than crossing rapidly from side to side.

It goes without saying that a speech perception system should first be able to recognize, as well as human listeners do, the consonants and vowels on which it has received training. But the strength of the system is based on its success in

generalizing this perceptual ability. Several of the critical forms of this generalization are described here.

The criterion that most sharply distinguishes existing word recognition systems from human listeners is recognition of the speech of novel speakers. A major effort of this project is to identify the neural computational basis of this human capability. A standard criterion for good generalization is an index of recognition accuracy based on the speech of 100 speakers that is close to the index for the independent speakers used to train the system. Additional aspects of cross speaker generalization were studied by testing for phoneme recognition in sets of speakers not represented in the training set, such as the alternate gender, or a set of speakers with an Hispanic accent, or rapid speaking rate. In general, artificial intelligence word recognition systems can generalize only within the envelope of speech characteristics that they have encountered in their training set.

Generalization

In order to perform a classification task with novel events, a network must first learn how to classify a set of known exemplars of each class. The back approach to this problem is to 'teach' the network the characteristics of each class by presenting many examples of each class. Training may be continued until the

network is able to classify the training utterances perfectly. However, it is not the performance of the system on its training data that is important, but rather its ability to classify previously unseen speech.

The composition of data in the training group should always be representative of the intended user population. The different classes of system for speech recognition are described below.

A speaker dependent system is one where the training data is provided by its one eventual user.

A speaker independent system is one which is trained by several speakers, but eventually used or tested by speakers who have not contributed to its training. If the training data is truly representative of the eventual user population then recognition performance can be expected to approach that of multi-speaker system.

For this dissertation, the study was concentrated on the speaker independent system. Graphs for consonants and vowels are evaluated to determine whether the pattern of generalization errors matches that of human listeners. Root mean squared errors are computed for all words and graphs are drawn with the performance (RMS) versus the frame numbers.

4.2.1 Experiment One

Problem: 'bad'/ 'pad' / 'spad' discrimination.

To demonstrate the importance of time sequence information in a speech recognition system, the problem of discriminating between the utterances 'bad', 'pad', and 'spad' was considered. The stop consonants which have been a subject of research in evaluating neural networks for phoneme recognition were used (Lippmann 1989). These stops possess several common features, but only two distinguishing phonetic features, place of articulation and voicing (Table 1).

The words 'bad' and 'pad' vary in voice-onset time (VOT), defined as the relative onset of periodic pulsing to signal onset. The schematic spectrograms of /ba/, /pa/, and /spa/ are shown in Figure 8. The syllable /spad/ shares with the syllable /pad/, the phoneme /p/. But they are by no means identical. The syllable initial /p/ of /pad/ is aspirated and voiceless. The /p/ of /spad/ on the other hand, is not aspirated and is voiced, when the vocal cords begin to vibrate (Klatt, 1975; Davidsen-Nielsen, 1974). The acoustic structure of /spad/, is such that it is composed of /s/, followed by silence, followed by a 10-ms VOT, which is identical to /bad/.

The spectra at the onset of voicing for a labial and an alveolar consonant differ not only with respect to the frequencies of the spectral peaks at voicing onset,

but also with respect to the overall shape of the spectrum as determined by the relative amplitudes of the spectral peaks (Stevens, & Blumstein, 1978). Just prior to the onset of voicing, a weak burst of turbulence noise is generated at the constriction of the syllable /bad/, resulting in a burst spectrum.

The spectrum at the release of a stop consonant can be influenced by the duration and shape of the window that is selected for examining the spectrum. Wide-band 2-octave units with a window size of 11-ms are used. The spread of spectral energy for the /bad/, /pad/, and /spad/ spectra is more diffused and the overall shape of the spectra is quite different. The syllables /bad/ and /pad/ contains both bursts and transitions.

The reason for choosing the syllables used as the database required an adapting syllable to share its acoustic structure with one end of the syllable, while sharing its phonetic identity with the opposite end of the syllable.

To study speech recognition the traditional approach is to run experiments with mixed speech data, male speech data, and female speech data separately, and then test the recognition system on mixed, male and female data respectively. Though the power spectrum of male and female speech look quite different, there is undoubtedly a great deal of similarity between the male and female pronunciation of a given syllable and vowel.

In order to force LearnNet to pass a difficult generalization test, it was trained with only male speech data and tested on male speech data and female speech data. LearnNet trained on male speech data performed slightly better when tested on male speech data than on female speech data (Table 3). These tables give a quantitative summary on the interaction among different training and testing procedures.

4.3 Experiment Results

The HearNet solutions to various problem set are summarized in Table 3. The performance of HearNet for speakers who vary in gender and age and its ability to generalize to speakers not in the training set is described below.

1.	Training set:	20 males	2.	Training set:	20 males
	Number of Trials:	200		Number of trails:	200
	Testing set:	10 males		Testing set:	35 females
	Recognition rate:	100%		Recognition rate:	99.55%

Conclusions and Summary

5.2 Conclusions

The goal of this research has been to build a speech recognition model by and to investigate the adequacy of neural networks for acoustic phoneme recognition by studying the properties of the auditory system. Based on the results of the research conducted over the past 3 years, presented and discussed in this dissertation, it may be concluded that HearNet is a prototype of a robust, computationally simple and accurate speech recognition system, and that by simulating some of the auditory filter properties of the biological auditory system, neural networks can be efficiently designed for speech recognition.

5.3 Summary

A speaker independent speech recognition system, HearNet was developed consisting of cascade of neural feature detection layers. Each feature detection layer was designed to detect several acoustic feature of the speech signal, such as the voice onset time, increments in loudness, burst, silence, the slope and the duration of the formant transition and the duration of the steady part. The signal was being constantly transformed as it moved through the network so that different transforms

of the signal were available at successive points in time. The variation in the dynamic range of different speakers were also identified.

The output of HearNet consisting of a feedforward FilterNet and a backpropagation LearnNet was the activation of sequences of phoneme-class units that were appropriate to the acoustic speech input with a recognition rate of 100%.

Table 1. Two distinguishing phonetic features between the stop consonants.

	Place of Articulation		
	Velar	Alveolar	Labial
Voiced	/g/	/d/	/b/
Unvoiced			/p/

Table 2. Teacher code for problem set 1.

	First Consonant				Vowel		Last Consonant	
	s	p	b	d	æ	a	g	d
dog	0	0	0	1	0	1	1	0
bad	0	0	1	0	1	0	0	1
pad	0	1	0	0	1	0	0	1
spad	1	1	0	0	1	0	0	1

Table 3. Performance of Hearnet model on problem set 1.

		Performance on Male Data	
Training data	Trials	Training data	Test data (10 speakers)
Male (20 speakers)	200	100 %	100 %
		Performance on Female Data	
Training data	Trials	Training data	Test data (35 speakers)
Male (20 speakers)	200	100 %	99.95 %

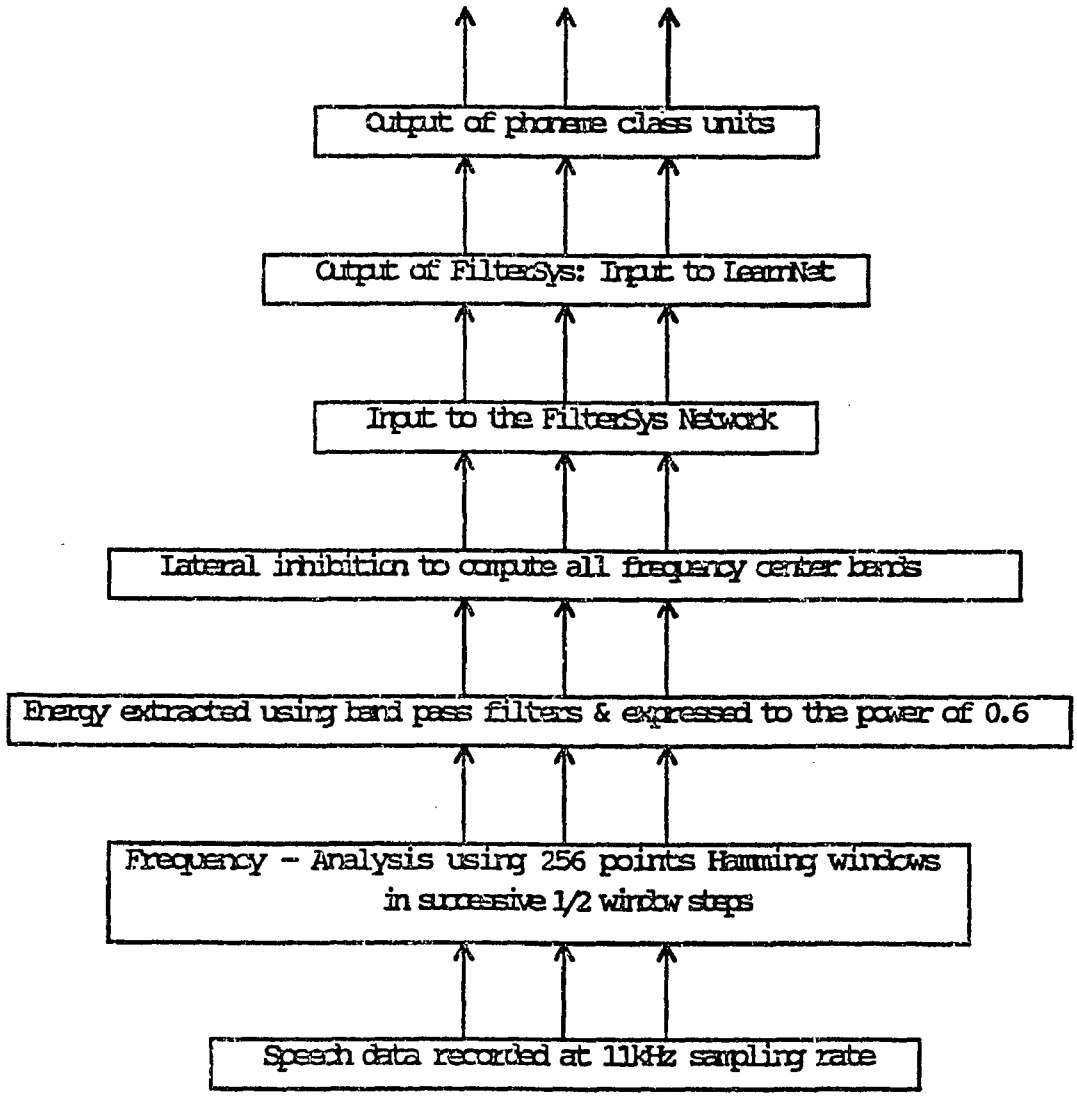


Fig. 1: Flow chart of the HearNet Model.

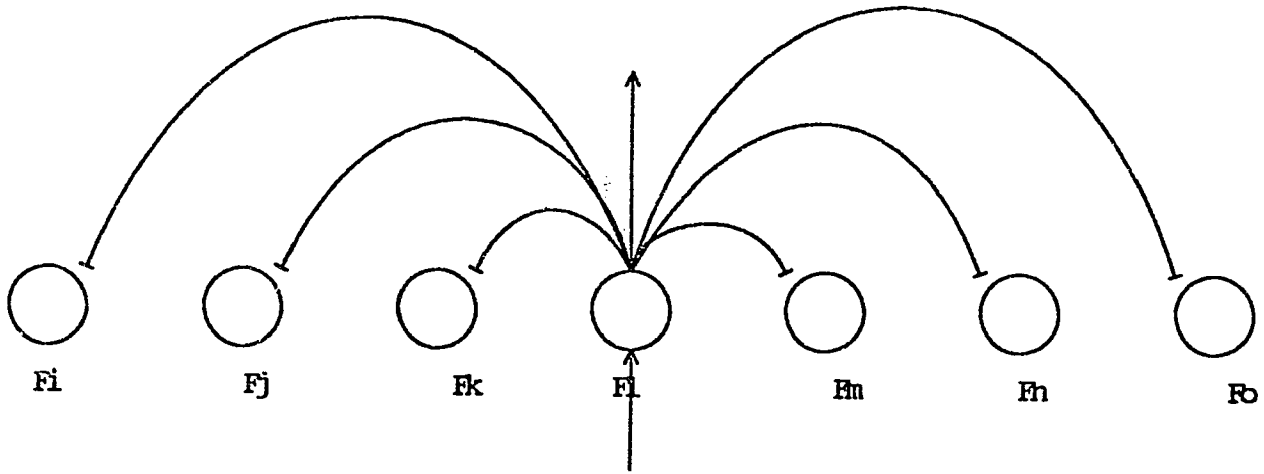


Fig. 2: Lateral inhibition to compute all frequency centre bands.

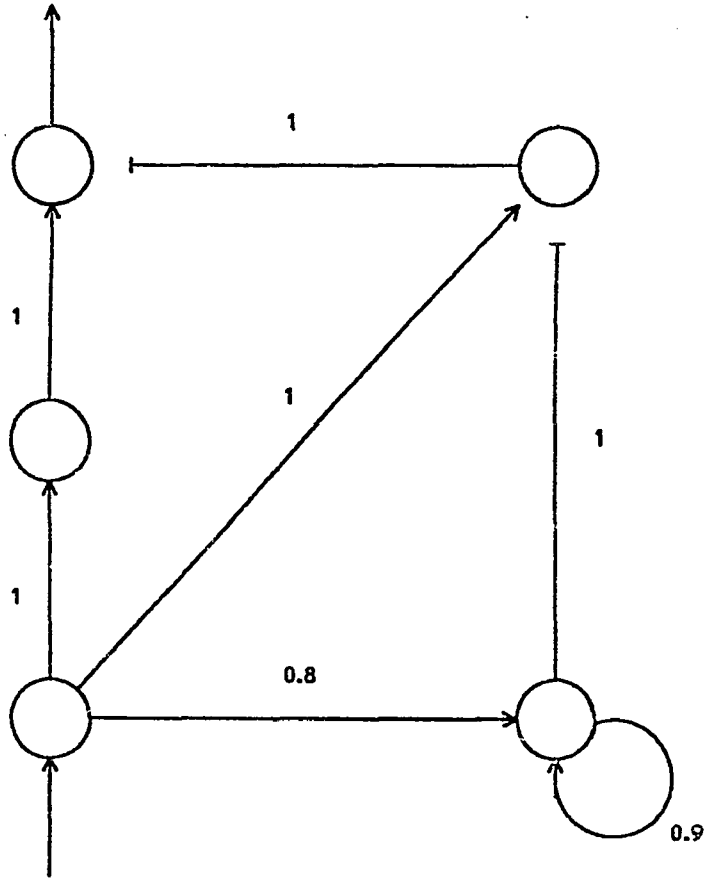


Figure 3. Pauser Unit - for timing. By changing the feedback weights, the pauser unit can be used to time a duration.

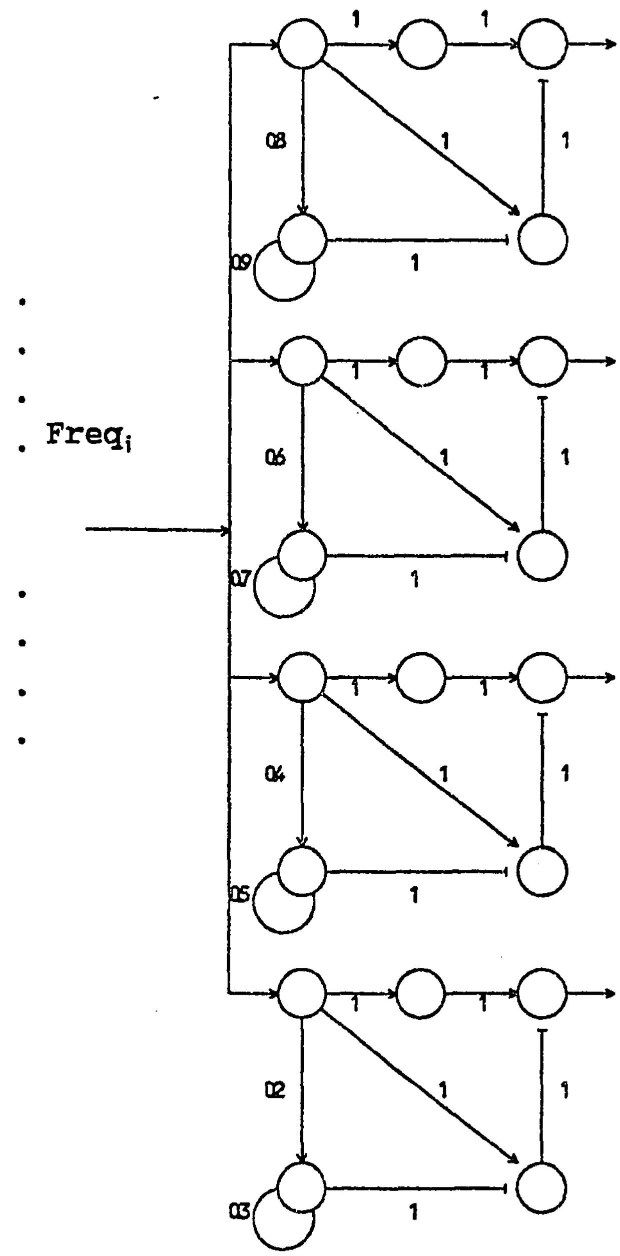
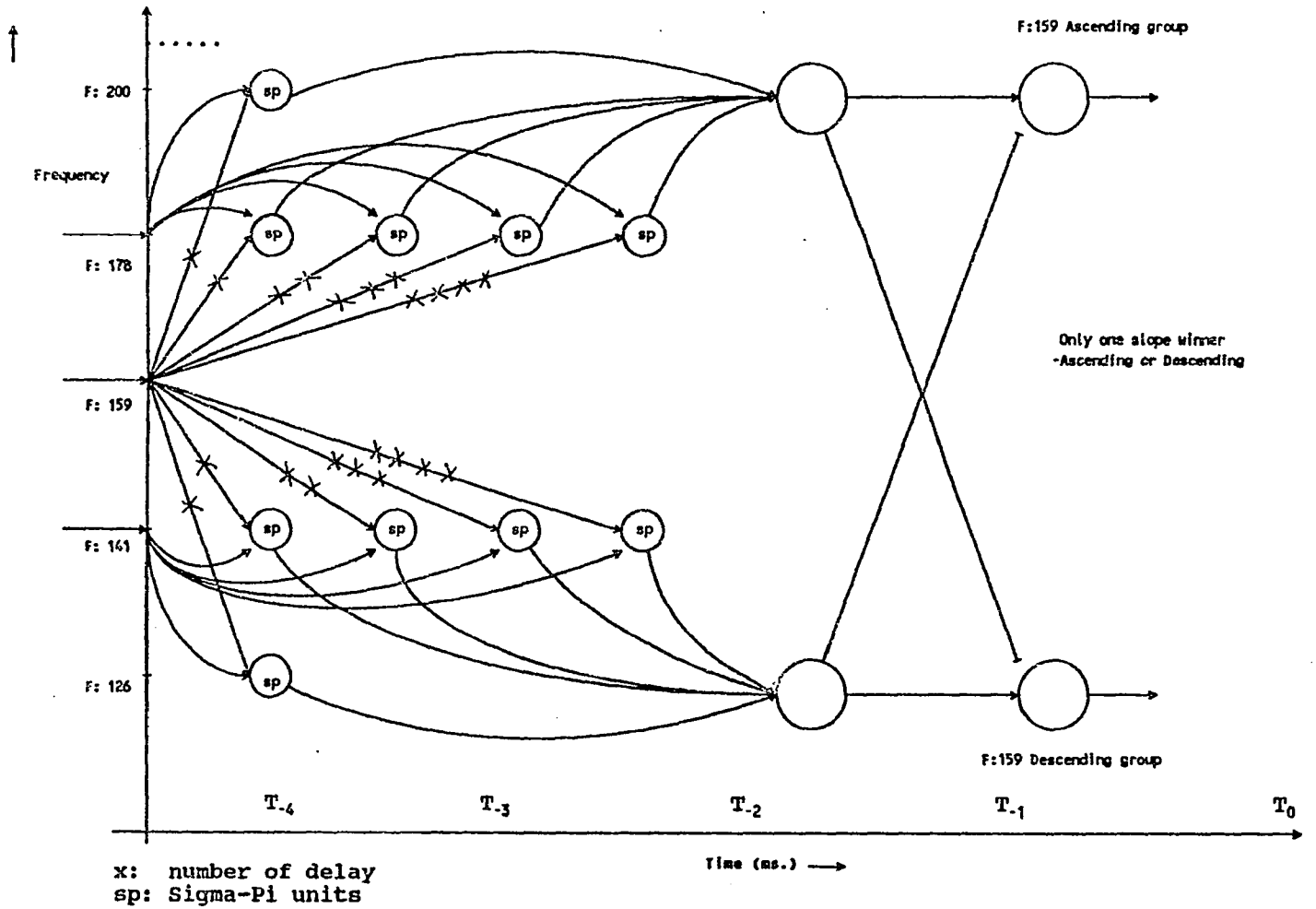
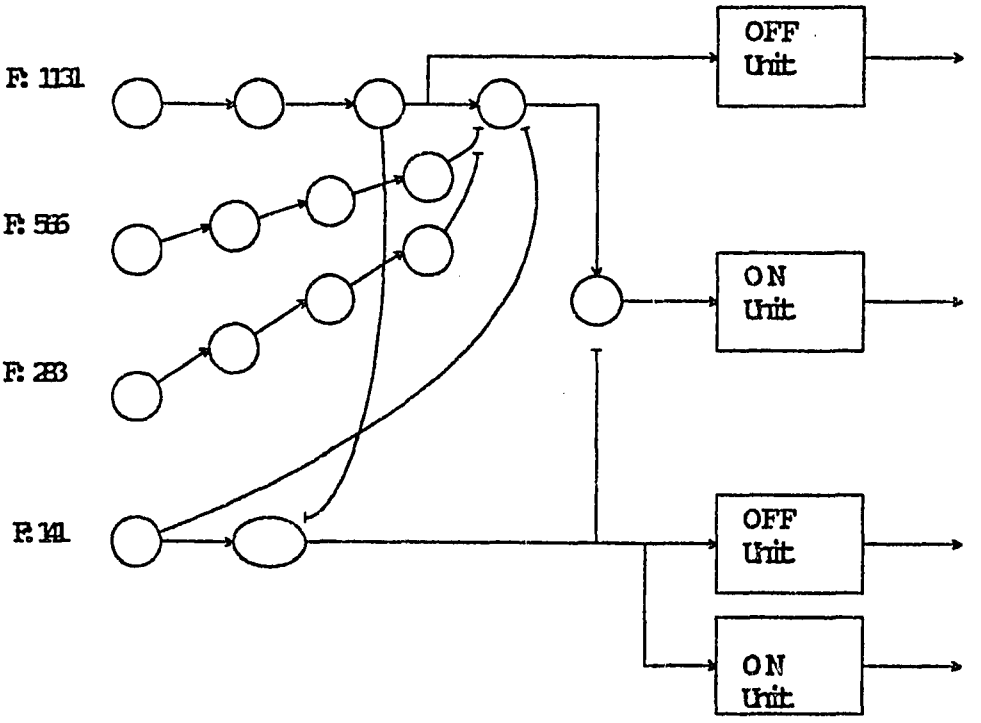


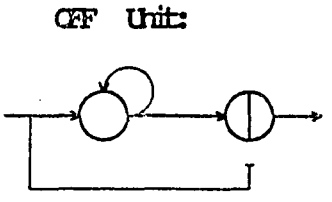
Figure 4. Interval frequency center band units linked to a set of 4 Pauser units with different feedback weights to compute the duration of the stationary part of the vowel.

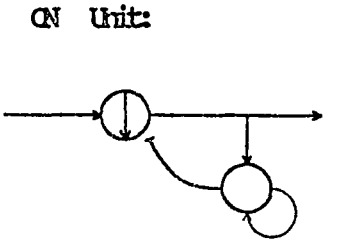


x: number of delay
sp: Sigma-Pi units

Figure 5. Ramp pitch units using delay units of up to 4ms responding as Sigma Pi units, selectively to a pitch change of the 1/3 octave scale Slopes are grouped as ascending or descending.







: fires only when its threshold is greater than 800.

Figure 6-1. Delta loudness units with 2 octave band units, ordered in 1 octave steps to 4.5 kHz. The array of ON units compute small increments in loudness.

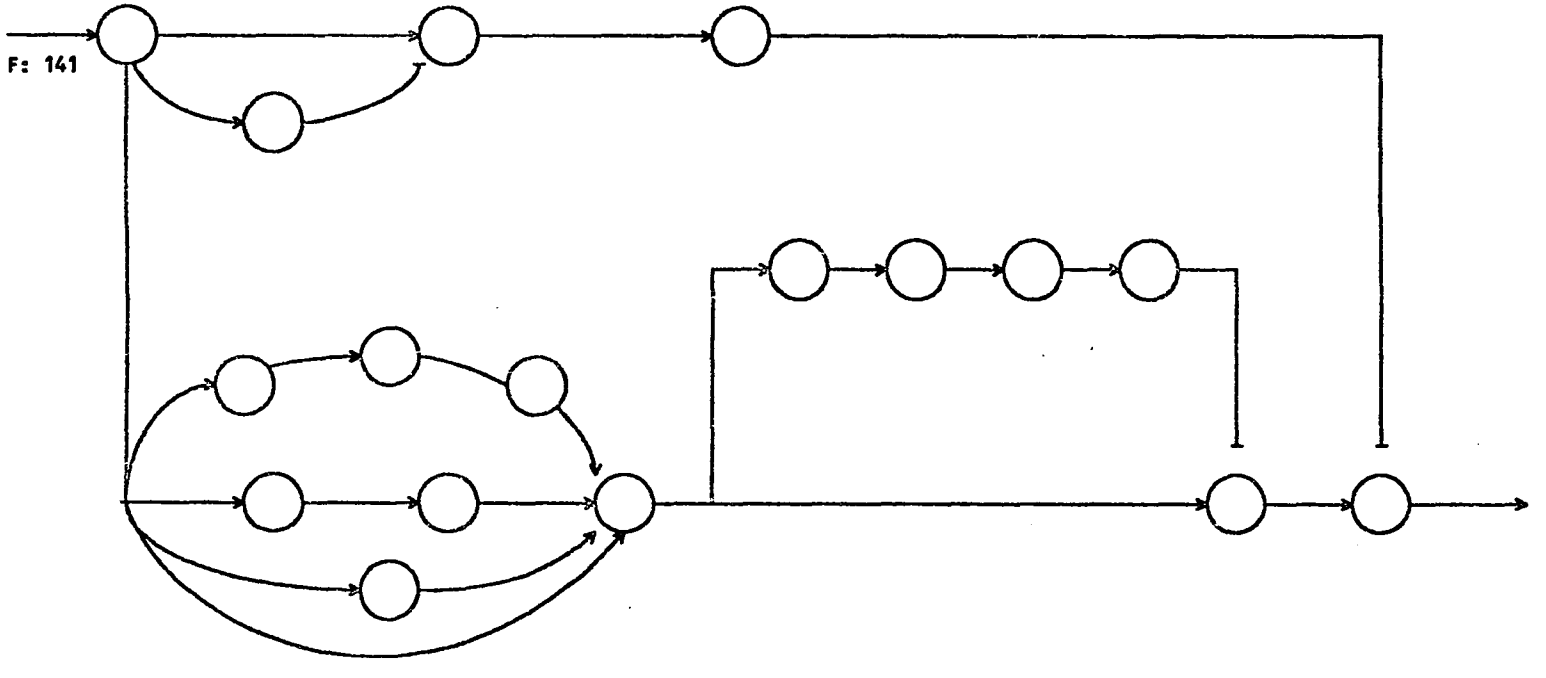


Figure 6-2. Delta loudness units with 2 octave band units, ordered in 1 ocatve steps to 4.5 kHz, computing the increments in energy.

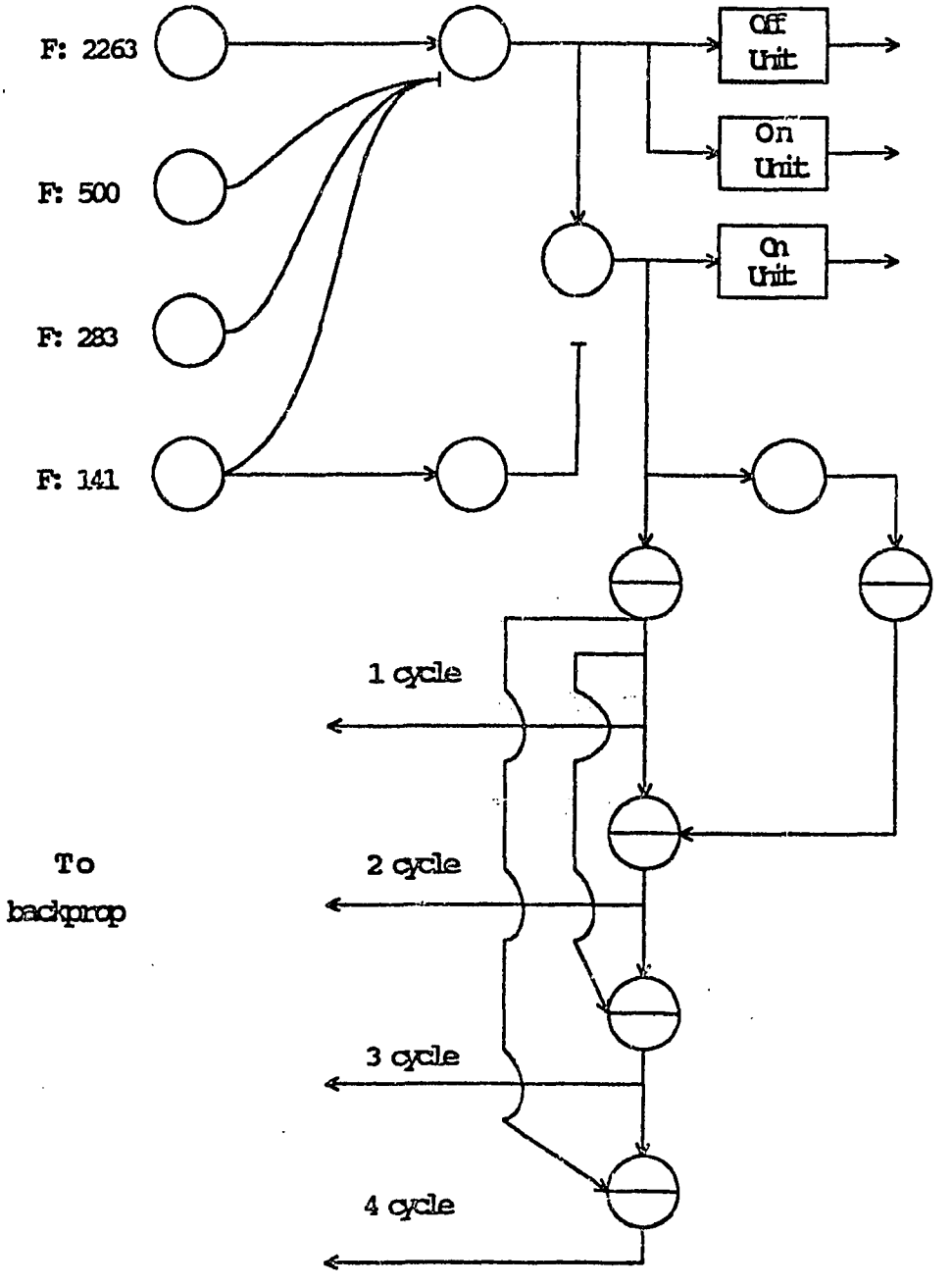


Figure 6-3. Delta loudness units with 2 octave band units, ordered in 1 octave steps to 4.5 kHz. The array of ON units compute large increments in loudness.

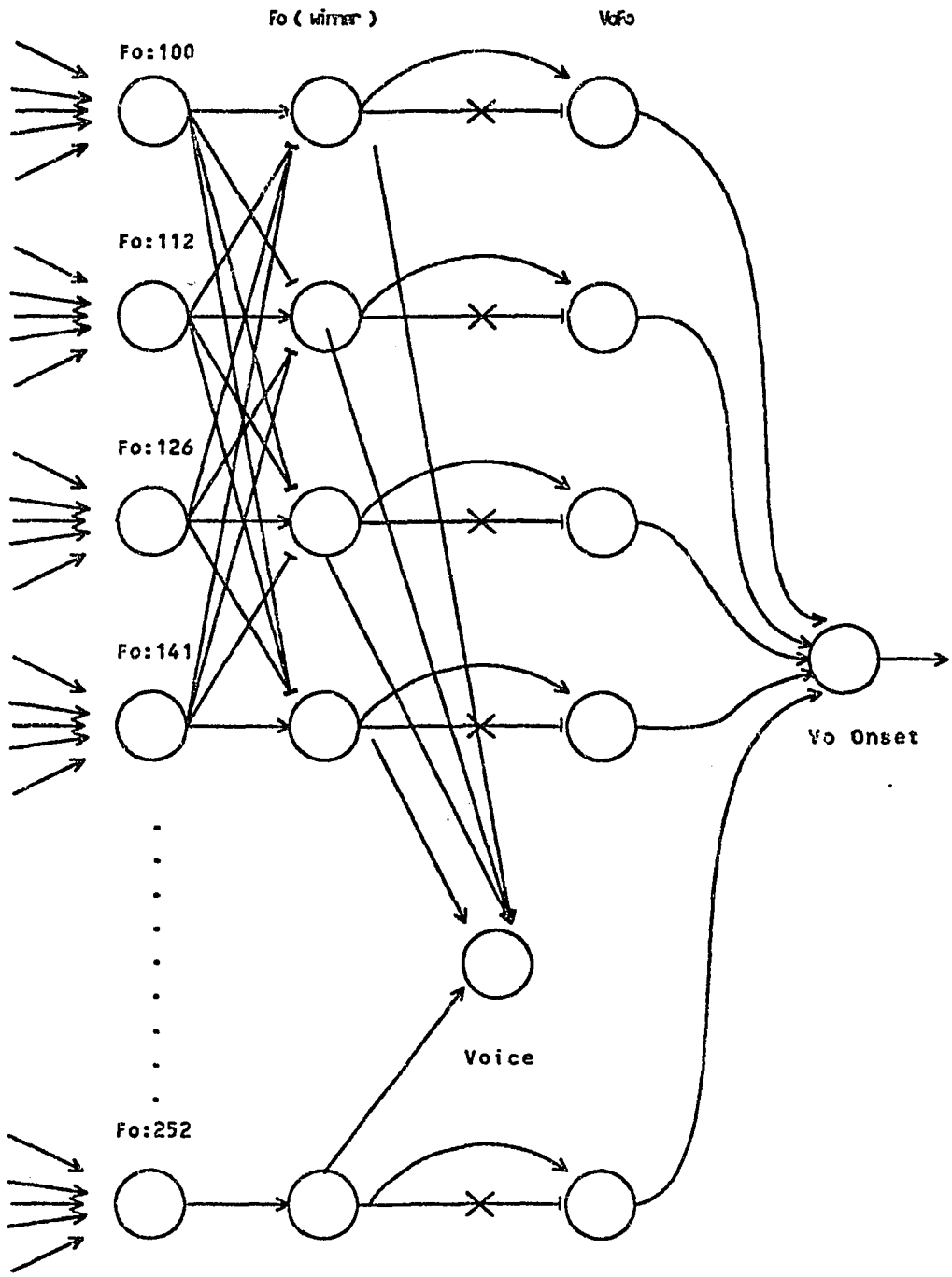


Figure 7. Sieve-Network: Sets of harmonically ordered intervals within the 0.1-2kHz range. Computes the value of the fundamental frequency Fo.

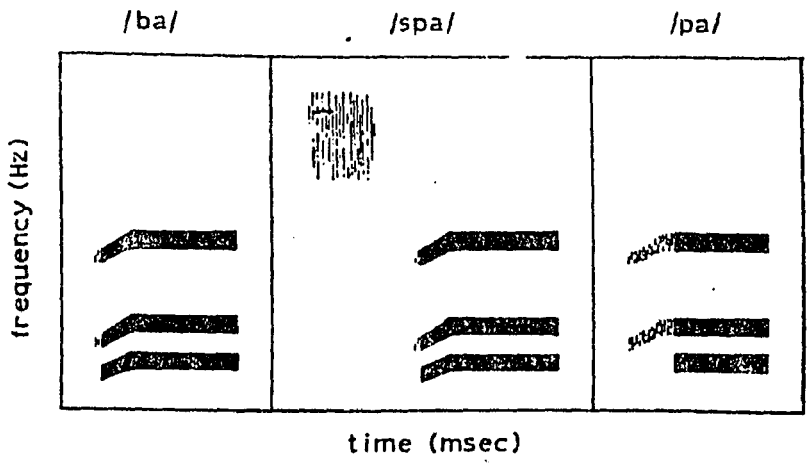


Figure 8. Schematic spectrograms of the syllables /ba/, /spa/, and /pa/.

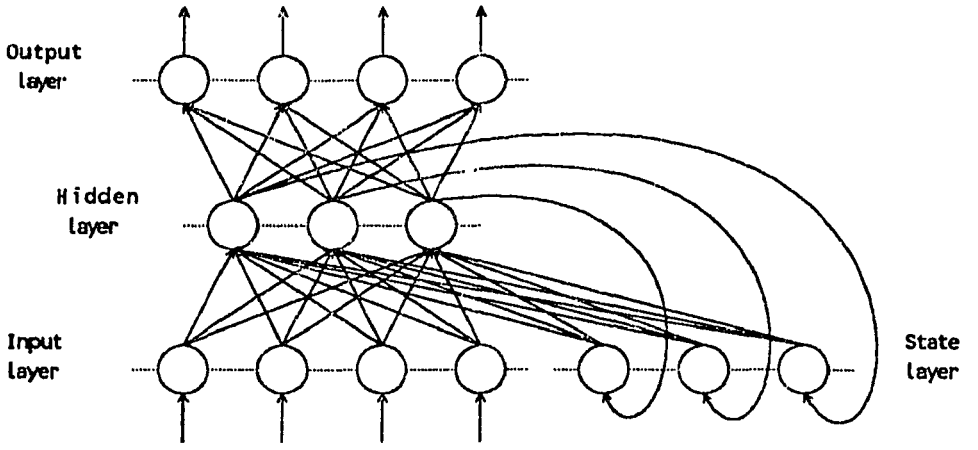


Fig. 9A: Loop from hidden layer back to state layer back to hidden layer.

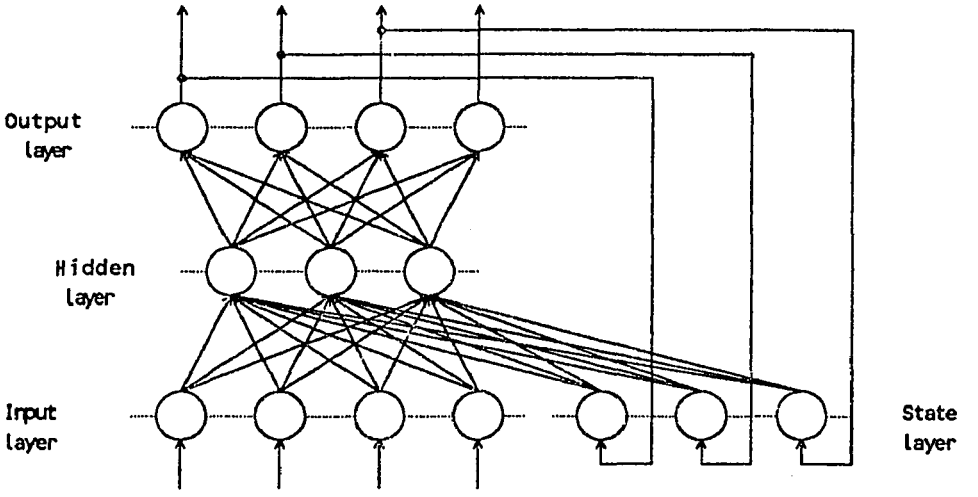


Fig. 9B: Loop from output layer back to state layer back to hidden layer.

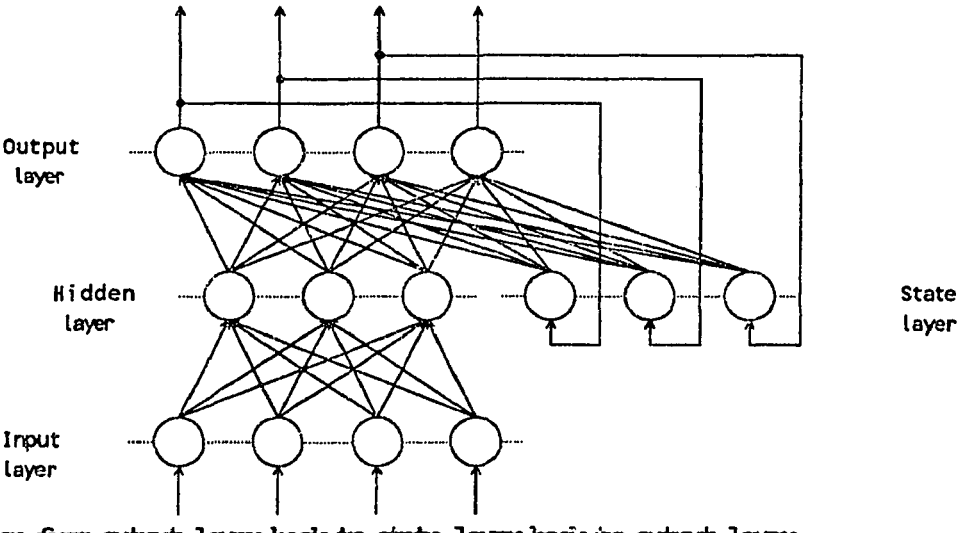


Fig. 9C: Loop from output layer back to state layer back to output layer.

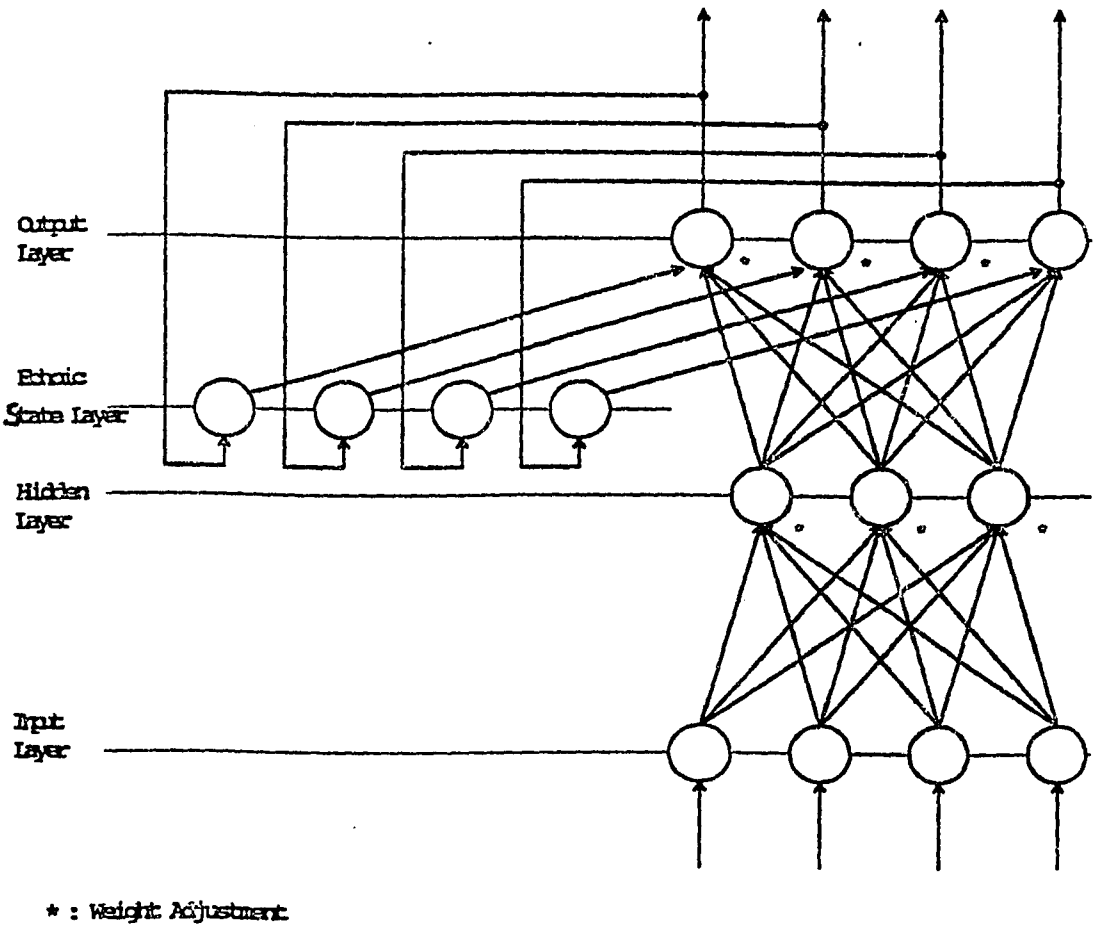


Figure 10. Backpropagation network with recurrent "echoic layer".

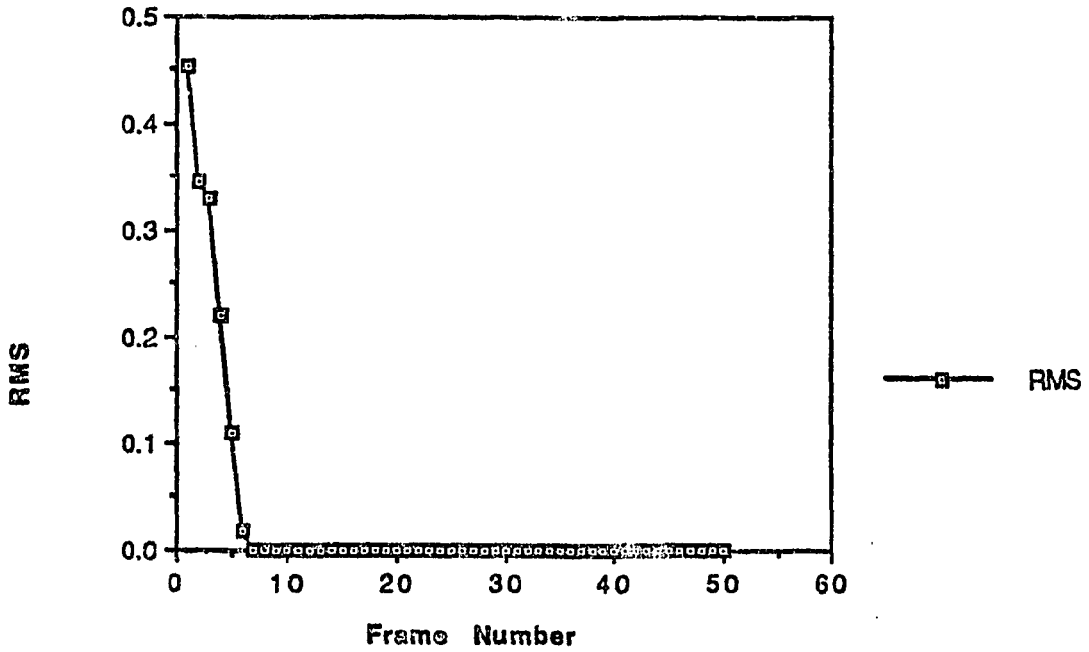


Figure 11. Root mean square error for recognition of /bad/ across ordinal time frame position, averaged over 10 male speakers in Generalization set.

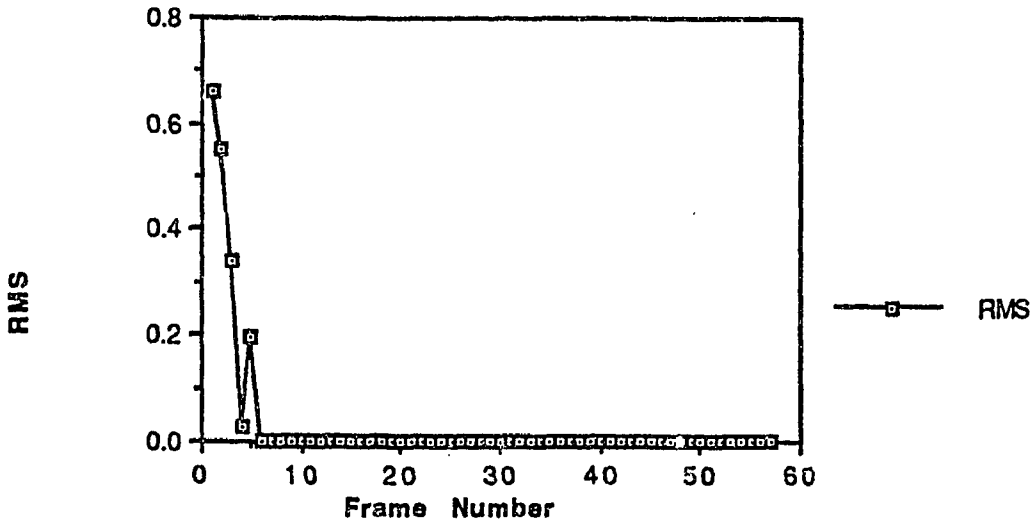


Figure 12. Root mean square error for recognition of /pad/ across ordinal time frame position, averaged over 10 male speakers in Generalization set.

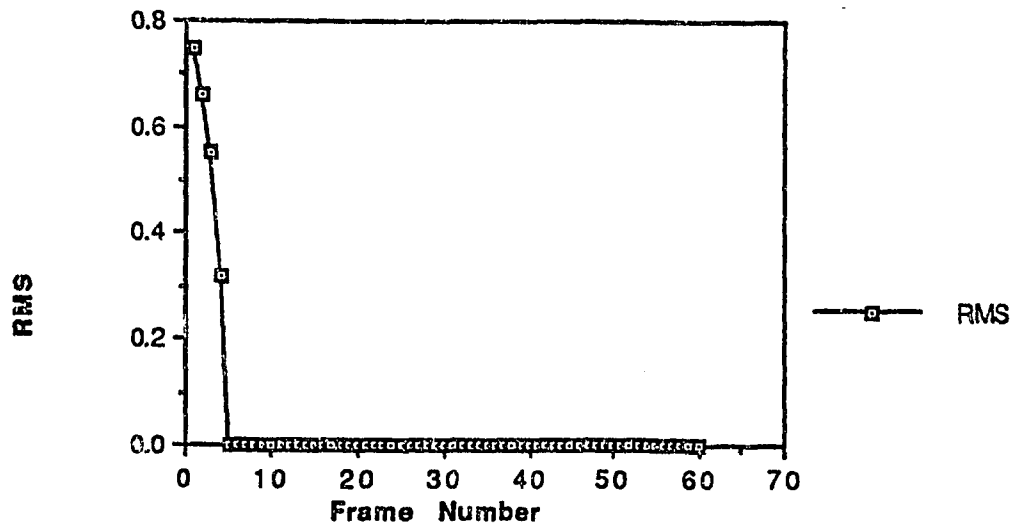


Figure 13. Root mean square error for recognition of /spad/ across ordinal time frame position, averaged over 10 male speakers in Generalization set.

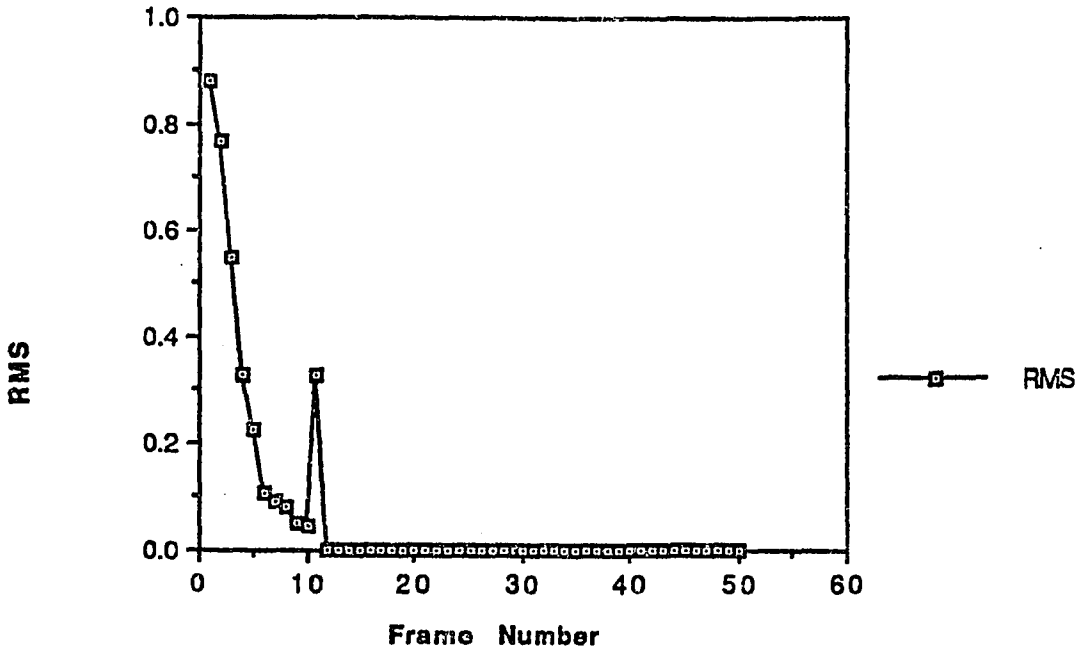


Figure 14. Root mean square error for recognition of /bad/ across ordinal time frame position, averaged over 35 female speakers in Generalization set.

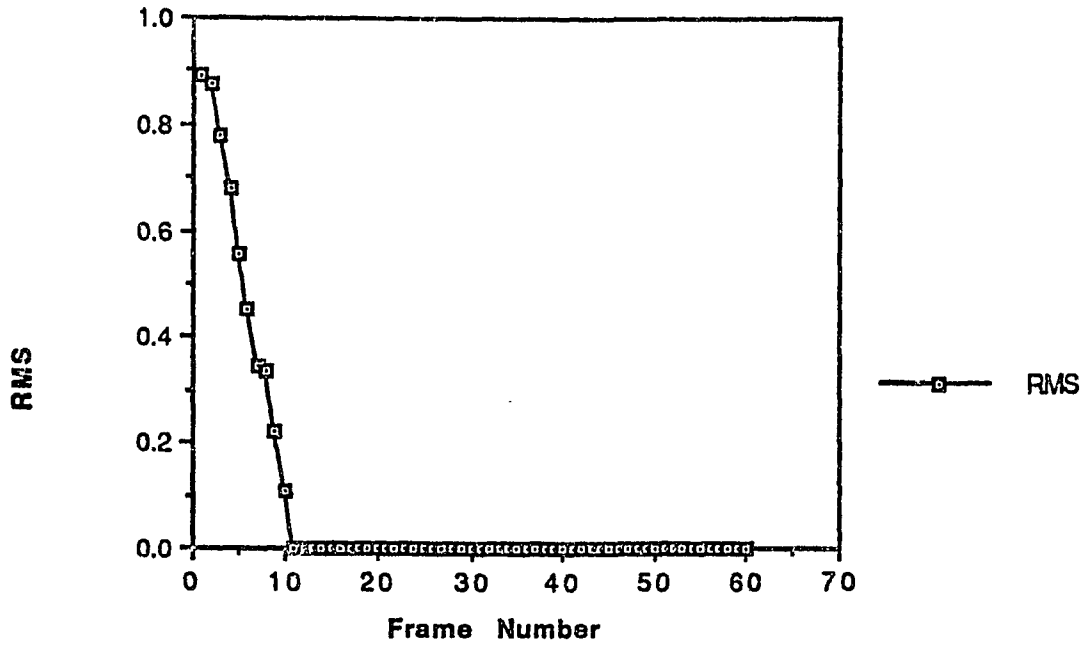


Figure 15 . Root mean square for recognition of /pad/ across ordinal time frame position, averaged over 35 female speakers in Generalization set.

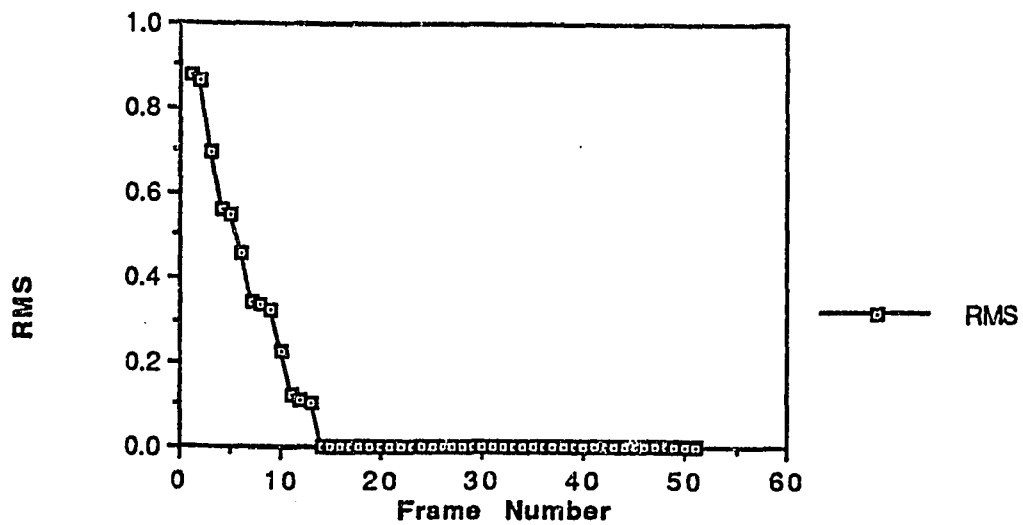


Figure 16. Root mean square error for recognition of /spad/ across ordinal time frame position, averaged over 35 female speakers in Generalization set.

Bibliography

Alankar, S. & Antrobus, S. (1991). Neural Network Tutorial. New York: City University of New York. 'C' Software.

Allon, N., Yeshurun, Y. & Wollberg, Z. (1981). Responses of single cells in the medial geniculate body of awake squirrel monkeys. Experimental Brain Research, 41, 222-232.

Altschuler, R. A., Bobbin, R. P., Clopton, B. M., & Hoffman, D. W. (Eds.) (1991). The neurobiology of hearing. N. Y.: Raven.

Antrobus, J,. Alankar, S., & Fookson, J. (1993) Phoneme recognition using nonlinear filters that simulate nuclei in the auditory pathway. Submitted for publication.

Antrobus, J., Alankar, S. & Yang, C. M. (1992). Speech recognition: Backpropagation with continuous compacting of sequentially-ordered information. Submitted for publication.

Antrobus, J., Bushell, C., Alankar, S., and Tartter, V. (1990). FORMANTR1: A Neural Net detector of second formants using cochlear nucleus transformation of cochlear signals. Proceedings of the 1990 Long Island I.E.E.E. Student Conference on Neural Networks.

Cohen, M. A., Grossberg, S. & Wyse L. (1992). A neural network model of pitch detection and representation. Technical Report CAS/CNS-92-024, Boston University Center for Adaptive Systems and Dept. of Cognitive and Neural Systems.

Davidsen-Nielson, N. (1974). Syllabification in english words with medial sp, st, sk. Journal of Phonetics, 2 (pp. 15-45).

Eimas, P. D., Tartter, V. C., & Miller, J. L. (1981). Dependency relations during the processing of speech. In P. D. Eimas & J. L. Miller (Eds.) Perspectives on

the study of speech (pp. 283-309).

Goddard, H. N., Lynne, K. J., Mintz, T. & Bukys, Liudvikas. (1989). Rochester Connectionist Simulator. Computer science technical report, 233.

Green, M. D., (1988). Audition: Psychophysics and Perception. Stevens Handbook of Experimental Psychology, 1, 366.

Hampshire, J. B. & Waibel, A. H. (1991). A novel objective function for improved phoneme recognition using time-delay neural networks. IEEE Transactions on Neural Networks, 1, 216-228.

Hubel, D. H., & Wiesel, T. N. (1963). Receptive fields of cells in striate cortex of very young visually inexperienced kittens. Journal of Neurophysiology, 26, 994-1002.

Jacobs, R. A. & Jordan, M. I. (1992) Computational consequences of a bias toward short connections. Journal of Cognitive Neuroscience, 4, 323-336.

Jordan, M. (1986). Serial order: A parallel distributed processing approach. Institute for Cognitive Science report 8,604, University of California, San Diego.

Klatt, D.H., (1975). Voice-onset time, frication and aspiration in word-initial consonant clusters. Journal of Speech and Hearing Research, 18, (686-706).

Kandel, R. E. & Schwartz, H. J. (1990). Principles of Neural Science. North - Holland.

Kiang, N. Y., & Peake, W. T. (1988). Physics and physiology of hearing. In Atkinson, R. C., Herrnstein, R. J., Lindzey, G., & Luce, D. R. Steven's handbook of experimental psychology. pp. 277-326, N. Y. Wiley Interscience.

Kohonen, T. (1984). Self-Organization and Associative Memory. New York: Springer.

Kohonen, T. (1988). The "neural" typewriter. Institute of Electrical and Electronics Engineers: Computer, 21, 11 - 24.

Landy, Michael. (1990). Image analysis software. SharpImage Software.

Lippmann, P. R. (1987). An introduction to computing with neural nets. IEEE ASSP 4 - 18.

Makhoul, J., Jelinek, F., Rabiner, L., Weinstein, C., & Zue, V. (1989). White paper on spoken language systems. In Speech and natural language: Proceedings of a workshop held at Cape Cod, Mass., October 15-18, pp. 463-479, DARPA/ISTO. San Mateo: Morgan Kaufman.

Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. Journal of the Acoustical Society of America, 69, 559-567.

Miller, J. D. (1989). Auditory-perceptual interpretation of vowel, Journal of the Acoustical Society of America, 85, 2114-2134.

Mullennix, J. W., Pisoni, D. B. & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. Journal of the Acoustical Society of America, 85, 365-378.

Norris, D. (1990). A dynamic-net model of human speech recognition. In Altman, G. T. M. Cognitive models of speech processing (pp 87 - 104). Cambridge, Mass.: MIT Press.

Pisoni, D. B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. Proceedings of the 1990 International Conference on Spoken Language Processing, Kobe, November 18-22.

Phillips, D. P., Reale, R. A., & Brugge, J. F. (1991). In R. A. Altschuler, R. P. Bobbin, B. M. Clopton, & D. W. Hoffman, D. W. (Eds.) The neurobiology of hearing (pp. 335-365). N. Y.: Raven.

Romani, G. L., Williamson, S. J., & Kaufman, L. (1982). Tonotopic organization of the human auditory cortex. Science, 216, 1339-1340.

Rhode, W. S. (1991) Physiological-Morphological properties of the cochlear nucleus In R. A. Altschuler, R. P. Bobbin, B. M. Clopton, & D. W. Hoffman, D. W

(Eds.) The neurobiology of hearing (pp. 47-77). N.Y.: Raven.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. (pp 318-362). In J. L. McClelland, D E. Rumelhart (Eds.) Parallel distributed processing: Explorations in t h e microstructure of cognition. Cambridge, MA.: MIT Press.

Searle, C.L., Jacobson, J.F., & Kimberley, B.P. (1980). Speech as patterns in the 3 space of time and frequency. (pp 73-102). In R.A. Cole(Ed), Perception and Production of Fluent Speech. Hillsdale, NJ.: Erlbaum.

Seidenberg, M. S. & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. Psychological Review, 96, 523-568.

Steinschneider, M., Arezzo, J. C. & Vaughan Jr., H. G. (1990). Tonotopic features of speech-evoked activity in primate auditory cortex. Brain Research, 519, 158-168.

Steinschneider, M. (1992). Speech and auditory processing in the auditory cortex of the monkey. Paper presented at City College of New York, Neurocognition Colloquium.

Stevens, N.K., & Blumstein, E. S., (1978). Invariant cues for place of articulation. Journal of Acoustical Society of America, 64(5).

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. Journal of the Acoustical Society of America, 85, 2135-2153.

Tartter, V. (1986) Language processes. New York: Holt, Rinehart and Winston.
Van Essen, D. C., Anderson, C. H. & Fellerman, D. J. (1992)

Information processing in the primate visual system: An integrated systems perspective. Science, 255, 419-422.

Waibel, A., Sawai, H, & Shikano, K. (1989). Consonant recognition by modular construction of large phonemic time-delay neural networks. ICASSP, 112-115.

Yost, W. S. (1992). Auditory perception and sound source determination. Current Directions in Psychological Science, 6, 179-184.