July 1988

# Speech Recognition Using Connectionist Networks Dissertation Proposal

Raymond L. Watrous
*University of Pennsylvania*



University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-88-44.

# Speech Recognition Using Connectionist Networks Dissertation Proposal


## Abstract

The thesis of the proposed research is that connectionist networks are adequate models for the problem of acoustic phonetic speech recognition by computer. Adequacy is defined as suitably high recognition performance on a representative set of speech recognition problems. Seven acoustic phonetic problems are selected and discussed in relation to a physiological theory of phonetics. It is argued that the selected tasks are sufficiently representative and difficult to constitute a reasonable test of adequacy.

A connectionist network is a fine-grained parallel distributed processing configuration, in which simple processing elements are interconnected by scalar links. A connectionist network model for speech recognition has been defined called the *temporal flow model*. The model incorporates link propagation delay and internal feedback to express temporal relationships. The model is contrasted with other connectionist models in which time is represented explicitly by separate processing elements for each time sample.

It has been shown previously that temporal flow models can be 'trained' to perform successfully some speech recognition tasks. A method of 'learning' using techniques of numerical nonlinear optimization has been demonstrated. Methods for extending these results to the problems selected for this research are presented.



## Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-88-44.

# SPEECH RECOGNITION USING CONNECTIONIST NETWORKS DISSERTATION PROPOSAL

**Raymond L. Watrous**

**MS-CIS-88-44**
**LINC LAB 117**

**Department of Computer and Information Science**
**School of Engineering and Applied Science**
**University of Pennsylvania**
**Philadelphia, PA 19104**

**July 1988**

---

**Acknowledgements:** This research was supported in part by DARPA grant NOOO14-85-K-0018, NSF grants MCS-8219196-CER, IRI84-10413-AO2 and U.S. Army grants DAA29-84-K-0061, DAA29-84-9-0027.

# Speech Recognition Using Connectionist Networks Dissertation Proposal

Raymond L. Watrous*

June 24, 1988

**Abstract**

The thesis of the proposed research is that connectionist networks are adequate models for the problem of acoustic phonetic speech recognition by computer. Adequacy is defined as suitably high recognition performance on a representative set of speech recognition problems. Seven acoustic phonetic problems are selected and discussed in relation to a physiological theory of phonetics. It is argued that the selected tasks are sufficiently representative and difficult to constitute a reasonable test of adequacy.

A connectionist network is a fine-grained parallel distributed processing configuration, in which simple processing elements are interconnected by scalar links. A connectionist network model for speech recognition has been defined called the *temporal flow model*. The model incorporates link propagation delay and internal feedback to express temporal relationships. The model is contrasted with other connectionist models in which time is represented explicitly by separate processing elements for each time sample.

It has been shown previously that temporal flow models can be 'trained' to perform successfully some speech recognition tasks. A method of 'learning' using techniques of numerical nonlinear optimization has been demonstrated. Methods for extending these results to the problems selected for this research are presented.

## 1 Introduction

The research described in this proposal is designed to establish that

*Helpful comments by Gary Kuhn, Mitch Marcus, Lokendra Shastri and Alex Waibel on earlier versions are gratefully acknowledged.

**Connectionist networks are adequate models for acoustic phonetic speech recognition.**

This thesis will be established by demonstrating that particular connectionist networks can be trained to solve selected acoustic phonetic recognition problems. It will be argued that these problems are representative of other problems which could also be solved by similar methods. Thus, it will be reasoned, connectionist networks could solve any acoustic phonetic recognition problem.

The precise meaning of the statement of the thesis will be made clear by definition of its terms. The importance of the thesis is then discussed in order to justify its selection. The method by which it will be established is then presented, followed by a plan for the work.

# 2 Definitions

In this section, the terms comprising the thesis are defined.

## 2.1 Connectionist Networks

By connectionist networks (CNs) is meant a computational model in which simple processing elements, called units, compute a single output value that is a nonlinear function of the weighted sum of its inputs. Units are interconnected with other units by links, such that each unit broadcasts its output value to the units to which it is connected. The information content of this model is expressed in the processing element function, pattern of interconnection, unit association strengths and link propagation delays.

This model has been alternatively called neural networks (NN), parallel distributed processing (PDP), and multilayer perceptrons (MLP). No attempt is made to delineate the overlapping definitions of these terms.

For the purposes of this definition, it will be maintained that connectionist networks does not refer to biological neural networks. Otherwise, the thesis would be self-evident, since on this definition human beings communicate adequately by speech using connectionist networks.

## 2.2 Acoustic Phonetic Speech Recognition

Acoustic phonetic speech recognition (APSR) is defined as the transformation of an acoustic signal into a string of phonetic symbols, from a small, structured set of symbols.

Acoustic phonetic speech recognition is a subset of speech recognition which does not include recognition of words and phrases. That is, transforming a sequence of phonetic symbols into words is excluded from the problem under consideration. To this extent, it is independent of language and dialect.

Acoustic phonetic speech recognition is furthermore distinct from speech understanding, which has as its goal extraction of the meaning of an utterance. Speech understanding requires the representation of syntactic, semantic and other linguistic (pragmatic) knowledge, which is excluded from view in acoustic phonetic speech recognition.

## 2.3 Adequacy

The adequacy of connectionist models for speech recognition is defined in terms of computability and learnability.

By computability is meant that the connectionist network is able to compute a solution to the appropriate APSR task. For practical reasons, the solution network should be implementable on a finite computer, with finite (real) response time.

By learnability is meant the condition that there exists a training algorithm such that given a connectionist network appropriate to the APSR task, a solution can be obtained. A solution is defined as meeting the task requirements with acceptable accuracy. Accuracy is considered acceptable if it meets or exceeds the average human recognition capability under similar conditions. Thus, the adequacy of accuracy is made relative to human performance, which, contrary to the expectation of naive speakers, is not perfect under conditions similar to those under which the computer operates. Examples of these conditions are lack of linguistic context and meaning.

Operationally, acceptable accuracy will be defined as accuracy comparable to that of existing speech recognition techniques. That is, the accuracy will be judged acceptable if the results are within several percent of the best results obtainable using mature technologies. This seems reasonable in view of the fact that other approaches have been refined through the efforts of many people over the course of several years, something which cannot be expected of this research.

A further condition on learnability is that the solution be obtainable for any speaker of any language. Excluded here are heuristic methods, based on knowledge of the specifics of a particular language or speaker.

The thesis of this research is that connectionist networks are adequate

for the problem of speech recognition by computer. There are many other methods of speech recognition which make this claim, and there is not time, nor is it the goal of this thesis to perform carefully controlled experiments with other methods. That could be an endless task, as each method has variations and refinements which would greatly complicate the comparison. Furthermore, the motivation for this research is the conviction that there is something fundamentally correct about connectionist networks for speech recognition, so the focus of the work is on developing this approach. There are qualitative differences and advantages over other methods which will be discussed in the process.

# 3 Significance of Thesis

This section defends the significance of the thesis by showing that the problem of speech recognition it addresses is important and difficult, and that the method of connectionist networks is novel and effective.

## 3.1 The Problem

That speech recognition and understanding is an important problem will be taken for granted. The extent to which computer speech recognition would change (improve) many aspects of work and life is certainly of great magnitude. Acoustic phonetic recognition is a well-defined and substantial subproblem of speech recognition.

Acoustic phonetic speech recognition is undeniably a difficult problem. One measure of its difficulty is the length of time it has been researched, factored by the number and skill of the researchers.

Furthermore, partly as a result of the large ARPA speech understanding project of the 1970s [4], it is known that accuracy in APSR is crucial to the success of speech understanding. Although higher level linguistic constraints were used effectively in overcoming ambiguity in acoustic phonetic identification, these constraints were most effective where the linguistic constraints were unrealistically strict [6]. This indicates that for speech understanding of natural language high accuracy is demanded at the level of acoustic phonetics.

Another advantage of APSR is that it is relatively language independent. As phonetic recognition, no attempt is required to address problems of phonology or word recognition. Thus, the methods developed in the thesis will be applicable to almost every language.

## 3.2 The Solution

The choice of connectionist networks as a method for solving the problem of acoustic phonetic speech recognition is defended in this section.

First of all, CNs derive as a computational model from biological neural networks, which have been successfully designed to compute and learn those functions necessary for speech recognition. Thus, the success of the biological networks inspires hope that solutions exist using synthetic networks. In fact, the synthetic solutions may be even more powerful along certain dimensions: airplanes can fly faster than birds (although with less grace).

It is known that connectionist networks can be used as pattern classifiers, and some work has been done investigating the relationship between connectionist networks and well-known classification methods, such as Bayesian and k-nearest neighbor [7, 5].

Second, learning algorithms exist which are able to optimize connectionist networks performance, as measured by an appropriate error function. This again holds out promise that, *if the network architecture and functionality are carefully chosen*, the optimization process can lead to a solution.

Third, CN provide a knowledge representation framework in which all levels of linguistic knowledge can be integrated. The small acoustic phonetic networks which will be developed in this research could be incorporated into larger networks which solve larger problems.

Fourth, connectionist networks offer a qualitative advantage over several other important methods of speech recognition. This advantage lies in the fact that networks only form internal representations for the purpose of discrimination; other methods form representations without reference to the purpose of discrimination.

Fifth, good preliminary results have been obtained on moderately difficult speech recognition problems using connectionist networks [3, 21, 22, 20, 15]. The results are sufficient to encourage further research.

Finally, connectionist networks were chosen because their suitability for speech recognition has not yet been adequately demonstrated. The initial studies have been promising, but solutions to many difficult problems have not yet been demonstrated. Several of these difficult problems have been selected for this research and are described below.

In conclusion, the choice of thesis is defensible since it proposes a novel solution to an important and difficult problem.

# 4 Basic Method

The basic method for establishing the thesis is to show that there exist connectionist networks and associated learning algorithms such that these networks can be trained on real speech data to solve several difficult speech recognition problems. The problems are chosen to be representative of speech recognition problems in general, and the argument is made that connectionist networks can, by extension, solve other problems as well.

In demonstrating that connectionist networks are adequate for acoustic phonetic speech recognition it is obviously impractical to require that every speech recognition problem be solved. The number of languages, speakers, phonemes and contexts to consider would make this an intractable problem. Consequently, some selection must be made. This selection is made with reference to a theory of phonetics so that the coverage of the acoustic phonetic problem space can be made clear.

In addition to selecting specific limited acoustic phonetic test cases, the approach taken in the dissertation will be to construct a network solution for each test case separately. Network solutions will be sought to the task of discriminating items within limited subsets of the complete acoustic phonetic space. This approach was chosen in order to decompose a large problem into several smaller ones, each of which is more tractable and illustrates different aspects of the total problem.

One difficulty with this approach is that failure to demonstrate the existence of a connectionist network which solves a particular speech recognition problem does not mean that no such network exists; it merely indicates that it has not been found among those that have been tried. There is good reason to hope that this difficulty can be avoided, by careful choice of the network functions and patterns of connectivity, and careful analysis of successful and unsuccessful network designs.

This approach also leaves open the question of how a single CN could solve the test cases combined. The answer to this question is affected by the response of a particular network to tokens not from its problem space. Although it is not planned to answer this question in the dissertation, approaches suggested by reflection on the experimental results will be discussed.

Finally, it would be most satisfying to be able to present a complete analysis of the solutions discovered by CNs to AP problems. Understanding a particular network solution would allow consideration of the question of whether simpler equivalent solutions exist. It would also shed light on pos-

sibly new phonetic discriminatory mechanisms and possible neural mechanisms, which in turn might suggest biological theories for experimental confirmation.

A complete mathematical characterization of the response of a network of nonlinear units with recurrent links would be extremely difficult. The calculus of mass action in neural nets is an area of current research. Solving this problem of analysis is clearly beyond the scope of the thesis. A descriptive approach will be taken instead, involving individual unit responses to speech data. Such analysis should provide appropriate information about the acoustic phonetic features formed by solution networks.

## 4.1 Preliminary Considerations

The choice of problems for acoustic phonetic speech recognition will be considered relative to a theory of phonetics in order to make explicit the extent to which the selected problems cover the space of possible problems. The theory is useful in tracing the outlines and showing the principal axes of the phonetic space.

For purposes of discussion, reference is made to the physiological theory of phonetics [13]. This theory proposes a model for phonetics which meets the requirements of applicability, completeness, consistency, and simplicity. Specifically, the model is intended to provide the means to describe all spoken languages using a relevant and optimal set of physiological parameters.

The physiological parameters define axes in articulatory phonetic space, and permit the specification of a point in that space by the parameter values. Certain parameters are more significant than others, and define major axes in the space. These parameters, which form a small set, can be considered to span the phonetic space. Consequently, the spanning axes will inform the selection of most of the experiments.

The parameter axes are not necessarily bases for the space; this fact complicates the choice of a direction along which to test discrimination. The proposed approach is to investigate variations along the dimensions of one parameter holding the others fixed. Because of non-orthogonality, discrimination along one axis may depend on the values of the other parameters. Thus, it is claimed that discrimination along one axis, with values along other axes fixed, is not identical to but is as difficult as discrimination along that axis for other values of other parameters.

Thus, for example, a network solving a manner of articulation discrimination problem is not intended to represent a manner detection network. It

is only a solution along the manner dimension for a particular combination of place and voicing parameters. The argument is that since a solution can be found along this dimension for one set of parameters, a solution could be found also for other values of the place and voicing parameter; but the solution is not necessarily the network already discovered.

### 4.1.1 Language

Languages vary in how they make distinctive use of different physiological parameters; consequently the choice of parameter axis for testing is linked to the choice of test language. It is claimed that any human language (and dialect) is sufficiently rich in phonetic diversity that it may serve as an adequate test of the sufficiency of connectionist networks for speech recognition, along the appropriate physiological phonetic directions.

Thus, for example, tonal languages[1], where tone (pitch) is used phonemically, may require tests different from those required for languages using clicks[2] or whistles[3].

For this research, the test language is chosen to be American English. This choice was made for the reasons that native speakers are numerous and available, and because a large body of acoustic phonetic research on American English is at hand.

It is postulated that other languages, all of which are characterized by acoustic differences of frequency and time, are not impossible to recognize using CNs, provided that a suitable set of acoustic phonetic problems can be solved using CNs for a single language.

### 4.1.2 Speaker

It is claimed that the speech of any speaker is sufficiently difficult to recognize that an adequate test of a CN is to show performance for a single speaker. Thus, successful use of CNs to recognize the speech of a single talker will not be considered idiosyncratic. This will be confirmed by tests on multiple speakers, taken, however, singly. That is, for each task, one network will be optimized separately for each speaker. The problem of a single network which works for all speakers (speaker independence) will not be addressed.

[1] Thai, Chinese
[2] South African Hottentot and Bushman
[3] Mazeteco

### 4.1.3 Context

The physiological theory of phonetics defines phonetic context as left and right (in time) phonetic events. It is known that the effect of phonetic context on phonetic parameter values can extend across multiple phonemes. These longer range contextual effects will not be considered in this research.

It is not claimed that the choice of context, as discussed in conjunction with the test problems below, represents the most interesting or difficult cases. It is also not claimed that a solution in one context will generalize to a solution of the same discrimination task in another context. It is noted that a solution to a phonetic discrimination task in *any* context is a solution to that task. Conversely, the failure to find a solution in a particular context is a failure of the method.

## 4.2 Problem Set

A major division is made in the physiological theory of phonetics between articulatory phonetics and prosodics. Each division is characterized by articulatory parameters which serve as dimensions of the corresponding subspace.

The articulatory parameters of the prosodic division are average laryngeal frequency, average speech production power and phonetic duration.

The parameters which characterize the articulatory phonetic space are divided into primary and secondary parameters. The primary parameters are manner, and place (horizontal and vertical). The secondary parameters include aspects of air flow (mechanism, direction, turbulence, pressure, release, and path), lingual air path, laryngeal action, and the shapes of the pharynx, tongue and lips.

Along the dimension of manner, which includes nasal, stop, flap, trill, sibilant, fricative, sonorant and vowel, there is a fundamental distinction between vowel and consonant. Although this is not explicitly part of the theory, it is a well-established and long standing distinction.

The acoustic phonetic recognition problems are classified in terms of these divisions of the physiological theory, as follows: articulatory phonetic, consisting of consonant and vowel, and prosodics.

### 4.2.1 Consonant Recognition Problems

The consonants are generally characterized by closure or obstruction in the vocal tract which results in lower signal energy. The excitation of the vocal

tract can either be from vocal fold vibration, or frication at the constriction, or both.

In the categorization of the phonemes in the articulatory phonetic space, the three primary parameters, manner, horizontal place and vertical place, are related in such a way that the space can be collapsed from three dimensions to two. Thus, for the consonants the major axes of the phonetic space are manner and horizontal place, whereas for the vowels, the axes are horizontal and vertical place.

Thus, two experiments were chosen to investigate the consonant dimensions of manner and place. An additional experiment is described in which the voicing dimension is explored. Voicing, or laryngeal control, is a secondary parameter which is of major importance for the consonantal subspace of American English.

The first consonant problem (Problem C1) involves discrimination of horizontal place of articulation. The manner of articulation was chosen to be the stop consonants, and the laryngeal action selected as voiced. The context was chosen as consonant initial with a variety of following vowels. Thus, Problem C1 consists of discriminating the voiced stops in CV syllables for a number of vowels. For American English this specifies the phonetic set [b,d,g]. This is an important and well-studied set, which will serve as a benchmark for performance evaluation, and possibly comparison with other methods.

The second consonant problem (Problem C2) involves discrimination of manner of articulation, with fixed horizontal place and voicing. The manner dimension is explored for voiced alveolar consonants, [n,d,z,l] and the palato-alveolar [dz]. This covers the manners of articulation of nasal, stop, sibilant, affricate and sonorant. The context again is chosen to be CV words, for following vowels [i,a,u].

The final consonant problem (Problem C3) involves a voicing contrast for particular values of manner and place. The bilabial stop consonants [p,b] were chosen for this experiment, in intervocalic context. For this purpose, the pair "rapid/rabid", was selected. This choice was made because this problem has been previously studied [8], and several factors are known to contribute to the discrimination, such as vowel duration and voicing onset time. Thus, it will be of interest which mechanism the CN will exploit in solving the discrimination task. This might involve generating features to detect burst/aspiration, spectral trajectories or vowel duration.

### 4.2.2 Vowel Problems

Vowels are generally distinguished from consonants by the open, unconstricted vocal tract, and vibration of the vocal folds, which produces phonation. Vowels are typically of much higher energy than consonants, and relatively easy to identify. In general terms, vowel identification is known to involve detection of formant values, duration, amplitude, and pitch. Each of these vowel parameters are variously affected by context.

As mentioned above, the major axes of the vowel articulatory space are horizontal and vertical space. However, the distribution of vowels within the matrix of horizontal and vertical place of articulation is somewhat different than that for the consonants. The shape of the distribution is roughly triangular, with the most densely packed dimension along the diagonal ([i,I,e,ae,a]). Therefore, the selection of experiments is guided by slightly different considerations than in the case of the consonants.

The first vowel problem (Problem V1) was chosen to investigate the detection of place of articulation within a subset of vowels which span the full range of horizontal and vertical place. The vowels were chosen to be [i,a,u]. The context was chosen to be initial voiced stops in CV words, relating this experiment to Problem C1.

The second vowel problem (Problem V2) was chosen to investigate the effect on detection of place of articulation of consonant context in CVC words. For this purpose, two adjacent vowels were selected for recognition in the context of voiced and voiceless stop consonants. The vowels [e,æ] were chosen [9, 10]. The solution to this problem is expected to require the formation of a formant-based vowel classifier which is context sensitive.

The final vowel problem (Problem V3) involves diphthongs and is concerned with the problem of formant trajectory. To this end, the diphthong pair [eu,ue] was selected. This may also be considered a problem involving the continuants [j,w] as in [ju,wi]. The solution to this problem is expected to require attention to sequence; there is not expected to be a single point in formant space which could be used to solve the discrimination. This is considered a potentially difficult task for connectionist networks of a certain type (temporal flow model: see below).

Additional dimensions of the vowel space such as lip rounding and nasality are not addressed by this study. Lip rounding is ignored because it covaries to some extent with place, and is not used phonemically in American English, as it is, for example, in German. The problem of nasality is also omitted from consideration, as it is less important in American English than

in French, for example.

### 4.2.3 Prosodic Parameters

In the physiological theory of phonetics, the parameters of laryngeal frequency, phonetic duration and average speech power are separated from the articulatory parameters under the category of prosodic parameters.

It was decided to explore duration as a prosodic parameter, to a limited extent. One reason for this choice was that spectral and spectral-temporal cues were considered in the consonant and vowel tasks, and duration would complete this set by including purely temporal cues. A second reason was that the prosodic parameter of pitch is not represented in the spectral analysis performed by the bandpass filters in view for the research and the development of robust pitch extraction software to augment the filter bank representation would delay the completion of the research.

There are serious difficulties with the choice of duration as a parameter, and so the investigation will be very constrained. Durational cues are relative to speaking rate, and so a reference point is required, against which relative changes can be measured. For the current research it was determined to consider durational effects at a given speaking rate in order to limit the complexity of the task. This problem of normalization is an important and interesting problem for the use of connectionist networks in speech recognition, which will be deferred to subsequent research.

A linguistic discrimination for English in which purely durational cues were employed was therefore sought. The distinction between single and geminate [n] in Arabic (bana - banna) is said to have purely temporal cues[12]. Consequently, single-geminate [n] pairs were sought in English. The pairs "synapse/sin naps" and "spinach/spin niche" were identified as potential candidates. This task will be identified as Problem P1. It needs to be demonstrated that the cues for this discrimination are purely durational, and that the pairs lack spectral cues for discrimination. Connectionist networks may then be developed which are based on purely temporal factors. This will be done for a single speaking rate as indicated.

These seven acoustic phonetic problems comprise the test set for demonstrating the speech recognition capabilities of connectionist networks.

# 5 Plan of Work

The scope of the work has been outlined above; the plan for carrying out the work of the dissertation is sketched below, beginning with certain aspects which have been completed already.

## 5.1 Network Model

Considerable consideration has already been given to the architectural and functional requirements of a connectionist network for speech recognition. Various alternative designs are discussed here.

### 5.1.1 TRACE Model

The TRACE model for speech recognition is the most fully developed connectionist model to date [2, 11]. A key characteristic of the model is that time is represented as an index across a set of identical feature detectors. Thus, time is mapped to space. Many other models adopt this time-spatializing approach [14, 1]. It will be argued in the thesis that this approach has advantages for research in discovering temporal relationships, but that it has serious problems as a solution method.

It is clear that such a model must encode explicitly the fact that relative time and not absolute time is important. This is referred to as the symmetry problem, since shifts along the time axis must have invariant results. This means that the link weights of subsequent time slices must be kept equal. There are other problems with this model having to do with segmentation and time alignment.

However, the most serious problem is that space and time are coreferenced. Consider a trajectory in formant space for example. The spatialized network which can detect a formant trajectory involves a set of spectral profile detectors at points along the trajectory, which are spaced according to time. Consequently, the same trajectory with a different time course will result in a mismatch along the time and spectral axis, which must be compensated for by reduced precision in the spectral dimension. It seems better to decouple time from space, and allow one stage to specify the spectral trajectory and another stage to specify the time course along that trajectory.

### 5.1.2 Temporal Flow Model

The temporal flow model [20] was motivated by the computational properties of biological neural networks[4], which include the integrative properties of the synapse, the transmission delay of axons, the propagation time of post-synaptic potentials, as well as the standard connectionist properties of nonlinear unit function, synaptic connection strength and interconnection patterns.

In the temporal flow model, time is represented indirectly, rather than directly. The model is characterized by internal localized feedback and delay to provide temporal dynamics, as data flows from input to output units in synchronous fashion, as input units are constantly being updated. The temporal flow model overcomes some of the weaknesses of spatialized models such as TRACE, such as segmentation and time alignment.

The dissertation will include an analysis of the computational properties of the temporal flow model, showing the effect of recurrence, delay, connection path and unit function on the network time response.

## 5.2 Learning

The problem of learning can be stated as an optimization problem, for which many solution techniques exist. Various optimization algorithms including back-propagation, steepest-descent, Davidon-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) have been implemented in conjunction with a general purpose network simulator written in C that is able to process networks with delays and recurrent links. The complexity and convergence characteristics of these optimization algorithms have been analyzed for various problems [17, 18].

It has been shown that gradient methods methods of optimization can be used for networks with recurrent links [16]. A method of gradient estimation for networks with recurrent links has been developed and used successfully [20].

The formulation of an objective function is necessary for optimization. The explicit target function method [20, 19], in which expected output values are expressed by parameterized standard functions has been used successfully.

[4] Particular inspiration was drawn from the biological system for sound localization, in which the neural network computes precisely those parameters which vary physically with the sound source location.

### 5.3 Speech Experiments

Connectionist network models of the temporal flow type have been successfully optimized for a consonant discrimination task in distinguishing the word pair "no/go" [20, 19]. Twenty-five word pairs were correctly discriminated using a network that had been trained on a single training utterance.

Stop consonant discrimination was demonstrated for the voiced stops in CV words for three vowels [22] and six vowels [21]. Vowel discrimination has also been demonstrated using the temporal flow model, in the context of voiced stops [21, 22]. This work has shown that networks with recurrent links can learn consonant discrimination in various vowel contexts, and *vice versa*. Practical experience with network architectures, target functions and optimization algorithms has been gained.

### 5.4 Computational Resources

The computing facilities available for this research include several Sun 3/260 workstations with floating-point accelerators, and the Cyber 205.

The optimization software has been partially modified for operation on the Cyber 205 Vector Processor, at the John Von Neumann Computing Center, Princeton, NJ. The vectorized code is nearly operational, but needs to be completed.

## 6 Experiments

This section discusses in more detail the discrimination tasks and the associated requirements.

### 6.1 Data

The speech data for the experiments pertinent to the seven discrimination tasks will be recorded digitally for further analysis. The digitized speech data will be played back and the resulting filterbank channel energies recorded from a Siemens CSE-1200 speech recognition device. The filterbank outputs are log-compressed, full-wave rectified and sampled every 2.5 milliseconds.

For certain experiments, it may become necessary to process the digital data differently. For example, the manner discrimination task may require some time domain measurements to represent frication.

The number of speakers will be set at two. The approach will be to demonstrate the method for one speaker, and use a second speaker to confirm the results for the original speaker. The choice to limit the number of speakers in this way was made for reasons of time.

The number of repetitions of each test token should be large enough to provide sufficient training and test data. One hundred repetitions was chosen as a reasonable number. With two speakers, and 100 repetitions per test utterance, 200 repetitions per token are required.

## 6.2 Network Design

The general temporal flow model will be the primary network design for all experiments. Various configurations will be considered in order to allow computation of spectral/temporal features as appropriate to the task. The general approach will be to start with the simplest network, and add capability in the case of failure.

## 6.3 Particular Experiments

The consonant and vowel problems defined above are discussed below in terms of their data requirements, and initial solution strategies.

### 6.3.1 Problem C1

The place discrimination task involves three voiced stops [b,d,g] in initial position (CV). Three following vowels are considered adequate variation in context for this task, and have been chosen as [i,a,u]. This amounts to 9 test tokens, for a total of 1800 utterances. It is anticipated that the temporal flow model may be adequate for this task, since it has been used in previously reported studies [21]. Performance measurements have not yet, however, been obtained.

### 6.3.2 Problem C2

The manner discrimination task involves 5 consonants in initial position. It is considered sufficient for this task, that three vowels be used, [i,a,u]. This results in 15 test tokens, for a total of 3000 utterances.

It is expected that spectral/amplitude features may not be sufficient for this task. Consequently, time domain features may need to be developed to indicate the presence of frication.

### 6.3.3 Problem C3

The embedded stop consonant problem will involve repetitions of a carrier phrase including the words "rapid" or "rabid". This would result in a data set of 400 items.

It is anticipated that the network may need some representation of delay in order to solve this problem. In the event that the problem is unsolvable using a temporal flow model with unit delay, a solution to the problem in stages will be considered, by learning to identify separate acoustic phonetic events, and then using edge operators and delays to allow the network the possibility of forming a duration feature.

### 6.3.4 Problem V1

The place of articulation discrimination task covering the vowel space uses the voiced stop consonants in initial position for phonetic context. This task uses the same data set as Problem C1.

This problem is expected to be solvable with a temporal flow model with one hidden layer, as reported previously [21]. Performance results and network analysis is required to complete this problem.

### 6.3.5 Problem V2

The vowel discrimination task for [e,æ] in stop consonant context is expected to be solvable with a temporal flow model with one or two hidden layers. The data should consist of CVC words in a carrier phrase, with voiced and unvoiced stop consonants. In order to reduce the data requirements, only symmetric combination of [p,t,k,b,d,g] will be used for the initial and final stops. This would result in a set of 12 test tokens and a complete set of 2400 utterance for the experiment.

### 6.3.6 Problem V3

The diphthong discrimination problem will require only a single pair of test tokens, [ju,wi], and a total of 400 utterances.

It is of interest whether the temporal flow model with unit delays will be able to solve this problem. The model may require some representation of formant trajectory or delay to solve this problem. This need will be addressed in the event of a failure of the initial model.

### 6.3.7 Problem P1

The single-geminate [n] task will require repetitions of the pairs "synapse/sin naps" and "spinach/spin niche", for total of 800 utterances. The lack of spectral cues will be tested by a method of local spectral distance measure along nonlinear temporal alignment path. Networks will be designed for this task which have the capability to represent duration and learn durational differences.

A total of 8800 utterances is required for these problems.

# 7 Conclusion

This dissertation will accomplish the goal of demonstrating that connectionist networks can compute complex spectral/temporal features of speech and that these characteristics can be learned from real speech data. It is expected that the work will show which connectionist network architectures and functionalities are required to solve specific acoustic phonetic speech recognition problems.

# References


[1] David J. Burr. A neural network digit recognizer. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, pages 1621–1625, October 1986.

[2] Jeffrey Elman and John McClelland. Exploiting lawful variability in the speech wave. In Joseph S. Perkell and Dennis H. Klatt, editors, *Invariance and Variability in Speech Processes*, chapter 17, pages 360–380, Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[3] Jeffrey L. Elman and David Zipser. *Learning the Hidden Structure of Speech.* Technical Report ICS Report 8701, UCSD Institute for Cognitive Science, February 1987.

[4] M. F. Medress et. al. Speech understanding systems. *IEEE Transactions on Professional Communication*, PC-20(4):221–225, December 1977. Reprinted from SIGART Newsletter, No 62, pp 4-8, April, 1977.

[5] Stephen Jose Hanson and David J. Burr. Knowledge representation in connectionist networks. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, July 1987. submitted.

[6] Dennis H. Klatt. Review of the arpa speech understanding project. *Journal of the Acoustical Society of America*, 62(6):1345–1366, December 1977.

[7] Richard Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4–22, April 1987.

[8] Leigh Lisker. Closure duration and the intervocalic voiced-voiceless distinction in english. *Language*, 33:42–49, 1957.

[9] Leigh Lisker. The distinction between [æ] and [e]: a problem in acoustic analysis. *Language*, 24(4):397–407, 1948.

[10] Leigh Lisker. *Reconciling Monophthongal Vowel Percepts and Continuously Varying F Patterns.* Technical Report SR-79/80, Haskins Laboratories, 1984.

[11] John L. McClelland and Jeffrey L. Elman. Interactive processes in speech perception: the trace model. In J.L.McClelland D.E.Rumelhart and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume II Psychological and Biological Models*, chapter 15, MIT Press, Cambridge, MA, 1986.

[12] D. Obrecht. Three experiments in the perception of geminate consonants in arabic. *Language and Speech*, 8:31–41, 1965.

[13] G. Peterson and J. Shoup. A physiological theory of phonetics. *Journal of Speech and Hearing Research*, 9:5–67, 1966.

[14] David C. Plaut, Steven Nowlan, and Geoffrey Hinton. *Experiments on Learning by Back Propagation.* Technical Report CMU-CS-86-126, Carnegie-Mellon University, 1986.

[15] R. W. Prager, T. D. Harrison, and F. Fallside. Boltzmann machines for speech recognition. *Computer Speech and Language*, 1(1):3–27, March 1986.

[16] David E. Rumelhart, Goeffrey Hinton, and Ronald Williams. Learning internal representations by error propagation. In J.L.McClelland D.E.Rumelhart and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Volume I Foundations*, chapter 8, MIT Press, Cambridge, MA, 1986.

[17] Raymond L. Watrous. *Learning Algorithms for Connectionist Networks: Applied Gradient Methods of Nonlinear Optimization.* Technical Report MS-CIS-87-51, University of Pennsylvania, June 1987.

[18] Raymond L. Watrous. Learning algorithms for connectionist networks: applied gradient methods of nonlinear optimization. In *Proceedings of the First International Conference on Neural Networks*, pages 619–627, June 1987.

[19] Raymond L. Watrous and Lokendra Shastri. Learning phonetic features using connectionist networks. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, pages 851–854, August 1987.

[20] Raymond L. Watrous and Lokendra Shastri. *Learning Phonetic Features Using Connectionist Networks: An Experiment in Speech Recognition.* Technical Report MS-CIS-86-78, University of Pennsylvania, October 1986.

[21] Raymond L. Watrous and Lokendra Shastri. Learning phonetic features using connectionist networks: an experiment in speech recognition. In *Proceedings of the First International Conference on Neural Networks*, pages 381–388, June 1987.

[22] Raymond L. Watrous, Lokendra Shastri, and Alex Waibel. Learned phonetic discrimination using connectionist networks. In *European Conference on Speech Technology*, pages 377–380, September 1987.