8-1-1989

Adaptive Networks as a Model for Human Speech Development

M. Fernando Tenorio
Purdue University

M. Daniel Tom
Purdue University

Richard G. Schwartz
Purdue University

Adaptive Networks as a Model for Human Speech Development

M. F. Tenorio
M. D. Tom
R. G. Schwartz

TR-EE 89-54
August, 1989

School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

Adaptive Networks as a Model for Human Speech Development

M. Fernando Tenorio
M. Daniel Tom
School of Electrical Engineering
and
Richard G. Schwartz
Department of Audiology and Speech Sciences

Parallel Distributed Structures Laboratory
Purdue University
West Lafayette, IN 47907

August 1989

TR-EE-89-54

TABLE OF CONTENTS

Page

Abstract....1

1. Introduction....1

2. Text-to-Speech Conversion....3

3. Representing Text and Phonemes....4

4. Preparing the Training Input....5

5. Training Procedures....6

6. Difference with NETtalk....6

7. Performance....7

8. Summary of Statistics....8

9. Observations....9

9.1. Orthographic Transparency....10

9.2. Learning Different Languages....10

9.3. Capturing the Characteristics of Speech....11

9.4. Training with a Second Language....12

9.5. Contextual Dependency....13

9.6. Number of Hidden Units....13

9.7. Cluster Analysis....14

9.7.1. Clustering Algorithm....14

9.7.2. English Training....15

9.7.3. Spanish Training....15

9.7.4. English Developmental Cluster Analysis....16

9.7.5. Spanish Developmental Cluster Analysis....17

9.7.6. Spanish-English Training Cluster Analysis....17

9.7.7. English-Spanish Training Cluster Analysis....18

10. Output Decision Algorithms....19

11. Problems, Solutions, and Potentials....19

References....20

Appendices....22

Adaptive Networks as a Model for Human Speech Development

M. Fernando Tenorio
M. Daniel Tom
School of Electrical Engineering
and
Richard G. Schwartz
Department of Audiology and Speech Sciences

Parallel Distributed Structures Laboratory
Purdue University
West Lafayette, IN 47907

Abstract

Unrestricted English text can be converted to speech through the use of a look up table, or through a parallel feedforward network of deterministic processing units. Here, we reproduce the network structure used in NETtalk. Several experiments are carried out to determine which characteristics of the network are responsible for which learning behavior, and how closely that maps human speech development. The network is also trained with different levels of speech complexity, and with a second language. The results are shown to be highly dependent on statistical characteristics of the input.

1. Introduction

Connectionist networks are arrays of simple highly interconnected computing elements. Two important properties emerge from this computing paradigm. First, these networks serve as an alternative form of massively parallel knowledge representation. Second, these networks can learn the input-output relationships by simple modification of the connection strengths. Because of their potential and simplicity, feedforward networks using the Generalized Delta Rule for learning have received the greatest attention [2,4,5,8].

These networks are particularly useful in building applications involving adaptive mappings. In this work, we will discuss the application of multilayered feedforward networks to the problem of text-to-speech conversion. This general application is studied before by Sejnowski and Rosenberg [9, 10]. Here, we examine the same problem and study it in relation to human speech development. We would like to understand the characteristics of such a network, and discover which characteristics are similar to human behavior, which are not, and what aspects of the algorithm is responsible for these similarities and differences. Our ultimate goal is to understand to what extent feedforward networks can be used for the study of human speech development and behavior, and what changes, if any, would improve their predictability of this behavior.

The particular algorithm used in our study is adapted from Sejnowski and Rosenberg [10]. This combination of a multilayer feedforward network and a learning algorithm is commonly referred to as the "back propagation algorithm" (or as the Generalized Delta Rule for adapting weights) discussed below. The feedforward network is composed of three layers. (The input layer units are fan-out units.) Connections exist between each element of a layer and all the elements of an adjacent layer. Each unit is composed of a quasi-integration unit at the inputs, and the output of the unit is found by applying a nonlinear (sigmoid) function to the result of the integration.

$$A_i = \sum_j w_{ij}\ p_j$$

$$p_i = S(A_i) = \frac{1}{(1 + e^{-A_i})}$$

The signals activate the input units (first layer), and are being changed by this function of the units as they are propagated toward the output layer (third layer). Notice that for a given number of input units there is a synergy between their activation values, which is produced by the full connectivity to the second layer (hidden layer). This synergy will vary for different elements because of different sets of weights associated with each element. An error correcting algorithm modifies the values of the weights to better represent the relationship between the pairs of input and output patterns presented to the network. This supervised learning algorithm, called back propagation, works as follows. First, the error signal between the activation of the output layer (layer N) and the desired pattern is calculated:

$$\delta_i^{(N)} = (p_i^* - p_i^{(N)})\ S'(A_i^{(N)})$$

The error signal is then transmitted to the adjacent layer (back propagated toward the input layer), by using the following recursive rule:

$$\delta_i^{(n)} = \sum_j \delta_j^{(n+1)} w_{ji}^{(n+1)} S'(A_i^{(n)})$$

where $S'(A_i)$ is the first derivative of $S(A_i)$, p_i^* is the desired value of unit i in the output layer, and $p_i^{(N)}$ is the output value of unit i.

The error gradient of each weight is calculated according to:

$$\Delta w_{ij}^{(n)} = \alpha\ \Delta w_{ij}^{(n)} + (1 - \alpha)\delta_i^{(n)} p_j^{(n)}$$

where α is a parameter that regulates the smoothness of the gradient descent, known as the momentum constant. Each weight is updated as follows:

$$w_{ij}^{(n)}(t+1) = w_{ij}^{(n)}(t) + \epsilon \Delta w_{ij}^{(n)}$$

where ϵ is a parameter that controls the rate of learning, or the step size in the gradient descent in the weight space (the effective step size is $\epsilon * (1 - \alpha)$).

Before training, the weights are initialized to small uniformly distributed random values between -0.3 and 0.3. Initial weights cannot be set to zero, otherwise learning would not occur.

2. Text-to-Speech Conversion

The problem of text-to-speech conversion is certainly a very interesting one. First, there is a wealth of research done in the area of speech synthesis which can be used as a basis for comparison. Second, text-to-speech conversion, especially in languages that are not orthographically transparent, is a very difficult many-to-many mapping problem. Third, the problem of reconciling rules and exceptions shares some of the characteristics of other difficult problems approached by traditional artificial intelligence technology [10]. Fourth, the size of the pattern space, number of phonetic rules, the number of input-output units, and the number of training presentations scale up the network close to the limit of effective tractability for this type of algorithm. The algorithm would certainly fall in the category of functions with a high-order of complexity for connectionist supervised algorithms [13]. The text-to-speech conversion problem gives us some practical insight into the use of such algorithms under scaled up, straining conditions. Finally, text-to-speech conversion may be analogous to aspects of early speech development. One basic task facing a child learning to speak is to match output to input. In fact, it is suggested that the errors made by this type of network during training are comparable to those made by children [9]. We hope to evaluate this claim further, with the goal of improving the simulation characterisitics of the network. A better simulation would allow us to gain insight into the behavior of the algorithm as it relates to human development. It may also provide insights into human development, particularly the relationships between input and output. Finally, such simulations should elucidate the process of learning second complex phonetic mappings, thus shedding light on first and second language learning, or synthetic language formation, and on speech recognition. The following simulations represent first steps in this direction.

3. Representing Text and Phonemes

In order to present the text and speech pattern pairs to the network, a suitable representation is necessary. Although there may be more suitable and simpler representations, we implement the network in a manner similar to NETtalk [10]. There are 203 units in the input layer and are divided into 7 groups of 29 units each. Each group represents one character of the input text in a window of 7 characters. Each unit in a group represents a character in the alphabet. The 26 letters in the English alphabet, plus the Spanish letter "ñ", the period, and the space are used. The number of input groups are varied in one experiment to determine the optimal size of the context window, and eventually to gain insight into the problem's predicate order k [7] (a measure of the complexity of the mapping).

Text enters one end of the window, with one character in one group of input units, and it is shifted to the other end. At each presentation, the input units corresponding to the characters are clamped, and the activation is fed forward.

The desired output corresponding to the central character in the input window is presented to the output units. There are 32 units in the output layer. Each unit represents an articulatory feature. Each phoneme is encoded by three articulatory features representing: manner, place, and voicing. The output vector consists of the 32 real output values of the output units. The phoneme vectors, with values ranging from 0.0 to 1.0, are compared to the output vector.

Two decision rules are used to define the output phoneme. The first rule compares the output vector with the phoneme vectors one by one. The phoneme vector making the smallest Euclidean angle with the output vector is considered to be the "best guess" phoneme. The second involves a dichotomous decision. An output value between 0.9 to 1.0 is assigned to 1.0, otherwise the value of 0.0 is assigned. If the truncated version of the output vector is one of the vectors in the phoneme dictionary, then this matching phoneme is considered to be a "perfect match" phoneme.

4. Preparation of Training Input

We employ three different sets of training inputs, English-speaking children, English-speaking adults, and Spanish-speaking adults. A separate network is used for each input set. The English-speaking children speech used for training is obtained from Carterette and Jones [1]. The first 1023 words from the transcription of first grade children are used as a training text. The text includes correctly spelled target words attempted by the children and a phonetic transcription of the children's production. These productions are not always accurate. They include connected speech simplifications (e.g., "an-" for and) as well as some sound substitution errors. The phonemes are aligned with the text, consonant with consonant, and vowel with vowel. In case a letter in the text is not pronounced, a special elide symbol /-/ is inserted. Some arbitrary decisions are made (e.g., for th's in the or in bath tub , the phoneme symbol is aligned with the "t" and the elide symbol is aligned with the "h"). If a letter in the text represents more than one phoneme, a special compound symbol is used (e.g., the word one is represented as /*n-/, where /*/ is a compound phoneme for /w^/). We added stress markers to reflect the assumed pattern of stress.

The adult speech used for training is prepared in a similar manner. The first 1008 words from the adult speech transcription in Carterette and Jones are modified to fit the transcription system employed.

The Spanish training input is from a difference source (the CHILDES database from Carnegie-Mellon University [6]). It is a 918 word orthographic transcription of two parents talking to a child. Because there is no accompanying phonetic transcription, we have to create a phonetic transcription with the assistance of a native speaker. Unlike the other input sets, this transcription does not include any speech errors or connected speech simplifications.

In Carterette and Jones, there is no suprasegmental information accompanying the phonemes. The stresses are inserted according to the natural intonation when the words are spoken.

5. Training Procedure

The text, phonemes, and stress markers for each corpus are used as training input for the neural network. Random initial weights are used. The number of passes employed varies, depending, in part, on the leveling off of performance. Statistics are gathered for the English children speech after 25 passes for a best guess model and after 50 passes for a perfect match model. Statistics for the English adult's speech are gathered after passing the training text and phonemes through the neural network 50 times.

The Spanish training input is prepared differently. A 918 word Spanish text transcription of two parents talking to a child is used. The text transcription has no accompanying phonemes. The phonemic transcription is obtained by applying the pronunciation rules in Spanish and then checking the transcription with a Spanish speaker. After 10 passes through the Spanish training input, statistics is gathered for the best guess phoneme output. The neural network trained with adult speech is trained with Spanish as a second language. The same Spanish training input is used. Statistics is gathered after 10 passes through the training text. The neural network trained with Spanish as the first language is trained with adult speech to observe its learning behavior. The adult speech text and phoneme transcription is presented 25 times and statistics is compiled.

The effect of context on learning is studied by varying the window size. Networks with window sizes 1, 3, 5, 7, 9 and 11 letters are trained with 10 passes of Spanish and 10 passes of adult speech separately, starting from random initial weights. Their performance is compared to the statistics of the original network with a window of 7 letters.

6. Differences with NETtalk

In our network, the output layer and the hidden layer consist of nonlinear units. The input layer is a group of linear fan-out units. In NETtalk, the input layer units are nonlinear units that perform sigmoidal transformations on the inputs. The binary input is used instead. It is unnecessary to normalize or compress the range of the inputs. There are several more units in the output layer than NETtalk has. These extra units are essential for representing the articulatory features that are present in Spanish phonemes but not in English. In the input layer, we use a unit that represents the Spanish "ñ". Although there are other letters in Spanish that are not in the English alphabet ("ch", "ll", "rr" are single Spanish letters), we use two characters to represent each of these Spanish letters. These 2-character representations for Spanish letters created difficulty in learning, as we will discuss later in Section 9.

In one of our experiments with bilingual language acquisition, we use an extra language unit in the input. The language unit is clamped when Spanish text is presented to the network. Otherwise, the language unit's activation remains low. This unit is used to indicate which language is presented to the network.

As indicated in Section 4, we put stress marks in words according to natural intonation. Our approach of representing suprasegmental information is different from that in NETtalk. In NETtalk, each word has a heavy accent (for single dictionary word training purposes), which would be unnatural in informal speech. We do not insert stresses in the adult speech. We find that the learning characteristics of phonemes can be separated from the learning characteristics of the the stresses.

7. Performance

The simulations are performed on a Gould NP-1 computer. The throughput is 6 phonemes per second during learning and 20 letters per second when the neural network is reading text and producing phonemes without learning. Vectorizing the simulator to run on the vector processor/arithmetic accelerator would speed up the process by a factor of 10.

Phonemic and suprasegmental outputs are translated to the phoneme representations of TI Speech (a speech generation package by Texas Instruments). The best guess output of the network during training as well as the input used for training is recorded as TI Speech synthesizes speech from the phonemes.

The following 15-second samples are recorded:

Database	Training Passes
Adult	1, 2, 10, 25, 50
Adult-Spanish	1, 2, 3, 4, 10
Spanish	1, 2, 3, 4, 10
Spanish-Adult	1, 2, 3, 10, 25
Children	1, 2, 3, 10, 25
Children-Spanish	1, 2, 3, 4, 10
Spanish	1, 2, 3, 4, 10
Spanish-Children	1, 2, 3, 10, 25
Children	Training Input
Adult	Training Input
Spanish	Training Input

8. Summary of Statistics

The neural network gains 50% accuracy in phonemes and 75% in stresses by the 10th pass when the perfect match criterion is used for determination of phonemes and stresses. By the 50th pass, the accuracy reaches 69% for phonemes and 87% for stresses.

When the best guess decision criterion is used, the accuracy reaches 55% in phonemes and 87% in stress in just 2 passes through the training text. Accuracy continued to rise at a slower rate. By 25th pass, 87% of the phonemes and 93% of the stress are correct. Out of the 1784 consonants, 89% are correct, 84% of the 1192 vowels are accurate, and 56% of 1023 words are pronounced as trained.

The best guess decision criterion is used in the rest of the simulations. For the adult speech training, statistics at the 10th pass is 81% in phonemes, 87% in consonants, 63% in vowels, and 34% in words. By the 25th pass, the accuracy is 87% in phonemes, 92% in consonants, 77% in vowels, and 54% in words. By the 50th pass, the accuracy is 91% in phonemes, 93% in consonants, 87% in vowels, and 65% in words. The accuracy of phonemes is saturated at 93.53% at the 48th, 49th, and the 50th passes.

Training with an orthographically transparent language is very different. In 2 passes through the Spanish training corpus, the accuracy is already 78% in phonemes, 65% in consonants, 86% in vowels, and 33% in words. By the 10th pass, the accuracy reaches 94% in phonemes, 94% in consonants, 99% in vowels, and 71% in words. A 525 word corpus of the continuation of the Spanish training corpus is used as a test for performance in novel words. The accuracy is 92% in phonemes, 93% in consonants, 98% in vowels, and 67% in words, which shows a high degree of generalization.

The neural network trained with adult speech picks up Spanish almost as fast as the neural network trained with Spanish as the first language. The performance is 72% in phonemes, 63% in consonants, 73% in vowels, and 27% in words in the first pass through the Spanish training corpus. By 10 passes, the accuracy reaches 92% in phonemes, 90% in consonants, 99% in vowels, and 66% in words.

The neural network trained with Spanish does not learn English very readily. The performance is 74% in phonemes, 80% in consonants, 39% in vowels, and 21% in words at the end of the 10th pass. By the 25th pass, the accuracy only reaches 81% in phonemes, 84% in consonants, 63% in vowels, and 37% in words at the end of the 25th pass. This is about the accuracy achieved by the neural network trained with only adult speech for 9 passes through the training input.

9. Observations

One objective of the above experiments is to study how closely a multilayer feedforward network trained with the back propagation alogrithm models human speech learning behavior. Besides just gathering the accuracy statistics from the simulations, the output of the netwòrks is examined in greater detail. In particular, we are interested in the specific errors during the course on training, and how those errors related to the characteristics of the input. In general, the networks makes the kinds of errors we see in young children's speech and, sometimes, in adults' speech.

For example, the neural network fills in consonants and vowels where the speakers elide (text: "that", phoneme: /--@-/, output: /D-@t/; text: "and", phoneme: /-n-/, output: /^n-/). In other cases, the neural network elides consonants and vowels where the speakers occasionally pronounce (text: "around", phoneme: /^rW-nd/, output: /^rW-n-/; text: "could", phoneme: /k-^-d/, output: /k---d/).

The networks also makes substitutions for target consonants or vowels that are related in one or more articulatory feature to the actual target (text: "swim", phoneme: /swIm/, output: /swim/; text: "teacher", phoneme: /ti-C-^r/, output: /ti-S-^r/). Sometimes the neural network pronounces a stressed vowel unstressed, and vice versa (text: "and", phoneme: /@n-/, output: /^n-/; text: "and", phoneme: /^n-/, output: /@n-/).

Other types of substitutions and deletions occur. For example, sometimes a stressed vowel is substituted with a more neutral, unstressed vowel, and vice versa (e.g., input text: "and", phoneme: /@n-/, output: /^n/). The network also seems to overcorrect by substitution in the case of flapped, intervocalic stops (e.g., text: "water", phoneme: /wad^r/, output: /wat^r/). The network also deletes consonants in a manner resembling connected speech simplifications (e.g., text: "and he", phoneme: /^n- hi/, output: /^n- -i/). Some of the input productions are appropriately corrected (e.g., text: "jumping", phoneme: /C^mp^n-/, output: /j^mpIn-/). Numerous other errors are observed (e.g., text: "much", phoneme: /m^C-/, output: /m^k-/; text: "edge", phoneme: /E-j-/, output: /E-g-/) that may be attributed either to characteristics of a orthographic phonemic mapping or to the characteristics of networks. The critical questions are the extent to which the output errors are a copy of errors in the input and the extent to which these errors are comparable to those seen in the course of speech development in children. With regard to the second question, three types of errors are important: (1) those that occur in the network output but not in the children speech, (2) those that occur in children speech but not in the network performance, and (3) those that occur in the output of the network and in children speech.

Due to the orthographic transparency of Spanish, the network learns the orthographic-phonemic mapping very readily. This fact is not of particular interest in text-to-speech processing, since a computer program can easily be written to transcribe Spanish text to phonemes. However, the general case of learning coupled with the few specific difficulties observed reveals a great deal about the nature of an adaptive algorithm. By 10 passes through the training input, the English-trained neural network acquiring Spanish makes only a few errors consistently: it does not recognize the Spanish letters "ch", "ll", and "rr". These are single letters in the Spanish alphabet, but each has a two letter representation in the text input line. The neural network outputs /l-/ instead of the correct /L-/ for "ll", and /[[/ flaps instead of the strongly trilled /]-/ for "rr". The neural network is confused with the letter "c" and "ch" in Spanish. It gives the phonemes /s-/ or /k-/ instead of /C-/ for the letter "ch". The Spanish letter "ñ" is entered as "~", but the neural net gives the phoneme /y/ as the output instead of the correct phoneme /˜/.

One solution to the above problem would be to assign special characters for "ch", "ll", and "rr" in Spanish. However, there is a need of preprocessing the Spanish text before training and actual text-to-speech processing. The neural network could also be trained with more passes through the training corpus until the output converges to the desired phonemes. The preprocessing is then incorporated into the network architecture.

9.1. Orthographic Transparency

We transcribe the Spanish text using Spanish pronunciation rules. The Spanish trained network is capable of pronouncing text more accurately than the English trained network. The English trained network retains the errors of connected speech, and appears to generalize them. Children are exposed to corrupted continuous speech signal. This interesting effect could be an explanation for some of the children speech errors.

9.2. Learning Different Languages

We train the network with two different English databases and a Spanish database. For the English database, it takes three times more word presentations for the same level of competence of Spanish. The final Spanish performance of over 90% is achieved after less than 6,000 words of presentation. The English network is slightly over 80% correct phonemes after 24,000 words. Spanish is orthographically more transparent than English. This allows the neural network to learn Spanish words in the second pass of the training set. Unfortunately, the general orthographic transparency of Spanish is confounded with "errorless" characteristics of Spanish input set. Both of these factors together account for the rapid convergence of the networks and its high level of accuracy. However, each clearly contributed to performance. For example, the phonetic coding employed for Spanish created some variation in the orthographic-phonetic mapping (e.g., sometimes "l" orthographically is the phoneme /l/ in the transcription, but sometimes it is the first character in the orthographic "ll" with a correct output of /l-/). Any instance of one-to-many, many-to-one, or many-to-many mapping would make learning more difficult (i.e., slower and, at least temporarily, lower in accuracy). In the long run, some variability may lead to a mapping that is more robust. However, the nature of the units chosen for the code is such that the variation leads to incorrect output, not a more robust category with fuzzy boundaries. Thus, even adding a small number of instances of non-transparent (i.e., inconsistent) mapping of orthographic symbols to a phoneme leads to errors.

The contribution of the errors or the absence of errors in the input set can be seen in another way. Many of the errors that occur in the English input sets are reflected in output, even after performance is at its peak. Furthermore, there is some apparent generalization of errors to words that do not contain the error in the input set and to orthographic characters that are accurately represented in the input. We are currently conducting experiments to determine the specific nature and extent of the generalization.

9.3. Capturing the Characteristics of Speech

Initial consonants are more frequently accurate in the output. In contrast, final consonants tend to be deleted, reflecting the characteristics of the input. Of course, in the Spanish trained network, these connected speech characteristics are not present because they are not present in the input. Thus, the statistical characteristics of the input are captured by the back propagation model. Just as the errors in input are reflected in output, the frequency of character-phoneme mappings is reflected as well in the overall performance and in the relative accuracy of individual mappings. For problems such as speech development in various languages, where the process involves a gradual approximation to the characteristics of the input, this type of characteristic is obviously desirable.

9.4. Training with a Second Language

Statistics compiled from Spanish after a network is trained with adult English input demonstrate the apparent ease with which a network can learn a second mapping. Of course, unlike second language learning, the second language mapping alters the weights for character-phoneme links in any case where there is an overlap. The second language mapping experiment thus yields mixed results. When Spanish is trained as a second language, the final performance is as good as when Spanish is the first language trained. This is not true when English is the second mapping trained. In both cases though, network performance indicates that the mapping of the first language is not retained. At the beginning of the second language training, the errors reflect the initial mapping. For example, the network produces the Spanish "v" incorrectly as a labiodental fricative rather than as a bilabial fricative. The letter "h" is produced as it is in English, rather than deleted as it is in Spanish. With enough training on a Spanish input set, these errors are generally eliminated. Surprisingly, the accuracy of the vowels exceeds that of the network trained with a Spanish as a first language. In contrast, the network trained initially with the Spanish input and then with English performs poorly. After 25 passes, the output accuracy is comparable to the network trained with English after only just 9 passes. The Spanish pronunciation of "v" is retained, and the letter "I" as the phoneme /e/ instead of the correct phoneme /A/.

In an attempt to retain some of the first language mapping, we add an extra language unit in each input letter group. The unit indicates which of the two languages is being presented to the network. Unfortunately, a single unit is inadequate for this purpose. It is simply overwhelmed by the number of other units. The network regards the extra unit's activation as noise. Thus, the results are similar to the training without the language unit.

The network could not sustain two separate mappings with the simple mechanism we use here. In general, English pronunciation rules with all the variation and the semantic dependencies may be beyond the memory capacity and the structure of the network we use.

The differences in the order of language training can be explained in two ways. First, the generally consistent, variation-free mapping of Spanish creates a series of local minima for English that are difficult to overcome. In contrast, the variation of the English mapping precludes these local minima. Thus, the subsequent convergence to Spanish occurs rather readily. An alternate characterization concerns the nature of the mapping in each language, which is reflected in the degree of distribution of the weights.

Specifically, the observed difference in training order may reflect the difference in making the transition from many-to-many mapping in English to one-to-one mapping in Spanish. This may resemble some aspects of the interference that occurs in second language learning.

It seems that the regular mapping of Spanish creates a series of local minima for English that are difficult to erase from the memory. This almost suggests that the patterns of one language cover another. This is an extreme version of the interference that is normally observed in a second language learning process. Care must be taken when training a back propagation network. If the order of pattern presentation is biased, this will be reflected in the mapping. We call this a ***recency order bias*** and it is a complex issue that warrants treatment elsewhere. This phenomenum will pose severe difficulties when trying to add new knowledge to a pretrained network.

9.5. Contextual Dependency

The results of the experiments concerning different window sizes show that a window size of 5 letters is sufficient for the Spanish mapping. This indicates that the context of a phoneme in Spanish is around 5 letters, which agrees with the maximum size of a syllable in Spanish. In Spanish, the vowel values do not depend on other syllables in the word. Spanish vowels are not elided, because Spanish is a syllable-timed language. The pronunciation of Spanish consonants depends mostly on the vowel following the consonant.

In English, a window size of 7, 9, and 11 letters give comparable performance for the neural network. The statistics from the simulations with various window sizes all show that the larger the window, the more accurate the neural network can achieve in a given number of passes through the training text. There are many ways of pronouncing one word (e.g. "read"). Vowels have different values in different words spelled similarly (e.g. "five" and "give"). Consonants and vowels are sometimes elided. The correct pronunciation of a word depends highly on the context, the grammar, and the semantics. Increasing the window size increases the contextual information available to the neural network for determining the correct phoneme.

9.6. Number of Hidden Units

We study the effects of varying the number of hidden units in the network. According to Kung et al. [5] there is a dramatic change in performance in the back propagation algorithm when the number of hidden units surpasses the number of training patterns minus one. This experimental limit, in our opinion, divides the two possible modes of operation of the network. One is a pure look up table mechanism, by dividing the space of M patterns with M-1 hyperplanes. The other is a generalization mechanism, using each hyperplane more than once for a decision boundary.

When the number of hidden units is sufficient to allocate one pattern boundary per hidden unit, basically a look up table is obtained. It has been shown that the training necessary for a given performance decreases to as much as 2 orders of magnitude. Kung et al. hypothesizes that theoretically there must be a learning scheme capable of maximizing class separability with the number of hidden units approximately equal to $\log_2 M$. With the above observation, the optimal point for the algorithm in this particular application occurs with a number of hidden units in the tens of thousands, and would practically be intractable. Computationally, each hidden unit increases the number of operations by the number of input units plus one.

9.7. Cluster Analysis

In order to study the excitation pattern of the hidden-layer units, cluster analysis on their activation levels is performed. First a network is trained to the desired level of accuracy. Then all the weights are frozen. Each training pattern is presented to the network, and let it propagate forward to the hidden-layer and the output-layer. The activation of each hidden-layer unit and the input-output pair are saved. Thus for each input-output pair, a (dimension 80) vector of hidden-layer unit activations is created.

When all the training patterns are presented, if their corresponding central window letter-output phoneme pair of the vectors are the same, they are grouped and averaged. The number of groups or letter-to-phoneme correspondences obtained is more than the number of phonemes. This is due to the fact that many central letters can be mapped to the same phoneme (e.g. "i" and "e" can be mapped to /i/,) and a single central letter can be mapped to different phonemes (e.g. "c" can be mapped to /k/ or /s/) in different contexts.

9.7.1. Clustering Algorithm

With these averaged hidden-layer unit activation vectors, we can perform hierarchical cluster analysis. The Lance and William General Algorithm with complete linkage [3] is used. The clustering algorithm begins with the calculation of a table of distances between any two vectors. Then the algorithm repeatedly finds the smallest distance in the table. The two vectors giving this smallest distance are renamed as a cluster. The distance table is revised by deleting the distances between any other vector and any of these two vectors, and adding the distances between the newly formed cluster and the other vectors. The distance between a vector and the newly formed cluster is the larger of the distances between the vector and the two old vectors that have just been clustered. The iteration stops when the distance table contains only one entry, i.e. when a single cluster is left.

9.7.2. English Training

From the cluster plot of the final (25th) pass of the first grade English speech training, we can observe some striking relationships with the characteristics of the input. The clearest clustering properties we observe are the following. All letter-to-phoneme correspondences are in one cluster, those of the letter "y" in another cluster, and those of the letters "a", "e", "i", and "u" in different clusters. All these vowel letters are in one big cluster different from all other consonant-letter-to-phoneme correspondences.

For the non-vowel-letter cluster, the distinction is not so clear, although we can make the following observations. The most distinct cluster is the word and the sentence boundaries. We also observe that all letter-to-elide-phoneme correspondences are in one cluster. The other smaller clusters are the letter-"t"-to-phoneme correspondences, and the letters-"s","v","f"-to-phoneme correspondences. The letter-"j"-to-phoneme correspondences form a cluster. The letters-"m","n"-to-phoneme correspondences form another cluster.

9.7.3. Spanish Training

The cluster plot of the final (10th) pass of the Spanish training set shows even more striking hidden-layer unit activation patterns. The vowel letters "a", "e", "i", "o", "u", and "y" are in one cluster different from the rest. Moreover there are two letter-"o"-to-phoneme correspondences ("o"-to-/o/ and "o"-to-/c/), and two letter-"e"-to-phoneme correspondences ("e"-to-/e/ and "e"-to-/E/). The other letters "a", "i", "u", and "y" all have one-to-one letter-to-phoneme correspondences. These correspondences follow the Spanish pronunciation rules closely. As the input is an exact transcription of Spanish text from the rules, we can see that the hidden-layer unit activation pattern captures the characteristics of the input.

Apart from the vowel-letter-to-vowel-phoneme correspondences, we also observe that the word boundary and the letter-to-elide-phoneme correspondences form a cluster. The sentence boundary cluster is distinct from the two above.

Within the consonant letter cluster, we can distinguish some sub-clusters from the phoneme part. The voiceless alveolars, voiced alveolars, voiced velars, voiced glides, voiced bilabials, /p,b,k/-stops, flaps, laterals, and nasals each form a cluster.

Comparing with the first grade English speech, the clusters formed by training with Spanish are more distinct, and the mappings are less in number. Not too many letters are mapped to the same phoneme, and a single letter is not mapped to too many phonemes. The distinctness of the clusters may also be considered to reflect the input. The Spanish input is an exact rule-based transcription, while the English input is a transcription of the production of some first grade children. The English transcription contains a significant number of speech errors.

9.7.4. English Developmental Cluster Analysis

As we are interested in understanding the behaviors of the network from the developmental point of view, we perform hierarchical cluster analysis of the hidden-layer unit activations at the end of a number of passes. In the English-trained network, we perform cluster analysis at the end of the 5th, 10th, 15th, 20th, and the 25th passes.

We observe some major characteristics of the cluster development process. The distance between the two most distinct clusters grows from 2.8 to 3.4. The inter-cluster distance is a measure of distinctness. As the network is being trained, it is becoming more effective in distinguishing the letter-to-phoneme correspondences. The elide phoneme cluster gradually becomes grouped with the word boundary and sentence boundary. The vowel letters become more and more separate from the non-vowel letters.

There are some other characteristics of particular interest. At the end of the 5th pass, the word boundary and the sentence boundary form the most distinct clusters, separating themselves from the other non-silence letter-to-phoneme correspondences. The same can also be observed from the phonemic output of the network. We observe that after just one pass through the training set, the word and sentence boundaries are the only correct letter-to-phoneme correspondences. The word and sentence boundaries are the most frequently appearing characters in the input (about 1 in every 6 characters). In the output representation, their articulatory representations are also far different from the other phonemes. The word boundary has pause and elide as features. The sentence boundary is represented by the pause and fullstop features. The elide phoneme has silent and elide as features. It is understandable that the elide phoneme becomes grouped with the word boundary.

The letter-to-elide-symbol correspondences are first grouped with some consonants and the vowel letters, as we can observe from the 5th pass. At the 10th pass, the network is eliding less letters, showing its gradual regularizing of letter-to-phoneme correspondences. From that point on, the number of letters that are elided by the network increases. Also, such letter-to-elide-phoneme correspondences are migrating toward the non-elide groups (e.g. "t"-to-/-/ groups with "t"-to-/t/).

9.7.5. Spanish Developmental Cluster Analysis

Developmental cluster analysis on the 2nd, 4th, 6th, 8th, and the final 10th pass of the Spanish-trained network is performed. Our observation is, in some aspects, similar to those observations drawn from the English-trained network. The distances between clusters gradually increase. The vowel letters and non-vowel letters become separate clusters. The word and sentence boundaries become distinct early in the training.

In addition, we observe that the number of correspondences decreases from 109 to 53. This reflects that the network is organizing its letter-to-phoneme correspondences according to the characteristics of the training input. Early in training, the network mapping is a chaotic, many-to-many mapping. At the end of the 10th pass, the mapping converges to somewhere around 1-to-1 and 1-to-2.

The migration of the vowel letter cluster is also worth noting. At the 2nd pass, the vowel letters "a", "e", "i", and "o" form a cluster, with letters "u" and "y" being clustered with some other consonants. At the 4th pass, "i" leaves "o", "e", and "a" to join "u" and "y" among the consonants. Letters "o", "e", and "a" become the most distinct cluster second to the word and sentence boundaries. At the 6th pass, all the vowel letters are in one cluster distinguished from the other consonants, but are grouped with the word and sentence boundaries. At the 8th pass, the vowel-letter-to-phoneme correspondences are reduce to a compact, almost 1-to-1 mapping. At the 10th pass, the vowel-letter cluster is no longer grouped with the word and sentence boundaries, but is a cluster all by themselves. The consonants become grouped with the word and sentence boundaries.

9.7.6. Spanish-English Training Cluster Analysis

When the Spanish-trained network is trained with first grade English informal speech as a second language, a new developmental cluster analysis is performed. The most significant observation is the increase in the number of letter-to-phoneme correspondences from 53 to 86 in just 5 passes. This number gradually increases to 105 at the 25th pass of English. Although the final number of correspondences is about the same as that of the English-trained network at the 25th pass, close scrutiny reveals that the mapping or correspondences are not identical. The cluster hierarchy is also different.

The other significant change with the second language training is the increase in distance between the most distinct clusters, which grows from 2.5 at the beginning to over 3.5 at the 25th pass. We also observe that the vowel-letter-to-phoneme correspondences merge with some consonant clusters to a small extent. Specifically, the correspondences involving letters "u" and "i" leave the main cluster of vowels and join a cluster with the nasal letters "m" and "n", and stop letters "t" and "d".

9.7.7. English-Spanish Training Cluster Analysis

The English-trained network exhibits different clustering characteristics when trained with Spanish as a second language mapping. Neither are the developmental characteristics similar or an exact reverse of the Spanish-English counterpart.

We can see that the distance between the two largest clusters grows from 3.4 to 3.6 before settling down to 3.4 again in the 10th pass of Spanish. This suggests that the English mapping is a local minimum with respect to the Spanish mapping. Training with 10 passes of the second language provides energy to upset the mapping before settling down to the global minimum.

The vowel-letter cluster migrates in a manner different from that exhibited by the Spanish network trained with English as a second language. The vowel letters are originally a major cluster separate from the consonants and the word and sentence boundaries. With the training of Spanish, they first join the word boundary. Then some correspondences of "u" leave to join the consonants. Finally the letter-"u"-to-phoneme correspondences rejoins the vowel cluster. The word boundary rejoins the sentence boundary and the consonant cluster.

There is also a reduction in the number of correspondences in the vowel cluster. There is one mapping from each of the letters "a" and "u", two from each of "e", "o", and "y",, and three from the letter "i". In the network trained with Spanish only, there are two mappings from the letters "e" and "o", and one from each of the letters "a", "i", "u", and "y".

Overall, the number of correspondences drops from 106 to 53. However, these are not the same 53 correspondences found in the Spanish-trained network. As noted above, some vowel-letter correspondences are not erased. Some consonant-letter correspondences never appeared.

For example, the second r of the double Spanish letter "rr" has an elide symbol associated with it in the training input. This correspondence is learned in the Spanish-trained network but not in the English-trained network with Spanish as a second mapping. The English network also retains some English mappings when trained with Spanish. For example, the letter-"n"-to-nasal-phoneme-/G/ correspondence and the letter-"v"-to-phoneme-/v/ correspondence are retained. Moreover, the English mapping helps retain the letter-"f"-to-phoneme-/f/ correspondence and the letter-"k"-to-phoneme-/k/ correspondence. These two correspondences are not learned by the original Spanish-trained network.

10. Output Decision Algorithms

The "best guess" and the "perfect match" strategies are used as ways of making output phoneme decisions based on the output vector and the phoneme dictionary. The best guess is a computationally more expensive strategy using the minimal Euclidean distance between the output and a phoneme vector. It performs between 20 to 30% better than the perfect match, which discards all values that do not represent the phonemes. We derive the back propagation algorithm as a function to minimize the expectation of the square of the difference between the output and the target value. In the derivation, a simple 50% thresholding scheme for the output emerges. The "perfect match" is modified to accommodate this change and it performs as well as the "best guess" with minimal computing.

11. Problems, Solutions, and Potentials

Possibly, the most important conclusion from this work is the dependency of the network mapping on the statistical characteristics of the input data. This has the following consequence. All the interesting behaviors of the network can be accounted for by the statistics of the input, including the input order, and the difficulty in recovering from a previous mapping. In so far as the application can be modeled by mapping and generalization of input characteristics, neural networks are satisfactory. Back propagation networks do not scale well [12, 13] and some scaling study of the data mapping should be carried out before an application can be trained.

For studies of speech development, several interesting aspects could be captured by the back propagation algorithm. Other mechanisms are still to be accounted for, and will require additional subsystems. Among those features that require attention are: top down processing of speech, grammatical constraints, etc.

References

[1] Carterette, E. C. and M. G. Jones, *Informal Speech,* University of California Press, Los Angeles, 1974.

[2] Denker, J., "Neural Network Models of Learning and Adaptation," AT&T Internal Communications, 1986.

[3] Diday, E. and J. C. Simon, "Clustering Analysis," in *Digital Pattern Recognition,* ed. K. S. Fu, Springer-Verlag, Berlin, 1980.

[4] Hopfield, J. J., "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Science USA, Biophyics*, vol. 79, pp. 2554-2558, April 1982.

[5] Kung, S. Y., J. N. Huang, and S. W. Sun, "Efficient Modeling of Multilayer Feedforward Neural Nets," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing,* New York, 1988.

[6] MacWhinney, B., "Manual for the Codes for the Human Analysis of Transcripts (CHAT Manual)," Carnegie Mellon University Department of Psychology, September 1987.

[7] Minsky, M. and S. Papert, *Perceptrons: An Introduction to Computational Geometry,* MIT Press, 1969; Expanded Edition, 1988.

[8] Rumelhart, D. E., G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* ed. D. E. Rumelhart, J. L. McCelland, and the PDP Research Group, vol. 1, Foundations, pp. 318-362, MIT Press, Cambridge, 1986.

[9] Sejnowski, T. J. and C. R. Rosenberg, "NETtalk: A Parallel Network that Learn to Read Aloud," The Johns Hopkins University Departments of Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01, 1986.

[10] Sejnowski, T. J. and C. R. Rosenberg, "Parallel Networks that Learn to Pronounce English Text," *Complex Systems,* vol. 1, pp. 145-168, 1987.

[11] Tenorio, M. F., M. D. Tom, and R. G. Schwartz, "Adaptive Networks as a Model for Human Speech Development," *Proceedings of the IEEE 2nd International Conference on Neural Networks*, vol. 2, pp. 235-242, San Diego, CA, July 1988.

[12] Tesauro, G., "Scaling Relationships in Back-Propagation Learning: Dependence on Training Set Size," *Complex Systems*, vol. 1, pp. 145-168, 1987.

[13] Tesauro, G. and R. Janssens, "Scaling Relationships in Back-Propagation Learning: Dependence on Predicate Order," TR-CCSR-88-1, Center for Complex Systems Research, University of Illinois at Urbana-Champaign, 1988.

[14] Tom, M. D., "A Neural Net that Simulates Human Speech Development," Purdue University School of Electrical Engineering Internal Report, December 1987.

[15] Tom, M. D., R. G. Schwartz, and M. F. Tenorio, "NETtalk as a Simulation of Human Speech Development," *Neural Networks*, vol. 1, sup. 1, p. 319, Pergamon Press, 1988.

Appendices

Articulatory Representations of Output Units for Consonants

Phoneme	Example	Articulatory Features		
/p/	*p*et	Unvoiced	Labial	Stop
/b/	*b*et	Voiced	Labial	Stop
/t/	*t*est	Unvoiced	Alveolar	Stop
/d/	*d*ebt	Voiced	Alveolar	Stop
/k/	*c*at	Unvoiced	Velar	Stop
/g/	*g*et	Voiced	Velar	Stop
/B/	*v*en	Voiced (Spanish)	Labial	Fricative
/f/	*f*in	Unvoiced	Labiodental	Fricative
/v/	*v*est	Voiced	Labiodental	Fricative
/T/	*th*in	Unvoiced	Dental	Fricative
/D/	*th*is	Voiced	Dental	Fricative
/s/	*s*it	Unvoiced	Alveolar	Fricative
/z/	*z*oo	Voiced	Alveolar	Fricative
/S/	*sh*in	Unvoiced	Palatal	Fricative
/Z/	lei*s*ure	Voiced	Palatal	Fricative
/V/	ti*g*re	Voiced (Spanish)	Velar	Fricative
/C/	*ch*in	Unvoiced	Palatal	Affricative
/J/	*g*in	Voiced	Palatal	Affricative
/m/	*m*et	Voiced	Labial	Nasal
/$/		Voiced (Spanish 'mf')	Labiodental	Nasal
/M/	absy*m*	Voiced	Dental	Nasal
/n/	*n*et	Voiced	Alveolar	Nasal
/N/	butto*n*	Voiced	Palatal	Nasal
/G/	si*ng*	Voiced	Velar	Nasal
/l/	*l*et	Voiced	Alveolar	Lateral
/L/	bott*le*	Voiced	Palatal	Lateral
/r/	*r*ed	Voiced	Alveolar	Liquid
/R/	bi*r*d	Voiced	Velar	Liquid
/w/	*w*et	Voiced	Labial	Glide
/y/	*y*et	Voiced	Palatal	Glide
/?/	*uh*	Glottal	Stop	
/h/	*h*ead	Unvoiced	Glottal	Glide
/H/	*j*aletina	Voiced (Spanish)	Glottal	Glide
/[/	ki*tt*y	Voiced	Alveolar	Flap
/]/	a*rr*iba	Voiced (Spanish)	Alveolar	Trilled

Articulatory Representations of Output Units for Vowels and Compounds

Phoneme	Example	Articulatory Features				
/i/	h*ee*d	High	Tensed	Front1		
/I/	h*i*d	High	Front1			
/|/	log*i*c	High	Front1	Front2		
/E/	h*ea*d	Medium	Front1	Front2		
/@/	h*a*d	Low	Front2			
/a/	h*a*	Low	Tensed	Central2		
/^/	h*u*d	Low	Central1			
/x/	*a*bout	Medium	Central2			
/O/	h*o*y	Medium	Tensed	Central1	Central2	
/c/	h*o*d	Medium	Back1			
/o/	h*o*ld	Medium	Tensed	Back2		
/U/	h*oo*d	High	Back1			
/u/	wh*o*'d	High	Tensed	Back2		
/e/	h*a*te	Medium	Tensed	Front2		
/A/	h*i*gh	Medium	Tensed	Front2	Central1	
/W/	h*ou*se	High	Medium	Tensed	Central2	Back1
/Y/	h*u*ge	High	Tensed	Front1	Front2	Central1
/*/	*o*ne	Voiced	Labial	Glide	Low	Central1
		(Compound:	/w/ + /^/)			
/X/	e*xc*ess	Unvoiced	Velar	Stop	Alveolar	Fricative
		(Compound:	/k/ + /s/)			
/K/	se*x*ual	Unvoiced	Velar	Stop	Palatal	Fricative
		(Compound:	/k/ + /S/)			
/Q/	*qu*est	Velar	Stop	Voiced	Labial	Glide
		(Compound:	/k/ + /w/)			
/!/	Na*zi*	Unvoiced	Alveolar	Stop	Fricative	
		(Compound:	/t/ + /s/)			
/#/	e*x*amine	Voiced	Velar	Stop	Alveolar	Affricative
		(Compound:	/g/ + /z/)			
/_/	Word Boundary	Pause	Elide			
/./	Period	Pause	Fullstop			
/-/	Continuation	Silent	Elide			

Symbol	Suprasegmental Features
/1/	Primary Stress
/0/	Secondary Stress
/2/	Tertiary Stress
/(/	Syllable Left Boundary
/)/	Syllable Right Boundary

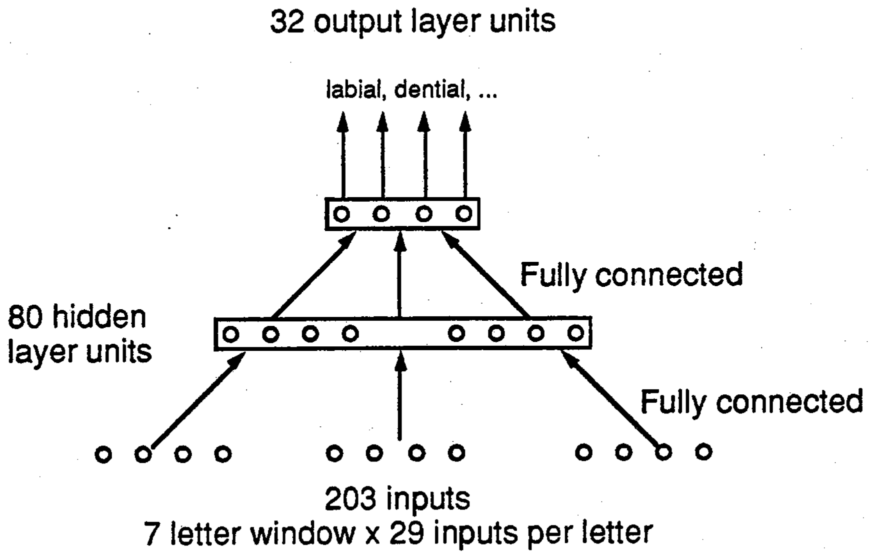

Figure 1. NETtalk architecture for informal children, adult, and Spanish speech training.

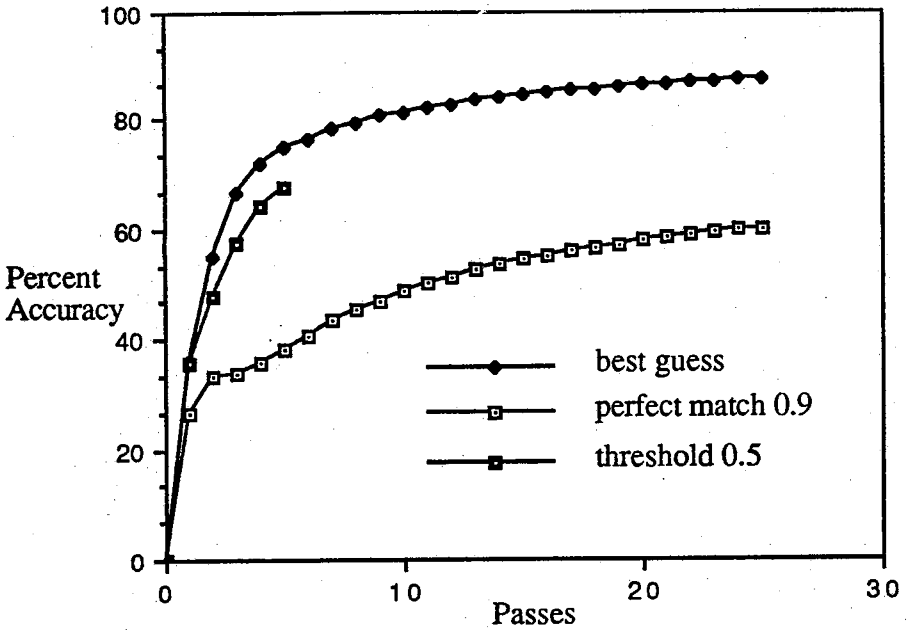

Figure 2. Learning curves for phonemes during training on children informal speech.

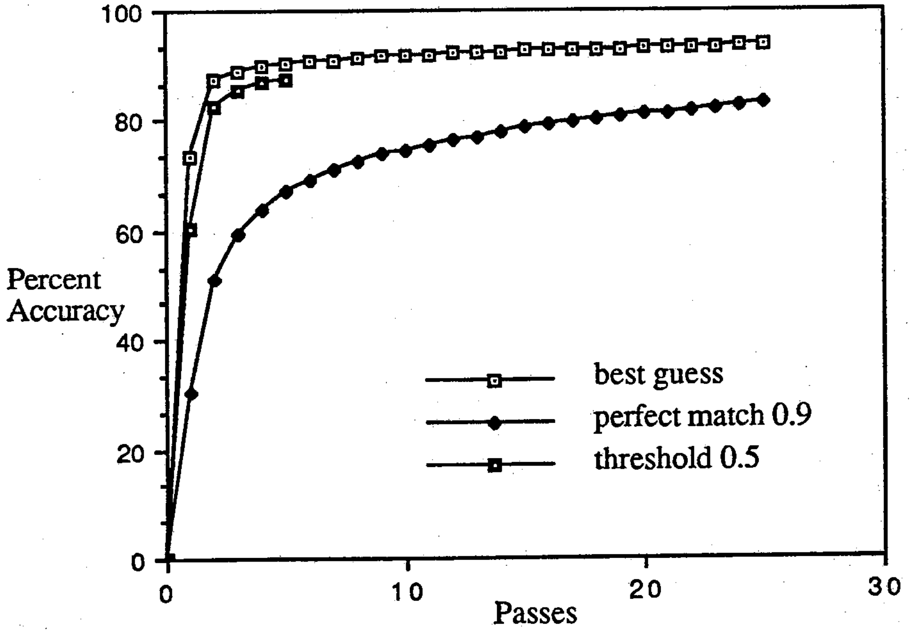

Figure 3. Learning curves for stresses during training of children informal speech.

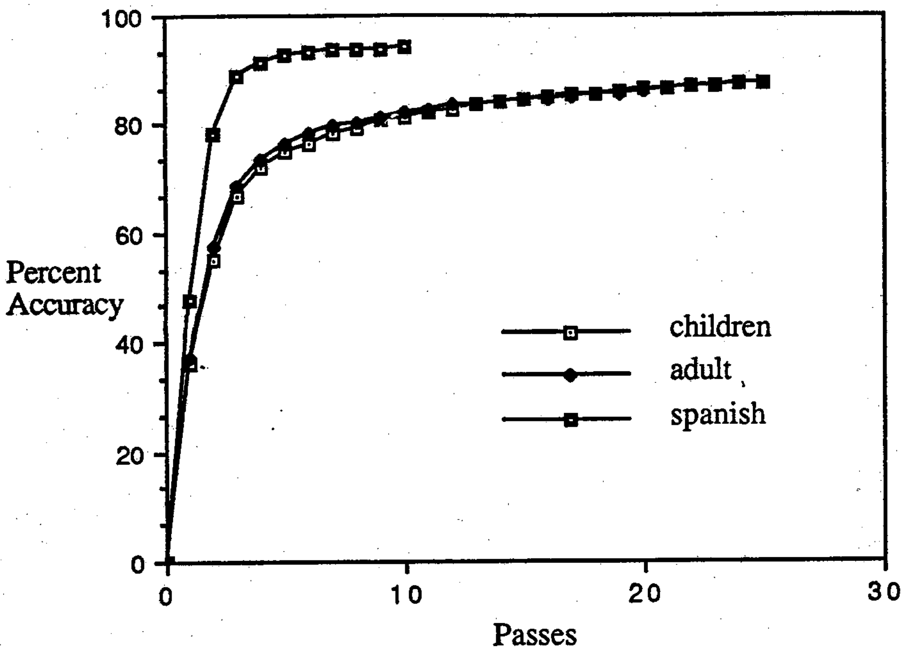

Figure 4. Learning curves for phonemes during training on children informal speech, adult speech, and Spanish.

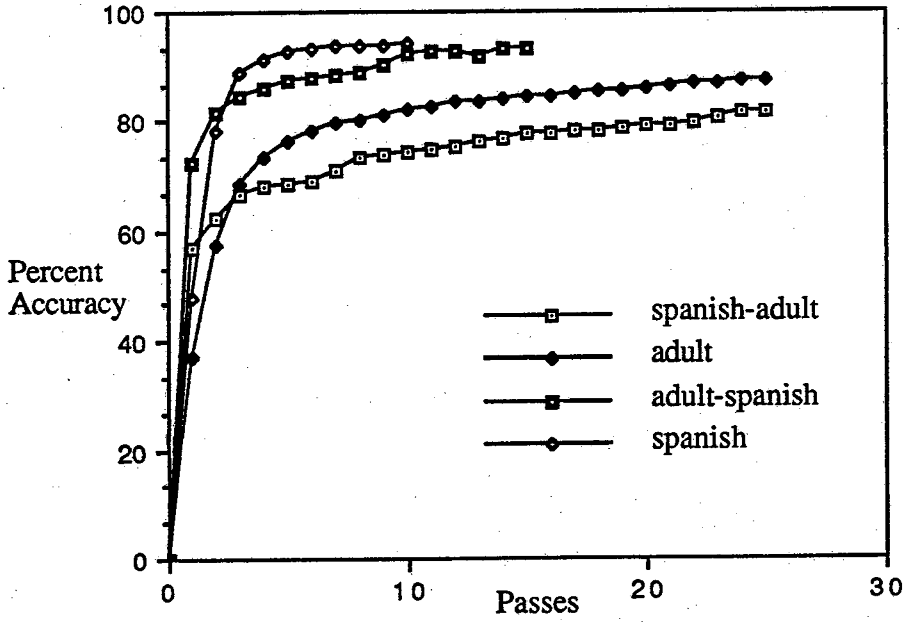

Figure 5. Learning curves for phonemes during training on Spanish and adult speech for first and second mappings.

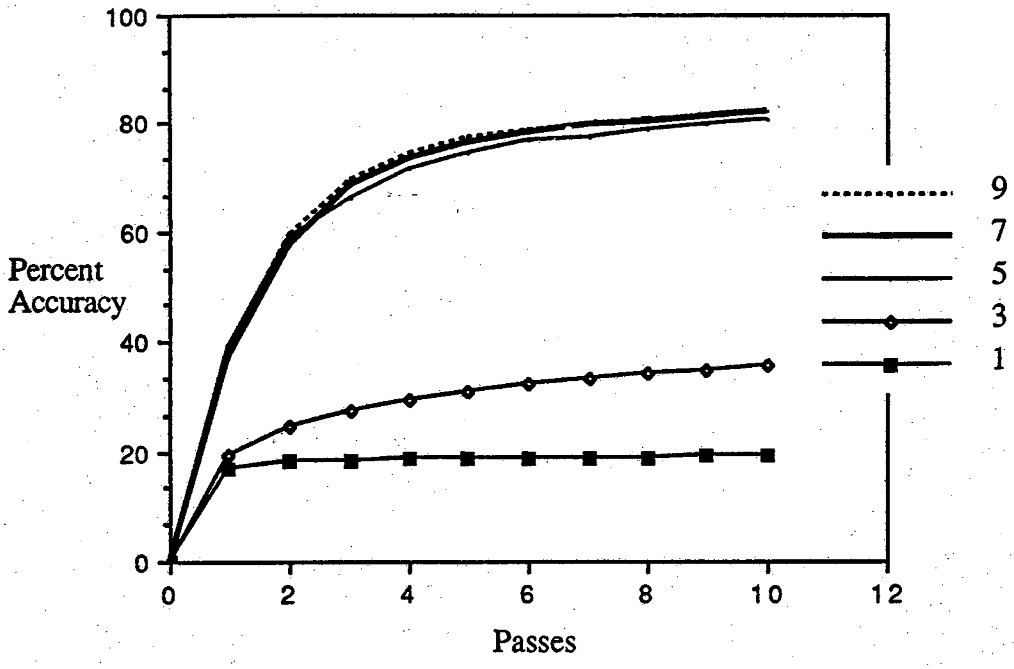

Figure 6. Learning curves for phonemes during training on adult speech using different window sizes.

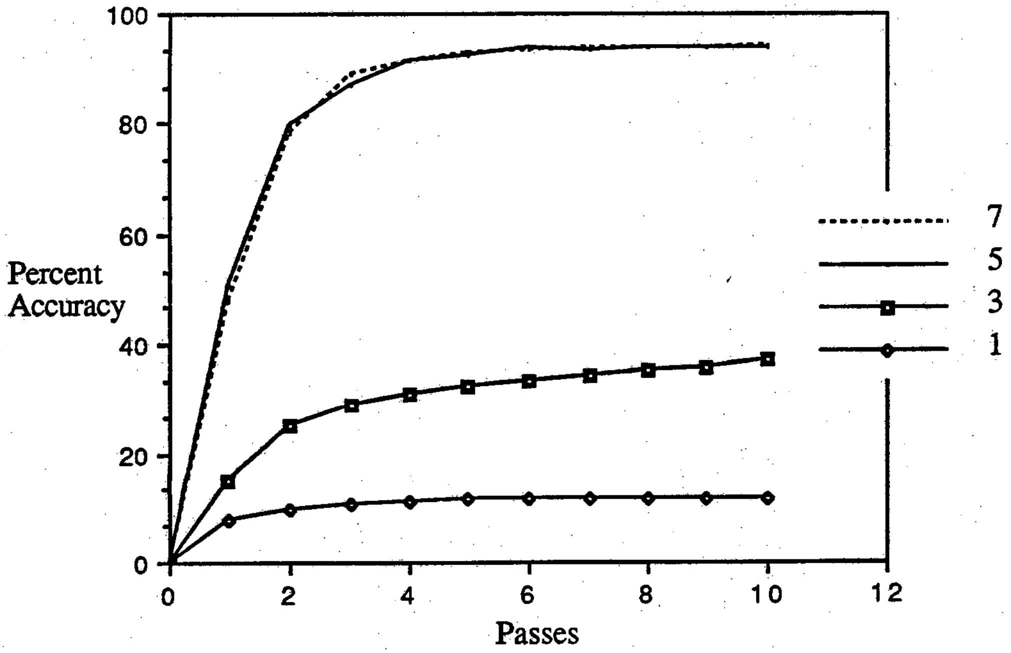

Figure 7. Learning curves for phonemes during training on Spanish using different window sizes.