A Government Phonology Approach To Automatic Speech Recognition

Simon Ahern

MSc in Speech and Language Processing
1999

Table of Contents

Abstract

The theory of Government Phonology has been incorporated in a rule-based way into the Speech Recognition process in recent years. This project automates the application of aspects of this theory through the use of Artificial Neural Networks. The system of Government Phonology primes is detected from acoustic signals, and recognition performance shows meaningful results when tested on the TIMIT corpus of read speech (61.3% correct). The phenomenon of prime headedness is also incorporated into this frame-by-frame recognition. Finally, the importance of a speaker's gender was found to be significant to the performance of speech recognition within this framework, while variation in a speaker's dialect region was not found to be so crucial in developing acoustic models.

List of Figures

List of Tables

Introduction

Human listeners have very impressive abilities when it comes to understanding speech. They seem to be unconcerned by adverse acoustical conditions and are able to resolve a wide range of allophonic variations like assimilations and deletions with apparent ease (Lahiri, 1999; 715). Although there are no identifiable boundaries between sounds or even words at normal speaking rates (Owens, 1993; 138), people can subdivide and recognise units as part of the whole, without any apparent effort.

While humans can produce a large number of sounds, each language has a small set of abstract linguistic units called phonemes that describe its sounds. A phoneme is the smallest meaningful contrastive unit in the sound system of a language. Phonemes may be associated with linguistic features or articulatory configurations; for example the phoneme /s/ has the features unvoiced, fricative and alveolar. Producing an /s/ involves an open glottis, a raised velum and a single narrow constriction in the alveolar region. Each phoneme results from varying the shape of the vocal tract, through muscular control of the speech articulators (Appendix A), the lips, tongue and jaw (Owens, 1993; 138). Considering the variation between speakers' vocal tract lengths and their articulators, it is impressive that phonemes can be so easily identified and classified, despite their range of possible pronunciations.

Added to this, environmental factors such as background noise, rate of speech, changes to phonemes in the context of other phonemes (coarticulation) or even a cold can bring about interference and auditory changes to speech. Differences in dialect often include leaving out certain sounds or replacing one sound with another (Owens, 1993; 138). Yet, addressees have

minimal difficulties in most domains and have mechanisms for filtering relevant speech and hypothesising words based on often partial information about phonemes.

When it comes to machines understanding human speech, there is a limited attempt to emulate these human abilities. Automatic speech understanding is the process by which a computer maps an acoustic speech signal to some form of abstract meaning of the speech. Samples of each of the relevant units are acquired, analysed and a parametric representation, such as formant frequencies of speech or LPC coefficients, is stored in the system (Owens, 1993; 107).

Many recognisers pass directly from the parameterisation of the speech signal to similarity measures without any further data reduction or linguistic interpretation. Others, however, prefer to further hone the speech representation to 'acoustic features' such as formant frequencies and/ or 'phonetic features' such as 'silence', 'vowel', 'frication' etc. (O'Shaughnessy, 1990; 422). The greater computational power of modern computing systems means that several alternative analyses of a particular sound segment can be offered. The evidence can also be weighed to derive estimates of the relative probabilities about the occurrence of different phonemes. A higher level can then interpret the alternative phonetic sequences to give valid words and phrases. Studies tend to focus on specific acoustics derivable from the waveform and spectral parameters, such as formant frequencies and bandwidths, F0 average and range, spectral tilt, differences in final lengthening and differences in phonetic realisation.

There have been many linguistic studies into the nature of human speech. Several theories about sound systems and the rules governing them (phonology) have been suggested. If any of the knowledge about human speech behaviour can be captured, formalised and implemented into an automatic speech recognition framework, it ought to provide greater improvement to their performance. This is the hypothesis that will be explored in the present study.

The central aims of the current project are to investigate the application of phonology to automatic speech recognition systems. In particular, the approach known as Government Phonology has been proposed by some researchers (Brockhaus and Ingleby, 1998; Williams, 1998) as containing some of the solutions to the need for greater phonological knowledge in these engineering applications.

Government Phonology contains speaker-independent phonological features, called primes, which correspond to acoustic patterns formed in human speech segments. The current study sets out to investigate the saliency of these primes, by constructing a working speech recognition application using primes as the form of representation. This will be achieved through the use of artificial neural networks, a computational tool that performs complex parallel pattern matching. This is seen to be an extension of the present rule-based approaches (Brockhaus and Ingleby, 1998, Lahiri, 1999, Reetz, 1999) in that the networks can train and learn from example automatically. The expectation is to attain an acceptable level of frame-by-frame phone (speech segment) recognition.

Provisional to the completion of this aim, the extended aims of the project are to determine which factors are influential on the operation of Government Phonology-based speech recognition. The performance of a speech recognition system on different speakers should be robust to factors such as the gender of the speaker and his/ her dialect region or accent type.

Chapter One of the present study will show how the limitations of automatic speech recognition could be improved with a theory that included a simpler method of representing sounds and capturing changes that occur in certain environments. The theory of the classification of sounds will be shown to be inadequate and the introduction of a new approach may be the best hope of yielding greater accuracy in speech recognition results. Chapter Two will examine the theory of Government Phonology, discussing its approach and the advantages to this perspective. Chapter Three is related to the materials and methods used in the present investigation. This will include an outline of previous work undertaken in the field and a description of each experiment undertaken. Chapter Three will also consist of a summary of Artificial Neural Networks, the tool that will be employed in all of the experiments. The results for each experiment, comprising graphs and tables, will be contained in Chapter Four. Finally, Chapter Five will be reserved for a discussion of the major findings and the implications that arise.

Chapter One Automatic Speech Recognition

1.1 Introduction

This chapter is intended to outline in broad terms the current approach to Automatic Speech Recognition (ASR). It should become apparent that the method that has become most successful is that based upon sheer statistical power, involving gauging the highest probability for the likely identity of each sound segment. Such probability distributions are estimated from large samples of data.

The strength of this method is that the information is gleaned through real-life speech samples, which do not always conform to abstract linguistic theories of speech production. However, there are also weaknesses that arise due to the peculiar aspects of language and the limitations of the current theory of the representation of sounds. The Chapter will conclude by examining the predominant theory used in many current speech applications and by highlighting the areas in which it could be improved.

1.2 How ASR works

Standard speech recognition systems consider words to be composed of sub-word units called *phones*. The acoustic features are considered to be relatively invariant in the middle of a phone and the co-articulation effects are captured in the transition between two or three phones (Owens, 1993; 107). A Hidden Markov Model, which characterises the speech signal as a parametric random process can join these phones units, and a language model then assigns probabilities to word sequences. The most probable word sequence is calculated and the sentence is then 'recognised'. There are actually several levels of processing. At the lower lever, there is acoustic input, which feeds into a higher level where the modelling of words occurs. All of this information is incorporated into a model of language (See Figure 1.1). There is a reliance on statistical techniques, in order to deal with the inter-speaker and token variability that occurs. The

advantage to this approach is that statistical methods can take real data, spoken in everyday speech, as their input. Linguistic theories of speech may often fail in engineering applications due to their idealisation of the way that people *ought* to use language. To rely solely on actual utterances in determining probabilities has been found to be more reliable. Very wide acoustic variations occur when different people speak the same phoneme, because of differences in vocal tracts (Owens, 1993; 138). Therefore, computers are better suited to discovering the probabilistic values that are assigned to a sound's identity.

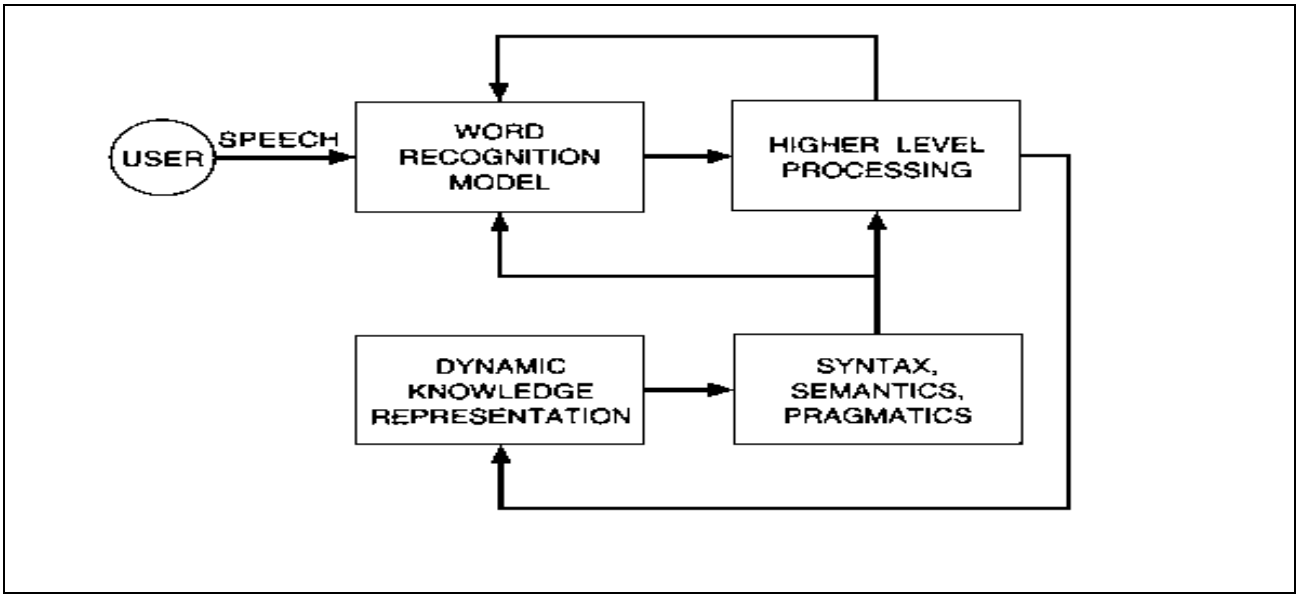


Figure 1.1 General block diagram of a task-oriented speech-recognition system

(From Rabiner and Juang, 1993; 3)

1.2.1 Drawbacks to Current ASR approaches

There are several disadvantages to this purely statistical approach as well. In recent years, it has been suggested that improvements in the accuracy of automatic speech recognisers may depend more on a solid linguistic theory that deals with such factors as the plausible combinations of sounds, their structure in syllables and words and their changes in differing contexts (allophonic variation). The practice of using context-dependent phone models to represent allophonic variation is essentially a pattern-matching exercise, lacking a theoretical basis. Also, Hidden Markov Models (HMMs) are still speaker and transmission-line dependent, or operate only with a restricted vocabulary, syntax and semantics (Lahiri, 1999; 715). By continuing a stream of

research based on pure statistics alone, without the benefits of phonological knowledge, the subtleties of consistent language rules may be missed. It is foolhardy to attempt to solve the complexities of ASR with little or no theoretical understanding of the issues involved in natural language.

> *Phonological knowledge tends to give way to brute-force pattern matching, with little or no use of 'abstract' linguistic knowledge* (Williams, 1997; 79).

It follows that there ought to be a solid linguistic foundation upon which to postulate the correct output. To reduce the situation to black and white standpoints; whereas a linguist is interested in the underlying structure of language that explains how meaning is encoded, or why only certain sound and sentence structures occur, the engineer is only interested in successful communication and building machines that work. Systems built on underlying structures of linguistic competence have very poor engineering performance. Examples include speech recognition systems with only 15% word accuracy and parsing systems that only provide acceptable syntactic structures for 60% of sentences found in newspapers (Magerman, 1994).

There are several aspects of recognition that could be aided by incorporating abstract linguistic knowledge into the underlying representation of language. For example, listeners use syllabic and metrical phonological cues, whereas ASR systems typically do not. Listeners do not seem to have much difficulty adjusting to different speakers, whereas engineering systems typically require training for each speaker, for best performance. A continuous speech system operates on speech in which words are connected together, i.e. not separated by pauses. Continuous speech is more difficult to handle because of a variety of effects. It can be difficult to find the start and end points of words. A further problem is *coarticulation*, whereby the production of each phoneme is affected by the production of surrounding phonemes. Similarly the start and end of words are often affected by the preceding and following words. At present, such regularly occurring phenomena are not captured formally in the framework of speech recognition systems.

1.2.2 How to remedy the problem

These problems all seem to require a better understanding of speech and its universal nature. It would be beneficial if the study of human speech could lead to better chances of identifying some approaches to automatic phonetic segmentation. Currently, the sound representations underlying ASR work require that a separate language-dependent phone model must be generated for each language. A more general yet efficient way of representing the possible phone inventories in all languages would be preferable. In addition, the search time for a given segment is proportional to the size of a system's vocabulary, and a better conception of sounds and their constraints would lead to quicker elimination of unlikely possibilities and more accurate identification overall.

There seems to be an obvious solution to this situation. Linguists have been studying the nature of language for many years and have developed extensive and comprehensive theories of many aspects of language. One area of linguistics that may have the greatest benefit for practical research is that of acoustic modelling. It is an area that is well developed, with the potential for application in technology. If successful, this would aid in the classification of each sound to be identified. It would also increase the reliability of partial information about segment identities (Frazier, 1987; 160). The aspect of linguistics that can provide the most insight into acoustic modelling is phonology, which operates in conjunction with phonetics in this domain.

1.3 Phonology

In linguistic theory, phonology is the study of 'the system of sounds that are manifested by natural language' (Bird; 1995). One role of phonology is to predict the likelihood that a phonetic representation corresponds to a possible realisation of a word. Phonological rules operate on the phonological units of a language in a predictable manner. This type of information is of great importance for capturing any known regularity of the language. If embedded properly within the recognition system, it should aid in the better performance of that device (Shoup, 1980; 125).

Some machines attempt to capture the effects of coarticulation in continuous speech via *ad hoc* phonological rules, which note how the standard phone composition of words can change due to the context of adjacent words (O'Shaughnessy, 1987, 473). Phonological rules can also apply across word boundaries (e.g. geminate reduction shortens the duration of successive identical phones). Recognition performance can be improved if such phonological rules are incorporated into hypothesis generation.

The current understanding of phonology in an engineering context has roots in the work of Chomsky and Halle, as outlined in the Sound Pattern of English, in 1968 (hereafter SPE). Rather than considering phonemes to be the minimal units of a language, a small set of orthogonal (often binary) properties or features are used to classify phonemes. Examples of these features include sonorant, anterior, continuant, rounded, syllabic, consonantal, etc. and are based on acoustics and articulation. These are used to describe either consonants or vowels, and they are represented with concepts such as markedness and linking conventions. Features are combined in a loose structural way, as in Figure 1.2. The variables α, β, γ and δ are filled with either a plus or minus to indicate presence or absence.

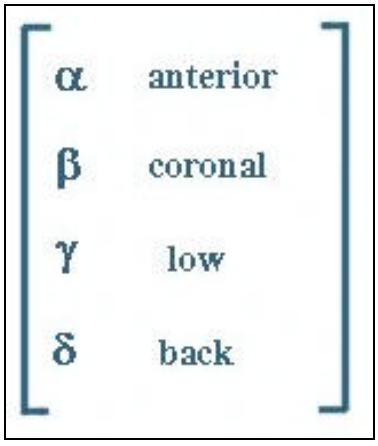


Figure 1.2 An example of SPE features

(Anderson and Ewen, 1987; 16)

> *Much of the phonological research work of the past twenty years has focused on phonological representations: on the make-up of individual segments and on the prosodic hierarchy binding skeletal positions together* (Ingleby and Brockhaus, 1998; 579).

This approach to phonology has enjoyed much attention in linguistic circles and has established itself as the predominant paradigm in the theoretical examination of phonological phenomena. However, there are certain limitations to this framework, which have resulted in criticism and a re-examination of the explanatory power of the theory.

1.3.1 Criticism of the SPE approach

According to Anderson and Ewen (1987; 16), "Chomsky and Halle's classification of classes lacks independent motivation". Anderson and Ewen argue that classes of phonological segments are not random. Rather, "phonological classes and the regularities into which they enter have a phonetic basis" (Anderson and Ewen, 1987; 8). Others feel that part of the central dogma of phonology since SPE is the assumption that any two forms that are both phonologically and morphologically similar must be derived (at least in part) from a common source (Kaye, 1996; 1). This restricts the range of possible interpretations. In addition, the realisation of a phonological segment varies according to producer, context, environment and repetition, but this approach does not yield a comprehensive understanding or motivation for this phenomenon.

Thus although in principle a speech recogniser might only need a few features of a series of segments to be able to identify an underlying word uniquely, in practice the fallibility of the features gives rise to multiple word candidates. The feature system overgenerates and permits feature combinations such as [+high, +low] which cannot occur in any natural language (Rennison, 1986; 282). Linear Generative Phonology (SPE) is a segmental theory that still maintains the notions of derivations, rule ordering and the cyclicity of rule applications. It is inherently procedural and unrestricted and for this reason cannot be considered representationally adequate for computational phonological modelling (Carson-Berdsen, 1998; 25).

According to Jelinek;

> *"The efficacy of the methods … depends crucially on the accuracy of the underlying phonology, which we have here assumed to consist of a lexicon of base forms. It has been lately recognised that the base form lexicon approach to phonology is inadequate, since it does not take into account any speech dynamics, such as rate, emphasis, prosody or speech mode"* (Jelinek, 1996; 217-218).

In a modern speech recognition system, the theory of phonology is undertaken by a pronunciation dictionary, the theory of syntax by N-grams, and a semantic theory by post-processing of the N-best utterance interpretations. In a knowledge-based system, the role of phonology is to parse transcriptions; the role of the lexicon is to find words that match the phonological sequence, while the role of syntax is to erect phrase structure. In modern ASR the task of generating hypotheses is left to the search engine. A system of constraints is required in order to aid the search engine. There are several constraints within the phonological theory extended from SPE, but they are basic concepts (e.g. a preference within languages for symmetrical structure, enforced either by phonological rules or by general principles), lacking independent motivation.

With the failings of SPE becoming apparent, a new phonological approach is required, which does not require an inventory of phonemes made up of many underlying classes for each language. Unlike the linear segmental model, theories of non-linear and dependency phonology have an architecture which can easily be modelled as a computational system. This has influenced the development of the theory that has come to be known as Government Phonology, discussed in Chapter Two.

Chapter Two The Theory

2.1 Government Phonology

Although it was intended as a contribution to phonological theory, SPE was also directly implementable on computer and it was an important foundation for work in speech technology (Bird, 1995; 2). It would be advantageous if any new linguistic theory could be utilised in computer applications. Bearing this in mind, the Chapter examines the theory of Government Phonology (GP) and its motivation. The system of subsegmental primes and their combinatorial constraints is explained. Then the incorporation of sounds into syllable-type structure follows, which leads into the central tenets of the theory – that of headedness, governing relations and licensing. Finally, there will be a summary of the main points and an exposition of why this theory is so suited to the task of Automatic Speech Recognition.

2.1.1 The Motivation

Phonological elements are motivated by the existence of natural classes (Crane, Aaron, 1997; 3). Several researchers (Kaye, Lowenstamm and Vergnauld, 1985; Harris 1990) have proposed a 'small set of subsegmental primes which may occur in isolation but may also be compounded to model the many phonologically significant sounds of the world's languages' (Ingleby and Brockhaus, 1998; 579). The claim is that phonological *primes* (elements) which constitute segmental representations are directly interpretable from the input signal. The primes are also assumed to have abstract articulatory associations. For example, Advanced Tongue Root (ATR), which accounts for the articulatory system of Tense vs. Lax is a "fundamental dichotomising property of the English Vowel System" (Harris, 1994: 114). A prime that can capture this division will be included in the set that makes up all the possible speech sounds.

2.2 The Primes

The set of primes that has been developed derives from the examination of acoustic and articulatory data. 'Elements are cognitive entities which are associated with certain physical manifestations' (Brockhaus and Ingleby, 1997; 4). They are acoustically realisable and have several characteristics. For example, the set of tense and lax vowels in English can be divided according to the [@] prime, which should reflect an acoustic model and relate to a detectable pattern. Those primes that will be used in the current study now follow. The set can be subdivided into three sections; resonance, manner and source.

2.2.1 Resonance

The resonance primes are essentially those primes that capture vowel sounds (see 2.3.1). Vowels can be differentiated because the configuration of the oral tract affects their resonance frequencies. Therefore knowing the formant structure of each vowel describes the articulators' position and helps to identify the vowel (Olive *et al*, 1993; 104). In auditory terms, there are spectral features that can be said to apply to each GP prime. The prime [A] consists of a spectral peak, while the prime [I] shows both a low first formant and a spectral peak. A spectrogram of the prime [U] should contain a spectral peak at the bottom of sonorant frequency zone, whereas prime [@] is neutral. In phonemic terms, the neutral prime [@] is most often thought of as being realised in the mid, central unstressed vowel schwa / ∂ /. Examples of vowels associated with each prime are /a/, represented by [A], /i/ is equivalent to [I] and /u/ is captured by [U].

2.2.2 Manner

The manner of articulation corresponds to the *way* in which a sound is produced. There are three major ways to produce a sound. First, as a *stop* consonant, such as /p/, /t/, /k/, /b/, /d/, /g/, individuated by voicing quality and place of articulation. Secondly, *fricatives* are produced when the vocal tract constricts enough to cause turbulence but not enough to completely stop the airflow. The fricatives of English (in IPA transcription) are / v /, / ð /, /z /, / ʒ /, / f /, / θ /, / s /,

/ ʃ / and / h / (Appendix B). The four voiceless fricatives can be differentiated by their characteristic distribution of energy in their spectra. The third manner class is that of the *nasal* consonants. The three nasal consonants in English (bilabial /m/, alveolar /n/ and velar //) are produced with a complete closure of the oral cavity. Spectrograms show that nasals have a prominent low frequency F1, referred to as the nasal formant. One of the indicators of a nasal sound is a clear and marked discontinuity between the formants of the nasal and those of adjacent sounds (Olive *et* al, 1993; 97). Figure 2.1 shows a waveform and spectrographic example of all three types of manner acoustics. In the phrase 'Open Sesame', there is a sequence of a bilabial stop consonant (/p/) an alveolar nasal (/n/), two alveolar fricatives (/s/) and a bilabial nasal (/m/).

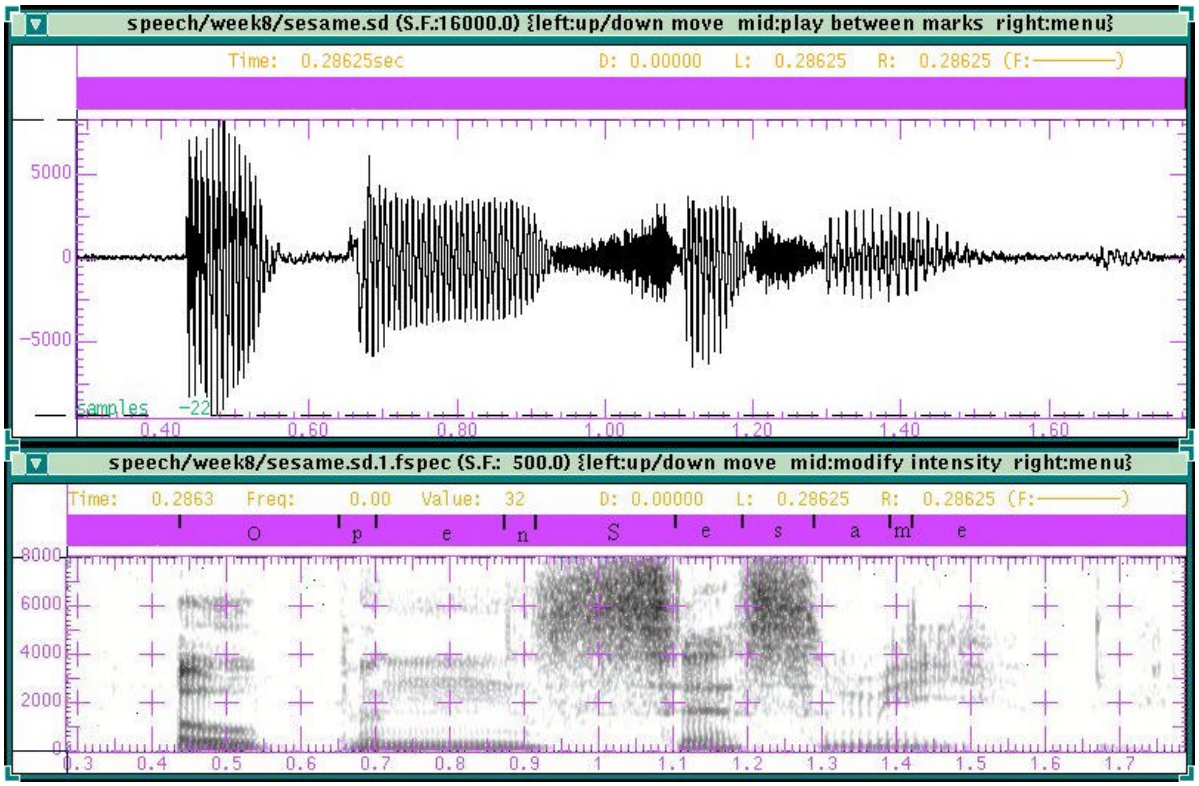


Figure 2.1 A waveform and spectrogram of the phrase 'Open Sesame'

A manner *prime* constitutes a feature that can capture these classes of articulatory or acoustic gestures. The acoustic features of the prime [?] essentially corresponds to a stop or any segment showing abrupt and sustained decrease in overall amplitude, while the prime [h] is present in a sound that contains noise or frication, or which shows up in a spectrogram as aperiodic energy.

[N] is the prime for nasals, articulatorily involving a closure of the oral tract and acoustically showing nasal zeros in the spectrum.

2.2.3 Source

There is only one source prime used in the version of the theory adopted in the present investigation, that of voicelessness. In vowels which are always voiced (with slack vibrating vocal cords), this prime is considered to always be absent. A voiced speech signal is quasi-periodic. It is only when a sound is produced with stiff vocal folds (i.e. voicelessly), that the [H] prime need be invoked, detectable by its nonperiodicity. These 8 primes are considered sufficient to describe all phonetic sound segments, with further expansion, which will be described below Table 2.1 shows the system of primes and the articulatory labels that they are considered to encompass.

Element Type	Element Name	Articulatory Label
Resonance	[I]	palatal
	[U]	labial
	[A]	low
	[@]	neutral
Manner	[?]	occlusion
	[H]	noise
	[N]	nasality
Source	[H]	voicelessness

Table 2.1 The Primes and their Corresponding Articulatory Labels

(From Brockhaus and Ingleby, 1997; 5)

2.3 Compound Expressions

All primes are called ‘phonological expressions’ (Kaye, 1997; 6). They enjoy stand-alone phonetic interpretability. However, any given vowel can be composed either of a single element or a fusion of two or more elements. The position of the tongue-body and lips is calculated by adding together components (Rennison, 1986; 282). Thus an /i/ and an /a/ component can combine to give a mid front unrounded vowel (e.g. fusing the primes [A] and [I] will produce /e/, while the primes [A] and [U] gives /o/). “Any given vowel is composed either of a single expression (a simplex expression) or a fusion of two or more elements (a compound)” (Harris, 1994; 97). These elements make up all vocalic and consonantal phonological segments.

2.3.1 Vowels

The vowel elements are composed chiefly of the 3 corner vowels (/a/, /i/, /u/) in the model of the oral cavity's articulatory space, which correspond to the primes [A], [I] and [U] (see Figure 2.2).

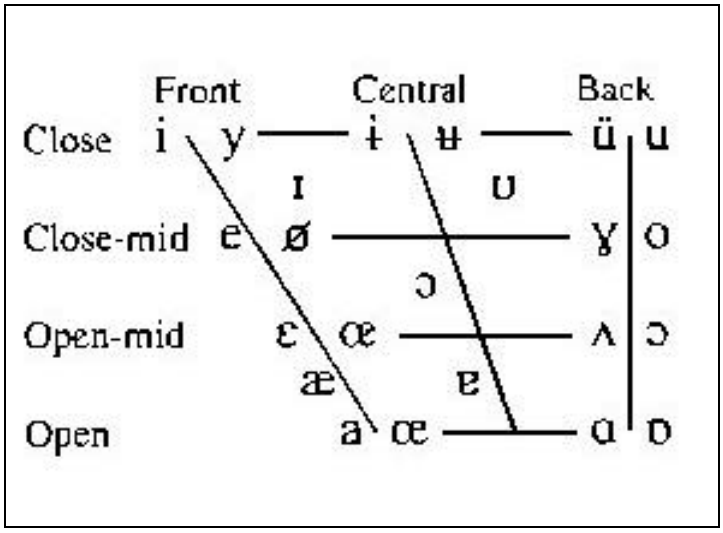


Figure 2.2 The vocalic system chart (/i/, /u/ and /a/ at the corners)

Any given vowel is composed either of a single element or a fusion of two or more elements (simplex v. compound). (E.g. [A, I] = /e/ [A,U] = /o/ [U,I] =/ü/). The ordering of primes is irrelevant. [U, I] is the same as [I, U]. The combinatorial possibilities reflect a conception of the articulatory aspects of each sound segment. For example, adding [A] gives a more open version of a certain vowel, while fusing a prime with [U] will produce as rounder vowel. Nasality and tenseness are included by incorporating the [N] and the [@] primes respectively. This neutral element [@] is considered to be the feature responsible for vowel reduction. The three main vowels, /a/, /i/ and /u/ represent 'extreme departures from a neutral position of the vocal tract' (Harris, 1994; 108). Combining the schwa-like [@] prime accounts for the 'neutralisation of peripheral vowel qualities under a centralised reflex' (Harris, 1994; 109).

2.3.2 Consonants

Consonants are represented with the same elements as vowels, plus some additional features (tenseness, nasality, frication, voicelessness, and occlusion). Consonants contain the manner and source primes ([?], [H], [N]), which vowels do not. Glides are the elements [I] and [U] occurring in non-nuclear positions (see Section 2.4).

2.3.3 Composition and Decomposition

The primary mechanism of change in GP is that of composition and decomposition. This involves the fusing or splitting of primes within a sound, yielding the inventory of both vowels and consonants. See Figure 2.3 for examples of a fused vowel and a vowel split into separate primes. This process occurs between segments too and is considered to be the sole explanation for every phonological phenomenon. No other process may be invoked.

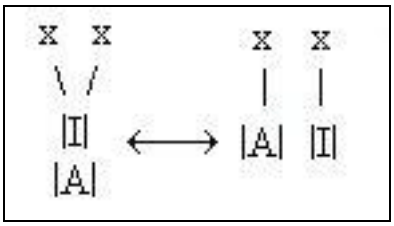


Figure 2.3 A fused vowel /e/ and a decomposed diphthong /ai/

(From Harris, 1994; Chapter 2)

This system of composition can be seen to produce some of the vowel inventory of English (Figure 2.4). A further factor, that of headedness will account for the rest (see Section 2.5).

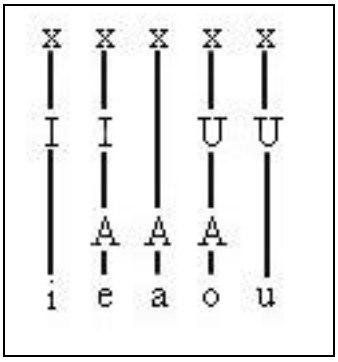


Figure 2.4 The composition of primes to form vowels

(Harris, 1994; Chapter 2)

2.4 Structure

> *Knowledge of constituent structure is a useful area of phonology. It introduces a level between the segment and the word into the speech recognition process*
>
> (Williams, 1997; 80)

To be audible, an element must be contained within a syllable structure of some sort. This section outlines the approach that GP has adopted. In the theory, every segment must be licensed and attached to a skeletal position, dominated by the Onset, Nucleus or Rhyme. This creates a system of nuclear vs. non-nuclear positions. This section will explain the terminology and the possible structural relationships.

2.4.1 Onset, Nucleus and Rhyme

There are three prosodic constituents, the Onset (O), Nucleus (N) and Rhyme (R). The nucleus is the head of the rhyme (left branch) (Brockhaus, 1995; 193). The Rhyme is therefore the first projection of the nucleus (Figure 2.5).

Figure 2.5 Onset, Nucleus and Rhyme

GP is a non-linear system that examines the relationships between Onset-Rhyme constituents in words. It takes the nucleus as the syllabic constituent. Nuclei are maximally binary. In fact, all the nodes can project singly or to a maximum of two nodes (see Figure 2.6).

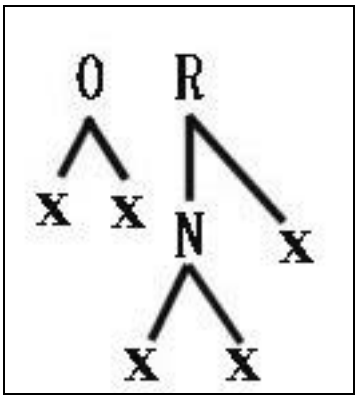


Figure 2.6 Maximally projected nodes

The prosodic hierarchy that results from this system holds the *word* at the top, containing the *foot*, which holds the *Onset*, *Rhyme* and *Nucleus* (Harris, 1994; 153). The theory is not one of syllables. Rather, it takes sequences of onset-rhymes stitched together in governing relations (Charette, 1990, 6), (see Section 2.6.1).

Under this framework a nuclear vs. non-nuclear system has been created. This can actually explain the phenomenon of palatal assimilation (see Figure 2.7) between adjacent consonants and glides. E.g.,

bet you did you kiss you faze you

t → tʃ d → dʒ s → ʃ z → ʒ

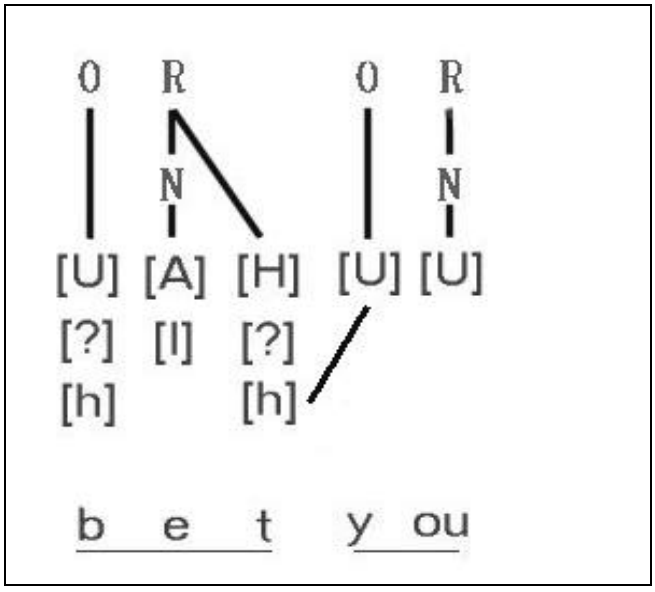


Figure 2.7 Palatal assimilation. Composition with non-nuclear prime [U]

The phonological expression corresponding to the /t/ of ‘bet’ is made up of the phonological elements [H, ?, h]. The glide [j] in the word ‘you’ is actually the vowel element [U] in non-nuclear position. The composition of both adjacent forms results in a new phonological expression made up of [H, ?, h, U]. This corresponds to the alveopalatal voiceless affricate / tʃ/. This GP account explains the forms that will be produced in rapid assimilated speech.

2.5 Headedness

As we have seen, GP states that 'all human phonological segments consist of one or more of a set of 6 to 8 elements’ (Kaye, 1996; 1). These elements may be pronounced in isolation or in combination and are called phonological expressions. In addition, they can be headed or headless. This means that "if elements are combined with each other, to form compound expressions, one element is usually predominant, the so-called head of the expression" (Brockhaus and Ingleby, 1997; 4). There is a maximum of one head. A phonological expression may contain zero or more non-heads, called operators (Kaye, 1996; 1).

Headedness means that one designates an element in a compound as the head of that expression and any other elements that may be present as dependents. This serves to highlight the more significant role played by one prime over the others. “In a simplex expression, the lone element can be the head” (Harris, 1994; 105). In the following examples, the headless vs. headed distinction is drawn. Headedness is indicated by an underlined prime:

[A]	(pat)	[<u>A</u>]	(part)
[I]	(pit)	[<u>I</u>]	(pete)
[U]	(put)	[<u>U</u>]	(boot)
[A I]	(pet)	[A <u>I</u>]	(boat)
[A U]	(pot)	[<u>A</u> U]	(bought)

The prime [@] is considered in some versions of the theory to be present underlyingly and when it surfaces, it will always be because it is the head of a phonological expression. The example (pat) above can either be seen as a headless simplex prime [A], or a compound expression

containing [A], with [@] as its head (i.e. [A @]. The latter approach will be adopted in the current paper.

2.6 Constraints

There is free association of the primes to compose and decompose in all the world's languages, but the difference between language inventories is seen to be the result of language specific restrictions that are adopted, and vary between languages. The English Nuclear system comprises several constraints. In English, the licensing constraints that limit the combination of elements are as follows;

1. Branching Nuclei are headed
2. Non-branching nuclei are headless
3. U and I cannot combine
4. Nothing can license I ([I] is not an operator) (Kaye, 1997; 6)

In comparison, the constraints adopted in Continental French create a different inventory.

1. All expressions are headed
2. Nothing can license U ([U] is not an operator)

This yields the following contrastive examples in the vowel sets of English and French (Table 2.2).

English	**Examples**	**French**	**Examples**
[A]	part	[A]	pas
[I]	pete	[I]	cri
[U]	boot	[U]	fou
[A I]	bait	[A U]	faux
[A U]	boat	[A I]	fé
[U A]	bought	[I A]	fait
		[I U]	fût
		[A I U]	feu

Table 2.2 Different Licensing Constraints for English and French produce separate vowel inventories

Given the structure of this theory in terms of phonological expressions with a headedness component being attached to positions such as Onset, Nucleus and Rhyme, what regulations

constrain the placement and behaviour of the expressions within this structure? The answer relates to the theory of phonological *government* for which this approach was named.

2.6.1 Phonological Government

"Government is an asymmetric binary relationship obtaining between two skeletal positions, and which is strictly directional. In addition, certain substantive conditions must be met"(Crane, 1997; 9). It is this feature of the theory that makes it substantively different to any other approach in phonology. 'A constituent is defined as an ordered pair, governor first on the left and governee second on the right' (Ingleby and Brockhaus, 1998; 579). Government is based on c-command. This is defined by the following rule: "Node **a** *c*(onstituent)-commands node **b**, *iff* the branching node most immediately dominating **a** also dominates **b**" (see Figure 2.8).

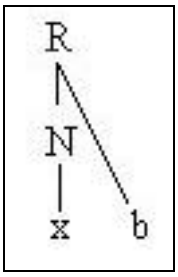


Figure 2.8 Node *a* (N) c-commanding node *b*

The theory recognises different forms of government. One form is between skeletal positions within a constituent (Figure 2.8). Another regulates inter-constituent skeletal positions between two contiguous constituents (Brockhaus, 1995; 192). For both forms, government is always directional. (Intra-) constituent government requires an initial, or left, headedness, while inter-constituent government takes head-final contexts as its domain. Constituent government is also strictly local - the governor must be adjacent to the governee. This implies a maximally binary branching constituent constituting an internal governing domain where the head governs a complement from left-to-right (Charette, 1990; 11).

Governing relations must have the following properties:

(i) Constituent government: the head is initial and government is strictly local.
(ii) Inter-constituent government: the head is final and government is strictly local.

It is apparent that there are several criteria in the GP theory that postulate the form of government that must be met. One of the main aspects of government which must be adhered to is licensing.

2.7 Licensing

This Section examines the concept of licensing, which according to Brockhaus, 1995; 203, is 'the motor which drives phonology'. Licensing constraints are language-specific constraints on elements' combinatorial powers (Crane, 1997; 2). Government is one form of licensing. However, there are several more constraints adopted in the theory, which regulate the permissible patterns of sound.

> *GP also has very stringent requirements for the licensing of all aspects of phonological representations. Unlicensed material remains inaudible*
>
> (Brockhaus and Ingleby, 1997; 4)

2.7.1 The Licensing Principle

Phonological Licensing is a principle of GP which states that 'all phonological positions except the head of a domain must be licensed within that domain' (Brockhaus and Ingleby, 1997; 5). This occurs both within and between constituents. This is a principle also requiring that licensing relations be local and directional (Harris, 1994; 156).

2.7.2 Constituent Licensing

Within constituents, licensing relations are head-initial. Projection Licensing dictates that between the projections of nuclear heads, licensing relations must be parametrically head-initial or head-final (Harris, 1994; 157). Languages can select either variety, but the licensing must apply. There are several forms of licensing, the most important of which are Onset and Coda Licensing. Briefly, Onset Licensing regulates that the head of an onset must receive its power to govern a complement (i.e. its license to govern) from the nucleus governing it. While for Coda Licensing, a rhymal adjunct position must be licensed by an onset position (Harris, 1994; 160).

2.7.3 Proper Government

'Proper Government is a form of internuclear government from right to left where no governing domain intervenes between the governor and the governee' (Scheer, 1998; 43). Proper Government plays a vital role in the phonological phenomenon of empty nuclei. When empty nuclei are properly governed, they remain inaudible (so, long vowels will never be inaudible). 'If they escape proper government, they are subject to a language-specific epenthesis' (Figure 2.9B) (Scheer, 1998; 43). A properly governed position cannot govern any other position.

Phonological effects are now seen to have a cause related to government. (As seen, an *empty nucleus* is not manifested when it is *properly governed* by a following nucleus which itself is *ungoverned*). For example, an empty nucleus is seen in the French example of [mɛn] v. [mne] (Charette, 1990; 3) (*mène* vs. *mener*) in Figure 2.9.

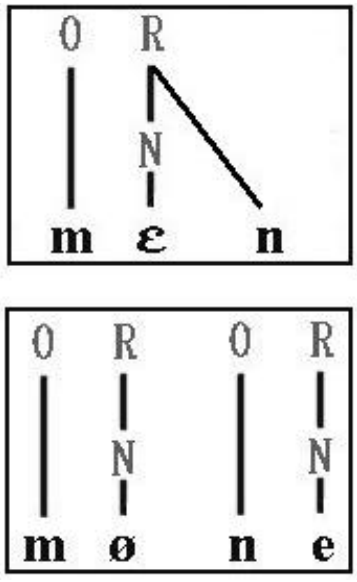


Figure 2.9 A) Phonetically realised B) Properly governed unrealised nucleus

In Figure 2.9B), the licensed empty nucleus has no phonetic realisation. This situation is now explained in terms of the proper government of underlyingly empty nuclei. "When an empty nucleus is properly governed by a following nucleus which itself is ungoverned, it is realised as null" (Charette, 1990; 2). If proper government fails to apply, the empty nucleus must be

phonetically expressed. Apart from parametrically licensed domain final positions, only properly governed positions may remain empty (Brockhaus, 1995; 198).

So, the empty nucleus is properly governed and not manifested when it is followed by a final vowel. When there is no proper governor to govern the empty nucleus, it is phonetically realised. This process is presumed to occur in all languages. According to Charette, (1990), vowel shortening in closed syllables is too widespread to be accidental or language specific.

2.8 Summary

"Government Phonology is a principles and parameters approach where principles are inviolable and language-specific facts are expressed by parameter" (Polgárdi, 1998; 1). Languages select parameters (Charette, 1990; 3). There is a small set of subsegmental primes. They may be pronounced in isolation or in combination with each other. GP has only two possible phonological operations: *composition* (the building up of structures from elements) and *decomposition* (the loss of elements). The phonological constituent structure is the Onset, Nucleus and Rhyme. The primes are attached to these positions on a skeletal tier, and their placement in either Onset or Nucleus position will affect their realisation and possibility of being subject to phonological change. For example, glides are viewed as vowels in non-nuclear positions. Nuclei are central to syllabic constituents. They are maximally binary (1 or 2 nodes only). This accounts for the differences in short vowels, diphthongs and long vowels (see Figure 2.10).

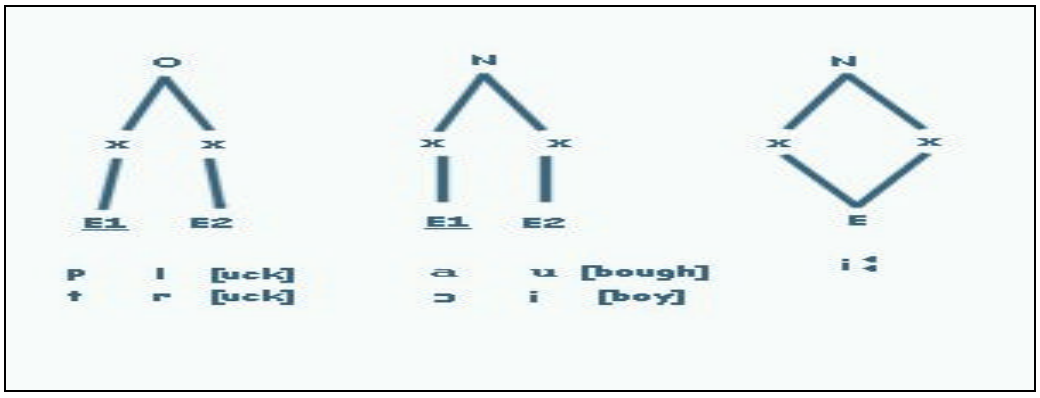


Figure 2.10 Branching Onset, Diphthongs and Long Vowel

Additionally, 'the skeletal positions to which elements may be attached (alone or in combination) enter into asymmetric binary relations with each other; so called Governing Relations' (Ingleby and Brockhaus, 1998; 579). The theory recognises governing relations at three levels: between syllabic constituents (constituent government), between contiguous syllables (inter-constituent government) and between syllabic nuclei which are heads of metrical structures (nuclear projection government) (Carr, 1993; 290). Government is also subject to the following principles:

(i) Only the head of a constituent may govern

(ii) Only the nuclear head may govern a constituent head. (Charette, 1990; 27)

2.9 The advantages of a GP approach to ASR

In the current study, a set of GP phonological expressions will be used, rather than a set of phonemes or binary features. In accomplishing the task of automatic speech recognition, there are several advantages to a theory that involves picking out elements, especially from an engineering standpoint:

1. Elements are subject to less variation due to contextual effects of the preceding and following segments than phonemes.
2. Elements are much smaller in number than phonemes. (8 primes v. 60 phonemes).
3. Elements, unlike phonemes, have been shown to participate in the kind of phonological processes that lead to variation in pronunciation.
4. Although there is much variation of the phoneme inventory from language to language, the element inventory is universal' (Ingleby and Brockhaus, 1998; 579).

> *Cross-linguistic variation is captured by differences in the allowed combination of these units, rather than completely arbitrary sets of segment inventories (admittedly with some degree of overlap)* (Williams, 1998; 96).

Chapter Three Researching the Question

3.1 Introduction

The intention in the current investigation is to achieve ASR using a GP framework. This requires a way to represent phonological constituents computationally. Previous studies have been carried out to determine the best methods of representation and application. In the first half of this Chapter, the background to the present study will be laid out, including the related literature and various methodologies. This will lead to an explanation of Artificial Neural Networks, the method that will be adopted here to achieve the aims of the project. The second half of this Chapter will contain the methods utilised in order to test the hypothesis that GP primes are a good representation of linguistic knowledge in ASR environments. Hopefully, this will "demonstrate the feasibility of using unary primes in speech-driven language processing" (Ingleby and Brockhaus, 1998; 578).

3.2 Background

Although the application of the GP theory to technology is relatively new, there have been several attempts to incorporate a system of phonological elements into speech and language processing. The primary research was carried out in Geoff Williams' 1998 Ph.D. Thesis. He derived acoustic parameters for each GP prime, essentially by extracting spectral properties such as formant positions and amplitudes. He used neural network classifiers on the TIMIT database of spoken English sentences to recognise targets as the basis of research into a multi-lingual recogniser (Williams, 1998; 173). He also investigated developing a GP-based phonological parser. This combined a broad class analysis of the acoustic data with a symbolic parser, which implemented the phonological constraints of a language in question. He tested the parser using simulated data with encouraging results. With a 3 class network (silence, nuclear, non-nuclear), with percentage

figures based on tokens rather than frames, he obtained 90% accuracy on TIMIT (Williams, 1998; 173).

In another recent study, Brockhaus and Ingleby (1998) examined an extended approach to locating GP primes. They also felt that 'it is possible to define the acoustic signatures of individual elements, so that the presence of an element can be detected by analysis of the speech signal' (Ingleby and Brockhaus, 1998; 579). Acoustic cues such as 'energy ratio, width, fall, change, duration, F1 trajectory and formant transitions' were used by the authors (Ingleby and Brockhaus, 1998; 580).

They adopted a stage-wise approach, where at each stage, segmentation and lexical access are launched. At stage 1 they check for the presence of the manner elements [h] and [?], by examining acoustic data. In stage 2 they explore the phonatory properties of those segments identified at stage 1. This means that the elements [H] and [N] are searched for. If successful identification has not occurred at this point, they proceed to stage 3, and check for their set of resonance elements ([U], [I], [A], [R]). This method has also shown good experimental results.

However, Ingleby and Brockhaus' method relies heavily on a handcrafted rule-based approach. While this approach is not without merits, it would instead be preferable to enable the recognition process to learn these methods itself, through a training process.

Besides the GP-driven studies into prime recognition, there is a convention of using the existing phonetic labels for the task of pure phoneme recognition (e.g. Robinson 1991, Ström, 1997b,). In Ström 1997b, artificial neural networks were utilised on the TIMIT database. After a training process, the output units approximated *a posteriori* probabilities for phonemes given input feature vectors. By use of Bayes' rule, the *a posteriori* probabilities are converted to phoneme likelihoods

to be used in a HMM framework. The lowest achieved phone error-rate, 26.7%, is very close to the lowest published result for the core test set of the TIMIT database. Robinson (1991) found ~70% typical phoneme recognition rate. Additionally, Hübener and Carson-Berdsen (1994) have achieved 90% frame-based accuracy on data from a single speaker.

Fallside *et al*, (1990) took four types of neural networks and applied them to the TIMIT database. The networks included a recurrent network phoneme recognizer and a compositional representation phoneme-to-word recognizer. The major result was for the recurrent network, giving a phoneme recognition accuracy of 57% from the *si* and *sx* sentences, a subset of the TIMIT database.

Koreman *et al*, (1999) mapped acoustic parameters onto phonetic features as part of an attempt to explicitly address the linguistic information in the signal (Koreman *et al*, 1999; 719). They used Eurom0, a multilingual database. The acoustic parameters are used in conjunction with IPA and SPE phonetic features for neural network training. The baseline experiment using acoustic parameters alone yielded an identification rate of 15.6% correct for phonemes. The acoustic phonetic mapping raised the rate to 42.3% for IPA and 31.7% with SPE features (Koreman *et al*, 1999, 720).

> *The likelihood of confusing (elements) is relatively small, making a system based on GP ... more robust than any searching for phonemes.*
>
> (Brockhaus and Ingleby, 1997; 4)

In the current project, the search for the presence and absence of phonological primes will also be achieved by use of Artificial Neural Networks, commonly known as “ANNs”, “neural networks” or “neural nets”. They are a useful tool for testing the various hypotheses that are being investigated. Section 3.3 examines the operation of this type of model.

3.3 Artificial Neural Networks

An ANN is a network of many very simple processors (“units” or “neurons”), each possibly having a (small amount of) local memory. The units are connected by unidirectional communication channels (connections), which carry numeric (as opposed to symbolic) data. The units operate only on their local data and on the inputs they receive via the connections. Most neural networks have some sort of “training” rule whereby the weights of connections are adjusted on the basis of presented patterns. A feature of ANNs is their high degree of interconnection, which allows a large degree of parallelism. Each neuron is pre-programmed and continuously active. This is a system modelled upon the brain's interconnected system of neurons.

3.3.1 Uses

ANNs are well suited for pattern recognition and parallel processing implementations of conventional processing tasks. In practice, ANNs are especially useful for classification and mapping problems which have lots of training data available, but to which hard and fast rules cannot easily be applied. Some real world applications of ANNs include foreign language translation, process control, medical data interpretation, quality control, financial forecasting, health care cost reduction, targeted marketing, bankruptcy prediction, machine diagnostics and securities trading. In the present project, they will be used for speech and pattern recognition.

3.3.2 ANN Architecture

ANNs are made up of an input layer, a certain number of hidden units, and an output layer. The units and their degree of connectivity determine the actual output. Values are presented to the input, the hidden units perform the computations and the output is produced (Figure 3.1). The choice of input representation as well as the structure of the ANN (e.g., the number of hidden units and the connectivity between layers) represents *a priori* knowledge in the ANN, because it puts constraints on the relations that the ANN can learn.

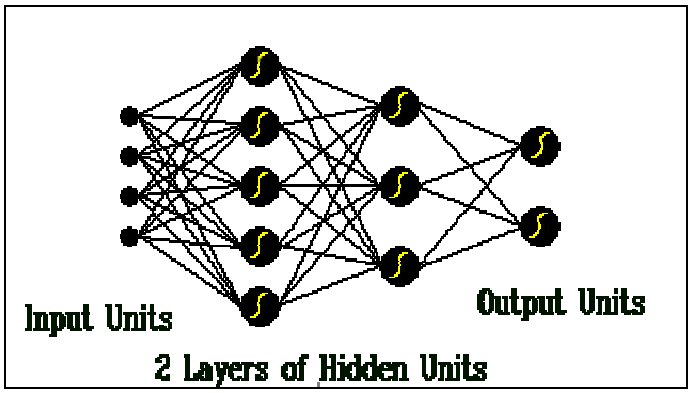


Figure 3.1 Sample Architecture of a Neural Net, showing hidden units and connections

The behaviour of a single processing *unit* in an ANN can be characterized as follows: first, the unit computes the total signal being sent to it by other processors in the network. Second, the unit applies an activation function to this total signal, in order to adopt a particular level of internal activity. Third, the unit sends a signal to other processors in the network; this signal is a function of the unit's internal activity. The signal that one processor sends to another is transmitted through a weighted connection, which is typically described as being analogous to a synapse. A neural net is basically a dense interconnection of simple, non-linear, computation elements of the type shown in Figure 3.2. "It is assumed that there are *N* inputs, labelled x_1, x_2, …, x_N, which are summed with weights w_1, w_2, …, w_N, thresholded, and then nonlinearly compressed to give an output *y*, defined as..

$$y = f\left(\sum_{i=1}^{N} w_i x_i - \phi\right)$$

where ø is an internal threshold or offset, and *f* is a nonlinearity" (Rabiner and Juang, 1993; 56). The output of each processing unit is some non-linear function of the weighted sum of the outputs from the previous layer.

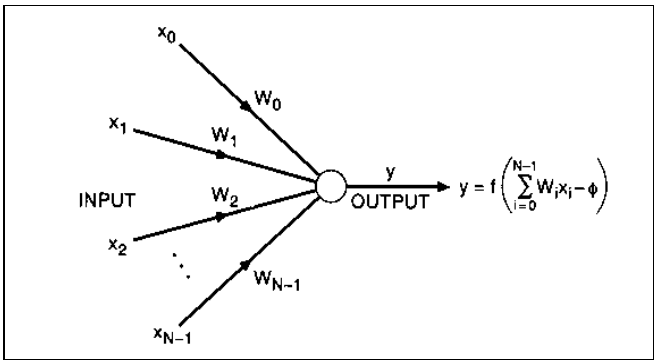


Figure 3.2 Simple Computation Element of a Neural Network

From Rabiner and Juang, 1993; 57

An ANN can be viewed as a system that generates a desired response to an input stimulus. The pattern of connectivity in an ANN (i.e., the strengths of the connections between various processing units) defines the causal relations between the network's processors, and is therefore analogous to a program in a conventional computer. However, in contrast to a conventional computer, the ANN is not given a step by step procedure to perform some desired task. Instead, the network teaches itself to do the task, as a function of its training.

ANNs imitate the brain's ability to recognise and classify patterns and learn from trial and error, discerning and extracting the relationships that underlie the data with which it is presented. Most ANNs have some sort of “training” rule whereby the weights of connections are adjusted on the basis of data (Sarle 1997). In other words, ANNs *learn* from examples and exhibit some capability for generalization beyond the training data. The networks are trained by an ‘error minimisation’ technique. Almost any vector function on a compact set can be approximated to arbitrary precision if there is enough data and enough computing resources. So, the training set is simply a set of examples used for learning, in order to fit the parameters (weights) of the classifier.

The adjustment of weights within the units is *hidden*, occurring between the nodes themselves, meaning that the training process is not readily accessible. The number of units and degree of connectivity can vary, which makes it difficult to pinpoint the exact repercussions of altering one of these parameters. Therefore, several networks must be trained, each using different parameters, and tested independently. Since the goal is to find the network having the best performance on new data, the simplest approach to the comparison of different networks is to evaluate the error function using data which is independent of that used for training (Bishop, 1995; 372). A validation set is a set of examples used to tune the parameters of a classifier, for example to choose the number of hidden units in a neural network. The performance of the selected network should be confirmed by measuring its performance on a third independent set of data called a test set. The test set, therefore is a set of examples used only to assess the performance (generalization) of a fully specified classifier (Ripley, 1996; 354). A test set, by definition, is never used to choose among two or more networks, so that the error on the test set provides an unbiased estimate of the generalization error (assuming that the test set is representative of the population, etc.). Any data set that is used to choose the best of two or more networks is, by definition, a validation set, and the error of the chosen network on the validation set is optimistically biased.

The number of hidden units an ANN possesses determines the size of that network. This can number in the hundreds, potentially the thousands. Once performance above a given number of hidden units begins to plateau, then it makes it possible to locate a best fit size for dealing with the training data. The larger the number of units, the longer the training time. A manageable number would be beneficial, given the number of alternative parameters that can be varied.

3.3.3 ANN Summary

> A neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes. *(*DARPA Neural Network Study, *1988; 60, cited in Sarle, 1997)*

Two key concepts in artificial intelligence are automatic knowledge acquisition (learning) and adaptation. One way in which these concepts have been implemented is via the neural network approach (Rabiner and Juang, 1993; 54).

In pattern-matching applications the network is trained by presenting a pattern vector at the input layer and by computing the outputs. The output is then compared with some desired output. The error between the actual output and the desired output is computed and back propagated through the network to each unit. The input weights of each unit are then adjusted to minimise this error. This process is repeated until the actual output matches the desired output to within some predefined error limit (Owens, 1993; 170).

The advantages to using ANNs include the fact that they can implement a massive degree of parallel computation. Also, they intrinsically possess a great deal of robustness or fault tolerance, due to the nature of ‘spread’ information within the network. This makes them less sensitive to noise or structural defects, because the weights on the connections are not fixed. They can be adapted in real time to improve performance, a fact that is inherent in the neural network structure (Rabiner and Juang, 1993; 62).

Chapter Four Experimental Design

4.1 Introduction

There were six experiments in total. This Chapter will explain the methods used to accomplish the experiments. Briefly, Experiment 1 examined the performance of the prime [A]. Experiment 2 studied the performance of the remaining primes ([I], [U], [@], [?], [h], [H] and [N]). In Experiment 3 the operation of all 8 primes concurrently was considered by detecting all the streams simultaneously using single networks.

Next, the GP concept of headedness was tested. This is the feature that removes ambiguity between combinations of the same elements to produce differing phonemes. It is necessary to distinguish just the vowels essentially, so in Experiment 4 the respective headedness of the primes [A], [I] and [U] were incorporated. Finally, Experiments 5 and 6 investigated the impact of gender-specific and dialect-specific data on the theoretical findings. The theory was tested to see whether the phonological expressions could be detected as easily if they were trained on male data and tested on female data, and vice versa. The hypothesis is that with the balance of male to female data at a ratio of 2:1 in the database that the male-trained networks will perform better, given their exposure to more data. The division of the TIMIT database into different dialect regions of the US (see Section 4.8) allows several of the dialect types to be used in the training stage, but with two of the northern dialect regions excluded, to test the performance of the network on this unseen dialect region.

4.2 Materials and Methods

To achieve the aim of this project and test the benefit of the GP theory for ASR, it required for each experiment that the phonological expressions relate to phonetic properties. Therefore a large amount of phonetically labelled speech was prepared, and values for the prime combinations associated with each label were adopted, replacing the former phoneme-based set. This newly prime-labelled data was then used for training ANNs.

This machine learning technique needed many examples in order to be able to generalise its findings. The best source for this data is a database of speech, such as the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, used in the present set of experiments.

4.2.1 The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT database) is a corpus of read text speech. TIMIT contains a total of 6300 sentences consisting of 10 sentences spoken by each of 630 male and female speakers from 8 major dialect regions of the United States. This corpus will be used in the present study for several reasons. It is clean speech, meaning that it was recorded in a studio environment with a high quality microphone and each sentence selected for inclusion shows good pronunciation. The corpus has also been carefully manually phonetically transcribed throughout to a high standard. The TIMIT corpus includes several files associated with each utterance. In addition to a speech waveform file there are associated transcription files showing an orthographic transcription of the words the person said, a time-aligned word transcription and a time-aligned phonetic transcription (see Appendix E).

4.2.2 Relabelling the database

To relate the features to phonetic properties, the phonetically labelled data in TIMIT was analysed for the values of the features associated with each label. The database was used to automatically generate GP labels from a set of phonemes. Each phoneme (see Appendix C) in a selection of sentences from the set of TIMIT files was manually relabelled with the corresponding primes from the GP theory, by examining several sample sentences and time labels (Appendix E and Appendix F). This led to the creation of a table containing the 60 TIMIT phonemes with their corresponding primes (see Appendix D). The entire database was then relabelled automatically using this prime table of phoneme-to-GP label correspondence.

4.2.3 Acoustic Input Data

The sound files from the database were converted to mel-cepstral coefficient files (MFCC). Mel-cepstral coefficients are parameters that are designed according to the way that the human ear perceives sound frequencies. They are standard speech feature vectors used as spectral parameters. These files capture the spectral shape of sounds rather than a fine domain waveform. Mel-based cepstral coefficients provide a representation in which a Discrete Fourier Transform (DFT) magnitude spectrum is frequency-warped (to follow the bark or critical band scale) and amplitude-warped (logarithmic scale) before the first 8 – 14 coefficients of an inverse DFT are calculated. They are reduced here to just 12 numbers, which provides a better representation for the model. According to Ström (1997a), ‘a small number of features make a good input representation because of the properties of the model and the statistical training methods’. The experience of pattern recognition shows that it is very hard to deal with large observation vectors, simply because there are too many possible combinations of fine-grained detail. Every 25 milliseconds, there are hundreds of samples of speech. The aim of this segmentation process is to partition the speech samples into sensible units.

The cepstral coefficients, which are the coefficients of the Fourier transform representation of the log magnitude spectrum, have been shown to be a more robust, reliable feature set for speech recognition.

(Rabiner and Juang, 1993; 115)

The corrected spectral parameters are used to compute phonological feature targets (see Section 4.4).

4.2.4 ANN Training

A number of different networks were trained since it is not known which network will produce the best result. A set of 3648 mel cepstral training files were used with a set of 3648 originally (phonemically) labelled training files and the table of GP primes (Appendix D) to create a set of 3648 GP-labelled training files. In order to train neural networks, a set of target training files were created from these label files. The official TIMIT training utterances were used for training, with a validation set removed, and the official core test set was used for the final evaluation.

The number of hidden units in the ANNs was varied, as too was the percentage of connectivity and the gain size adjustments. For the purposes of this experiment the intention was to uncover a minimum sized network of a size above which performance did not seem to improve significantly. For the current project, the number of hidden units that were attempted ranged from 10 to 80. Until an acceptable peak in performance was located, it was decided to adopt a practice of testing with 4 options, 25%, 50%, 75% and 100% connected. The results obtained indicated the optimum values that could henceforth be more confidently adopted.

The adjustment of weights along the connections between units can take varying step sizes. They can be large at first until performance improves and then with every iteration become gradually

smaller in step size, once the network has reached a level close to the optimum possible. The gain sizes used in this experiment were 10^{-04}, 10^{-05} and 10^{-06}.

With every iteration of the network, the weights are adjusted internally to reduce the level of error between the target files and the input cepstral coefficient files. The ideal is zero error, and while this will never be achieved, zero error is always the goal. While training a network, a logfile of each iteration records the value of error and these values can be plotted, indicating the rate of learning.

The number of iterations the networks makes can be increased if it is seen that the error minimisation process has not levelled off. If the error was still significantly dropping after 50 iterations, it was possible to allow the network to continue training in order to determine any degree of further improvement. The *back-propagation* training method typically took several days. The performance of the networks was then compared by evaluating the error function using an independent validation set to locate the network with the smallest error (from Bishop, 1995; 372).

4.2.5 ANN Testing

In order to obtain information about given parameters, the completed network was 'excited' by exposure to the test corpus of data contained in the TIMIT database, which had also been converted into a recognisable set of files. This pattern matching technique determined the degree of correspondence achieved by each ANN. A further operation extracted percentile results comparing the outputs of a network to the target test files. This indicates the overall accuracy of the recognition process by that network (see Section 4.2.6).

4.2.6 Statistical Methods

However, percent correct is not an evaluative measure for a prime, given that there is a need to take the difficulty of the problem into account. Speech recognition systems are used in many differing tasks. The probability of each task varies according to the probability of the most likely outcome occurring at random. This is known as the chance level and must be calculated for each task undertaken in an analysis. In the current research, the chance level for each prime was required in order to determine the success of the system in identifying that phonological element. This was achieved by examining the whole corpus of test data. Chance levels were calculated by dividing the number of target frames where a feature (for separate primes, sequences of given outputs, headedness, gender, dialect type) is considered to be off by the total number of frames. The target files that were compared to the output of these networks have a global average for the percentage of the time that a feature is *on* or *off*. The percentage of time it is off is always higher, and if the system were to score around this number, it would be deemed to be performing only at chance.

The results obtained from a network are calculated based on an average of the results for the number of times the prime was judged to be on, when it *was* on, and the number of times it was correctly considered off. E.g.

Results:

1	137099/164251 =	83.5%
0	213573/246669 =	86.6%
Frames correct	350672/410920 =	85.3%

In the evaluation of results from these experiments, there was a requirement for a test that compares actual values with expected chance values. A Chi-Square (χ^2) tests the hypothesis that row and column variables are independent, without indicating the strength or direction of the relationship. The Chi-Square Test is a nonparametric procedure that compares the actual values

with the expected values and computes a chi-square statistic. The Chi-Square test works out whether the scores are removed a sufficiently large amount from chance (Hatch and Lazaraton, 1991; 395). This is in the form of a probability of significance value (p-value), which reveals a significant difference from chance when it falls below 0.05.

4.2.7 Summary

The analysis of these results was used as a preliminary method for ascertaining the suitability of the chosen methodology and to limit the parameters required by further networks. This is a necessary objective given the computational and time constraints of the project. Training a neural network requires both large amounts of training data and very long training times.

Once the results from Experiment 1 were mapped to a contour showing the best performance with given values for certain parameters, it eliminated a number of network parameter options that would otherwise require training with the remaining primes. A description of each experiment performed in this study now follows.

> In the same way that the full spectrum of visible colours can be generated, and likewise detected, by reference to the three primary colours Red, Green and Blue, the recognition of the full set of linguistic expressions is achieved by determining the presence or absence of the individual components separately
>
> (Williams, 1998; 96)

4.3 Experiment 1 The Prime [A]

This preliminary experiment was devised to ascertain that a single GP prime [A] is capable of being automatically recognised from the speech signal. The first approach was to apply GP primes to acoustic models, using the NICO toolkit (Ström, 1996) for ANNs. The initial stages of recognition involve normalising the signal and extracting parameters and features from the data. Frame by frame information regarding when the prime [A] is on and when it is off, was presented. 39 networks were trained for this prime, with variable parameter values. The number of hidden units varied from 10 to 80. The Gain parameter ranged between 10^{-04}, 10^{-05} and 10^{-06}. The degree of connectivity between the nodes of the networks was set to either 25%, 50%, 75% or 100%. The number of iterations taken by all the ANNs was set to 50. The steps involved in the identification of GP primes as a whole are contained in Table 4.1.

4.4 Experiment 2 The Remaining Primes Separately

In order to extend the procedure to include the remaining primes, the procedure for the prime [A] was repeated. The steps in Table 4.1 were repeated for each of the following primes, [I], [U], [@], [h], [?], [N] and [H].

Step One	Convert MFCC and GP labels into GP-labelled files
Step Two	Take input GP label files and prime on/off files to create target files
Step Three	Use the NICO Toolkit to start a network, setting values for the number of hidden units, degree of connectivity, gain and number of iterations
Step Fours	Take trained networks and compare test set patterns to outputs
Step Five	Compare outputs to correct answers to output percentage correct results
Step Six	Plot the percentages, error minimisation iterations, chance levels etc.

Table 4.1 The Steps taken in the experiments

The search time for locating the best performing networks could be reduced in light of the findings from Experiment One. Certain poorly performing parameter values could be eliminated, and some values were considered to be more important than others. Again, the results were compared to their target values, and percentages extracted.

4.5 Experiment 3 All the Primes Concurrently

At this point, the primes were achieving a level of performance on networks trained on each prime alone. However, this makes overall generalisations about GP label adequacy difficult. Therefore, a further experiment took ANNs specified for when each of the primes *combined* were *on* and *off*, as its targets. Again, several alternating parameter values were used in training various networks. The expectation is that the performance on these networks will be lower than when all the primes were tested individually. This is because the 8 individual networks have free parameters and all the nodes are working on analysing a single prime. When using a single network with the same parameter specifications to extract all 8 primes simultaneously, there are cross connections between given levels of nodes, which influences the computational operation of the ANN. There are 8 outputs being produced. This makes it harder to get 8 things right simultaneously. Therefore the decision was made to attempt 300 Hidden Units, to test whether increasing the computational power of the network could improve the results. 300 Hidden Units is approximately the level adopted for many investigations into typical phoneme recognition experiments (Robinson 1991, Ström 1997), making it a likely region at which to set this value. The processing time of the ANNs was increased greatly however, which was a factor to be accounted for given the time constraints of the project.

4.5.1 Comparison to Concatenated Individual Primes

As noted, training a network with all the primes simultaneously requires that 8 input streams be matched to 8 output streams. This takes a lot longer and is a more difficult task to perform well on. The best way to compare the overall performance of this approach was to concatenate the results of networks trained on single primes– making them identical in format to a single concurrent prime network output - and extract results from these newly created files. In comparing *like* with *like*, the same parameter values were adopted in both contexts.

4.5.2 Mapping to Nearest Phonological Expressions

It was necessary to extend the results further, since the outputs of the networks were not at this point always producing frame hypotheses corresponding to valid segments from the table of primes (Appendix D). Therefore, it was decided to maximise the possibility of a network output at least postulating a valid phoneme, by mapping the outputs onto actual prime combinations. This was achieved using a Euclidean measure of distance. The sum of the squares of the differences between the output and the possible target phonemes was computed. The closest legal phonological expression was located and the frame was then mapped onto it.

Tables 4.2 and 4.3 show how the outputs of a network are mapped to the closest phonological expression by calculating the distance of each frame output from a valid phonological expression. In the examples shown, the frame in Table 4.2 is closer (0.5300) to the phonological expression [A, U] than the same frame in Table 4.3 is from [I, U], where the distance is greater, (2.2477).

Primes	[A]	[I]	[U]	[@]	[?]	[h]	[H]	[N]	
Output Frame	**0.7951**	0.4996	**0.7337**	0.4095	0.0066	0.0000	0.0001	0.0013	
Prime comb.	**1.0000**	0.0000	**1.0000**	0.0000	0.0000	0.0000	0.0000	0.0000	
Distance ^{2}	0.0419	0.2496	0.0709	0.1676	0.0000	0.0000	0.0000	0.0000	**0.5300**

Table 4.2 Sum of squared distances from a possible phonological expression

Primes	[A]	[I]	[U]	[@]	[?]	[h]	[H]	[N]	
Output Frame	**0.7951**	0.4996	**0.7337**	0.4095	0.0066	0.0000	0.0001	0.0013	
Prime comb.	0.0000	**1.0000**	**1.0000**	0.0000	0.0000	0.0000	0.0000	0.0000	
Distance ^{2}	1.7558	0.2504	0.0709	0.1676	0.0000	0.0000	0.0000	0.0000	**2.2447**

Table 4.3 Sum of squared distances from a different possible phonological expression

This in essence forces an output to become an official prime combination as defined in the prime table (Appendix D). The *results* procedure can be run on the mapped outputs to determine the effect. The expectation is that the performance of the simultaneously trained ANNs will improve, since it will eliminate illegal prime combination outputs. At this stage, the experiments will have achieved frame-by-frame phone recognition.

4.6 Experiment 4 Headedness

In previous studies, headedness measurements were achieved by plotting acoustic data to determine the degree of involvement for each prime (Brockhaus and Ingleby 1998, Williams 1998). Since in this experiment, the intention is to enable the recognition process to identify the primes by sheer exposure to sample data, the table of phoneme-to-prime correspondences was adjusted. Three additional columns were added, (see Appendix D) and three new head primes ([a], [i], [u]) signified when that prime was the head of an expression. Since the prime table had shown several identical labellings for certain phonemes (e.g. /iy/ as in 'bee' and /ih/ in 'bit' are both represented with the prime [I] alone). So, /iy/ is now relabelled [I] (Prime I, head i) to distinguish it from /ih/, which is now merely captured by [I] (Prime I, no head).

A set of networks was created and trained on the data showing when a headed prime was on and off, in a sequence of speech frames. And then, as before, these networks were tested, by exposing a trained network to previously unseen data to provide an assessment of the accuracy of the training.

4.7 Experiment 5 Gender

The opportunity to further extend the findings arose in the possibility of comparing the performance of ANNs when the gender of the training and the test set is altered. The TIMIT database is divided into an equal 2:1 ratio of male to female speakers in both the training and the test sets. This allowed networks to be trained on male voices alone and female voices alone. A gender-specific trained network was then tested with three data sets; the same gender, the opposite gender and both genders combined.

Two types of input were chosen; all of the primes together and the best performing single prime (prime [N]). This allows comparison with the original combined-gender results from experiments 2 and 3. The expectation is that networks that are trained with a specific gender will perform

better on test data of the same gender, given the increased exposure to that specific gender's vocal traits. A solely male-trained network will theoretically do better on male test input since in training it has been exposed only to male data.

Two networks for a male input signal and two networks for female speech signals were trained using different data sets. The first networks took the prime [N] as their target, since this prime is quite easily detected and should provide a measure more reflective of pure gender bias. The second type used the protocol of taking all the primes concurrently as input to the hidden unit layers.

A general-purpose ANN was also trained on both-gendered data. The expectation is that given the bias of twice as many male files to female files in the TIMIT database, the performance of this network on male test data will be higher than on female test data. Finally, the pattern-matched files produced from the male training data on male test data were combined with the files from the female-trained ANN after exposure to female test data only. Then results on these gender-specific files were derived as before. This could be compared to the results from ANNs trained and tested in a gender-insensitive fashion. The question that this subsection will hopefully answer: 'Is it worth having gender-specific acoustic models?'

4.8 Experiment 6 Dialect Type

In a similar experiment to gender, the dialect regions (*dr*) of the United States were separated to determine the impact of dialectal variation on the GP approach. The TIMIT database is subdivided within the database into 8 regions: New England, Northern, North Midland, South Midland, Southern, New York City, Western and Army Brat (moved around). Two dialect region types (*dr2 and dr3*), corresponding to Northern and North Midland dialect regions, making up approximately 30% of the total data along this parameter, were excluded from training. The remaining 6 dialect types were used to train both 'all the primes concurrently' in one network and

a ‘best performing prime singly’ (prime [U]), for comparison. These Northern *dr* files were also used in training two further networks with the streams ‘all the primes concurrently’, and the prime [U]. Finally, the networks for all the primes concurrently and for the prime [U] were tested on both set of dialect-specific data separately. The expectation is that testing will yield improved performance over and above normal performance, when both the training and test data are from the same dialect region.

Experiment One	The Prime [A]
Experiment Two	The Primes [I], [U], [@], [?], [h], [H], [N] Separately
Experiment Three	All the Primes Together
Experiment Four	The Headedness Factor
Experiment Five	Gender
Experiment Six	Dialect Region

Table 4.4 The Experiments Undertaken in the Current Study

The next Chapter will show the findings extracted from the six experiments using ANNs (see Table 4.4). This will include the progress of several ANN’s error minimisation through progressive iterations, as well as a performance assessment in the form of frame-by-frame phone recognition, achieved by testing trained networks on previously unseen data.

Chapter Five Experimental Results

5.1 Introduction

The following Chapter presents the findings of each of the experiments as outlined in Table 5.1. They are presented in the following sequence, with figures and tables for each subsection, reflecting the involvement of various parameters and several extended results of interest.

Experiment 1	A	Prime [A] Contour Map
	B	Prime [A] Complete Results
	C	Error Minimisation (10-40 Hidden Units)
	D	Error Minimisation (50-80 Hidden Units)
Experiment 2	A	The Other Primes Separately
	B	Gain Results
	C	Best Performance of each Prime
Experiment 3	A	All the Primes Concurrently
	B	Concatenated Single Prime Outputs
	C	Map to Nearest Phonological Expressions
	D	Error Minimisation
Experiment 4		Headedness
		Incorporate Headedness into Recognition
Experiment 5		Gender
Experiment 6		Dialect Region

Table 5.1 The Experimental Findings to be discussed in this Chapter

Experiment 1 (A) The Prime [A]

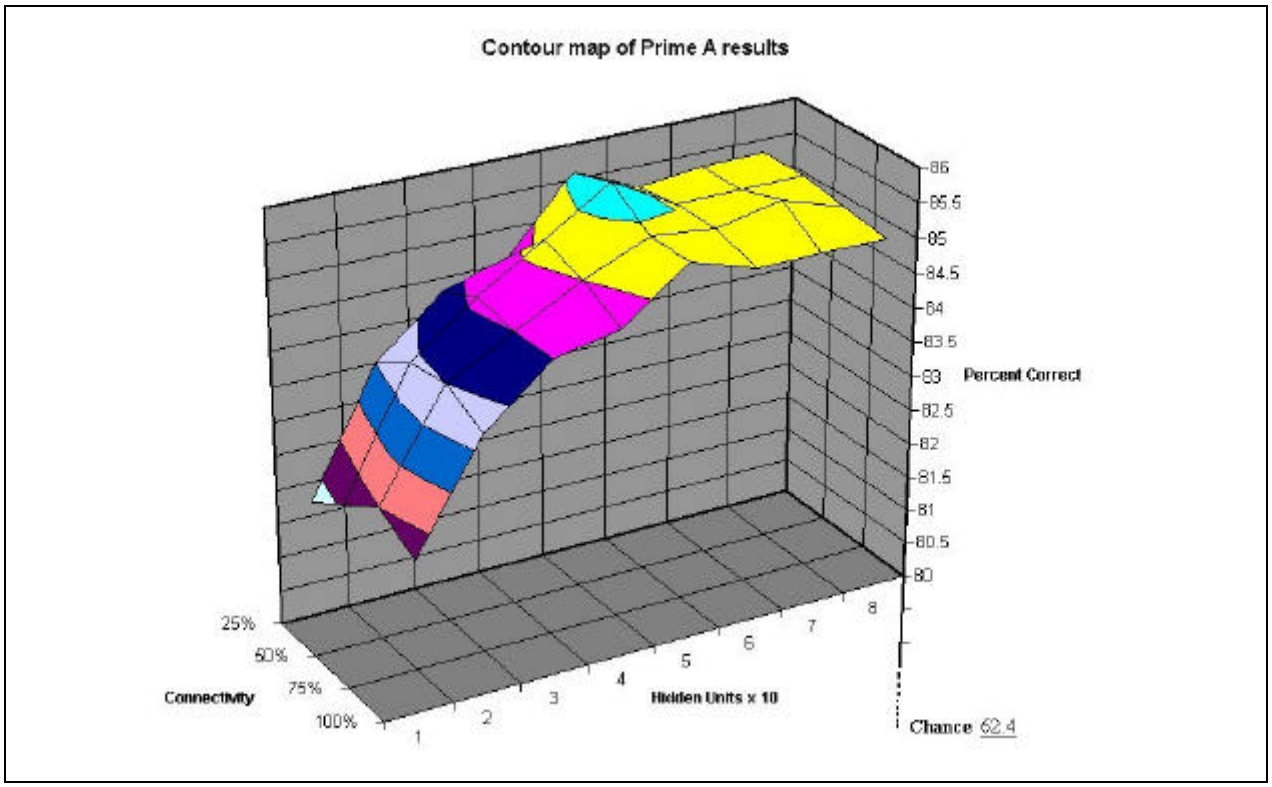


Figure 5.1 A 3D diagram of the results for the prime A at 10^{-05} Gain

Figure 5.1 and Table 5.2 show the performance of those networks that were set to a Gain of 10^{-05} along the axes of number of Hidden Units (10-80) and percentage of Connectivity (25%, 50%, 75% and 100%). The Chi-Square test shows that the observed performance of these networks is significantly above the expected chance performance level [χ^2= 248.24, $p < 0.01$]. There is a slight peak in performance on the network trained on the prime [A], with the parameters, 50 Hidden Units and 50% Connectivity (see Table 5.2).

	25% Connectivity	50% Connectivity	75% Connectivity	100% Connectivity
10 Hidden	81.7	82.1	82.5	82.2
20 Hidden	83.5	83.9	84	83.7
30 Hidden	84.3	84.6	84.5	84.5
40 Hidden	84.6	85.2	85	84.7
50 Hidden	85.6	**85.8**	85.4	85.4
60 Hidden	85.2	85.3	85.3	85.1
70 Hidden	85.3	85.3	85.5	85.1
80 Hidden	85.3	85.3	85.2	85.1

Table 5.2 Selected Results for the Prime [A] with 8 levels of Hidden Units and 4 levels of Connectivity. Gain is set to 10^{-05}

Experiment 1 (B) The Prime [A] (Complete)

Connectivity→	25%			50%			75%		100%	
Gain→	10^{-04}	10^{-05}	10^{-06}	10^{-04}	10^{-05}	10^{-06}	10^{-05}	10^{-06}	10^{-05}	10^{-06}
10 Hidden		81.7	80.1		82.1	81.1	82.5	81.6	81.7	81.6
20 Hidden		83.5			83.9		84.0		83.7	
30 Hidden	79.5	84.3			84.6		84.5		84.5	
40 Hidden		84.6		72.9	85.2	83.1	85.0		84.7	
50 Hidden		85.6			**85.8**		85.4		85.4	
60 Hidden		85.4			85.3		85.3		85.1	
70 Hidden		85.3			85.3		85.5		85.1	
80 Hidden		85.3			85.3		85.2		85.1	

Table 5.3 Complete Percent Correct Results of the Prime [A] from 39 ANNs, with Connectivity, Gain and Hidden Units as network parameters

Table 5.3 shows the complete percentage correct results for all the networks trained on the Prime [A]. The results show that above a setting of 50 Hidden Units and 50% Connectivity, the results from the networks plateau and do not show significant improvements (see Figure 5.1).

The accuracy results indicate that the optimum efficient setting to achieve meaningful recognition is around these peak values. This was of help in extending the results to the remaining primes, especially given the time constraints of the experiment. It can also be noted from the table, that a Gain set to 10^{-04} produced worse results than those networks with a Gain of either 10^{-05} or 10^{-06}, for several randomly chosen networks. Therefore, in the consequent experiments, the Gain was set to 10^{-05} and 10^{-06}, for the remaining networks for single primes, in the interests of reducing the number of variable parameters and obtaining the best performing networks.

Experiment 1 (C) Error Minimisation Prime [A] (10-40 Hidden Units)

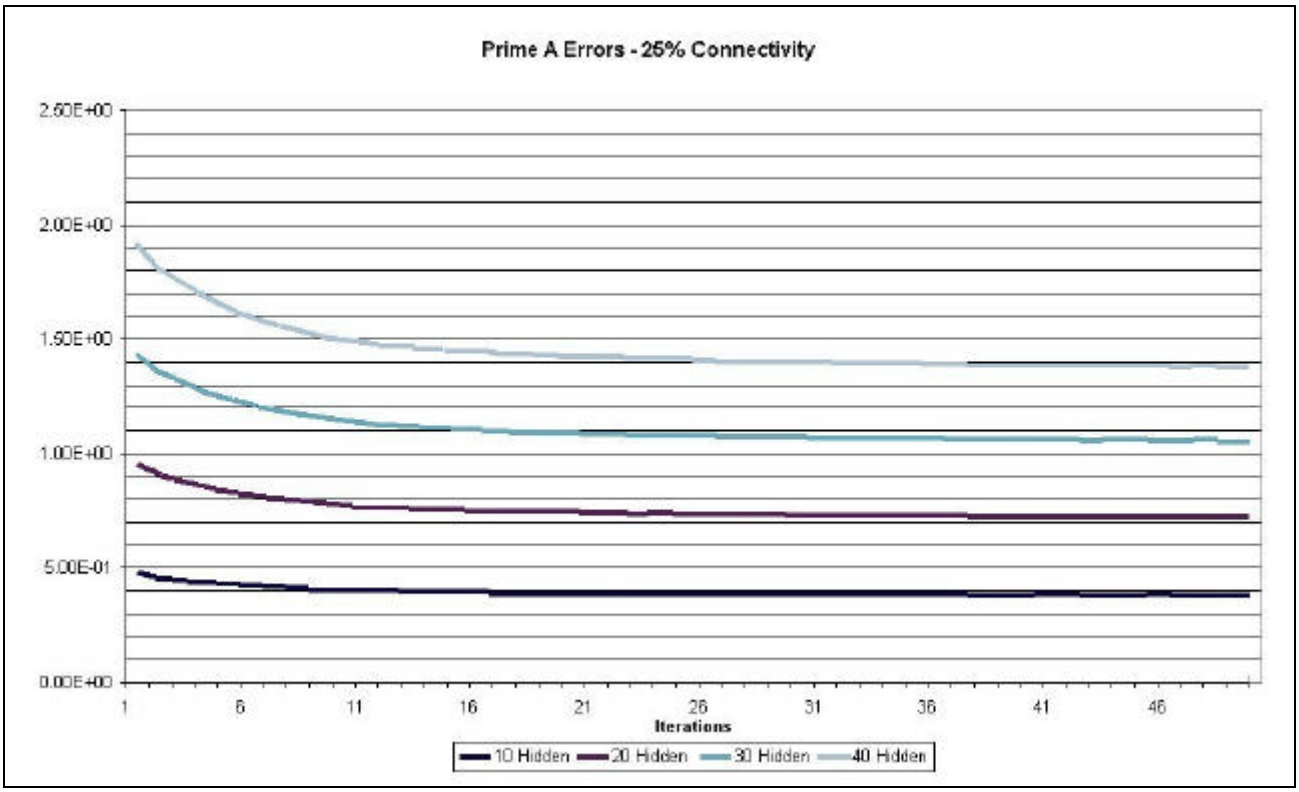


Figure 5.2 Error Minimisation results over 50 iterations of network training for Hidden Hnits 10-40

Figure 5.2 shows the results of the error minimisation technique. For the results given here, the number of Hidden Units varied from 10 to 40, and the Connectivity was set to 25%. These networks were set to perform 50 iterations and the Gain was set to 10^{-05}.

From the graph, we can see that there is an initial descent in the error performance and that after that the performance levels off and does not show further improvements. At about 20 iterations, the networks have achieved their maximum improvement in error minimisation.

Experiment 1 (D) Error Minimisation Prime [A] (50-80 Hidden Units)

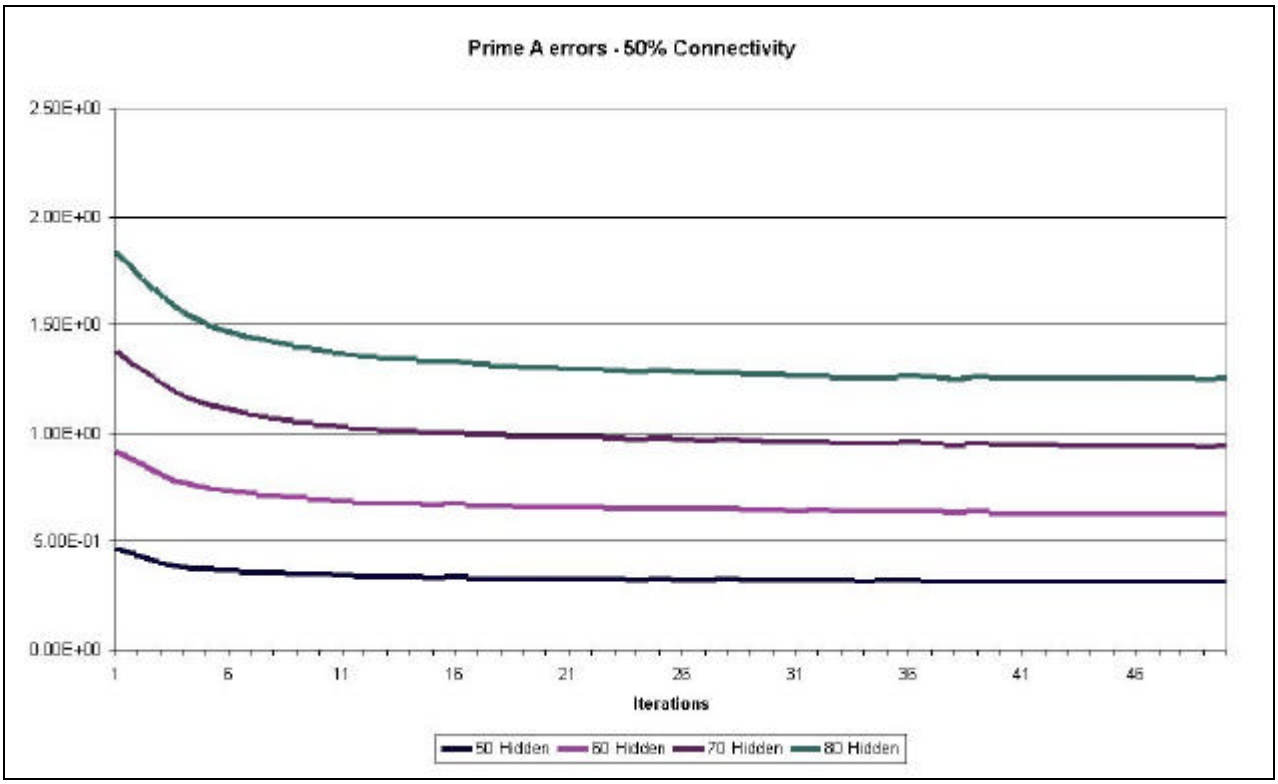
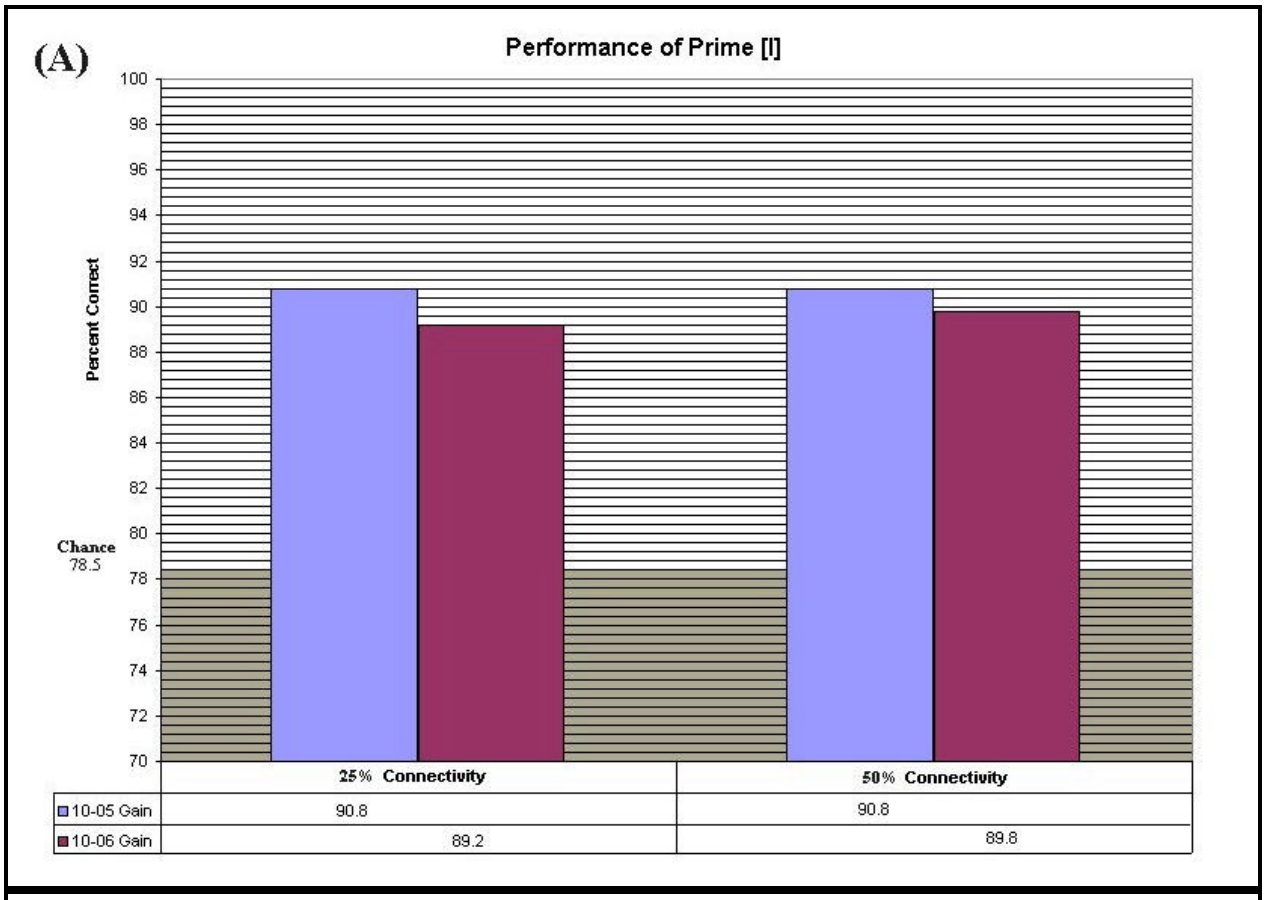
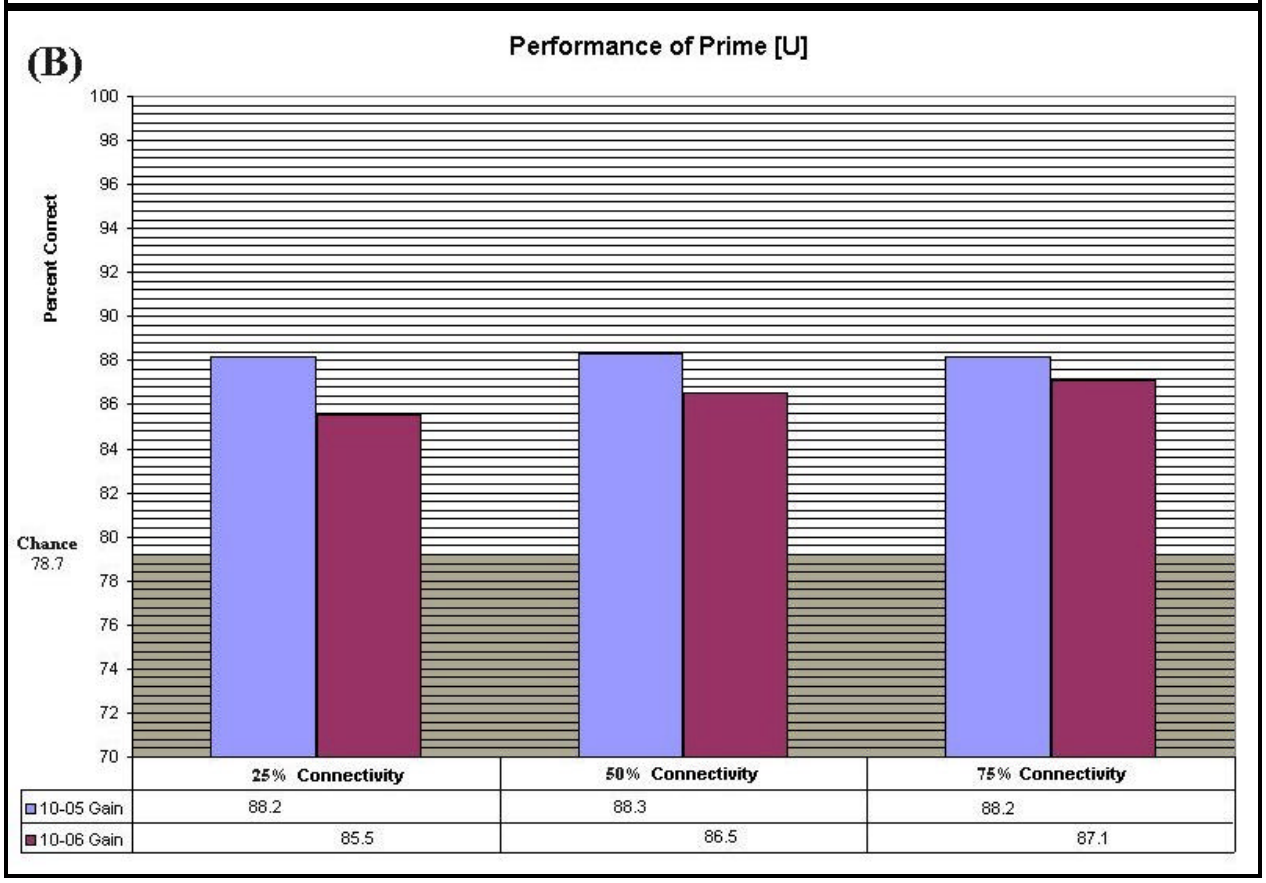


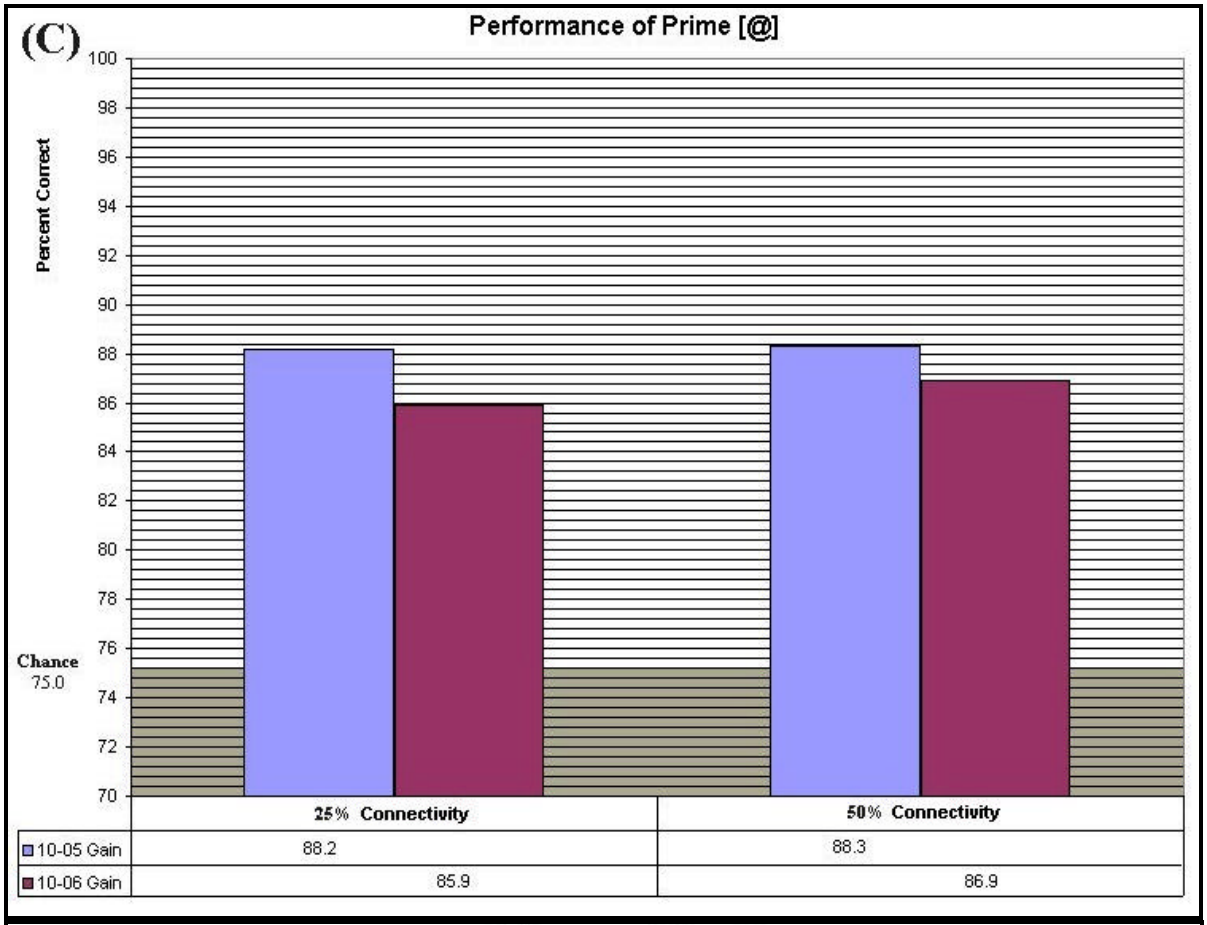
Figure 5.3 Error Minimisation results over 50 Iterations of network training for Hidden Units 40-80

In Figure 5.3, the networks shown are those containing 50, 60, 70 and 80 Hidden Units. For these four networks, the Gain was set to 10^{-05} and the degree of Connectivity between nodes was at 50%. Plotting along these axes, the pattern is seen to be consistent throughout. In a similar fashion to Figure 5.2, an initial sequence of large improvements slows to smaller and smaller error minimisations. The performance steadies after around 15 iterations.

Experiment 2 (A) The Other Primes Separately






(C)
Performance of Prime [@]
Percent Correct
100
98
96
94
92
90
88
86
84
82
80
78
76
74
72
70
Chance
75.0
25% Connectivity
50% Connectivity
10-05 Gain
88.2
88.3
10-06 Gain
85.9
86.9

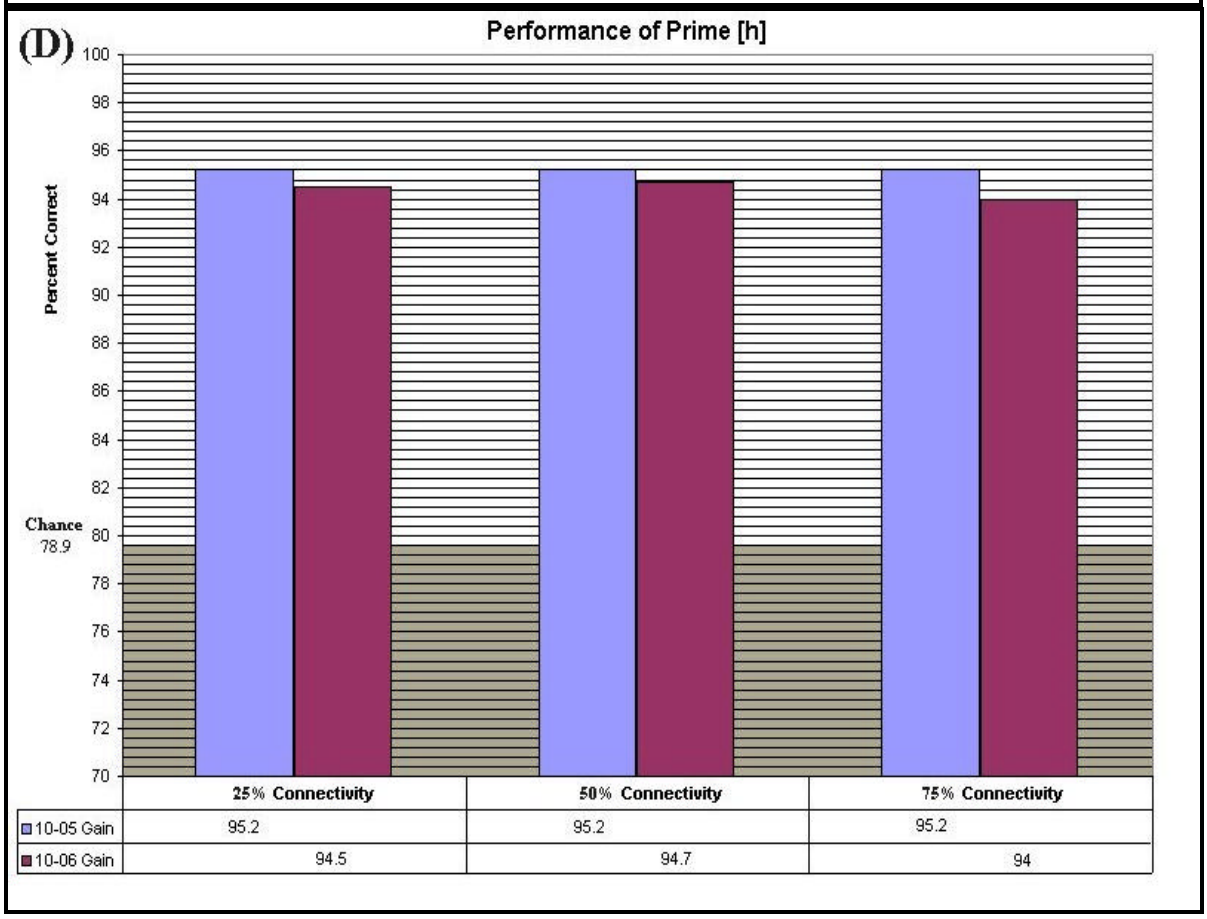

(D)
Performance of Prime [h]
Percent Correct
100
98
96
94
92
90
88
86
84
82
80
78
76
74
72
70
Chance
78.9
25% Connectivity
50% Connectivity
75% Connectivity
10-05 Gain
95.2
95.2
95.2
10-06 Gain
94.5
94.7
94

Figure 5.4 A) Performance of Prime [I] B) Performance of Prime [U] C) Performance of Prime [@] D) Performance of Prime [h]

Experiment 2 (A) The Other Primes Separately (Complete)

	[I]		**[U]**		**[@]**		**[N]**		**[?]**		**[H]**		**[h]**	
Gain	10^{-05}	10^{-06}	10^{-05}	10^{-06}	10^{-05}	10^{-06}	10^{-05}	10^{-06}	10^{-05}	10^{-06}	10^{-05}	10^{-06}	10^{-05}	10^{-06}
25% Connectivity	90.8	89.2	88.2	85.5	88.2	85.9	97.8	97.4	91.2	90.0	94.6	94.1	95.2	94.5
50% Connectivity	90.8	89.8	88.3	86.5	88.3	86.9	97.7	97.3	91.6	90.5	94.6	94.0	95.2	94.7
75% Connectivity			88.2	87.1					91.5	90.9			95.2	94.0

Table 5.4 Results for the Primes [I], [U], [@], [N], [?], [H] and [h] with two levels of Gain and 3 levels of Connectivity. There are 50 Hidden Units for these networks

Table 5.4 indicates that each individual prime was trained with a network set to 50 Hidden Units, with variable parameter values at 25% and 50% Connectivity and 10^{-05} and 10^{-06} Gain. For comparison, an additional 6 networks were trained with 75% Connectivity and 10^{-05}/ 10^{-06} Gain.

Four representative graphs of the performance of each prime individually are shown in Figure 5.4 (A-D). The graphs also show the different levels of chance for each prime. This is indicative of the level of involvement of each prime in the database (as described in Section 4.2.6).

A Chi-Square for Figure 5.4 A) shows that the total results for the prime [I] are not significantly higher than chance overall [χ^2=6.97, p< .10]. However, when the test is run only on those results that were derived from networks with 10^{-05} Gain, [χ^2=3.85, p< 0.05] these results *are* seen to be significantly better than chance.

Figure 5.4 B) shows the results of training on the prime [U]. The performance of each network is above the random level of chance. However, the Chi-Square test does not show statistical significance above the chance level with either the combined or separated results along the Gain parameter. The best performance was derived from the subset of networks trained with 10^{-05} Gain [χ^2=4.62, p< 0.10].

Figure 5.4 C), showing percent-correct performance for the prime [@] is significantly above chance levels overall on all the ANNs involved in the assessment of this prime [χ^2= 8.15, p< 0.05]. However, when divided according to Gain size, it can be seen that those networks with a parameter value of 10^{-05} Gain are performing significantly better than chance [χ^2= 4.68, p< 0.05]. In contrast, those networks trained with 10^{-06} Gain performed at only marginally-significantly better than the expected level [χ^2=3.47, p< 0.10].

The performance of the prime [h] is contained in Figure 5.4 D). This showed very good results, and overall the performance was a highly significant observation compared to the random expectation level [χ^2=19.24, p< 0.01]. Once again, by dividing this result into differing Gain size performances, ANNs with a Gain size set to 10^{-05} showed highly significant above-chance performance [χ^2=10.11, p< 0.01]. Those networks with a Gain size of 10^{-06} also achieved significant differences [χ^2=9.13, p< 0.05], to a lesser degree.

Experiment 2 (B) Impact of Gain Parameter

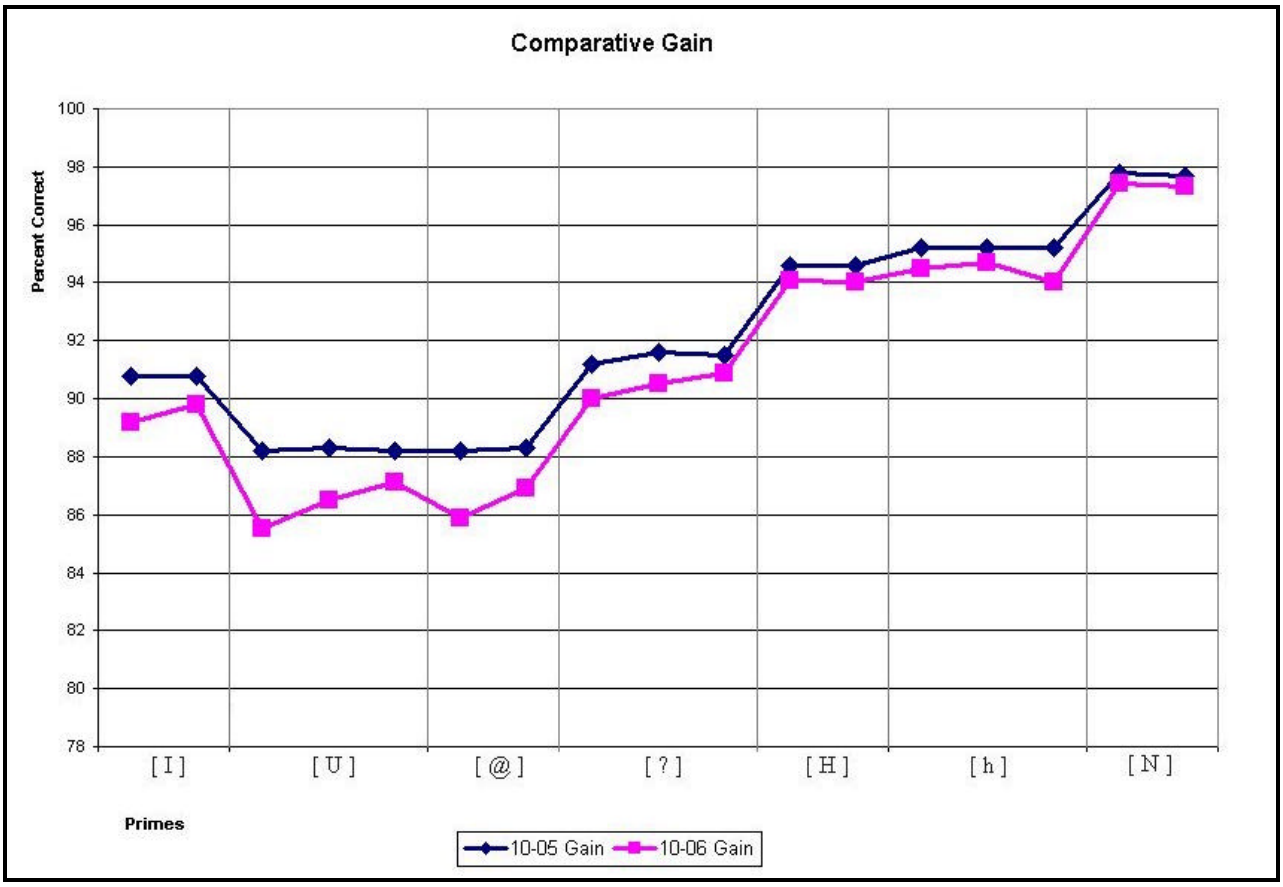


Figure 5.5 Comparison of the Performance of the primes with Gain as the variable parameter

In Figure 5.5 the performance of each prime is seen to be consistently lower when the Gain is set to 10^{-06} in the training of the networks. The Gain parameter produced higher results when its value was set to 10^{-05}. This discovery assisted in the extension of the experimental design when it came to adopting parameters for training all the primes concurrently.

Experiment 2 (C) Best Result for each Prime

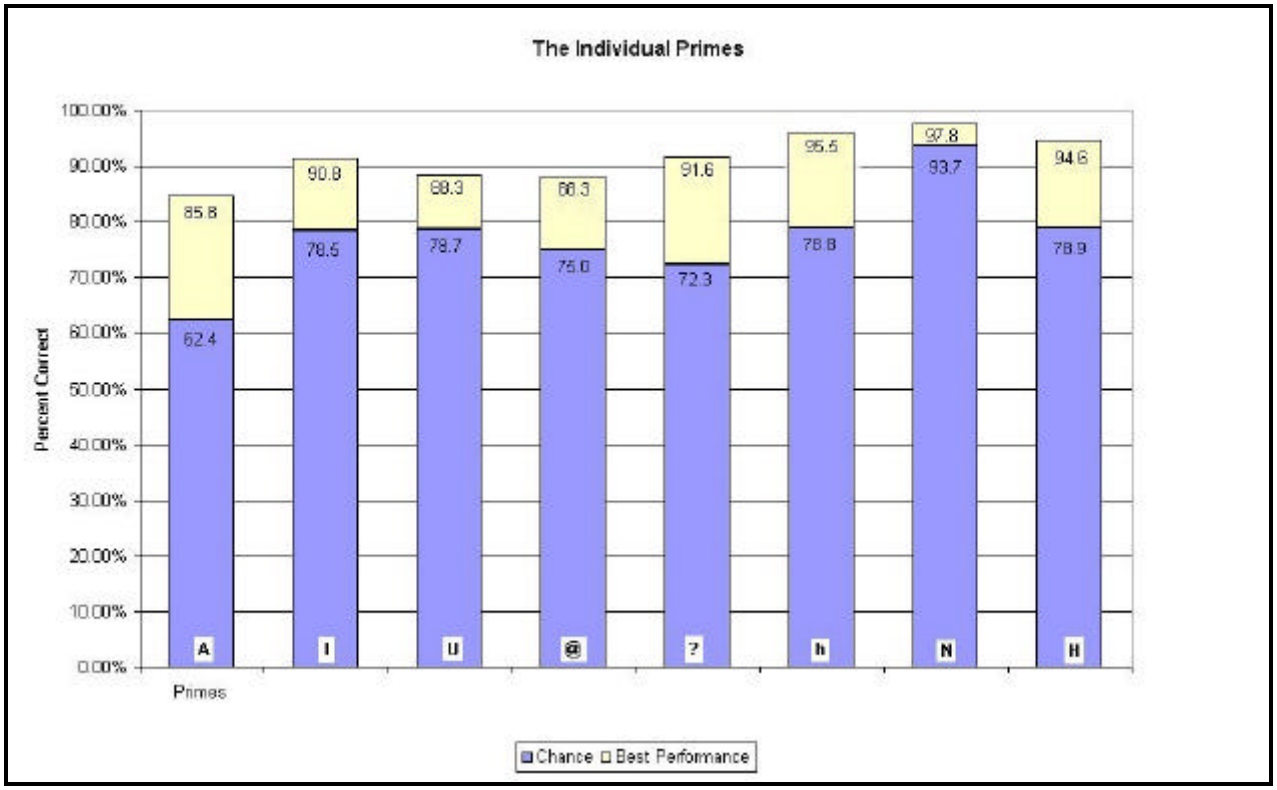


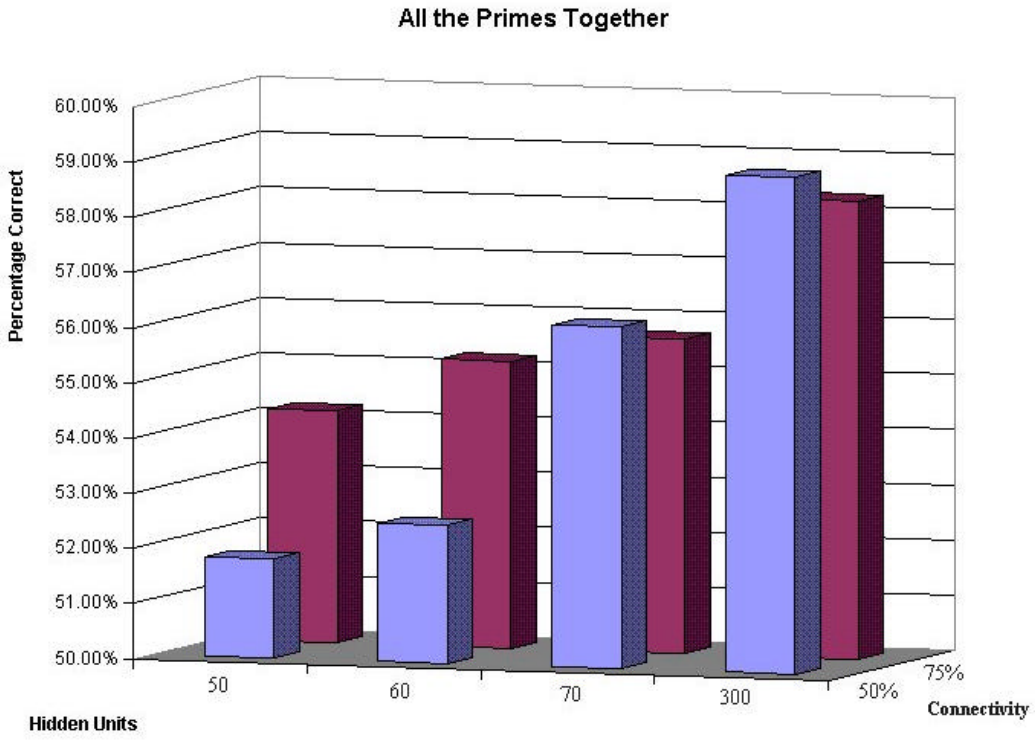
Figure 5.6 Best Results for each of the 8 primes showing individual chance levels

Prime	[A]	[I]	[U]	[@]	[?]	[h]	[N]	[H]
Best Score	85.8%	90.8%	88.3%	88.3%	91.6%	95.2%	97.8%	94.6%
Chance	62.4%	78.5%	78.7%	75.0%	72.3%	78.8%	93.7%	78.9%

Table 5.5 The Best Scores for each prime, together with their chance levels

Figure 5.6 and Table 5.5 show the best result for each prime and an indication of the chance level for each, obtained according to the procedure outlined in Section 3.5.6. The graph shows that all the primes perform above chance to varying degrees, and that the percentage correct result for each prime varies according to type. The first four columns in the chart represent the resonance elements, corresponding to vowel and glide elements, which can be compared to the last four columns, with primes that are classed as manner and source (consonants and source primes performed better overall).

Experiment 3 (A) All the Primes Concurrently



	50	60	70	300
50% Connectivity	51.80%	52.50%	56.20%	59.00%
75 % Connectivity	54.20%	55.20%	55.70%	58.30%

Figure 5.7 All the Primes together at 50, 60, 70 and 300 Hidden Units

Figure 5.7 contains the results of training and testing networks with 8 concurrent input and target streams, representing whole phonological expressions. After initial testing with 50, 60 and 70 Hidden Units, two networks were trained with 300 Hidden Units, since it transpired that the computational strain of 8 inputs required larger ANNs. The Gain size for all the networks were set to 10^{-05}.

The chance level for all the primes trained and tested together is the probability that the most common result be cited. Given the nature of the TIMIT database files, the highest frequency occurrence is that of silence, which is contained in 13.96% of the frames in the test set. If the chance of silence is removed, the chance of a network selecting the most frequent phonological element is the chance of identifying the prime [I], given no silence, which occurs in 12.19% of frames in the test files. The second-most frequent phonological expression is [A, I] at 9.29% The

Chi-Square reveals that the performance of all the primes compared to chance was very significant, [χ^2=871.57, p<0.01].

Experiment 3 (B) Concatenated Primes

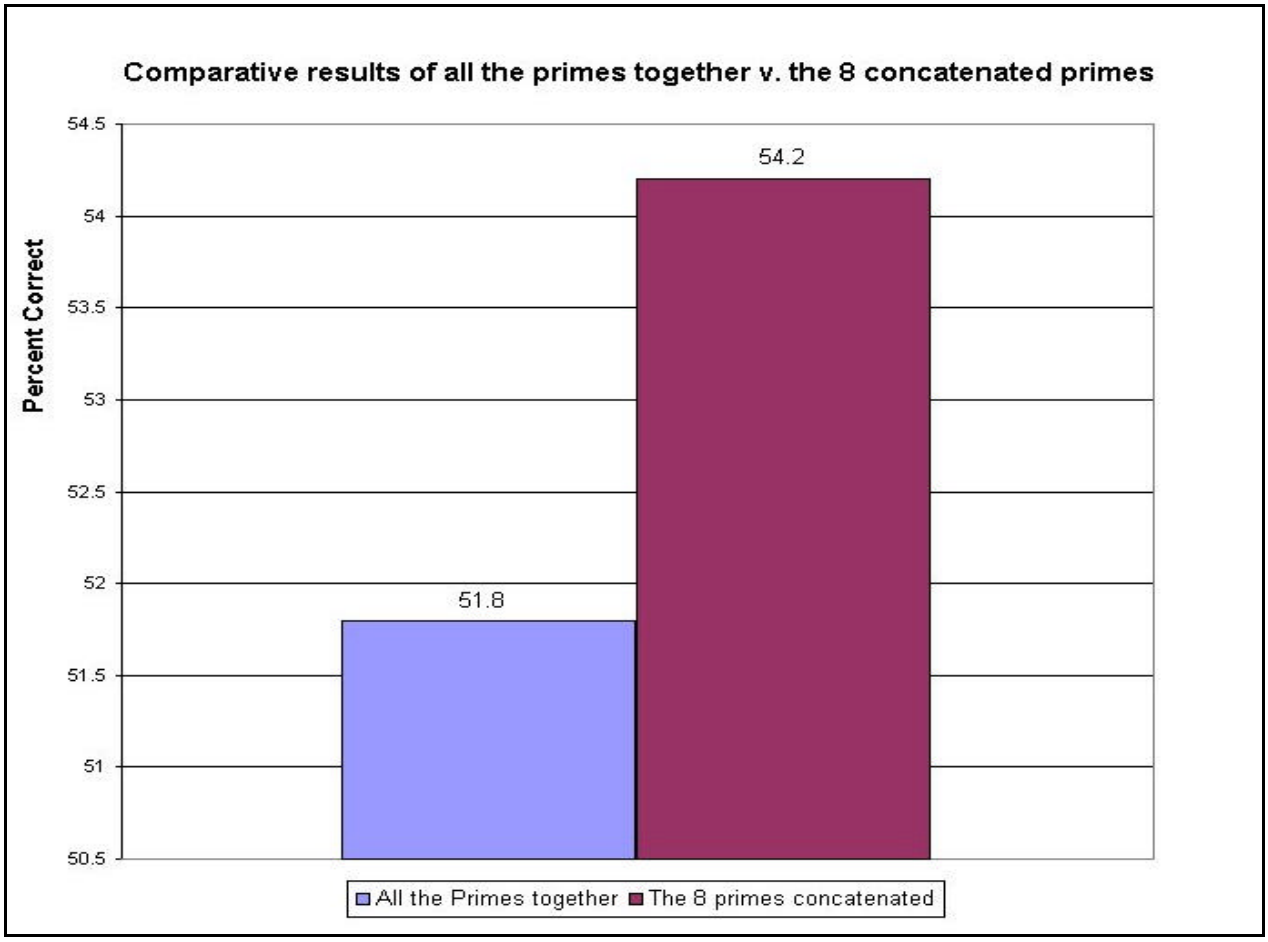


Figure 5.8 Performance of the 8 primes trained together v. concatenated results of 8 primes trained separately

Figure 5.8 contains a comparison in the performance of 8 concatenated individual prime outputs, which were trained with the parameters 50 Hidden Units, 50% Connectivity, 50 Iterations, 10^{-05} Gain. The original network result on all the primes together for comparison, was also trained with the same parameters (from Figure 5.7).

The graph shows an improvement of 4.6% on the result from the network trained on all 8 primes together. This reflects the difficulty in producing multi-streamed output as noted in Section 4.5.

Experiment 3 (C) Mapping onto the nearest phonological expression

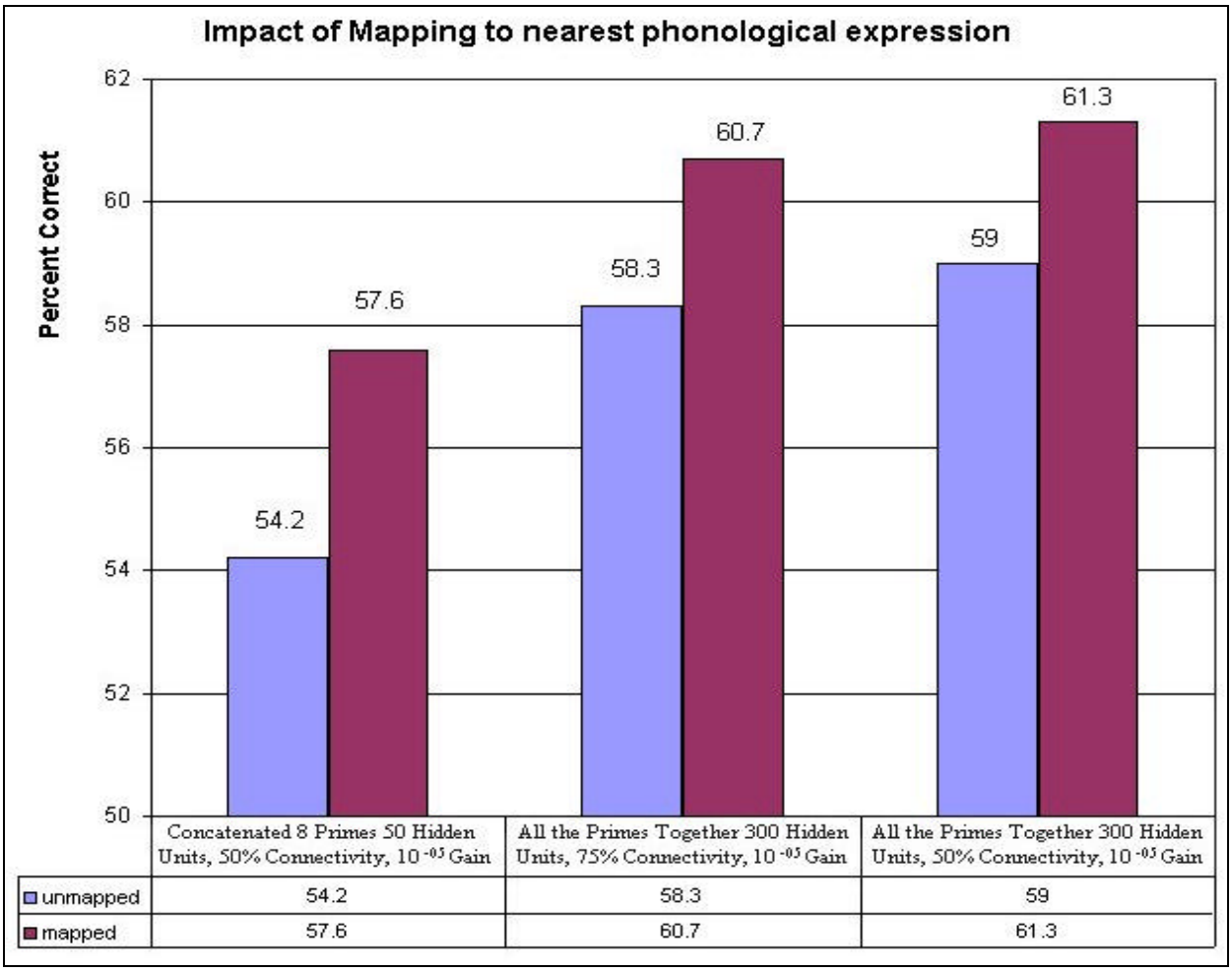


Figure 5.9 Comparison of unmapped outputs with outputs mapped onto the nearest phonological elements

Figure 5.9 shows how the performance was improved when the outputs of three networks were mapped to their nearest legal phonological expressions, as explained in Section 4.5.2. This procedure resulted in improved performance (3.4%) of the concatenated 8 network outputs (from Figure 5.8), which is a total 6.1% improvement on the original 8 streamed network based on all the primes simultaneously (from Figure 5.7).

The output of the network with 300 Hidden Units and 75% Connectivity was mapped, thereby forcing any invalid outputs, with perhaps only a single incorrect stream to become the closest legal phonological expression. Results derived on the new output showed an improved performance from 58.3% to 60.7% indicating the effectiveness of this simple procedure. The best performing concurrent network (300 Hidden Units, 50% Connectivity) was similarly mapped and produced the best performance achieved on the TIMIT database with 61.3% correct (Figure 5.9).

Experiment 3 (D) Error Minimisation

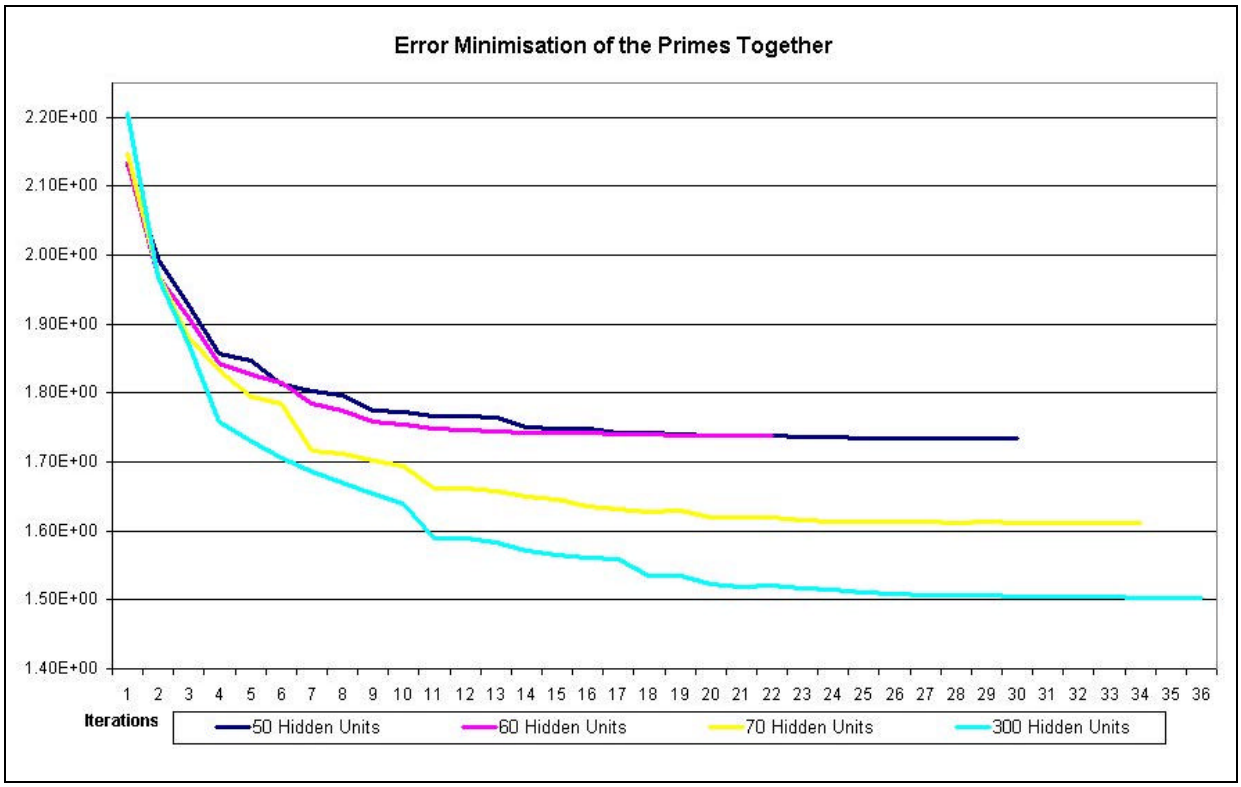


Figure 5.10 Error Minimisation graph of all the primes together at 50, 60, 70 and 300 Hidden Units

The plotting of the error minimisation in Figure 5.10 was derived from logfiles of networks trained on all the primes simultaneously at 50% Connectivity and 10^{-05} Gain. The networks were all set to perform 50 iterations. However, there are different numbers of iterations produced for each (30,22, 34,36 for 50, 60, 70 and 300 Hidden Units respectively) due to the impact of the error criterion. The error criterion takes the minimisation function and applies a reduction in gain step size to the network (see Figure 5.10 at 23 and 31 iterations), when the performance is out of the bounds of expected progress. These networks were set to a maximum of ten such reductions before the error criterion was met and training was halted. The graph shows that the computational strain of those networks with 50, 60 and 70 Hidden Units was too great and the Error Criterion halted their processes. The network with 300 Hidden Units showed a more comfortable descent and the greatest degree of error minimisation.

Experiment 4 Headedness

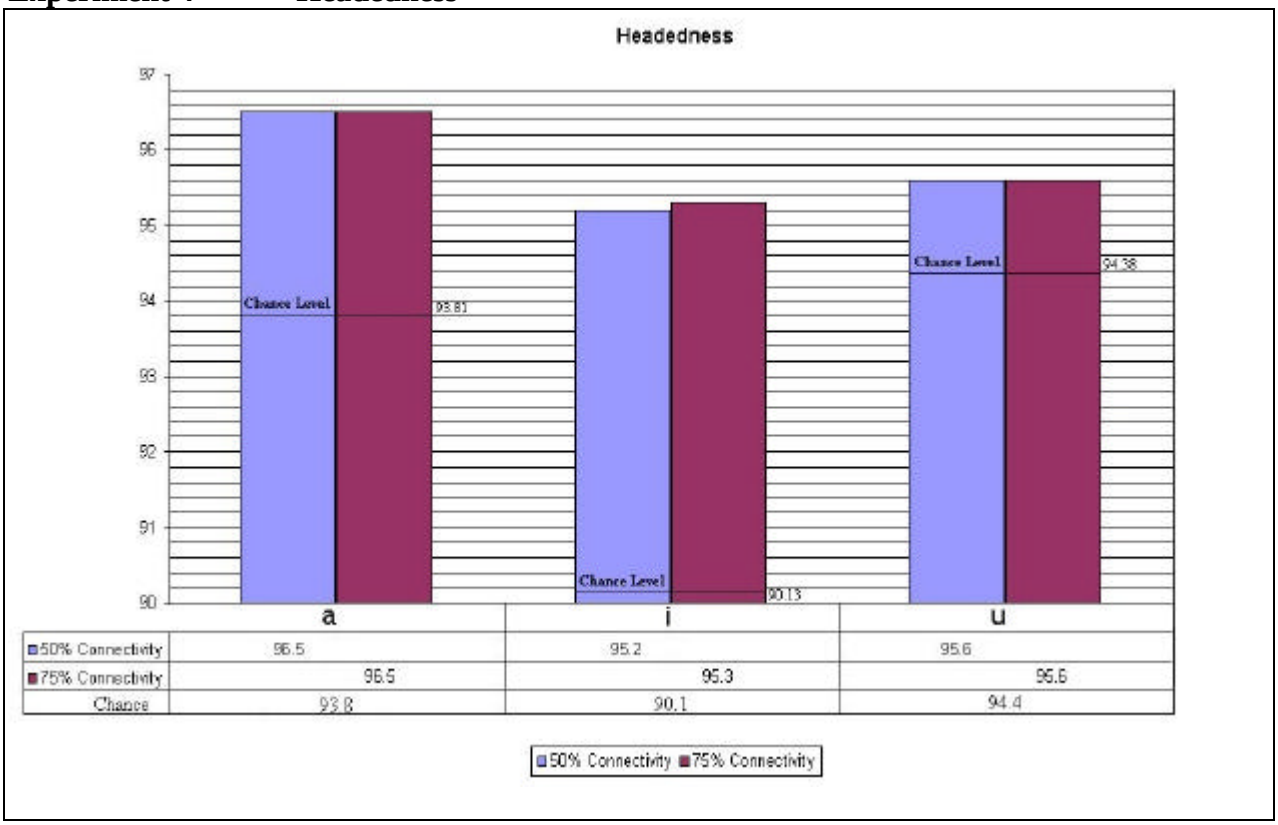


Figure 5.11 Headedness results for [a], [i] and [u]

In Figure 5.11, the graph shows the comparative performance of each of the headedness primes with their respective chance levels. The networks tested contained both 50% and 75% Connectivity and they were trained with 50 Hidden Units and 10^{-05} Gain. The results of a Chi-Square test do not show a significant improvement on chance levels. However, this is not unusual, given the high level of chance derived from the database. The Connectivity parameter does not appear to impact greatly on the performance of the networks on headedness. *The head prime* [i] shows the only variance in this aspect, but by only a fraction of a percent (95.2% v. 95.3%). Otherwise, the performance of networks with each type of Connectivity produced identical results.

Frame-by Frame Phone Recognition incorporating Headedness

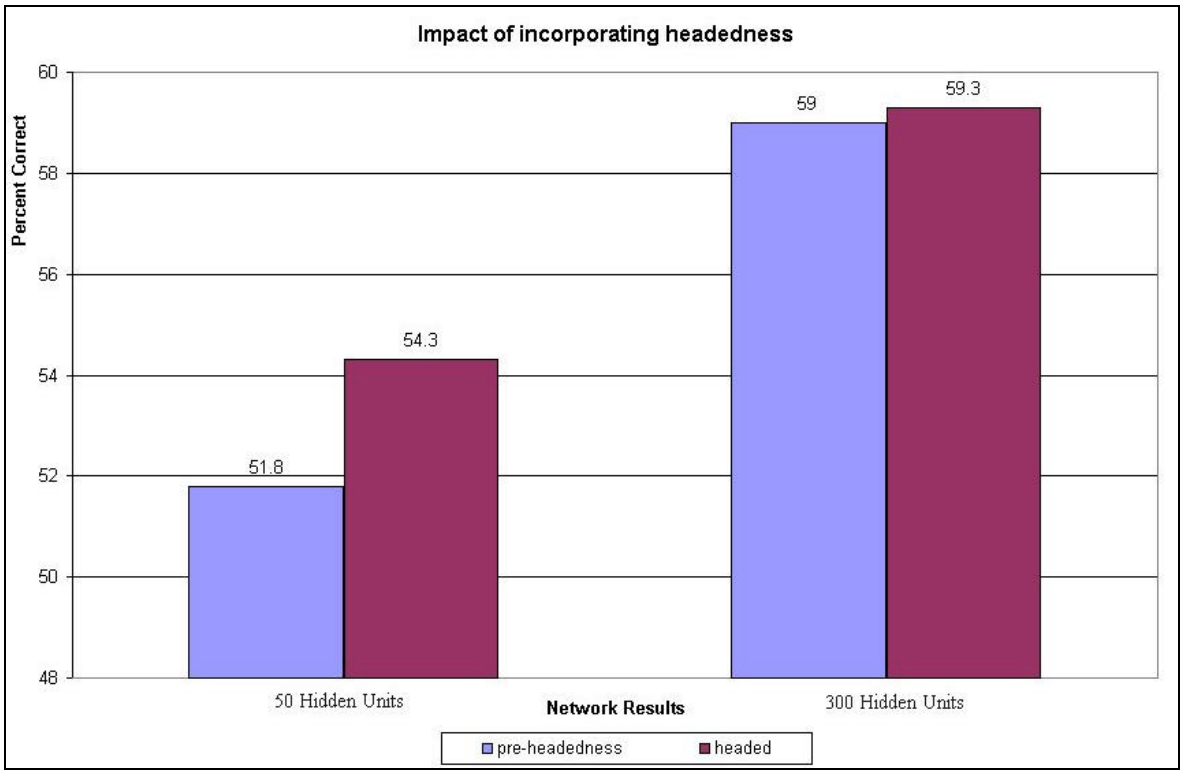


Figure 5.12 The Frame-by-Frame Recognition Results before and after the concatenation of headedness

The results of this subsection of the investigation of headedness are plotted in Figure 5.12 above. The first pair of columns reflect the improvement in performance when the network with 50 Hidden Units and 50% Connectivity was concatenated with the results from the networks trained on the head primes [*a*], [*i*] and [*u*] respectively. The concatenated output files were then matched to the concatenated target files, for percent correct results. In the second pair of columns, the findings are those derived from concatenation of the head primes outputs with the output of the 300 Hidden Units network. The Gain step size for all the networks, primes and heads, is 10^{-05}.

Experiment 5 Gender

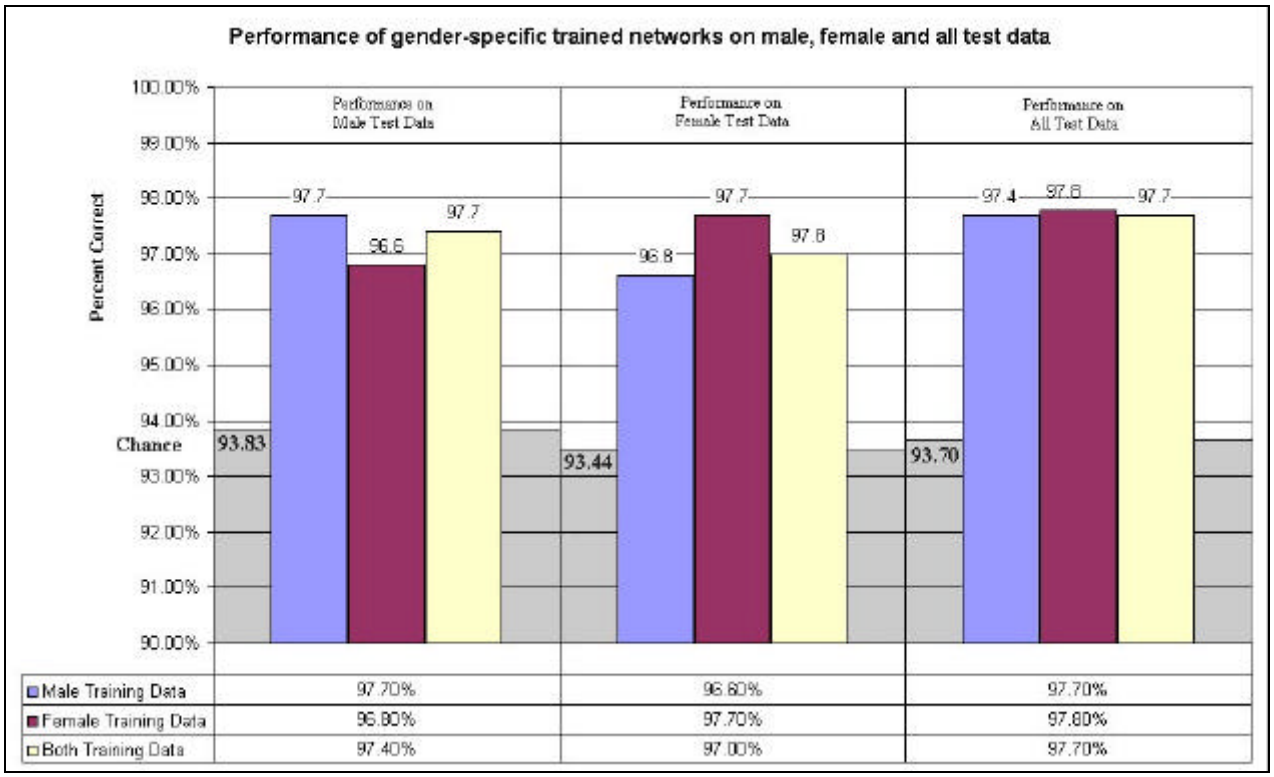


Figure 5.13 The actual and chance performance of networks trained on the prime [N] data that are either male, female or both. These networks were tested on each type of test set

In Figure 5.13, the following parameters were adopted in the gender-biased networks for recognition of the prime [N]. Hidden Units – 50, Connectivity - 50%, Iterations – 50 and Gain - 10^{-05}. The male-trained network, tested with male data achieved 97.7%. The score with female test data on the female-trained network was 97.7%. The performance when both data sets were exposed to both training data sets simultaneously in a gender-non-specific network is also 97.7%, which is expected since it is equivalent to training and testing with both genders individually. The graph shows that male training data performs best on male test data, that it performs worse on unseen female data and that exposure to both yields a performance in the middle. The same is true for female training data achieving its best performance on female test data. The both-gendered training data performs best on the female test set, which would not have been expected, given the male bias of the training and test data in the TIMIT database.

The impact of gender was further investigated by taking male test files as exposed to the male-trained network and the female test files after exposure to the female-trained network, to determine whether gender-specific networks are more powerful. The combined files from the male test outputs after recognition by the male-trained network and the female test outputs after exposure to the purely female data trained network were used to extract percentage correct results shown in Figure 5.14

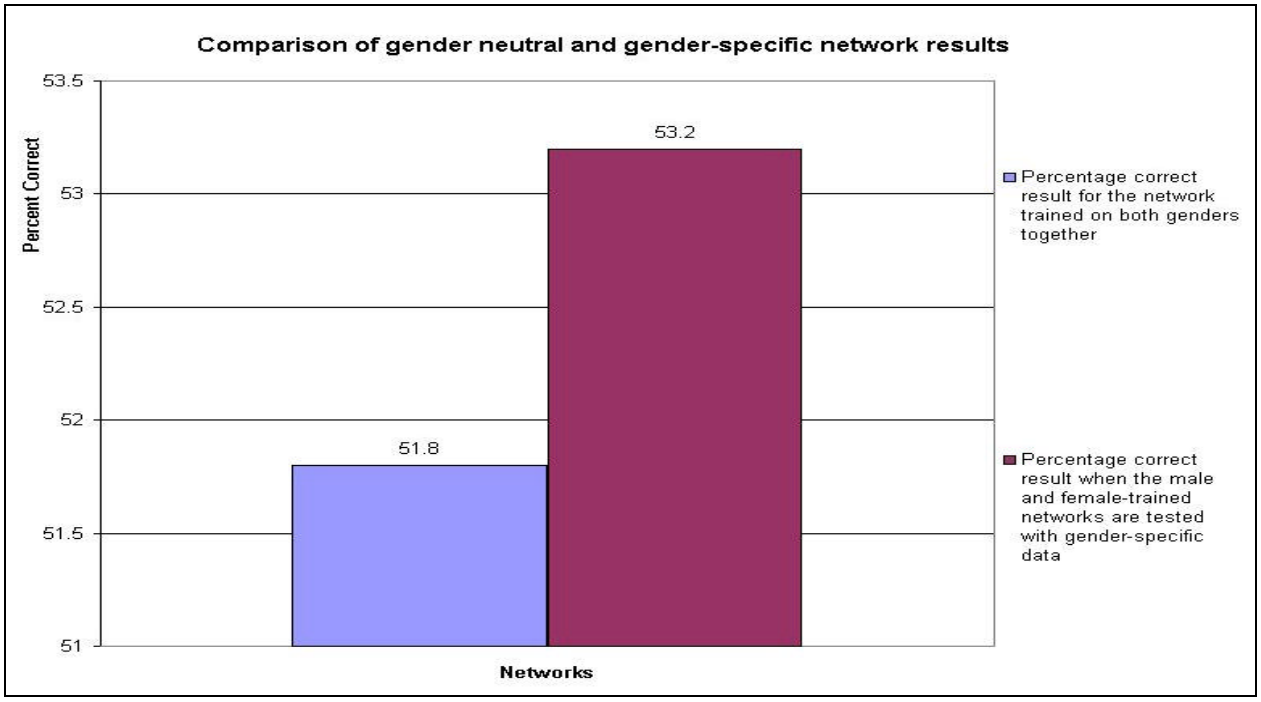


Figure 5.14 Comparison of gender-insensitive and gender-sensitive training

All the Primes Concurrently	**Male test data**	**Female test data**	**Test data for both**
Male trained network	55.8%	45.0%	52.1%
Female trained network	37.3%	48.4%	41.1%
Both-trained	52.2%	51.0%	51.8%

Table 5.6 Results of gender on networks trained with all the Primes together

Table 5.6 consists of the results from the network of all the primes together, trained with the following parameters, Hidden Units – 50, Connectivity - 50%, Iterations – 50 and Gain - 10^{-05}.

Experiment 6 Dialect Region

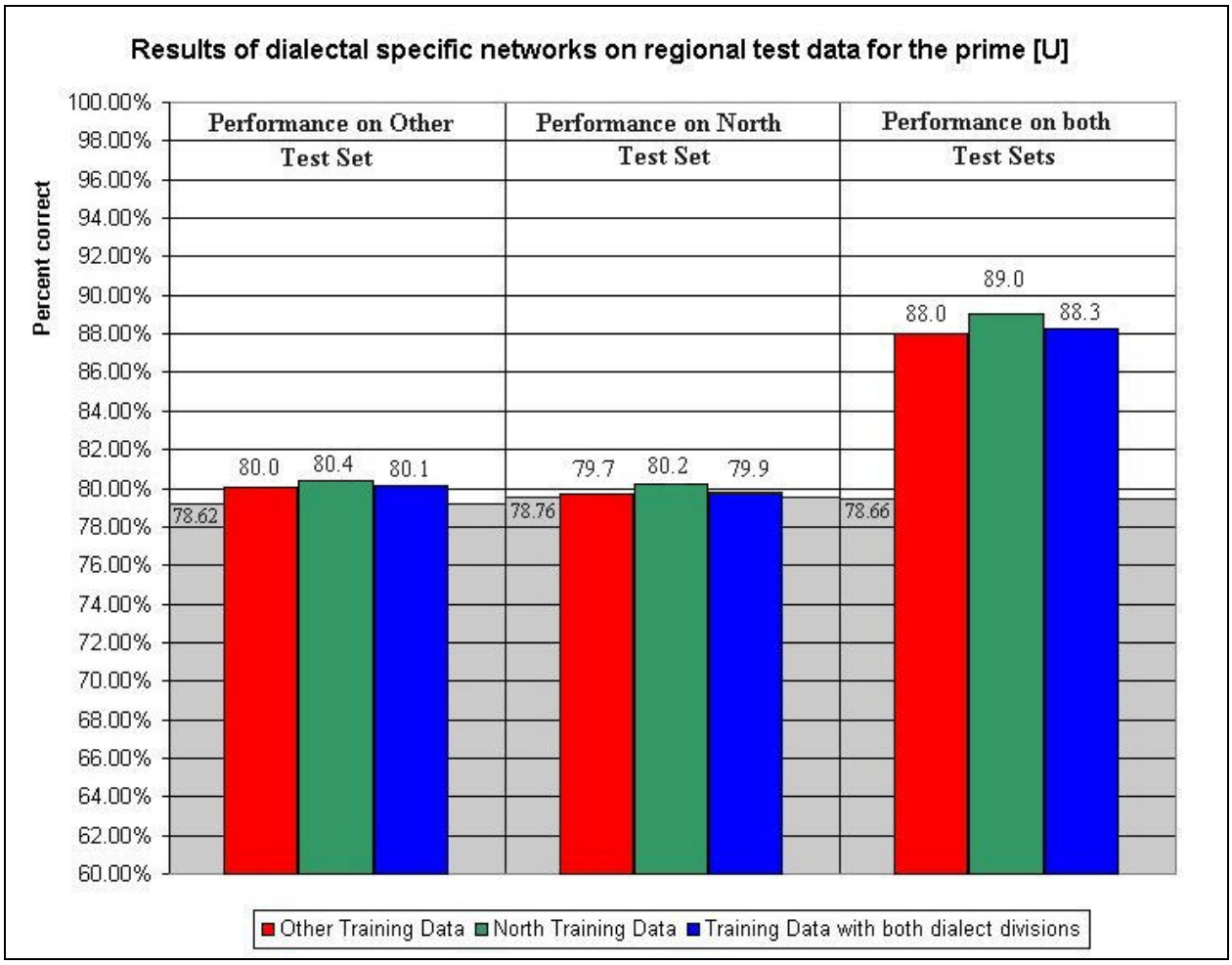


Figure 5.15 Results of dialectal specific networks on regional test data for the prime [U]

The results for the dialect region differences on the prime [U] are contained in Figure 5.15. The parameters of the network are 50 Hidden Units, 50% Connectivity, 10^{-05} Gain and 50 Iterations. Table 5.7 shows the results of training with dialect differences on all the primes together.

All the Primes together	**Other**	**North**	**Both**
Other	49.3%	50.8%	49.7%
North	49.3%	51.1%	49.9%
Both	51.3%	53.2%	51.8%

Table 5.7 Results of dialectal specific networks on regional test data for all the primes together

The results from Figure 5.15 based on the prime [U] are surprising in that they show very slight differences in performance for ANNs trained on the Other and Northern dialect groups, when

tested on each others data sets. The Northern-trained network actually performed better on dialects it had never seen than did the Other-trained ANN when tested on its own data type.

To determine whether dialect-specific models are an appropriate approach, a comparison was drawn between dialect-specific and dialect-insensitive networks.

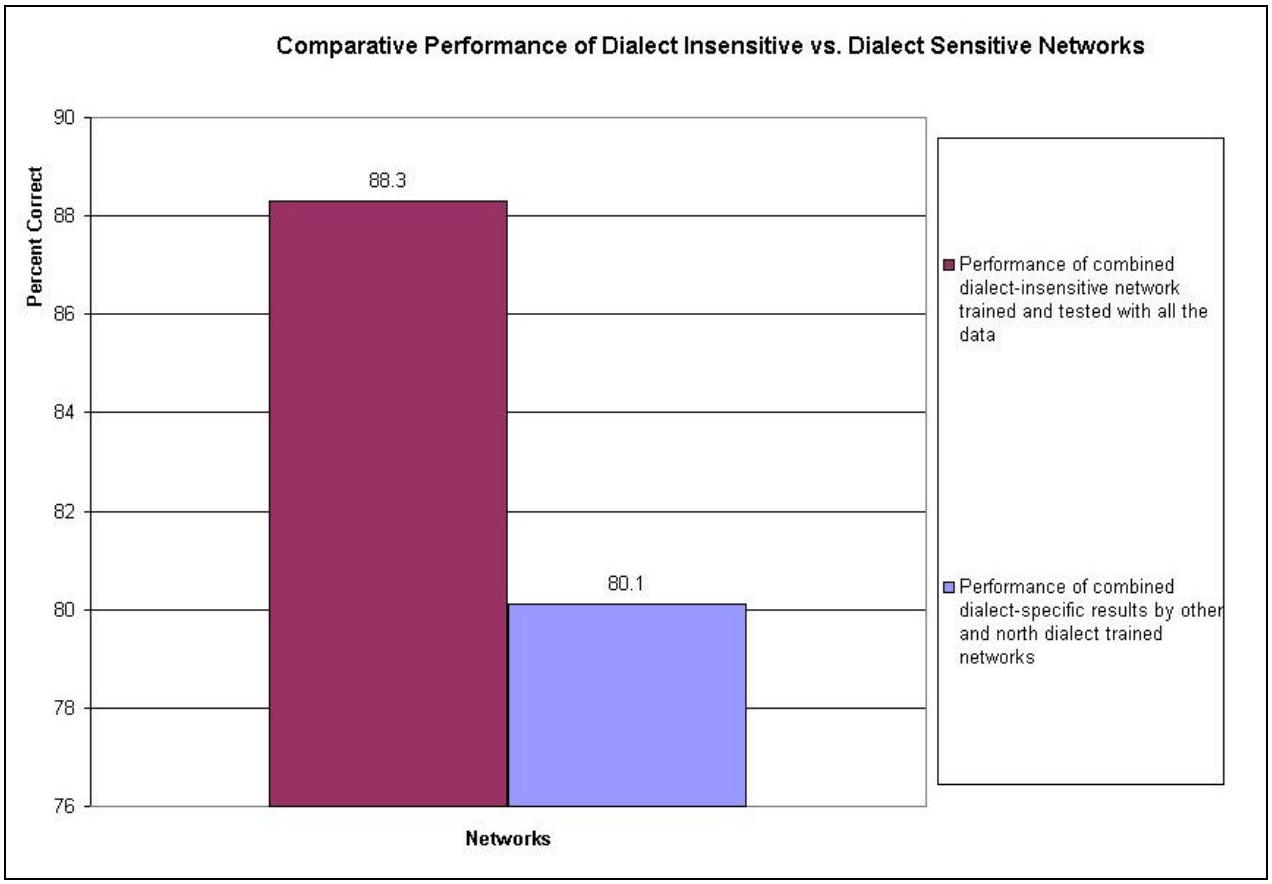


Figure 5.16 Comparison of a both-dialect trained and tested network with the results on both test sets from networks trained and tested on male data only and female data only.

The results of an *other*-trained network on *other* dialect test data were combined with the results from a *north* trained network tested with *north* dialect type test data. Then the comparison with the overall test patterns yielded percent correct results, as shown in Figure 5.16. It can be seen that there is no improvement in performance in training the dialects separately. The performance drops from 88.3% to 80.1%, when the dialects are split according to North dialect types and Other dialect types. There is therefore seen to be no practical benefit to the division of the training set according to dialect regions.

Chapter Six Discussion

6.1 Major Findings

In total, 95 different networks were trained, 39 for the Prime A, 34 for the remaining separate primes, 8 for all the primes together, 6 for headedness, 4 for gender and 4 for dialect region types.

The results of the preliminary experiment on Prime [A] were encouraging. The network trained well with the new GP labels. The contour map (Figure 5.1) of the results showed that the performance of the networks on this prime was well above the chance level for this prime.

The sensitivity of performances to each ANN parameter was also shown to fluctuate. The results so far indicate that some of the network parameters generate more variation than others. For example, Experiment 1 (B) showed that varying the Gain of a network away from 10^{-05}, either higher or lower, produces less impressive results (see Table 5.3). The level of Connectivity (e.g. Figure 5.11) had less of an impact on overall performance in the test stages.

Error minimisation results showed that the rate of learning seems to follow a general pattern throughout, with no one network outshining another in this capacity. A quick drop in the error function of a network levels off quickly and maintains a steady almost horizontal minimal decline subsequently (see Figure 5.2 and Figure 5.3).

It remained to attempt to locate the best parameters, in order to ascertain the best-equipped network for the task of individual prime recognition. The results obtained from testing with data not previously used in the training stage provided a gauge as to which networks had trained well.

However, there is no alternative when it comes to evaluating network parameters than adopting a range of values for each and then training networks with those specifications.

The results from Experiment 2 showed that the remaining primes also performed well to varying degrees, many significantly higher than chance. The manner and source primes ([?], [h], [N] and [H]) performed better overall than the resonance primes ([A], [I], [U] and [@]) (see Figure 5.6). This is possibly due to the distinctive nature of the acoustic cues to which they correspond. The vowel primes involve slighter degrees of variation, making them more difficult to confidently recognise. This is also shown in Williams, (1998) who states that certain elements are more easily recognised than others, notably the elements [H], [N], [?]). In the Williams study, lower scores were also obtained for the resonance elements ([U], [A], [I]). Detection of the same elements in consonants remains the greatest challenge. 'It is a notoriously difficult distinction to make from the signal alone' (Williams, 1998; 160).

The differences can also reflect the amount of training examples available to the networks in the database. The performance levels are more impressive when the actual system of prime combinations is considered. The phoneme /e/, for example, contains both the primes [A] and [I], yet may be acoustically realised in different ways by speakers of different dialect regions and different gender. In order to detect the prime [I] consistently, a network needs exposure to many examples of frames containing this prime. This includes the phoneme /e/ and also /iy/, /ih/, /y/ and /ai/, from which predictions about the presence of the prime can be made. Without many examples of each case in different contexts and environments, the power of a network is restricted.

The primes trained separately show a performance reflecting their individual degree of presence or absence, varying from prime to prime, dependent on its level of inclusion in the data set. The performance of the primes, all trained together, was undertaken in Experiment 3. The results came

out between 50%-60%, although this is not considered to be a drop in performance from that of the individual primes. When it is considered that the number of outputs is increased, the performance is actually quite impressive. There is an inherent difficulty in training on 8 or even 11 independent variables concurrently. The connections between the nodes link all the outputs internally, adding complexity across levels of nodes that is not present in individual prime networks. The results are lower on average than the findings produced by Williams (1998), 65.78% on TIMIT, although his investigation relied on a 5-class broad classifier using only the elements ([?], [H], [U], [A] and [I]), which differs from the present study.

There was a need for an increased number of Hidden Units to deal with the additional input data and to cater to the fact that the network is learning a more complex mapping operation. This point was evident in the improvement seen in the performance when the number of Hidden Units was increased to 300. Performance increased from 51.8% to 59.0% by varying this one parameter alone. Training time for such networks was a limitation however, with several weeks required to train a single network. The results do show the promise for improvements, given the availability of greater computational time and power.

By concatenating the eight best prime results together in Experiment 3 (B), a comparison could be drawn, more reflective of the potential performance of an optimally performing machine. The network outputs from all 8 primes separately were combined in the correct order and in such a way as to make them indistinguishable from the outputs of the ANNs trained on all the primes at once (see Appendix G). The parameters of 50 Hidden Units, 50% Connectivity, 50 Iterations and 10^{-05} Gain were the same in both the concatenated results and the results from the single network trained on all the primes together in order to compare *like* with *like*. Figure 5.8 shows how the concatenated results from the single primes can be equated to the performance of the network trained on all the primes in parallel. The concatenated performance measure (54.2%) came out

slightly higher than the *combined*-prime score for the same parameters (51.8%). This may be due to the difficulty in identifying all 8 primes correctly at the same time. In a sequence of outputs, if even one of the primes were incorrectly identified, then the whole frame would be considered to be wrong. Comparison to the concatenated single prime results shows that the performance of all the primes together is not far from this comparative level.

6.2 Headedness Results

The method of capturing headedness in the current study appears to have worked well. The results (Figure 5.11) *do* show that the headedness feature is capable of being recognised and can provide a good estimation of the particular segment's predominant element. This is especially important for identifying vowels, which require fine-grained distinctions. An extension of this aspect of the project could be achieved by a tighter conception of headedness. The theory is not altogether consistent with regard to headedness in consonantal segments. The proposals about which element is to be considered the head of a phonological expression are constantly being adjusted. This is compounded by disagreement over the actual number of primes. The neutral element [@] is not included in some versions and it remains to be seen in further experimental work whether it is a theoretically justified prime/ entity. For the purposes of the present investigation, it has been shown to be salient.

6.3 Frame by Frame Phone Recognition

Extending the results even further, it was noted that the hypothesised outputs of the networks did not necessarily correspond to an actual phonological expression, as defined in the prime table (Appendix D). There are 60 phonemes, reasonably well separated acoustically. This provided the chance that the output of a network may be close to one of them (see Section 4.5.2). The 'distance' from the training set could be calculated with the sum of the squares of the distance that each output frame was from a target phoneme (Appendix H). The output of this analysis would no

longer be an indication of how much a feature is in evidence, but rather a yes/ no decision on the presence of that phonological expression. This decided on a *winner* between neighbouring choices.

From the graph labelled Figure 5.9, the performance was seen to improve by virtue of this mapping procedure. The process produced the best recognition result of 61.3%, for the network with 300 Hidden Units and 50 % Connectivity. The improvement was due to the elimination of invalid outputs from the networks, allowing only accepted prime combinations. The incorporation of headedness (Figure 5.12) yielded an improvement from 59% to 59.3%. However, this slight improvement may be the result of concatenating headedness outputs from networks trained with 50 Hidden Units onto combined prime outputs from the 300 Hidden Unit network. Further training with totally specified networks may generate further improvements. Alternatively, there may be an alternative representation for headedness, which may be better suited to locating the more dominant prime in an expression. The current conception may need further consideration.

6.4 Gender Results

The experiment using data from speakers of differing genders proved that networks trained on male data did better when tested using male speech than when tested on female input or on both genders combined. The male and female-trained networks were tested with 3 different excitation signals; a set of male test data, a set of female test data, and the combined test set used previously throughout the experiments. The male-trained networks tested with male data show higher correct recognition percentages than the female-driven ANNs on female test data (55.8% v 48.4% on all primes together). This is perhaps due to a bias within the TIMIT database, in terms of the proportion of male to female speakers (66% male 33% female in both the training and test sets). With networks exposed to more male-oriented data, they have a stronger ability to develop

generalisations. This is also an indication of the importance of the amount of data required to train ANNs.

The improved performance on gender-specific test data is also true for those networks trained and tested with a female speech signal alone. It has been shown here that when the test gender is the same as the training data, the results are improved (see Figure 5.13). Although the female trained network performed marginally better on both-gendered data (97.8% vs. 97.4 for male-trained data) on the prime [N]. This was not expected, given the bias towards male data in the training and test sets. This finding may make it difficult to show a predictable impact of gender in GP prime recognition.

In order to determine whether this was the case, it was decided to test the concurrent primes male-trained network with male test data (55.8% correct) and similarly test the network trained on female speech input with only female data (48.4% correct). These output files were then combined and tested on both genders together to yield the aggregate result of two gender specific networks being exposed only to their own gender. The accuracy result came out at 53.2%, (a weighted average of both performances individually); a higher result than training and testing both genders together (51.8%) as seen in Figure 5.14. This finding that indicates that training and testing the genders separately yields enhanced results.

The theory of GP does predict such differences in the recognition accuracy of the male-trained and female-trained networks, since it is based on prime-driven processing. The theory of abstract speaker independent elements does not concretely relate to the natural acoustic differences between male and female speech. The physiological differences between the vocal tracts of men and women produce different fundamental frequencies and formant transitions, making the postulation of abstract elements into an idealisation, at least in present technological terms. In

fact, the implication of these findings is that it may be worth developing gender-specific networks, which deal only with training and test sets of specific genders separately.

6.5 Dialect Region Results

The experiment into the impact of dialectal variation followed a similar line of investigation but uncovered different results (see Figure 5.15). The performance of the dialect regions was not as consistent compared to the findings from the gender experiment. The finding was that by subdividing the speech samples by dialect region, no major improvement was uncovered. The performance of the Other Trained network (Figure 5.15) on the unseen North Test data was slightly higher than even the performance of the North-specific network on this same data. This may be caused by the choice of dialects to group together. North and North Midland are perhaps not a good class of dialects to take. The fact that the set of 'other' dialect types also contains a category for Army Brat, meaning that the speakers moved around, may also influence the outcome. Another alternative is that the theory is insensitive to dialect type and exposure to all types of dialect is the best approach to take. This was shown in the combination of dialect specific networks tested with dialect specific data. When the results were concluded on the complete test data, the finding was that training with all the dialect types together is a more successful approach to take. Figure 5.16 plots the decrease in performance from 88.3% correct, from training on all the dialects, to 80.1% when the dialects were subdivided, trained and tested separately. This is a converse result to that obtained in the gender experiment.

6.6 Extension possibilities

There are several directions in which the findings of this investigation could be extended.

The results could be incorporated into a superstructure of nuclear and non-nuclear positions. It would most likely improve phone recognition accuracy to introduce a level of GP phonological structuring into the speech recognition process. Phonological parsing of this type exists in the forms of applications such as ANGIE, hierarchical framework for capturing phonological phenomena, and PhonMaster, a GP-specific structurally based prototype. This would help in the

identification of segments such as glides, which are only distinguished from vowels under this approach, by their appearance in non-nuclear position.

Certain primes do not occur together. For example, [U] and [I] cannot combine in English. This is a language-specific constraint. The prime [H] (voicelessness) and the vowel primes [A], [I], [U] and [@] occurring together is also unusual since vowels are inherently voiced. This constraint is universal. In the mapping procedure stage it may be possible to tie two language-specific, mutually exclusive primes to constrain the networks from considering one particular prime as a potentially combining element when the other has been identified. Although the network should learn this automatically from exposure to data, it may still help to refine the system, and to develop language-specific models. It may further improve the performance of speech recognition if the labelling of the database could be enhanced, by incorporating these improved constraints and eliminating impossible output values before they occur would likely remove a lot of the incidental error. For example, the procedure in Experiment 3 (C) of mapping the networks to legal phonological expressions can be incorporated internally into any ASR system to remove any null output hypotheses, making front-end adjustments unnecessary.

The current study has shown the significance of varying certain parameters within an ANN. Given more time and computing resources, it would be interesting to test neural networks with higher numbers of Hidden Units and different levels of Connectivity and Gain size. The time constraints of the project required sufficient levels of performance, which were achieved, but there are more optimal settings, as yet uncovered in training the ANNs, which remain to be explored.

Given the need for examples of GP-labelled speech samples, one way to provide them would have been to take the TIMIT database and relabel it frame by frame with GP primes. However, this task is too time consuming, making it impossible for the current project. The table of primes-to-GP

labels (Appendix D) indicated a conversion that was implemented automatically by computer. It would be preferable to have a fully GP-transcribed database, or a more rigorous endorsement of the validity of the conversion of phonological elements into GP labels.

As is the case in any speech recognition study, alternative data could be found. The present study could be expanded to include further examples of American English, thereby improving the accuracy of the recognition in this domain. However, the advantages of an approach based on GP theory means that multilingual databases ought to be as easily recognised. Features such as /ö/, will merely require training examples and some possible constraints in order to achieve recognition accuracy for a language containing this phoneme. As noted, the constraints for each language can be set to accept only certain combinations of elements. The simplicity of the prime-based approach makes this relatively straightforward to implement.

Finally, there may be a need for an alternative acoustic input, if the procedure is to be applied to forms of data other than TIMIT. Current speech recognition systems use surprisingly few parameters, about 10 to 16 spectral envelope parameters per 10ms. For the purposes of this experiment, this was quite adequate, given the clean, noise-free speech of the TIMIT database. However, an alternative data source may be used in other investigations, requiring a better representation. There has been an investigation into alternatives to mel cepstral coefficients by Ström, (1997a) who utilised a tonotopic architecture, more representative of the way in which a human ear perceives sound. His preliminary results are encouraging and are worth considering for future research.

6.7 Comparison to other works

Koreman *et al* (1999) used Eurom0, a multilingual database, making their study into phonetic mapping broader than the current project, which is based on the TIMIT database of purely

American English. In their approach, the results of mapping acoustic parameters onto sets of phonetic features yielded 42.3% for the feature set based on the IPA chart and 31.7% for the SPE features. The baseline result achieved with cepstral parameters alone was 15.6% (Koreman *et al*, 1999; 720), which seems to be a very poor baseline result. It can be difficult to draw comparisons with phoneme-based research. The use of different databases and the reporting of results in different styles compound this. Even when the database is identical, some authors use tokens as their measure of recognition accuracy rather than frames, while others report results for single speakers or subsets of speakers.

Reetz (1999) and Lahiri (1999) are working on a similar approach to that adopted in the present study, but without the shift to GP features. The speech signal is converted into speaker independent sets of phonological features, which are compared with feature sets stored in the lexicon using a ternary logic (Reetz, 1999). A system of *matching, no mismatching* and *mismatching* of features leads to the selection of word candidates. 13 phonological features are used (consonantal, vocalic, continuant, RTR, voice, strident, abrupt, nasal, labial, coronal, dorsal, high and low). They are an extension of Chomsky and Halle's set of binary features and are acoustically related to formants. For example, [high] is defined by F1 < 450 Hz (Reetz, 1999). A rule based approach is taken that first adopts a special treatment for the feature [abrupt], due to its rapidly changing nature which may be considered to be errors in analysis of the other features. Then gaps shorter than 5ms are searched for and filled with a feature. Next, isolated stretches of features in a track that are shorter than 15ms are removed. This is to account for the slow movement of articulators relative to this time span.

Lahiri (1999) also explores the possibility of FUL (featurally underspecified lexicon). The speech signal is converted from the waveform into an online spectral representation made up of formants and a few parameters describing the overall spectral shape. The LPC and spectral parameters are

converted into distinctive phonological features which, in turn, are directly compared with all the entries in the lexicon.

The disadvantage to their approach is that they are still using a version of the SPE phonological approach, which has been shown to have many drawbacks in this field (see Section 1.3.1). Also, there have been no experimental findings produced in terms of evaluative percentages with which their study can be compared.

6.8 Conclusions

The field of speech recognition is constantly improving. This should involve developments both in the areas of technology and in the theory of speech underlying it. Human speakers have mechanisms as yet not fully understood for coping with speech input.

> *Although there are many sources of variability in speech (such as talker identity, manner of speaking – loud, soft, fast slow, muttered, etc. -, linguistic context, acoustic environment), we are able to robustly perceive the message in the sequence intended by a speaker*. (Kakehi *et al*, 1996; 125)

It should be noted that human listeners have a huge exposure to spoken training data every day of their lives. Conversation surrounds us, and we have a huge neural network architecture inbuilt, with which to process it. For artificial neural networks with limited parameters to be achieving a measure of human success rate is significant in itself. Although it is true that human low-level phonetic recognition performance is actually not that good. There is a difference in the identification performance between words in high-predictable and low-predictable contexts.

The advancement of ASR is curbed slightly by a lack of consistency in phonetic transcription, even by experts experienced in analysing databases. The best pattern recognition performance on the TIMIT database is still only 70% segments correct (Fallside *et al*, 1991). Hence it can be seen that ASR systems don't *analyse* the data. Instead the knowledge components evaluate all possible

hypotheses (Huckvale, 1996). By combining engineering expertise with the right kind of linguistic knowledge and allowing performance to be the ultimate judge, the competence of recognition systems will continue to increase (Huckvale, 1996).

In these experiments, the differences in the performance and chance levels for each prime are attributable to several factors, including the variability of the acoustic cues and the frequency of occurrence of each prime in the training files. Overall however, the GP approach has been shown to be successfully capable of detecting and mapping phonological expressions from acoustic input.

The main conclusions derived from this paper are that:

- The theory of Government Phonology has applications to the field of ASR, and it is possible to perform effective speech recognition with a framework based on phonological expressions.
- Primes are acoustically detectable from the input signal and can be pattern-matched to templates providing frame by frame phone recognition.
- Headedness is also detectable and can be incorporated into the effectual identification of phonological elements.
- Gender is an important criterion in the development of this theory. It may be worth training gender-specific models.
- Dialect Regions are less predictable and do not seem to benefit from subdivision. Exposure to a variety of dialectal variations is the best approach.

The final lesson we should take from speech recognition systems is that we should not miss the opportunity to apply *all* the knowledge we have at any one time, both phonological and structural, to the decoding of a single utterance.

References

Anderson, John M. and C.J. Ewen, Principles of dependency phonology, Cambridge University Press, Cambridge, (1987)

Bird, S., *Introduction to Computational Phonology*, University of Edinburgh unpublished draft. (1995) [http://www.sil.org/computing/comp-morph-phon.html]

Bishop, C.M, Neural Networks for Pattern Recognition, Clarendon Press, Oxford (1995)

Brockhaus, W. and M. Ingleby, *A concurrent approach to the automatic extraction of subsegmental primes and phonological constituents from speech*, Proc. COLING-ACL '98, Montreal, (August 1998) 578-582.

Brockhaus, W., *Skeletal and Suprasegmental Structure within Government Phonology* in Durand Jacques and Francis Katamba (Eds.), Frontiers of Phonology: atoms, structures, derivations, Longman, London, (1995), pp 180-221

Brockhaus, W. and M. Ingleby, *Unary Primes and Constituent Structure in language*, (1997), unpublished research, University of Manchester, Personal Copy.

Carr, P., Phonology, Macmillan Press, Basingstoke, (1993)

Carson-Berdsen, J., Time Map Phonology Finite State Models and Event Logics in Speech Recognition, Kluwer, Dordrecht, (1998)

Charette, M., Conditions on Phonological Government, Cambridge University Press, Cambridge, (1991)

Chomsky, N and M. Hallé, The Sound Structure of English, Harper and Row, New York, (1968)

Crane, A., *Licensing Parameters: Restrictions on Phonological Expressions*, School of African And Oriental Studies, (1997), [http://www.praeclarus.demon.co.uk/ling]

Fallside, F, Lucke, H, Marsland, T.P, O'Shea, P.J, Owen, M.S.J, Prager, R.W, Robinson, A.J, Russell, N.H, "Continuous speech recognition for the TIMIT database using neural networks", Proceedings ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, (1991), Vol.1, pp.445-448

Frazier, L., *Structure in auditory word recognition* in Frauenfelder, U. H. and Tyler, L. K. (Eds.), Spoken Word Recognition, MIT Press, (1987)

Harris, J., English Sound Structure, Blackwell, Oxford, (1994)

Harris, J., *Segmental Complexity and Phonological Government*, Phonology 7, (1990), pp. 255-300,

Harris, J. and G. Lindsey, *The elements of phonological representation* in Durand, J. and F. Katamba (Eds.), Frontiers of Phonology: atoms, structures, derivations, Longman, London, (1995)

Hatch, E. and A. Lazaraton, The Research Manual: Design and Statistics for Applied Linguistics, Newbury House Publishers, New York, (1991)

Hübener, K and Carson-Berdsen, J. *Phoneme Recognition using Acoustic Events*, in Proceedings of the 3rd International Conference on Spoken Language Processing, Vol. 4, 1919-1922 (1994)

Huckvale, M., *Learning from the experience of building Automatic Speech Recognition Systems*, (1996) [http://www.phon.ucl.ac.uk/home/sh19/markh/huckvale.htm]

Jelinek, F., Statistical Methods for Speech Recognition, The MIT Press, Cambridge, (1995).

Kakehi, K., Kato, K. and M. Kashino, *Phoneme/ Syllable Perception and the Temporal Structure of Speech*, in Takashi, Otake and Anne Cutler (Eds.), Phonological Structure and Language Processing Cross Linguistic Studies, Mouton de Gruyter, Berlin, (1996)

Kaye, J. D., *The Acquisition of Phonology*, HAL Trust Llc, School of African and Oriental Studies, unpublished manuscript, (1997), The Government Phonology Archive [http://jk.soas.ac.uk/wp5.htm]

Kaye, J. D., *Why this Article is not about the Acquisition of Phonology*, School of African and Oriental Studies, unpublished manuscript, (1996), The Government Phonology Archive [http://jk.soas.ac.uk/wp5.htm]

Kaye, J. D., Lowenstamm, J and Vergnauld, J. R., *The internal structure of phonological representations: a theory of charm and government*, Phonology Yearbook 2, (1985), pp. 305-28,

Koreman, J., Bistra, A. and H. Strik, *Acoustic Parameters versus Phonetic Features in ASR*, Proc. The XIVth International Congress of Phonetic Sciences, (August 1999), San Francisco, pp719-722

Lahiri, A., *Speech Recognition With Phonological Features*, Proc. The XIVth International Congress of Phonetic Sciences, (August 1999), San Francisco, pp. 715-718

Magerman, D. M., Natural Language Parsing as Statistical Pattern Recognition, Ph.D. Thesis, Stanford University, (1994)

Maxwell, M., *Parsing using linearly ordered Phonological Rules,* in Computational Phonology: First Meeting of the ACL Special Interest Group in Computational Phonology, Proceedings of the Workshop (1 July 1994), pp 59-70.

Morgan, N. and Bourland, H., *Continuous Speech Recognition*, IEEE Signal Processing Magazine, Volume 12 3 (May 1995), pp. 24 - 42

Morgan, N. and Bourland, H., *Neural Networks for statistical recognition of continuous speech,* Proceedings of the IEEE, Volume 83 5 (May 1995), pp. 742-772

Olive, J. P, Greenwood, A. and J. Coleman, The Acoustics of American English Speech, A Dynamic Approach, Springer, New York, (1993)

O'Shaughnessy, D., Speech Communication, Human and Machine, Addison-Wesley Publishing Company, Massachusetts, (1987)

Owens, F.J., Signal Processing of Speech, Macmillan, London, (1993)

Polgárdi, K., *Constraint Ranking, Government Licensing and the Fate of the Final Empty Nuclei*, UCL Working Papers in Linguistics 8 (1996)

Rabiner, L. and B.-H. Juang, Fundamentals of Speech Recognition, Prentice Hall, (1993)

Reetz, H., *Converting Speech Signals to Phonological Features*, Proc. The XIVth International Congress of Phonetic Sciences, (August 1999), San Francisco

Rennison, J., *On Tridirectional Feature Systems for Vowels* in Durand Jacques, (Ed.), Dependency and Non-Linear Phonology, (1986), Croom Held Ltd., New Hampshire, pp 281-303

Ripley, B. D., Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge, (1996)

Robinson, A.J, Hochberg, M. and S. Renals, *The Use of Recurrent Neural Networks in Continuous Speech Recognition* [http://svr-www.eng.cam.ac.uk/~ajr/rnn4csr94/rnn4csr94.html] (1995),

Robinson, A. J., *Several Improvements to a Recurrent Error Propagation Network Phone Recognition System*, Technical Report TR82, Cambridge University Engineering Department, (1991)

Sarle, W.S., *Neural Network FAQ, part 1 of 7: Introduction, periodic posting to the Usenet newsgroup comp.ai.neural-nets*, (1997), URL: ftp://ftp.sas.com/pub/neural/FAQ.html

Scheer, T., *A unified model of proper government* (Government Phonology, vowel-zero alternations, French and Czech), Linguistic Review, Vol.15, N.1, (1998), pp.41-67

Shoup, J. E, *Phonological Aspects of Speech Recognition* in Lea, W. A. (Ed.), Trends in Speech Recognition, Prentice Hall, Eaglewood Cliffs, New Jersey, (1980)

Ström, N., *A Tonotopic Artificial Neural Network Architecture for Phoneme Probability Estimation*, Proc. of the 1997 IEEE Workshop on Speech Recognition and Understanding, (1997a), pp. 156-163

Ström, N., *Phoneme Probability Estimation with Dynamic Sparsely Connected Artificial Neural Networks*, The Free Speech Journal (1997b), Issue 5 [http://cslu.cse.ogi.edu/fsj]

Ström, N., *Sparse Connection and Pruning in Large Dynamic Artificial Neural Networks*, Proc. Eurospeech, Rhodes, Greece, (1997c), pp. 2807-2810.

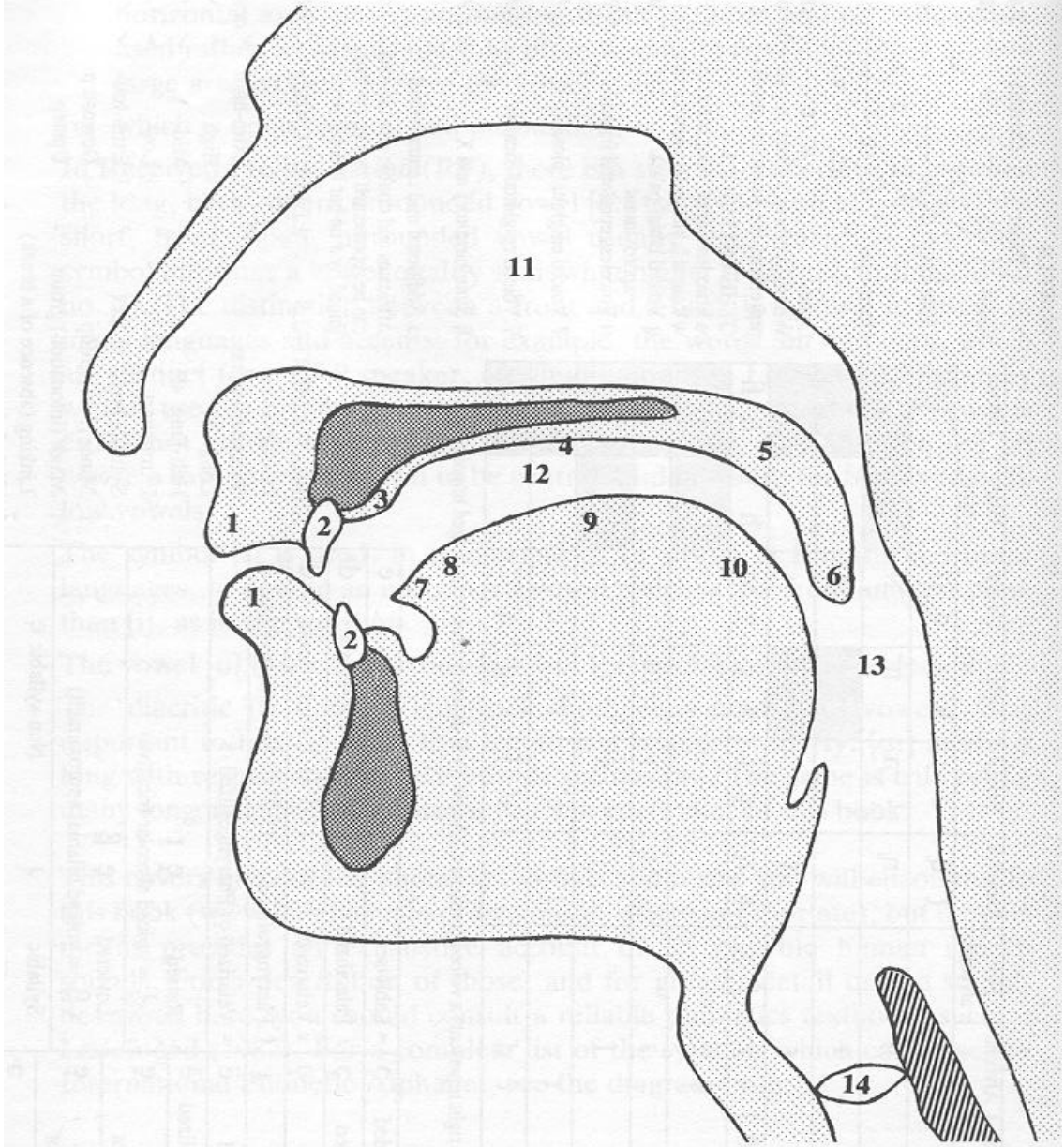
Ström, N., The Online NICO manual, [http://www.speech.kth.se/NICO/index.html], (1996)

Williams, G., *On the role of Phonological Parsing in Speech Recognition*, SOAS Working Papers in Linguistics, (1997)

Williams, G., *The phonological basis of speech recognition*, SOAS dissertations in linguistics, Thesis (Ph.D.), School of Oriental and African Studies, University of London, (1998)

Appendices

Appendix A



The Organs of Speech

1. Lips
2. Teeth
3. Alveolar ridge
4. Hard palate
5. Soft palate (velum)
6. Uvula
7. Tip of the tongue
8. Blade of the tongue
9. Front of the tongue
10. Back of the tongue
11. Nasal cavity
12. Oral cavity
13. Pharynx
14. Larynx

(From Carr, 1999; 12)

Appendix B

The International Phonetic Alphabet

CONSONANTS

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Phar-yngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			
Ejective stop	pʼ			tʼ		ʈʼ	cʼ	kʼ	qʼ		
Implosive	ɓ ɠ			ɗ ᶑ			ʄ ʃ	ɠ ʛ	ʠ ɢ		

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

(From Carr, 1993; 11)

Appendix C

This file contains a table of all the phonemic and phonetic symbols used in the TIMIT lexicon and in the phonetic transcriptions. These include the stress markers {1,2} found only in the lexicon and the following symbols which occur only in the transcriptions:

1) the closure intervals of stops which are distinguished from the stop release. The closure symbols for the stops b,d,g,p,t,k are bcl,dcl,gcl,pcl,tck,kcl, respectively. The closure portions of jh and ch, are dcl and tcl.

2) allophones that do not occur in the lexicon. The use of a given allophone may be dependent on the speaker, dialect, speaking rate, and phonemic context, among other factors. Since the use of these allophones is difficult to predict, they have not been used in the phonemic transcriptions in the lexicon.

- flap dx, such as in words "muddy" or "dirty" - nasal flap nx, as in "winner"
- glottal stop q, which may be an allophone of t, or may mark an initial vowel or a vowel-vowel boundary
- voiced-h hv, a voiced allophone of h, typically found intervocalically
- fronted-u ux, allophone of uw, typically found in alveolar context
- devoiced-schwa ax-h, very short, devoiced vowel, typically occurring for reduced vowels surrounded by voiceless consonants

3) other symbols include two types of silence; pau, marking a pause, and epi, denoting epenthetic silence which is often found between a fricative and a semivowel or nasal, as in "slow", and h#, used to mark the silence and/or non-speech events found at the beginning and end of the signal.

	SYMBOL	EXAMPLE WORD	TRANSCRIPTION
Stops:			
	b	bee	BCL B iy
	d	day	DCL D ey
	g	gay	GCL G ey
	p	pea	PCL P iy
	t	tea	TCL T iy
	k	key	KCL K iy
	dx	muddy, dirty	m ah DX iy, dcl d er DX iy
	q	bat	bcl b ae Q
	b	bee	
	d	day	
	g	gay	
	p	pea	
	t	tea	
	k	key	
	dx	dirty	
	q	bat	
Affricates:			
	jh	joke	DCL JH ow kcl k
	ch	choke	TCL CH ow kcl k
	jh	joke	
	ch	choke	
Fricatives:			
	s	sea	S iy
	sh	she	SH iy
	z	zone	Z ow n
	zh	azure	ae ZH er
	f	fin	F ih n
	th	thin	TH ih n
	v	van	V ae n
	dh	then	DH e n
	s	sea	
	sh	she	
	z	zone	
	zh	azure	
	f	fin	
	th	thin	
	v	van	
	dh	then	
Nasals:			
	m	mom	M aa M
	n	noon	N uw N
	ng	sing	s ih NG
	em	bottom	b aa tcl t EM
	en	button	b ah q EN
	eng	washington	w aa sh ENG tcl t ax n
	nx	winner	w ih NX axr
	m	mom	
	n	noon	
	ng	sing	
	em	bottom	
	en	button	
	eng	washington	

nx	winner	

Semivowels and Glides:

l	lay	L ey
r	ray	R ey
w	way	W ey
y	yacht	Y aa tcl t
hh	hay	HH ey
hv	ahead	ax HV eh dcl d
el	bottle	bcl b aa tcl t EL
l	lay	
r	ray	
w	way	
y	yacht	
hh	hay	
hv	ahead	
el	bottle	

Vowels:

iy	beet	bcl b IY tcl t
ih	bit	bcl b IH tcl t
eh	bet	bcl b EH tcl t
ey	bait	bcl b EY tcl t
ae	bat	bcl b AE tcl t
aa	bott	bcl b AA tcl t
aw	bout	bcl b AW tcl t
ay	bite	bcl b AY tcl t
ah	but	bcl b AH tcl t
ao	bought	bcl b AO tcl t
oy	boy	bcl b OY
ow	boat	bcl b OW tcl t
uh	book	bcl b UH kcl k
uw	boot	bcl b UW tcl t
ux	toot	tcl t UX tcl t
er	bird	bcl b ER dcl d
ax	about	AX bcl b aw tcl t
ix	debit	dcl d eh bcl b IX tcl t
axr	butter	bcl b ah dx AXR
iy	beet	
ih	bit	
eh	bet	
ey	bait	
ae	bat	
aa	bott	
aw	bout	
ay	bite	
ah	but	
ao	bought	
oy	boy	
ow	boat	
uh	book	
uw	boot	
ux	toot	
er	bird	
ax	about	
ix	debit	
axr	butter	
ax-h	suspect	

Others:

pau	pause
epi	epenthetic silence

```
h#      begin/end marker (non-speech events)
1       primary stress marker
2       secondary stress marker
```

Appendix D

prime	A	I	U	@	?	h	H	N	a	i	u
p	-	-	+	-	+	+	+	-	-	-	-
t	+	-	-	-	+	+	+	-	-	-	-
k	-	-	-	+	+	+	+	-	-	-	-
b	-	-	+	-	+	+	-	-	-	-	-
d	+	-	-	-	+	+	-	-	-	-	-
g	-	-	-	-	+	+	-	-	-	-	-
pcl	-	-	+	-	+	-	+	-	-	-	-
tcl	+	-	-	-	+	-	+	-	-	-	-
kcl	-	-	-	+	+	-	+	-	-	-	-
bcl	-	-	+	-	+	-	-	-	-	-	-
dcl	+	-	-	-	+	-	-	-	-	-	-
gcl	-	-	-	-	+	-	-	-	-	-	-
ch	-	+	-	-	+	-	+	-	-	-	-
jh	-	+	-	-	+	-	-	-	-	-	-
dx	+	-	-	+	+	-	-	-	-	-	-
q	-	-	-	-	+	-	-	-	-	-	-
m	-	-	+	-	+	-	-	+	-	-	-
n	+	-	-	-	+	-	-	+	-	-	-
ng	-	-	-	+	+	-	-	+	-	-	-
l	+	-	-	-	+	-	-	-	-	-	-
el	+	-	-	-	+	-	-	-	-	-	-
em	-	-	+	-	+	-	-	+	-	-	-
en	+	-	-	-	+	-	-	+	-	-	-
eng	-	-	-	+	+	-	-	+	-	-	-
nx	+	-	-	-	+	-	-	+	-	-	-
s	-	-	-	+	-	+	+	-	-	-	-
th	+	-	-	-	-	+	+	-	-	-	-
sh	-	+	-	-	-	+	+	-	-	-	-
f	-	-	+	-	-	+	+	-	-	-	-
z	-	-	-	+	-	+	-	-	-	-	-
dh	+	-	-	-	-	+	-	-	-	-	-
zh	-	+	-	-	-	+	-	-	-	-	-
v	-	-	+	-	-	+	-	-	-	-	-
r	+	-	+	+	-	-	-	-	-	-	-
w	-	-	+	-	-	-	-	-	-	-	-
y	-	+	-	-	-	-	-	-	-	-	-
hh	-	-	-	-	-	+	+	-	-	-	-
hv	-	-	-	-	-	+	-	-	-	-	-
iy	-	+	-	-	-	-	-	-	-	+	-
ih	-	+	-	-	-	-	-	-	-	-	-
ey	+	+	-	-	-	-	-	-	-	+	-
eh	+	+	-	-	-	-	-	-	-	-	-
aa	+	-	-	-	-	-	-	-	+	-	-
ae	+	-	-	-	-	-	-	-	-	-	-
ao	+	-	+	+	-	-	-	-	-	-	+
ah	-	-	+	+	-	-	-	-	-	-	-
ow	+	-	+	-	-	-	-	-	-	-	+
ax	-	-	-	+	-	-	-	-	-	-	-
uw	-	-	+	-	-	-	-	-	-	-	-
uh	-	-	+	+	-	-	-	-	-	-	-
ix	-	+	-	+	-	-	-	-	-	-	-
axr	+	-	-	+	-	-	-	-	-	-	-
ux	-	-	+	-	-	-	-	-	-	-	-
ax-h	-	-	+	+	-	+	-	-	-	-	-
er	+	-	-	+	-	-	-	-	-	-	-
aw	+	-	+	-	-	-	-	-	+	-	+
ay	+	+	-	-	-	-	-	-	+	+	-
oy	+	+	+	-	-	-	-	-	-	+	+
pau	-	-	-	-	-	-	-	-	-	-	-
epi	-	-	-	-	-	-	-	-	-	-	-
h#	-	-	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-

Appendix E

dr5/mgsh0/si1806.phn

0 2120 h#	2120 3960 sh	3960 5221 aw	5221 5973 l
5973 6725 w	6725 7720 ih	7720 9400 f	9400 9984 l
9984 10600 ih	10600 11800pcl	11800 12600 ax	12600 13827 kcl
13827 15475 k	15475 19467 oy	19467 20052 n	20052 21080 tcl
21080 22031 t	22031 22520 ix	22520 24480 s	24480 25720 iy
25720 26920 hh	26920 27346 w	27346 28200 ih	28200 28600 tcl
28600 29577 ch	29577 30120 ix	30120 31160 v	31160 32920 ah
32920 34200 s	34200 35000 gcl	35000 35499 g	35499 38040 ow
38040 39394 s	39394 40680 f	40680 44120 er	44120 45720 s
45720 46600 tcl	46600 47160 t	47160 50480 h#	

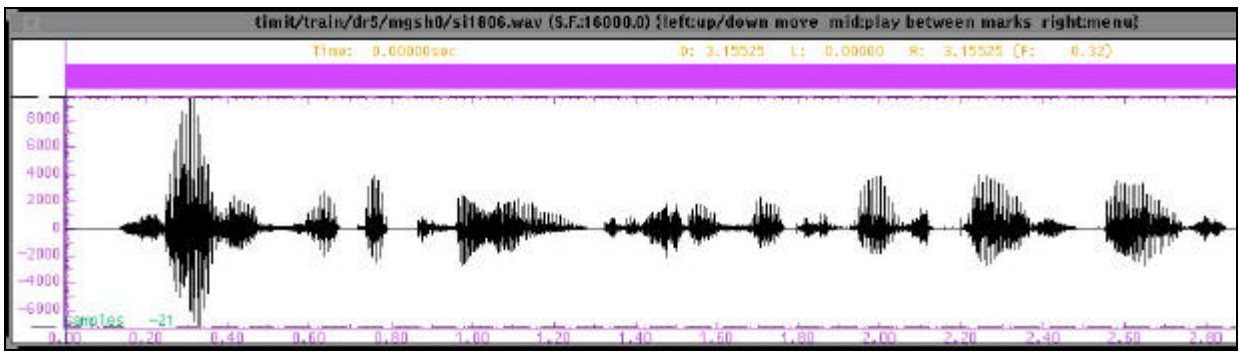
dr5/mgsh0/si1806.wrd

2120 5973 shall	5973 7720 we	7720 11800 flip
11800 12600 a	12600 20052 coin	20052 22520 to
22520 25720 see	25720 29577 which	29577 31160 of
31160 34200 us	34200 39394 goes	39394 47160 first

dr5/mgsh0/si1806.txt

Shall we flip a coin to see which of us goes first?

dr5/mgsh0/si1806 waveform (xwaves - Entropic Research Lab)



Appendix F Time labels

timit/train/dr6/majp0/si1704.phn

```
0 10400 h#          h#
10400 10969 y       I
10969 11860 axr     U @
11860 12600 ix      I @
12600 13911 tcl     h R ? H
13911 15148 t       h R ? H
15148 16226 ae      A I
16226 16790 kcl     h @ ? H
16790 17114 k       h @ ? H
17114 18178 s       h R H
18178 19210 pcl     h U ? H
19210 19820 p       h U ? H
19820 21932 ey      A I
21932 23960 er      @ R
23960 29480 pau     pau
29480 30704 hh      h
30704 32041 aw      A U
32041 33044 s       h R H
33044 34400 hh      h
34400 35300 ow      A U
35300 36121 l       R ?
36121 36720 dcl     h I R ?
36720 36840 d       h I R ?
36840 38080 axr     U @
38080 38527 epi     epi
38527 39700 l       R ?
39700 41907 ae      A I
41907 43414 n       N I
43414 44734 l       R ?
44734 45862 ao      A U @
45862 47160 r       R
47160 47710 dcl     h I R ?
47710 48360 d       h I R ?
48360 69520 h#      h#
```

Appendix G Section of Concatenated Results of dr5/msfh1/sx280

A	I	U	@	?	h	H	N
0.0102	0.0000	0.0024	0.0029	0.0027	0.2215	0.0016	0.0023
0.0124	0.0000	0.0029	0.0046	0.0049	0.3302	0.0015	0.0005
0.0162	0.0000	0.0013	0.0033	0.0030	**0.5587**	0.0016	0.0000
0.0165	0.0002	0.0016	0.0061	0.0086	**0.6381**	0.0076	0.0000
0.0152	0.0014	0.0108	0.0057	0.0415	**0.7026**	0.0380	0.0000
0.0268	0.0742	0.1107	0.0127	0.1261	**0.6115**	0.0326	0.0000
0.0343	**0.7557**	0.2582	0.0128	0.0782	0.0268	0.0011	0.0000
0.0253	**0.8783**	0.1475	0.0076	0.0362	0.0027	0.0001	0.0000
0.0464	**0.9064**	0.2114	0.0049	0.0514	0.0007	0.0002	0.0000
0.1211	**0.9112**	0.3242	0.0071	0.0207	0.0002	0.0004	0.0001
0.4011	**0.8830**	0.4826	0.0656	0.0058	0.0001	0.0003	0.0003
0.7202	**0.7421**	**0.7029**	0.2701	0.0033	0.0000	0.0002	0.0009
0.7951	0.4996	**0.7337**	0.4095	0.0066	0.0000	0.0001	0.0013

Appendix H Sample of Mapped Results from dr5/msfh1/sx280

A	I	U	@	?	h	H	N	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Silence
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	/h/
0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Silence
0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	/iy/ /ih/
0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	
0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	
0.0	1.0	**1.0**	0.0	0.0	0.0	0.0	0.0	/w/ /uw/
0.0	1.0	**1.0**	0.0	0.0	0.0	0.0	0.0	
1.0	0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	/ow/
1.0	0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	