Stochastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Care of Articulation in Spontaneous Speech

Matthew Peter Aylett

Thesis submitted for the degree of Doctor of Philosophy
University of Edinburgh
2000

Declaration

I hereby declare that I composed this thesis myself, and that the research which is reported herein has been conducted by myself unless otherwise indicated.

Matthew P. Aylett
Edinburgh
February 14, 2000

Acknowledgements

I've really enjoyed working on this PhD in the pleasant and helpful environment of HCRC and the Department of Linguistics. Well, more specifically, when I was stressed out, bored, or burnt out, the working environment was always supportive, pleasant and helpful.

People have often felt alone when doing a PhD. I have never felt like that here.

I won't give a long list of the people who have helped me over the years (you probably know who you are) as, like goodbyes, I prefer brevity to an outpouring of emotion (I'd like to save that for the pub if that's okay).

However I must especially thank Alice Turk who worked very hard correcting and advising me on my work. (Yeah I know it was her job but I mean VERY HARD as in especially hard). A special thanks also goes to Ellen Bard, not just for suggestions and help, but also for making sure my job did not conflict with my own research work.

Let's see where it goes from here.

Abstract

Within spontaneous speech there are wide variations in the articulation of the same word by the same speaker. Some words become extremely reduced while others seem to stand out more strongly in a phrase or sentence. This thesis explores these variations in articulation from two different but, arguably, related perspectives, prosodic structure and redundancy.

I argue that the constraint of producing robust communication while efficiently expending articulatory effort leads to:

1. An inverse relationship between language redundancy and care of articulation
2. The need for a strong 'checking' signal

The inverse relationship improves robustness by spreading the information more smoothly across the speech signal leading to a smoother signal redundancy profile. Checking in contrast leads to a more robust signal by ensuring that errors are detected and corrected.

I argue that smooth signal redundancy and a checking signal could be implemented by prosodic prominence and prosodic boundaries. Prosodic prominence increases care of articulation and appear to coincide with unpredictable sections of speech. In doing so prosodic prominence leads to a smoother signal redundancy. Prosodic boundaries cause syllabic lengthening and, by bounding self contained chunks of information (such as a word or phrase), signal that a listener should have a meaningful section of speech as well as offering a location for a listener to request clarification or re-transmission. In this way prosodic boundaries could be regarded as a checking signal.

The work presented here concentrates on the issue of smoothing redundancy. In order to explore this idea quantitatively, prosodic coding, metrics of language redundancy (word frequency, syllabic trigrams and givenness) and of care of articulation (normalised syllabic duration and vowel quality) are formulated and applied to a large corpus of English spontaneous task-oriented dialogue.

Results confirm the strong relationship between prosodic structure and care of articulation as well as an inverse relationship between language redundancy and care of articulation. In addition, when an opportunity for a checking signal is controlled for, in some circumstances language redundancy can predict up to 65% of the variance in raw syllabic duration. This is comparable with 64% predicted by prosodic structure. Moreover most (62%) of this predictive power is shared.

This leads to the conclusion that, within English, prosodic structure is the means with which constraints caused by requiring a robust signal are expressed in spontaneous speech. Finally it is argued that, if redundancy is indeed a driving force behind prosodic structure, notions of redundancy and predictability should be more formally included into prosodic theory.

Table of Contents

Chapter 1

Introduction

We often don't say the same word the same way in different situations. If we read a list of words out loud we say them differently from when we produce them, spontaneously, in a conversation. Even within spontaneous speech there are wide differences in the articulation of the same word by the same speaker. Some words become extremely reduced while others get longer and louder and seem to stand out more strongly in a phrase or sentence. This thesis explores these variations in articulation from two different but arguably related perspectives, prosodic structure and redundancy.

1.0.1 Prosodic structure

Phoneticians and phonologists have studied 'suprasegmental' effects, variation that appears to occur at the phrase or word level, for many years and proposed various theories of prosodic structure to account for them. They have shown that these variations are not random but often extremely systematic. In general, theories of prosodic structure concentrate on three distinct though clearly related phenomena:

1. **Prominence:** Some parts of the speech stream stand out more than other parts.
2. **Boundaries:** Speech is split up into chunks which are marked by suprasegmental phenomena. (For example pauses, differences in tone, amplitude, segmental duration and prominence.)
3. **Information Giving:** Changes in prosodic structure can alter the meaning of the message. (For example altering the topic of a statement by changing the prominence of certain words.)

Looking closely at the way prominence is realised in spoken language laboratory phonetics has found that prominent syllables are more clearly articulated (e.g. van Bergem, 1988). That is, the segments tend to be longer, the spectral characteristics are more distinct, they are louder and often marked with pitch change. Words with such prominence also tend to be easier for human subjects to recognise when excerpted from context.

In general:

prominence = more care of articulation = more noticeable = easier to recognise

1.0.2 Redundancy

Prosodic structure clearly affects care of articulation; however another factor, redundancy, also appears to have a major impact (Lieberman, 1963; Hunnicut, 1985; Wright, 1997, amongst others). More common words and words you can easily predict from context (more redundant) tend to be articulated less clearly. For example the 'nine' in the phrase 'a stitch in time saves nine' is less clearly articulated than the nine in 'the number you will hear is nine'.

Lindblom (Lindblom, 1990) in his H&H theory suggests that we put only as much effort into articulation as required for the listener to understand. He argues that we tend to under-articulate predictable (redundant) sections of speech and over-articulate difficult to predict (less redundant) sections of speech.

This change in articulation can be manifested both as an overall postural setting where the speech style becomes more careful overall and also locally where individual words and speech sounds are more carefully produced. There is substantial evidence that the phonetic effects we see in speech which are carefully articulated as a whole are similar to the phonetic effects we see within a speech style when an individual section of speech is carefully articulated. It is these local changes which appear to reflect differences in redundancy.

1.0.3 Motivation and Hypotheses

So we appear to have two quite different factors controlling the care with which we articulate speech. On one hand we have a complex prosodic structure which allows prominence and the chunking of speech and on the other we have complex interactions within the structure of language which makes some sections of speech predictable and others less so.

Unfortunately very little work has considered both these factors when examining care of articulation. A major criticism levelled at the Lieberman (1963) work is that prosody was not controlled for. This general lack of any prosodic control persists in much of the work reporting a redundancy effect (see chapter 4). Similarly work that has considered the impact of prosodic structure on care of articulation has not taken even basic redundancy effects such as word frequency into account.

This thesis will try to disentangle these factors. It explores the relationship between theories of prosodic structure, care of articulation and measurements of redundancy in a corpus of spontaneous spoken language. In doing so it aims to unite traditional phonological views of language structure with a stochastic, data driven approach to language analysis.

I will argue that a relationship between redundancy and care of articulation is desirable in speech because it leads to more robust communication. I will present strong evidence that much of the effect of redundancy is implicitly represented in prosodic structure (see chapter 6). This leads to the conclusion that prosodic structure is the means with which redundancy effects are implemented linguistically within language (see chapter 2). In turn this suggests that redundancy can be thought of as a *reason* why much prosodic structure is as it is within English. I will finally speculate on the extent this may also be true cross-linguistically.

Understanding these variations in articulation is of great importance for both engineers who wish to design effective speech recognition and synthesis software and also psycholinguists and phoneticians who wish to understand the human language system. Potentially such an investigation can help refine theories of suprasegmentals and allow us to not only predict articulation variation in the speech stream but use this variation to explore the internal state of a speaker's language system.

The central questions this thesis will address are:

1. Can we build an effective model of care of articulation that allows a quantitative analysis of large quantities of spontaneous speech? What are the problems and limitations of such a model?

2. To what extent does a modern theory of prosodic structure account for such changes in the care of articulation in contrast to some simple measures of redundancy?

3. How much interdependency exists between redundancy measurements and prosodic structure? Can concepts of predictability and prosodic structure

be integrated together to offer a stronger predictive framework of changes in care of articulation.

1.1 Brief outline of Methodology: A Corpus Approach

Most studies of prosodic structure and care of articulation have been carried out on carefully controlled read laboratory speech (for example van Bergem, 1988; Moon and Lindblom, 1994; de Jong, 1995). Such an approach allows the careful construction of the data set that a study wishes to explore so that any particular language feature can be carefully controlled for. In doing so the amount of material that needs to be analysed to address a particular question is kept to a minimum. Coding and measuring speech data by hand is a time consuming business. The traditional laboratory approach is able to minimise time spent coding and analysing while maximising the factors that can be studied so that cleverly selected materials can expose interdependencies between factors. While this approach has been extremely successful in speech research there is also a need for work based on more natural speech.

It has been shown that patterns in care of articulation vary significantly across speech styles. Read speech, although similar in many ways to spontaneous connected speech, is generally more carefully articulated (Fowler, 1988). Prosodic structure also differs from that in spontaneous dialogue (Silverman *et al.*, 1992). This means that you cannot necessarily generalise results across speech styles. Therefore, in order to address the main questions of this thesis we need to examine spontaneous speech. In turn, because spontaneous speech cannot be so carefully controlled, to cover the many different prosodic and redundancy contexts a lot of spontaneous speech is required. The more speech we have to consider the more impractical hand coding and hand measurement becomes and the more we need to rely on automatic methods. This in turn introduces noise which means yet more material is required.

This work is based on a large corpus of spontaneous task oriented dialogue collected by the HCRC at the University of Edinburgh - the HCRC Map Corpus (Anderson *et al.*, 1991). The corpus, comprising of about 15 hours of spontaneous speech, 64 speakers and around 200,000 syllables, gives sufficient scope for some hand coding as well as offering a very large data set with which to apply automatic methods.

In order to explore the relationships between care of articulation, prosodic structure and redundancy using quantitative techniques in this material it was necessary both to define more clearly what these terms mean theoretically and, to some extent, limit the scope of these terms to produce an operational metric.

Chapters 2,3,5 go into detail concerning the measurement and coding strategies of these factors and the thinking behind them. A summary is as follows:

1.1.1 Redundancy

A trigram measurement over syllables, word frequency and givenness are used as redundancy measurements. Chapter 2 goes into some detail concerning the issues in arriving at and using redundancy measurements.

1.1.2 Prosodic Structure

Chapter 3 discussed problems in applying prosodic coding to speech material, gives an overview of the theoretical background behind the coding used and goes into detail concerning the methodology of applying this coding to a large corpus of spontaneous speech.

1.1.3 Care of Articulation

A very large number of factors can be used to examine care of articulation. In this study vowel spectral clarity and syllable duration were used as operational measurements. (Chapter 5)

1.2 Structure of the Thesis

Chapter 2 introduces the concept of redundancy and addresses the question of why redundancy might be linked to care of articulation. Chapter 3 reviews literature in the areas of prosody and presents the coding system used to represent prosodic structure in this work. Chapter 4 reviews work that has looked at care of articulation in terms of prosodic structure and work which has looked at care of articulation in terms of redundancy. This chapter also goes into some depth concerning the acoustic factors which are connected with carefully articulated speech. Chapter 5 describes the method used in this thesis for measuring care of articulation in terms of syllabic duration and the spectral quality of vowels.

Chapter 6 presents results from the analysis of these materials looking at the interrelationships between these measurements. Finally Chapter 7 discusses the implications of these results, possible future work as well as some of the limitations in the approach used here.

Chapter 2

Redundancy

2.1 Introduction

This chapter aims firstly to give a brief introduction to the concept of redundancy and secondly to explore the reasons why redundancy might relate to care of articulation and prosodic structure. The aim here is to give the reader the necessary background for understanding the application of statistical techniques for measuring redundancy as used in this thesis. This chapter does not attempt to present a detailed appraisal of research in statistical language processing for the following reasons:

1. To a large extent the statistical techniques used to measure redundancy in this work are ’off the shelf’ and are relatively simple and uncontroversial.

2. Excellent textbook introductions to using statistical techniques in the study of natural language (e.g. Charniak, 1993) and to approaches in corpus linguistics (e.g. McEnery and Wilson, 1996) already exist.

The ideas that I will discuss in this chapter are fundamental to the approach of my work. They form the basis of why I believe care of articulation is related to redundancy as well as to prosodic structure. In order to explore these ideas it is crucial that the terms used in my argument are clearly defined and explained. In the first part of this chapter I will present these basic ideas. First I will discuss the concept of redundancy in language and in the acoustics of language. I will then consider how these notions relate to a noisy channel model of communication and give a definition of the three different types of redundancy considered here, language redundancy, acoustic redundancy and signal redundancy.

I will then consider why prosodic structure, redundancy and care of articulation should be inextricably linked given this framework. This in turn will lead to a number of testable hypotheses which I will return to in chapter 6. In the final part of the chapter I will describe in detail the method I use to represent and measure redundancy throughout this work.

2.2 Historical Background

In 1948 Shannon (Shannon, 1948) published a mathematical theory of communication. Although strongly mathematical, his approach was also very general. By expressing information in terms of choice or uncertainty it was possible to formally measure information in terms of *bits* (the number of 1s or 0s required to represent the information). In this way information theory can define how many bits of information can be sent per second over perfect and imperfect channels and it can specify how such information can be encoded efficiently. Parallels between Shannon's analysis of electrical communication and human communication were quickly drawn (Miller and Frick, 1949). Other work has varied from mathematical observations such as Zipf (1949), who noted that the number of occurrences of a word in a long text is the reciprocal of the order of frequency of occurrence, to specific experiments in psychology such as McGill (1954) which attempted to relate differences in entropy with a subject's response to stimuli.

For a broad non-mathematical introduction to the concepts within information theory and an overview of early psychology work related to information theory see Pierce (1961). What follows here is a non-technical explanation of how some of the concepts within information theory (in particular redundancy and the noisy channel) can be related to speech and how such a perspective forms the basis of the hypotheses examined in this work.

2.3 Redundancy

Redundancy means how predictable an observation is given its context. The more predictable the easier it is to guess and the more redundant the information. For example Lieberman (1963) used different contexts to produce high and low redundant words. One much quoted example is: "A stitch in time saves ..." and "The number you will hear is ..." to elicit redundant and non-redundant tokens of the word *nine* (for a detailed examination of this and other laboratory work relating redundancy to care of articulation see chapter 4).

In order to formalise this notion of redundancy one could generate a numerical probability that *nine* is the last word in these two sentences. This is non trivial because of the many different factors that govern natural language. Without being able to model all these factors it is not possible to generate the true numeric probability of guessing the word. What Lieberman did was instead use the response of human subjects to calculate probabilities. He asked 60 subjects to guess the word and the number that were correct out of the total number was used as the probability of predicting *nine* given these different contexts. However such an approach is infeasible when dealing with very large data sets. In this case, in order to produce a formal numerical probability of a word occurring it is necessary to build a statistical model which can generate these probabilities.

All such formal measurements require a model and in **all** cases, when a formal redundancy measurement is made, it is with regards to a model.

For example imagine throwing two dice. What is the most redundant result of adding the two numbers produced? The answer is 7. This is because of the thirty-six different possible outcomes six add up to 7 (1/6, 2/5, 3/4, 4/3, 5/2, 6/1) meaning the chance of the dice reading seven is about 16.67% whereas the chance of it adding up to 12 (with only one outcome 6/6) is only about 2.78%. Where is the model? The model is built on the assumption that each number on each dice has an equal chance of appearing.

A model of this nature can be built from two perspectives:

1. We can argue that it is a good model of two dice because we believe there is no more chance of one side appearing than any others when the dice is thrown normally. This is a theoretically led model.

2. We can roll the dice and observe what happens. Then we can collect the observations and build a probabilistic model from them. This is an observationally led model.

In practise most models are a combination of both approaches. A theoretical approach is first taken to build a prototype model which is then tested and adapted with regard to observations.

In speech, where such observations are the acoustic signal, we may wish to separate the acoustical observations from an underlying language model. For example when looking at speech we might choose to separate the signal into words. To do this we need to connect the acoustic observations with a particular word. In speech technology this is often carried out using a statistical acoustic model which produces a set of probabilities of different words occurring given the signal and a language model which given these different words calculates the most likely sequence of words. Splitting the models up like this has a profound effect on the meaning of redundancy in natural language.

2.3.1 Acoustic Models versus Language Models

In the dice example we know what the outcome of each dice throw is. Let's imagine that instead the person who rolls the two dice shouts out the sum that is produced. We then have a set of acoustic observations which are connected to an event. We can take these observations and build an acoustic model which connects them to each word that is spoken. If for example you observe high amplitude fricative noise which is mostly above 4Khz this indicates an /s/ has probably been produced by the speaker. Given an 's' it is unlikely the dice roller has rolled two, three, etc., and more likely they have rolled six or seven. If in contrast you observe lower amplitude broader spectrum fricative noise this indicates a 'f' or a 'th' has probably been produced by the speaker. Given this then a three, four or five is more likely to have been rolled.

We now have two statistical models. One represents the likelihood of a particular number appearing on the dice. This is, in effect, our language model because it models the likelihood of different words appearing. The other, the acoustic model, connects acoustic observations with these words. We can use the combination of both models to make the best guess of what number was rolled and what the dice roller said given the acoustic observations.

We can also calculate the redundancy (or predictability) of events and observations occurring with regards to these models. The result is a number of 'levels' of redundancy. We have the redundancy of the event 'seven' occurring but also of the sound 's' being produced given acoustic observations. We can combine these measures of redundancy in the same way as we can combine the statistical models to produce the final signal redundancy.

There is no limit on how much we might want to divide these two models further. In statistical natural language processing it is possible to build different models for the different features of language. For example redundancy in the sentence "I'm going to the beach" can be calculated with regards to a probabilistic model of making that statement given some situation (e.g. it's a sunny day), with regards to the syntax (e.g. more likely than "going I beach"), with regards to the lexicon (e.g. "beach" is a more common word than "zanja"[1]), with regards to the sounds (e.g. 'b' is a more common sound than 'ch'), and with regards to the acoustic observations (e.g. we wouldn't expect a fundamental frequency over 250Hz).

The fact we can calculate the redundancy of an event given a model does not of course mean the model is a good one or that the redundancy value it produces reflects the underlying system that produced the event. This is especially true in natural language where there is a great deal of dependency from one event to the next. Unlike the dice example, where our model regards each dice throw as independent, in language each word we produce, each sound, each message is very much dependent on what has gone before and what is expected to come after.

As I will explain in the next section, variation in redundancy at the level of the language model (language redundancy) has some important implications with regards to communicating in a noisy environment. These implications can help explain why prosody might be used for *checking* and why language redundant sections of speech might be attenuated by prosody.

2.3.2 Noisy Channel

Introducing noise into the signal has important considerations on the need for redundancy both in terms of smoothing redundancy over the whole signal (the signal redundancy) and in terms of introducing checks which are built into the signal (see below). To clarify what is meant by a noisy channel imagine a crowd is watching the dice roller roll his dice and they are shouting random encouragement. The noise they are making will degrade the acoustic observations and thus make the chances of guessing the correct number from the acoustic observations worse.

The effect this will have on different sounds is different. 'f' is normally quieter than 's' so this random noise is more likely to make 'f' indistinguishable from some other sound than 's'. Computing the redundancy of each event for the observer is now a combination of:

1. The likelihood of the event (the language model).
2. The likelihood of the acoustic observations representing this event being degraded by the random noise (the noise model).

[1] An irrigating canal according to Chambers English Dictionary

3. The likelihood of these degraded acoustic observations being associated with the event (the acoustic model).

A certain amount of redundancy in a noisy channel environment is a good thing. This is because it offers protection to loss of information. It is also good for such redundancy to be smooth in the signal so that the signal will degrade gracefully. Graceful degradation can be thought of as the relationship between loss of data in the message and loss of information carried by the data. An example of poor degradation is the loss of one binary instruction in a computer program. If one instruction is lost the entire program could well fail. An example of more graceful degradation would be the loss of a few random characters from a text file. The text file would probably still contain most of the useful information. A smooth signal redundancy profile can be regarded as not putting all your eggs in one basket; by distributing the information evenly a critical error is less likely to occur (see Pierce, 1961, chapter 8).

However an alternative approach to dealing with a noisy environment (and in some cases a more efficient one) is to build checks into the communication (also see Pierce, 1961, chapter 8). Rather than have a passive receiver which may fail to correctly decode the message the receiver and the transmitter have a built in structure of checks. Typically the transmitter sends a chunk of message and the receiver responds with an 'okay I received this' message. If the message is not received correctly then it is resent. Using checks complicates redundancy. We now not only need to send the message but also the checks. We therefore need to add a model representing these checks to the system. The structure of these checks needs to be predictable so that the checks themselves are unlikely to be missed in the noisy environment.

Both smoothing signal redundancy and checking could be associated with prosodic structure. The first because prominence, by making speech more distinct, affects its acoustic redundancy, the second because prosodic boundaries, by affecting the duration of speech, could act as a checking signal at the end of each prosodic constituent.

2.3.3 Dealing with Checking

The same arguments I use to justify the need for smooth signal redundancy can be used to justify the existence of a checking signal. However, although I go on to present evidence for the smoothing of signal redundancy I do not advance any model of checking or present any evidence of the existence of checking signals. As discussed briefly in section 2.6.2.3 and in more detail in section 7.3 such a model would require significant research in itself. The approach taken in this work is instead to accept that checking may occur and include this *possibility* in the hypotheses advanced in section 2.5. By controlling for the sites of a possible checking signal it is possible to address the central issue of this work, smooth signal redundancy, without requiring a checking model.

Despite this pragmatic approach, checking is still an important part of the framework used in this thesis. For this reason, the way a checking model may integrate with prosodic structure, smooth signal redundancy and care of articulation will be discussed at a theoretical level (see section 2.5).

2.3.4 Three Different Types of Redundancy: How Can Prosodic Structure, by Controlling Care of Articulation, Smooth Redundancy?

As discussed earlier redundancy only has a meaning with regards to a model. In language we can build different models for different levels of structure. The two models I have mentioned are the language model which is the likelihood of a word, syllable or phoneme appearing in the speech stream and the acoustic model which is the likelihood of specific acoustic observations being connected with a word, syllable or phoneme. For example, the likelihood of 'to' following 'going' might be included in a language model. In contrast the likelihood of the word 'to' being associated with 100ms of sound with most of the vocalic energy between 0 to 2500Hz with peaks at 310Hz, 870Hz and 2250Hz (typical formant values of the vowel /u/) might be included in an acoustic model.

The combination of these two models produces the final or *signal* redundancy in the speech stream. So we have three different types of redundancy. In order to avoid confusion let's look more closely at what I mean by these types of redundancy which I have termed *language redundancy*, *acoustic redundancy* and *signal redundancy*.

- **Language Redundancy:** This is the conventional use of the term redundancy which is used in work such as Lieberman (1963). It refers to how predictable a word, syllable or phoneme is given its context. All references to redundancy in this work, except where specifically noted, are to this conventional meaning. All the metrics I present later in this chapter are trying to measure this type of redundancy.

- **Acoustic Redundancy:** This is a less common use of the term redundancy. As I discussed earlier a word is expressed in the acoustic signal as a set of acoustic observations. Using these observations we can guess what the word may be. The easier it is to guess the word the more redundant the acoustic observations are. To a large extent our acoustic model is similar to a speech recognition model which ignores any other factors except the acoustic signal. It does not make use of any predictabilities in the structure of language, it simply looks at a set of signals and guesses what word, syllable or phoneme they represent. This idea, that the acoustic signal is analysed with regards to a probabilistic model is central to almost all modern speech recognition technology. In the work presented here I do not present such a model or deal with the implications of any such model other than in the broad sense of saliency and discriminability. The more salient and the more discriminable the less likely noise will degrade the signal and the easier it is to guess the word, syllable or phoneme from the acoustic observations. The easier it is to guess the identity of the language unit from such observations the more redundant these observations are. By looking at acoustics in this way saliency equates to acoustic redundancy.

- **Signal Redundancy:** Signal redundancy is the final redundancy in the signal which is a combination of the language model and the acoustic model. This is the final redundancy that any recognition system faces which knows something about the structure within language as well as the structure in the acoustics of speech. Because signal redundancy is the combination of these two previous models and because it is good for signal redundancy to be smooth to combat noise this leads to my central hypothesis. *For signal redundancy to tend to smoothness requires that sections of speech which are very language redundant will tend to be sections of speech which are less acoustically redundant and thereby less salient and distinctive.* The converse will also tend to be true. This is illustrated in figure 2.1. The graph shows the language redundancy, acoustic redundancy and combined signal redundancy of the phrase “okay, starting off we’re above a caravan park”. The least language redundant syllables “star” in “starting” and “park” also tend to be more acoustically redundant. By combining these values the standard deviation reduces from 1 to 0.65 suggesting a smoother less varying signal. In this way care of articulation can smooth signal redundancy and prosodic structure (by controlling care of articulation) can contribute to a robust noise resistant signal. The extent to which this occurs is an open research question that this thesis seeks to address as there is also evidence that *checking* and psycholinguistic constraints could undermine such a relationship (see section 2.5).

2.4 A Theoretical Relationship Between Redundancy and Care of Articulation

The need for a smooth redundancy pattern when transmitting in a noisy environment is directly at odds with the complex compositional structure of natural language. To start with the frequency of different words varies leading to concentrations of high and low redundancy. For example the word 'the' is very high frequency whereas 'zanja' is not. Parts of words vary enormously in how predictable they are. In general the second syllable of a two syllable word is a lot more predictable than the first syllable when you know the identity of the syllable that precedes it. Complex syntactic structure means that many words are predictable simply in order to produce grammatical sentences. There is indeed enormous redundancy in language but it is concentrated in certain areas of the message.

However, in general, within spoken language (ignoring visual cues) acoustic observations are the only clue to the contents of the message. The final redundancy of the message is the combination of the models representing the linguistic events (the language model) **and** the acoustic model (the model which maps parametric acoustic observations onto these linguistic events). Speakers may not be able to alter the redundancy of the message to make it smooth at the level of the lexicon and syntax but they can alter the acoustic signals produced and thus the final redundancy of the signal.

If, in the dice rolling example, the speaker didn't want to lose their voice they might only shout the less predictable dice results. By making the acoustic observations for 'seven' less distinct and for 'twelve' more distinct the final signal redundancy of these messages changes. This is because the final signal redundancy is a combination of the language redundancy (in the dice example 'seven' is more frequent and thus more redundant than 'twelve') and acoustic redundancy (the more distinctly articulated the more acoustically redundant the speech).

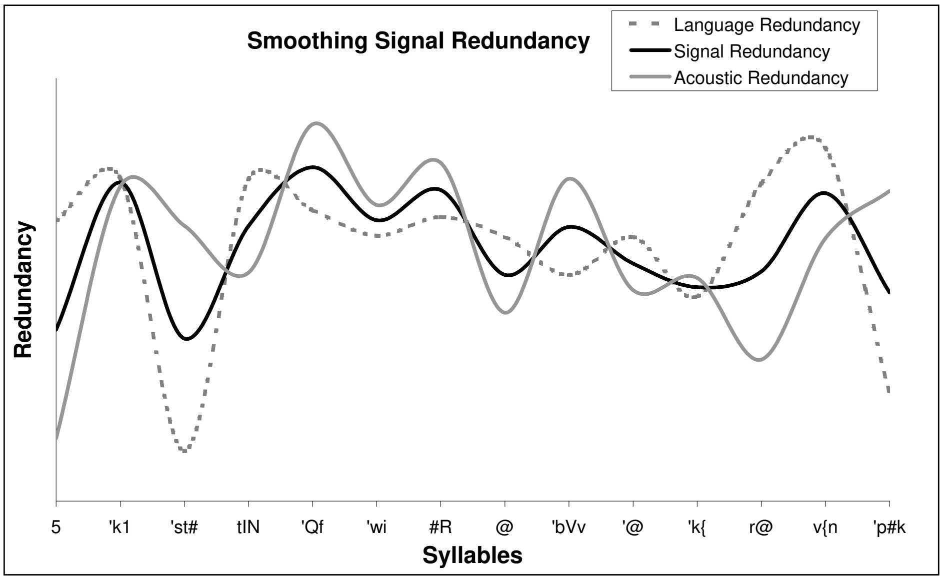


Figure 2.1: **Smoothing signal redundancy:** The graph shows the language redundancy, acoustic redundancy and combined signal redundancy of the phrase "okay, starting off we're above a caravan park". The x axis lists each syllable in CELEX DISC format (see appendix A) and the y axis shows the change in redundancy. No scale is used for redundancy because all language and acoustic redundancy measurements were normalised. The language redundancy was calculated on the basis of the trigram syllabic model described in section 2.6.2. The acoustic redundancy is, more controversially, calculated by normalising the *normalised duration measurement* (the k score described in chapter 5). The acoustic redundancy used here is purely for demonstrative purposes as this work does not offer an acoustic model on which to calculate it properly. However as we can see, the least language redundant syllables "star" in "starting" and "park" also tend to be more acoustically redundant (in this case longer). By combining these values the standard deviation reduces from 1 to 0.65 suggesting a smoother, less varying signal.

Altering the care of articulation would then be a direct means of making speech a more efficient means of communication. This is because it makes better use of articulatory effort. By over-articulating unpredictable sections of speech and under-articulating predictable sections of speech the same overall effort leads to a smoother signal redundancy profile which in turn makes speech more robust in a noisy environment. (In chapter 4 we look, in detail, at laboratory results which have shown that care of articulation is indeed reduced in many redundant contexts.)

However the more checks we use in communication the less important smoothing the redundancy in the signal becomes. We are left with an open question as to the extent (or even if) care of articulation is indeed used to offset language model redundancy and to what extent checks make this unnecessary. Secondly, even if care of articulation does relate to language redundancy, this does not mean that we are using language redundancy information directly when controlling care of articulation. It is possible that prosodic structure, both as it is represented implicitly in the lexicon and as it is realised in speech, may offer a linguistic system to effect these changes.

2.4.1 Is Prosody Related to Redundancy?

There are a number of observations which suggest that prosody is related to redundancy, and therefore, that prosody, both at a lexical and phrase level, may be a linguistic means of smoothing signal redundancy.

Lexical redundancy is caused by the different internal structure and frequencies of words. Prosodic structure at the lexical level appears related to these patterns.

1. Most open class words have metrically strong first syllables. It is the first syllable which is the least language redundant.

2. Open class words are, in general, less frequent than closed class (or function) words). Closed class words are often realised without lexical stress. Again realisation of lexical stress appears to mirror predictability at the lexical level.

3. Long words are spoken relatively more quickly than short words. Long words have more redundant information in them. This is because often with long words, once you hear the beginning part of the word the rest of

the word is very predictable. For example it is easier to guess the rest of the word 'televi..' as '..sion' than the rest of the word 'd..' as '..oor'.

We see a similar pattern at the phrase level with more informative and less redundant parts of a phrase being accented while less informative and more redundant parts of a phrase are de-accented.

"It is well known that accents tend *not* to be placed on elements that are repeated or 'given' in the discourse, or on elements that are vague or generic. For adherents of the radical FTA (focus-to-accent) view, this fact is a clear illustration of the general principles governing accentuation in any context: the speaker assesses the relative semantic weight or informativeness of potentially accentable words and puts the accent on the most informative point or points in a sentence." (Ladd, 1996, p175).

However these prosody/redundancy relationships are far from simple at either the lexical and phrase level. Firstly many words do not have stressed initial syllables. This suggests that even if a direct redundancy/prosody relationship exists it is a tendency rather than a rule. Secondly, as Ladd (1996) points out, there are cases at the phrase level when a simple accent/informative relationship does not occur as well as many examples of other languages where such a relationship appears to be absent.

Lengthening at the end of phrases (e.g. Price *et al.*, 1991) also appears to undermine any simple relationship between redundancy and prosodic structure. The ends of phrases are generally more predictable from context and thus more language redundant than the beginning phrases. In general prosody appears to attenuate redundant sections of speech yet here we have areas of speech which are in fact more redundant and prosodic structure seems to be making them longer[2]. The extent such boundary effects can be attributable to the checking described in section 2.3.2 remains an open question.

I will now consider how we can test the ideas discussed here more formally.

2.5 Hypotheses

The argument linking language redundancy, prosodic structure and care of articulation can be summarised as follows:

[2]Some care must be exercised when describing saliency or care of articulation purely in terms of lengthening. In chapter 4 this question is addressed in detail.

- Speech is an example of transmission over a noisy channel. In order to be efficient and robust the final redundancy in the signal (the signal redundancy) needs to be as smooth as possible.

- Care of articulation modifies the acoustic signal in terms of distinctiveness and saliency. By doing so care of articulation modifies redundancy in terms of an acoustic model (the acoustic redundancy).

- Redundancy in terms of the language model (the language redundancy) is far from smooth because of the constraints of semantic, syntactic and lexical compositionality.

- Signal redundancy is the combination of language redundancy and acoustic redundancy. To make signal redundancy smooth, acoustic redundancy compensates for extreme variation in language redundancy. Assuming there is a limit on the overall articulatory effort which can be expended the result is a tendency to poorly articulate language redundant sections of speech and to carefully articulate non language redundant sections of speech.

- There is a lot of evidence that prosodic structure not only affects care of articulation but does so in a way which seems associated with patterns in language redundancy.

- This leads to the hypothesis that, in order to achieve smooth signal redundancy linguistically, prosodic structure, as one of its functions, implicitly encodes much language redundancy variation both lexically and post lexically in terms of lexical stress, accent and boundary lengthening.

Put more bluntly:

The Smooth Signal Redundancy Hypothesis

Prosodic structure smoothes signal redundancy by controlling care of articulation.

Two arguments can be made to support the idea that prosodic structure would be a good means of encoding an inverse relationship between language redundancy and care of articulation.

1. Computing language redundancy is non trivial. Calculating the overall redundancy of a section of speech on the basis of lexical, syntactic, semantic and pragmatic factors is hard. In addition many of these statistics remain

independent of each other. It would seem sensible to encode such statistics into a simpler linguistic form especially at a lexical level. By using prosody at the lexical level the effects of word frequency and structure on redundancy can be encoded in terms of lexical stress and syllabic structure. In turn, effects caused by structure at the phrase level can be modelled using prosodic structure at that level, such as adding phrasal stress to semantically unpredictable open class words. The overall result would be to approximate the highly complex statistical patterns in language into simpler, prosodic building blocks.

2. Results from psycholinguistic experiments (see chapter 3) suggest that prosodic structure has psychological validity. By this I mean that naive human subjects can detect prosodic structure such as number of syllables, phrase boundaries and different levels of prominence. The extent that human subjects are directly aware of redundancy patterns in language is less clear.

However the claim that prosodic structure encodes language redundancy requires some qualification. There is considerable evidence that prosodic structure is also used as a form of chunking and *checking*. As Nooteboom points out "These (prosodic) cues... organize the message into chunks that are easily processed by the listener..." (p668 Nooteboom, 1997). There are two factors which need to be considered here:

1. Psycholinguistic processing factors such as memory, articulatory buffer size, and lexical access time will effect how long an utterance can be and what is a manageable chunk of speech.

2. Robust communication can be achieved by *checking*. Prosodic structure may fulfil this function by acting as a "I have finished did you receive something sensible" signal.

Both these factors could confound the *Smooth Signal Redundancy Hypothesis*. It is possible that restrictions on the human processing of language do not mirror redundancy in language. This would force chunking which was not predicted by redundancy. This is left as an open question. The aim in this work is not to present a psycholinguistic model of language production but to clearly establish whether redundancy, prosodic structure and care of articulation are linked and if so to what extent. The issue of *checking* is, however, more central to the arguments presented here. In this chapter I have argued that signal redundancy is smoothed because it makes speech communication more robust in noisy environments. Yet *checking* can also fulfil this role. Therefore *checking* must be taken into account as a possible confounding factor in this work and considered in any analysis.

This leads to a weaker hypothesis:

The Smooth Signal Redundancy Hypothesis: Weak Version

Prosodic structure smoothes signal redundancy by controlling care of articulation except when it acts as a checking signal

In order to examine this hypothesis we need to address the following questions:

1. To what extent does prosodic structure relate to and thus arguably control care of articulation? To what extent is any such control lexical or post lexical?

2. To what extent does language redundancy relate to care of articulation?

3. Does prosodic structure account for this relationship? If not, to what extent does language redundancy relate to care of articulation independently of prosodic structure?

4. To what extent does a checking signal confound the smooth redundancy hypothesis?

To help clarify these different arguments it is useful to compare what could be regarded as a traditional view of prosody with the models suggested by these hypotheses. Figure 2.2 is taken from Shattuck-Hufnagel and Turk (1996) and shows a traditional view of prosody. Here a whole set of different factors are controlling how prosodic structure is expressed in terms of phonetics.

In contrast figures 2.3 and 2.4 show how the strong smooth redundancy hypothesis and weak smooth redundancy hypothesis could be modelled. Rather than having a set of different factors affecting prosodic structure you have only language redundancy, and in the weak hypothesis as shown in figure 2.4 also checking. These two factors are then encoded into prosodic structure in order to make the signal redundancy smooth and the communication robust.

Despite the apparent fundamental differences in these models they can be related to each other. In figure 2.5 the traditional prosodic model is amalgamated with the weak smooth redundancy hypothesis.

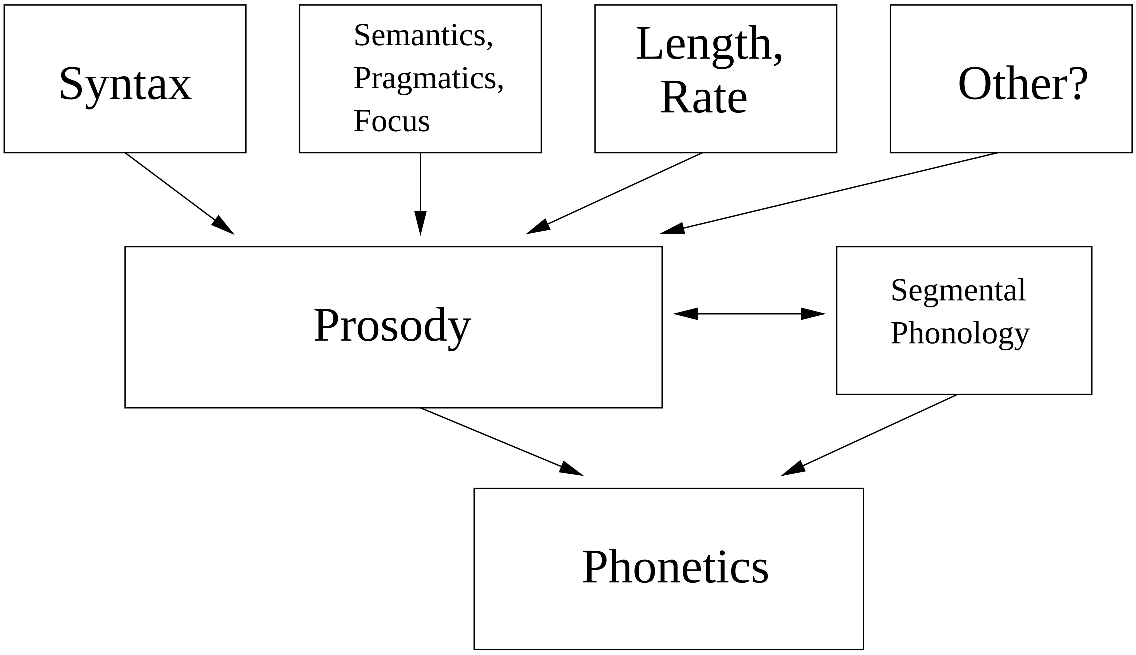


Figure 2.2: One view of the role of the prosodic component of the grammar (taken from Shattuck-Hufnagel and Turk, 1996, page 237).

In fact such an amalgamation is not quite as simple as it seems. In figure 2.2 the arrows represent the processes in a production model. For example, if a major syntactic boundary is produced the language system adapts the prosodic structure accordingly. In figure 2.3 and figure 2.4 the arrows represent more general conditioning processes. For example, lexical stress, a prosodic factor, will tend to be word initial because of redundancy factors. This is a result of the evolution of the lexicon and the English prosodic system, not a direct production model. In contrast, at the phrase level, prosodic factors, such as the location of phrase breaks and accent placement, are being conditioned by redundancy factors in a more similar way to the factors that are shown to condition prosody in figure 2.2. To what extent these factors directly alter prosodic structure during production and to what extent the phonology of prosodic structure has *already* evolved to take such factors into account is more unclear. For example, a sense of familiarity may be sufficient to cause de-accenting without the need to calculate, online, the actual redundancy of the repeated word in that context given the dialogue structure.

Despite these complexities the diagrams do help illustrate the potential relationship between a redundancy based model and a traditional model. Many of the effects attributed to the different factors in the traditional prosodic model can be

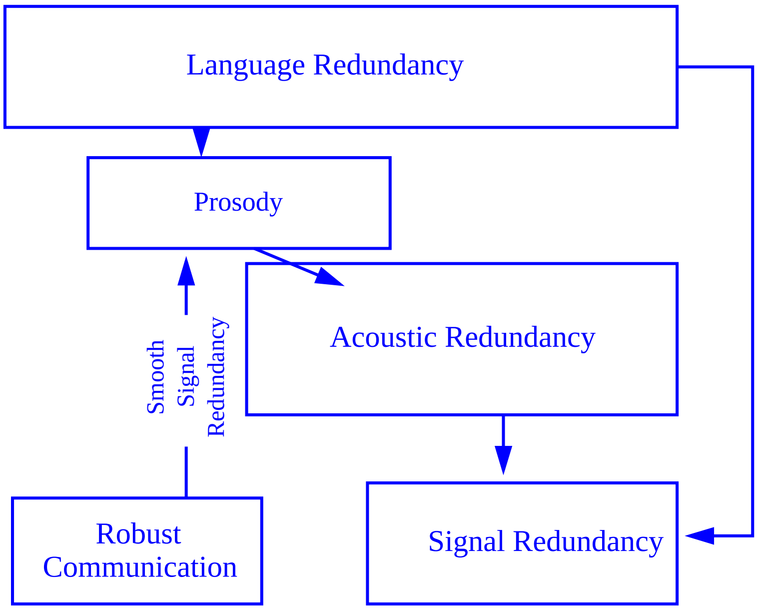


Figure 2.3: Strong smooth redundancy hypothesis.

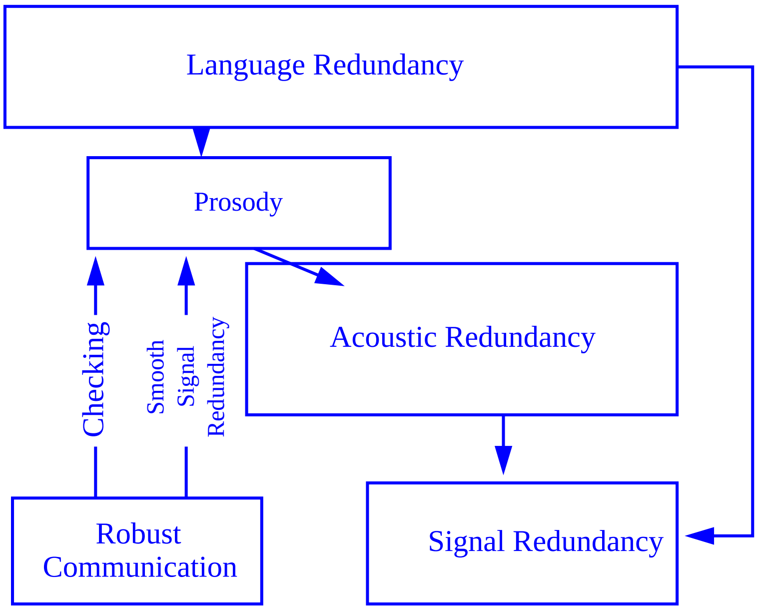


Figure 2.4: Weak smooth redundancy hypothesis.

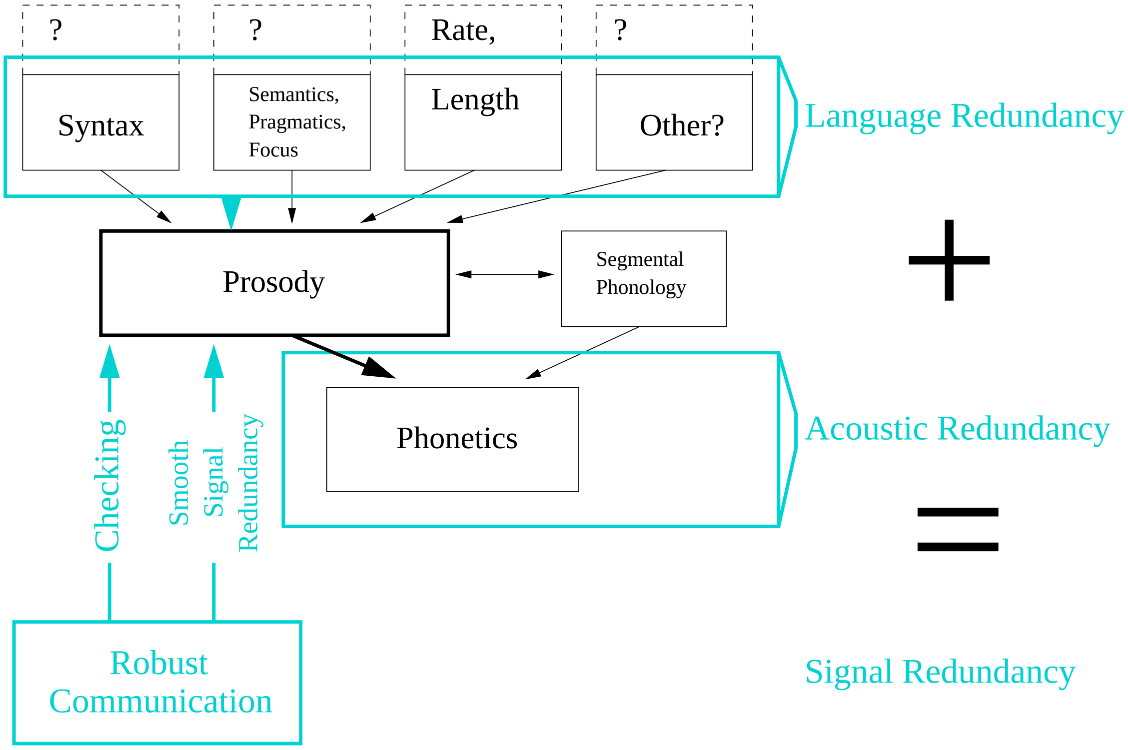


Figure 2.5: How the weak smoothing signal redundancy model could be amalgamated with more traditional views of prosody (based on the figure Shattuck-Hufnagel and Turk, 1996, p237). (Rate is shown outside language redundancy because rate is more closely linked to acoustic redundancy. In general the faster information is produced the less redundant it is. However the complex interaction between rate change and prosodic structure make it difficult to define exactly how such a relationship should be represented in the diagram).

regarded as contributing to language redundancy. For example:

- Function words are often very redundant from a syntactic perspective.
- Open class words are less predictable from a syntactic perspective but some times very predictable from a semantic perspective.
- The longer a section of speech the more predictable the end will generally become.
- In addition factors such as the matching up of prosodic boundaries and syntactic boundaries could be regarded as instances of checking.

However there are examples of changes in prosodic structure which are not easily attributable either to redundancy or checking. For example it is possible to phrase the sentence "Sesame Street was brought to you by the Children's Television

Workshop" as either one intonational phrase or two with either a phrase break before 'by', or after 'by'. In all three cases we are dealing with the same string of words with arguably the same syntax and semantics. Thus in all three cases language redundancy should be the same. Yet we have a variation in prosodic structure. Such variation suggests that prosodic structure cannot be completely conditioned by language redundancy.

In order to test the hypotheses represented by these diagrams we need to examine the relationship between prosodic structure, care of articulation and language redundancy. The more prosodic structure predicts the same changes in care of articulation as language redundancy, the more convincing the argument that prosodic structure is there to effect these changes. The first step in any such quantitative analysis is to produce metrics of the factors we wish to examine. In the next part of the chapter I will discuss the problems that exist in measuring redundancy and give details of how redundancy is measured in this work.

2.6 Measuring Redundancy

To explore the questions raised in the previous sections we need to be able to measure redundancy at different levels. Such levels could vary from the redundancy of a phoneme to the redundancy of a statement in a discourse. As previously stated any formal measure of redundancy at any level requires a model. There are however some difficulties in generating statistical models of natural language.

1. There is enormous interdependency in language. Unlike throwing a dice the production of each word is extremely dependent on words that have gone before and words that will follow. In many cases these interdependencies are 'long distance'. For example in the sentence, 'The man, who was wearing a red raincoat, crossed the road.' there is a relationship between *man* and *crossed* although there are six words in between. To further complicate matters the constraints on what words can or cannot be used in a particular location depend on pragmatic, semantic and syntactic constraints. Formulating such constraints in terms of probability theory is non-trivial.

2. Natural language is sparse. Even with a massive sample of language (millions), words that we can easily recognise and produce may not appear in the sample. For example in the written part of British National Corpus, (Containing over 89 million tokens), the word *zanja* does not occur. This

sparsity is even more pronounced for statistics that represent the chance of words co-occurring.

In order to address the first problem we will use three simple models of redundancy. Rather than suggest any of them offer a true picture of the complex redundancy profile of natural language I will argue they offer a solid basis with which to compare articulation and prosodic factors with redundancy measurements. The models address three different factors which contribute to the overall redundancy in natural language:

1. Word frequency: One of the easiest measurements to make and one of the factors most clearly associated with differences in predictability in language. See section 2.6.1 for a detailed description.

2. Trigram Syllabic Frequency: This measurement examines relationships between syllables within words and between syllables across word boundaries by predicting a syllable on the basis of the previous two syllables. See section 2.6.2 for a detailed description.

3. Reference: The number of times something has been talked about. This offers a higher level redundancy at the semantic, pragmatic and discourse level to compare with the other two measurements. See section 2.6.3 for a detailed description.

The main requirement of these measurements is that they reflect, to some extent, true redundancy in natural language. They also have the advantage of being simple measurements that relate to prosodic patterns. Word frequency relates to prosody in that frequent function words are often unstressed. Trigram syllabic measurements relate to prosody in the preference for word initial lexical stress on the least redundant first syllable of a word. Reference redundancy relates to prosody in a tendency for 'given' referents to be de-accented. They also give a broad coverage of several different levels of redundancy. Trigram syllabic measurements act within and across the word level. Word frequency acts at the word level. Reference redundancy acts at a semantic and discourse level.

The problem of sparsity (see section 2.6.2) was dealt with by using a combination of large corpora, mathematical smoothing techniques and in the case of reference by focusing only on a small easily defined data set. Below I will describe in detail the methodology used to make each separate measurement.

2.6.1 Word Frequency

The HCRC Map Task does not have a very large vocabulary (just over 2000 different words). The CELEX online dictionary was consulted to extract the COBUILD frequency for each word (Baayen *et al.*, 1995). These frequencies are taken from the COBUILD corpus of the University of Birmingham. The 1991 version was used, corrected by CELEX and contained 17.9 million words from different sources. The log of the raw string count was used as the word frequency measure for each string. So for example 'canoe' was given a different value to 'canoes'. The sheer size of COBUILD and the relatively small vocabulary size of The HCRC Map Task meant that coverage was extremely good with 93% of all syllables appearing in a word with a frequency score. The words not represented were mostly composed of disfluencies and cliticized forms (such as 'gonna').

For a detailed account of the relationship between word frequency and care of articulation see chapter 4 section 4.5.4.

As I will explain in chapter 3 the syllable was used as the primitive data point in this analysis. Each syllable was coded for prosodic values, care of articulation metrics and the three redundancy values. For word frequency this meant that syllables in the same word were given the same value (for example the 'moun' syllable as well as the 'tain' syllable in 'mountain' were both given the same value).

2.6.2 Syllabic Trigram Probability

2.6.2.1 What are n-grams?

Charniak says of n-grams:

"One of the least sophisticated but most durable of the statistical models of English is the n-gram model. This model makes the drastic assumption that only the previous n - 1 words have any effect on the next word. While this is clearly false, as a simplified assumption it often does a serviceable job. A common n is three (hence the term *trigram*)." (Charniak, 1993, p39)

N-grams are one of the most frequently used statistical models of natural language. As Charniak points out this is because, despite their simplicity, they often do a 'serviceable' job. N-grams capture some of the interdependency between words. In a word trigram model, although only the previous two words are taken into account when calculating the probability of a new word this is sufficient to capture a lot of structure. For example, trigrams give the probability of 'go to' being followed by a determiner such as 'the', of 'to the' being followed by a noun such as 'beach' and 'the beach' being followed by the end of the sentence. The effect is to make 'go to the beach.' Much more likely than 'go to beach the.' Thus although the trigram model does not **know** about the syntactic structure of noun phrases it can model them quite effectively.

N-gram models are generated by counting co-occurrences of three words. If for example the count showed 2000 instances of 'go to the', 1000 instances of 'go to a' and 1000 other instances of 'go to ...'(something else) then given 'go to' 'the' has a probability of 0.5 of following 'go to', 'a' of 0.25 and everything else of 0.25. In this example this makes 'the' the most redundant token to follow 'go to' because it is the most likely.

Sparse data presents a serious problem for n-gram models. The bigger n the more serious the problem. Let's say that in the data we look at there are no examples of the words 'Princes Street'. If we are using the model and come across this co-occurrence in other data we can make no assumptions as to what may be likely to follow it. There is also a problem if we only have one example of 'Princes Street' such as in the sentence 'I like Princes Street.' If we then use our model to examine the sentence in new data 'Princes Street is the main shopping street in Edinburgh.' Our model gives a probability of zero of 'is' following 'Princes Street' simply because it has not been exposed to this trigram. Even a very large corpus of data cannot cover all trigram probabilities. There are just too many words and many are just too infrequent.

Fortunately statistical techniques can be applied to raw n-gram data to smooth probabilities caused by rarely occurring tokens. In this work the CMU-Cambridge toolkit was used to calculate probabilities. This toolkit comes with a number of these techniques built in. One of these techniques, Good-Turing discounting, can be used to estimate and modify trigram probabilities which are unreliable or absent because of a very small number of observations. In the *Princes Street* example Good-Turing would adjust the probabilities. Because the number of observations of 'Princes Street' was very low, unknown trigrams such as 'Princes Street is' are given small probabilities based on unigram and bigram probabilities. At the same time the high probability of predicting the word following the sole example of 'Princes Street' (end of phrase in this example) is reduced (see section 2.6.2.3).

However, even with a solution to the sparsity problem, the actual complex structure of natural language can only be approximately modelled using n-grams. What does Charniak mean by a 'serviceable job'?

Figure 2.6 shows an example of using trigrams to generate language. Rather than word trigrams this example was produced using syllables as the units. For example 'okay' becomes two units 'o' and 'kay'. The language is generated as follows:

1. Start with two units (in this case 'o-kay').
2. Choose a random location in the corpus and search through until you find these two units.
3. Add the unit following them onto the string (In this example the first unit found was *silence*).
4. Increment the two units you are searching for. ('o-kay' becomes 'kay silence').
5. Go back to 2. Continue until you are bored.

If you compare this randomly generated trigram example with a real dialogue (See Appendix B) you can see what Charniak means by serviceable. The generated dialogue is gibberish but it is readable gibberish. There are some structural errors but in all it does look a lot like dialogue. Compare this to an example of dialogue produce by unigram syllables (Figure 2.7) and bigram syllable models (Figure 2.8). It is the combination of the simplicity of trigram models and the extent they do model language that make them attractive and why such a model was chosen to complement the other measures of redundancy in this work.

2.6.2.2 Why Use a Syllabic Model?

A syllable model was used rather than a word model for a number of reasons. The use of the syllable as the primitive data point meant that a trigram model suffered from less sparsity problems than a word trigram model. Although the number of different syllables used in the maptask does not differ greatly from the number of different words (1500 v 2000) in the British National Corpus which was used to calculate syllabic trigram probabilities the difference is enormous (8000 different syllables versus 90,000 different words). This led to two advantages:

okay
right I'm there
no
okay just draw a couple of centimeter from the left
and you'll pass
on the vertically
erm
do a wee bit from there
okay
yes
okay
north west
right above me bandit territory

Figure 2.6: Example of randomly generated map task using syllable trigrams.

Then a
I
The left
Oh
Erm tree it's
No said
on got the
Just the should *buv* down said

Figure 2.7: Example of randomly generated map task using syllable unigrams. (Syllables taken out of polysyllabic words are shown in italics with approximate orthographic spellings).

Okay
because your page
It's on I want to the
It's about half way
A river to go past the white mountain of that there's a dot there
You want to go east lake

Figure 2.8: Example of randomly generated map task using syllable bigrams.

1. A much larger proportion of the 225,000 syllabic trigrams were represented in the BNC corpus.

2. The probabilities gave a sense of within word and across word redundancy. For example common phrases such as 'go to the' would be represented as well as the increased redundancy of syllables following the initial syllable in a polysyllabic word.

Investigating the effects of word trigram models and comparing the results with the syllabic trigram model would be an interesting exercise. However the aim here was not to produce an exhaustive set of statistical models but a representative set. By including word frequency and a syllabic model it was hoped that this would represent differences at the word level but also at the syllable level. Representing redundancy at the syllable level is important because the other metrics, prosodic structure and care of articulation, are also represented at the syllabic level (see chapters 3 and 5).

2.6.2.3 Method

In order to build the language model speech data from the BNC (British National Corpus) Corpus was used. This consisted of over 10 million syllables taken from speech produced by a wide variety of speakers in a wide variety of speaking situations. Each word in the BNC speech corpus was looked up in the CELEX online dictionary (Baayen *et al.*, 1995) for phonemic content and syllabification. For detail on the syllabification technique see chapter 3 section 3.3.1.1. Words not found were marked as unknown.

This stream of syllables and silences was then used to build a trigram language model using the CMU-Cambridge Statistical Language Modelling Toolkit (version 2) (Clarkson and Rosenfeld, 1997). The CMU-Cambridge Toolkit is a set of Unix software tools to allow the construction and testing of conventional bigram and trigram models. The model was constructed using back-off and Good-Turing discounting.

Back-Off. Back-off is a process used to deal with unknown tokens or context markers in a corpus. A context marker might be a full stop in a written corpus or a silence in a spoken corpus. By using back-off you can decide not to take into account information before the marked context when calculating probabilities for a token after the marker. In effect the marked context becomes a boundary over which the trigram probabilities do not stretch. Because of this boundary the first token following it does not have the two token context that is required to calculate trigram probabilities. Thus either a modified bigram probability is calculated for a known context marker such as a 'sentence start' or a unigram for an unknown context marker such as an 'unknown word' token. The advantage of using back-off and context markers is being able to deal with unknown tokens (by ignoring them) and to build into the model the domain over which trigram probabilities will be considered. Looking back at figure 2.6 the trigram context produced better formed output within each stream of phonation than across silences. This is because the factors governing the production of words across silences are more strongly affected by high level discourse factors which are not modelled using this simple trigram technique. By using back-off at silences it is possible to ignore these transitions and produce probabilities for the more reliable 'within phrase' contexts.

By using back-off I have explicitly avoided trying to use trigrams to find phrase breaks on the basis of low transition probabilities. This demands an explanation given that I have regarded *checking* at such boundaries as a potentially confounding factor in this work. Instead of trying to build checking into the stochastic model, potential checking locations are instead explicitly marked by examining whether a pause occurs after the syllable. This allows a clear separation between checking and smoothing which is important in comparing the power of the weak and strong hypotheses discussed earlier. I also felt that a good stochastic checking model would require considerable investigation and was beyond the scope of the work presented here.

For further discussion on the issue of checking see chapter 7.

Good-Turing Discounting. As mentioned earlier one problem faced by trigram models is sparse data. Even a very large corpus such as BNC will not contain every example of every possible trigram. In order to produce better estimates of the probabilities of infrequent or unseen trigrams it is necessary to smooth the data. Good-Turing Discounting is the default smoothing method in the CMU-Cambridge took kit. Discounting methods are also required in conjunction with back-off to produce estimates of probabilities when data is unknown or missing. What Good-Turing does is to estimate probabilities for unseen trigrams based on unigram and bigram probabilities and to modify probabilities for examples where few examples exist (less than

```
P( 5 | 000 ) = 0.0294142 logprob = -1.531443 bo_case = 2
P( k1 | 000 5 ) = 0.196258 logprob = -0.707173 bo_case = 3
P( st0 | 5 k1 ) = 1.07294e-06 logprob = -5.969426 bo_case = 3-2-1
P( tIN | k1 st0 ) = 0.188013 logprob = -0.725812 bo_case = 3x2
P( Qf | st0 tIN ) = 0.0474383 logprob = -1.323871 bo_case = 3
P( wi | 000 ) = 0.015397 logprob = -1.812563 bo_case = 2
P( 0R | 000 wi ) = 0.0350345 logprob = -1.455504 bo_case = 3
P( @ | 000 ) = 0.0145099 logprob = -1.838335 bo_case = 2
P( bVv | 000 @ ) = 0.00267014 logprob = -2.573467 bo_case = 3
P( @ | 000 ) = 0.0145099 logprob = -1.838335 bo_case = 2
P( k{ | 000 @ ) = 0.00102698 logprob = -2.988440 bo_case = 3
P( r@ | @ k{ ) = 0.155882 logprob = -0.807203 bo_case = 3
P( v{n | k{ r@ ) = 0.761092 logprob = -0.118563 bo_case = 3
P( p0k | r@ v{n ) = 1.36341e-05 logprob = -4.865375 bo_case = 3-2-1
```

Figure 2.9: Example of output from the CMU-Cambridge toolkit when applying a syllabic trigram model produced using the BNC corpus and applied to the HCRC map Task. Each phoneme in each syllable is represented using the CELEX DISC set (see appendix A) where a single character is assigned to each phoneme. E.g. 5 is /əʊ/ k is /k/ 1 is /eI/ etc. The first line reads as follows: The probability of /əʊ/ following a silence (represented as 000) is 0.0294142 the log probability is -1.531443 and the back off is 2. Back-off is 2 because we have only a bigram context as the token is preceded by a silence. Good Turing would have been used to estimate this probability.

7) to take into account these unseen trigrams (See Clarkson and Rosenfeld, 1997, section 3.1.1. for details).

Once the language model was constructed it was then applied to the HCRC Map Corpus in order to calculate syllabic trigram probabilities. The HCRC Map Corpus was converted into a stream of syllables separated by silences. For detail on the syllabification technique see chapter 3 section 3.3.1.1. These syllables where then fed into this model and the probabilities were calculated. See figure 2.9 for the output for the phrase 'okay *silence* starting off *silence* we are *silence* above *silence* a caravan park'.

2.6.3 Reference Redundancy

As I argued above the intention of these redundancy measurements was not to produce a complete model but to produce adequate coverage of some main factors in redundancy. The word frequency measurement together with the syllabic trigram measurement both give a degree of coverage at the lexical and syllabic level. In order to contrast and compare results with these 'low level' factors it was also felt necessary to include a higher level factor which represented redundancy at a more structural and semantic level.

In the dialogues that compose the HCRC Map Task speakers commonly refer to items that are drawn on the map several times. For example:

> GIVER: Have you got **a rope bridge?**
>
> FOLLOWER: Uh-huh I've just up to sort of.
>
> GIVER: Uh-huh. So if you start just drawing... drawing a line up...
>
> towards **the rope bridge.**
>
> FOLLOWER: Up towards going diagonally across to **the rope bridge.**
>
> GIVER: Uh-huh. Just going up then veering off to the right,...
>
> up to **the rope bridge.**
>
> FOLLOWER: 'kay.
>
> GIVER: Then you're going to go across **the rope bridge.**
>
> FOLLOWER: Right, okay. So I draw a line through **the rope bridge.**
>
> GIVER: Uh-huh. You're going to go through **that.**
>
> FOLLOWER: Okay.

(Taken from dialogue Q4NC1 move 47-61 from the HCRC Map Corpus.)

The first reference to rope bridge is in the question 'Have you got a rope bridge?'. Rope bridge is then mentioned several times throughout this snippet of dialogue. The more 'rope bridge' is referenced the more predictable these references become. The first reference or 'introductory mention' is the least redundant because it is the most difficult to predict from context. In contrast, as the rope bridge is discussed, the following mentions become more predictable from discourse context and thus more redundant. Mentions to referents do not always have the same form. For example the final mention of rope bridge in this snippet of dialogue is 'You're going to go through **that.**' where **that** is referring to the rope bridge.

Repeated mention relates strongly to the concept of 'Givenness'. Given information is information shared by listener and talker. The concept of 'Givenness' and its treatment in discourse literature varies. Halliday (1967) uses the term with specific references to de-accenting and the ordering of information within an 'information unit'. Chafe (1974) uses 'Given' in a more restrictive sense relating it specifically to what is foreground in the listeners consciousness. Clark (e.g. Clark and Clark, 1977) suggests that 'Given' information is information that both listener and speaker agree upon (for an overview of these views see (Brown and Yule, 1983, chapter 5).

In this work it is mention which is coded as a redundancy measurement. The extent a mention is 'Given' relies more strongly on questions of what is happening in the speakers' minds as well as complex structure at the discourse level. However, in general, the more a reference is mentioned the more 'Given' it becomes, the easier it is to predict and thus the more redundant it is. Although mention is a crude measure of such information status, in this work, as a contrast to the lexical and syllabic measures, it serves as a metric of redundancy at the discourse level. As with the other models of redundancy it is used here as an approximation to the actual predictability of language and is not put forward as a theoretical account of this predictability.

There is extensive evidence that mention, whether reflecting 'Givenness' or not, is strongly related to prosodic structure in terms of de-accenting (see chapter 4 section 4.5.6) and to changes in articulation (see chapter 4 section 4.5.3).

2.6.3.1 Method

Reference coding was carried out on the HCRC Map Corpus by members of the dialogue group. The final coding was then thoroughly checked by another coder. Only references to landmarks printed on either of the maps were coded. Elliptical references and references to parts of landmarks were ignored. The order of mention was established by sequential time of mention within the dialogue.

The result was a set of just over 31,000 syllables coded for mention out of the total 200,000 or so syllables in the HCRC Map Corpus. Of these 1553 had also been hand coded for prosodic structure.

2.7 Summary

This thesis explores the idea that redundancy relates strongly to articulation. This chapter has discussed the term redundancy and its relationship to statistical models as well as the importance of a noisy channel model of communication. In order to examine relationships between redundancy and care of articulation three different metrics of redundancy have been presented. The aim of these measurements is not to present a theoretical model of redundancy in language but rather to approximate such redundancy. The metrics cover redundancy at the syllable level (syllabic trigram probability), at the word level (log of word frequency) and also at the discourse level (order of mention of referents). These measurements will give a representative, robust and simple measure of redundancy allowing a large scale quantitative corpus analysis.

Chapter 3

Prosodic Structure

3.1 Introduction

In this chapter I will review current literature and theory in the area of prosodic structure. I will then relate this to the approach used in this thesis. Finally I will describe the coding scheme and the methodology I used to describe prosodic structure in this work. This review will concentrate on work carried out on English. Research in other languages, except where directly relevant to English is beyond the scope of this thesis.

Including prosodic information in this work allows the exploration of the key question of this thesis:

- Does prosodic structure smooth signal redundancy by controlling care of articulation?

It will also allow us to look at a number of secondary questions including:

1. How accurate is automatic prosodic coding given word segmentation compared to hand coded prosodic coding?

   In order to code the large corpus of spontaneous speech used in this study automatic prosodic coding was carried out as well as hand coding. An evaluation of this automatic coding is presented in chapter 6.

2. Do results from spontaneous speech support laboratory results with regard to the effect of prosodic structure on care of articulation?

   As we shall see in chapter 4 the majority of the work examining the relationship between prosodic structure and care of articulation has been carried out on read speech. This thesis contributes to the field by examining these relationships over a large corpus of spontaneous speech.

There is clear laboratory evidence that prominence and constituent boundaries affect care of articulation both in terms of duration and spectral clarity (Price *et al.*, 1991; Beckman and Edwards, 1990; van Bergem, 1988, amongst others). Because this work relates prosodic structure directly to the surface structure in speech there is a need to examine these prosodic factors and to discuss prosodic theory relevant to them. For a clear introduction to many of the issues in prosodic theory outside the scope of this thesis I refer the reader to Ladd (1996), Couper-Kuhlen (1986), Hogg and McCully (1987) and to review papers by Shattuck-Hufnagel and Turk (1996) and Nooteboom (1997).

3.2 What is prosody?

Although a universally acceptable definition of prosody has been elusive (Shattuck-Hufnagel and Turk, 1996) there is much consensus on what we are dealing with when we are dealing with prosody.

Prosodic phenomena can be summarised as:

- **Being described by at least four acoustic parameters including: Duration, Amplitude, F0 and Pause.** Other acoustic parameters such as spectral clarity and spectral tilt also appear to be related to some extent (van Bergem, 1988; Sluijter, 1995; Campbell and Beckman, 1997). None of these acoustic parameters have a simple mapping onto prosodic structure for a number of reasons:

  1. A direct mapping is confounded by phonetic context, identity and inter-speaker differences. For example different phones are produced with different amplitudes by different speakers.
  2. The same prosodic result, such as an increase in perceived prominence, can be achieved by using different parameters. For example a speaker could make a word seem more prominent by either lengthening it or by making it louder.
  3. Different prosodic constituents affect the same acoustic parameters. For example accenting a syllable will make it longer but so will a phrase boundary. Thus lengthening may be an indication of a number of different prosodic influences.

- **Affecting domains larger than a single phonetic segment.** Prosodic acoustic parameters appear to signal constituent boundaries and prominences. These acoustic cues can extend over domains larger than a single segment or even single syllables. For example, an accent on the first syllable in a bisyllabic word affects the length of the subsequent syllable (Turk and Sawusch, 1997; Turk and White, 1999).

- **Requiring a degree of abstraction in its definition.** Different segment types are affected in similar ways. For example phrasal stress increases duration of a syllable whatever the contents of that syllable. In addition laboratory results suggest (Wightman *et al.*, 1992; Price *et al.*, 1991) that differences in the duration of the rhyme of a syllable can be explained by a hierarchical set of constituents with the edges of smaller constituents lining up with the edges of larger constituents.

In order to put modern work in context I would like to first clearly adopt Shattuck-Hufnagel and Turk's (Shattuck-Hufnagel and Turk, 1996) working definition of prosody and give a brief description of the key terms and concepts in prosodic research.

The definition of prosody proposed by Shattuck-Hufnagel and Turk is:

"(1) Acoustic patterns of F0, duration, amplitude, spectral tilt, and segmental reduction, and their articulatory correlates, that can best be accounted for by reference to higher-level structures, and (2) the higher level structures that best account for these patterns." (Shattuck-Hufnagel and Turk, 1996, p196).

Key concepts to most theories include notions of constituents, hierarchical structure and prominence. A brief review of these concepts follows.

3.2.1 Constituents

Constituents of various levels are posited within theories of prosodic structure. The extent one constituent is made up of others, whether recursive constituents exist and the number and type of constituents vary between different theories. However a great deal of common ground exists. For example most prosodic theory regards the syllable as a prosodic constituent. In general the following constituents are referred to in theories of prosodic structure (From the smallest to the largest):

- mora

- syllable
- within word foot
- prosodic word/ clitic group
- phonological phrases (major and minor)
- intonational phrases (full and intermediate)

There is much agreement on the definitions and domains of moras, syllables, feet and full intonational phrases. Prosodic words/clitic groups, phonological phrases and intermediate intonational phrases however have been the subject of some discussion. For a description of these different constituents and the role they play in different theories see (Shattuck-Hufnagel and Turk, 1996). In general such constituents are defined in several ways:

1. As the domain of phonological rules.
2. As the domain of an intonational tune or contour.
3. In some theories and for some constituents in terms of rhythmic prominence. For example a foot is a sequence of a strong syllable followed or preceded by a number of weak syllables.

Both phonetic and psycholinguistic evidence supports the existence of some constituents. An example of phonetic evidence is phrase final lengthening at the end of intonational phrases (Shattuck-Hufnagel and Turk, 1996). An example of psycholinguistic evidence is the listeners' preference for interrupting at constituent boundaries (Shattuck-Hufnagel and Turk, 1996).

Different theories present different hierarchies of constituents where each constituent is made up of smaller constituents (Hayes, 1989; Beckman and Pierrehumbert, 1986; Nespor and Vogel, 1986; Selkirk, 1978). For example Selkirk (1978) proposed a strict hierarchical structure with **intonational phrases** as the largest component in turn being made up of **major phrases** which in turn are made up of **minor phrases** which in turn are made up of **prosodic words** which in turn are made up of **feet** which in turn are made up of **syllables**.

Different constituents appear to have different effects on some acoustic parameters. For example boundary lengthening is greater at an intonational phrase boundary than at a minor phrase boundary. Also the relationship between constituents and prominence vary. Beckman and Edwards (1990) suggest that different types of prominence are associated with different constituents. In their theory the prominence associated with a particular constituent is termed its head. For example the head of an intermediate intonational phrase is a nuclear pitch accent. The work reported here does not explicitly link prominence with constituent structure in this way but does look at several levels and types of prominence.

3.2.2 Prominence

Prominence can be regarded as the extent a sound or syllable stands out from others in its environment. It is realised chiefly through three acoustic parameters, pitch, amplitude and duration (Fry, 1958). The term stress is often used to describe prominence. However the word stress is used in different ways by different researchers and can vary from meaning the potential for a syllable to be accented (lexical stress) to the realisation of such accenting (phrasal stress which is normally associated with a change in pitch). It can also be used to describe syllables which have longer durations and high amplitudes without any associated pitch change. Cruttenden (1986) and Ladefoged (1982) use the term degrees of stress and associate it with three phenomena:

1. Reduced versus Full Vowels, such as the /i/ in spongy /spʌnʤi/ in contrast with the /ə/ in after /ʌftə/.
2. Lexical Stress, for example the 1st and 4th syllable in “MUL-ti-pli-CA-tion” are lexically stressed. Here “MUL” is described as having secondary stress and “CA” as primary stress.
3. Phrasal prominence, for example “beach’ in “I’m going to the BEACH” which normally has a change in F0 associated with it as opposed to “beach” in “I’m going to the NUDIST beach” which would normally be unaccented.

Cruttenden also makes a distinction between nuclear pitch accents and non-nuclear pitch accents in English. In a normal intonational phrase a nuclear accent (or sentential accent) will be the last accent before the end of the phrase. This last accent often gives the impression of greater prominence than preceding pitch accents. Cruttenden (1986) argues that we need to distinguish four different types of stress:

1. Primary stress (Prominence caused by a nuclear pitch accent)

2. Secondary Stress (Prominence caused by other pitch accents)

3. Tertiary Stress (Prominence caused only by lengthening and loudness but no pitch change) For example a lexically stressed syllable which is realised without a pitch accent.

4. Unstressed

Ladefoged in contrast ignores the lexical stress/non-nuclear phrasal stress distinction and adds vowel type also giving four types of stress:

1. Tonic Accent (Prominence caused by nuclear pitch accents).

2. Lexical Stress (Prominence caused by lexical stress).

3. Vowel type. (Prominence caused by a full as opposed to a reduced vowel. For example Ladefoged would regard the /i/ in /spʌnʤi/ as more prominent than the /ə/ in /ʌftə/ although neither are lexically stressed).

4. Unstressed

The higher levels of stress require stress at all lower levels. For example a pitch accent must be associated with a lexically stressed syllable and a lexically stressed syllable must have a full vowel.

In this thesis a combination of the factors described by Cruttenden and Ladefoged will be adopted to describe prominence rather than the descriptions of prominence, such as metrical grids and trees, adopted in metrical phonology (Hayes, 1989; Beckman and Edwards, 1990; Nespor and Vogel, 1986; Selkirk, 1978). To a large extent this is a purely pragmatic approach as the factors described by Ladefoged and Cruttenden are relative easy to encode for a quantitative analysis.

In addition to these traditional prominence factors, syllables will also be coded for spillover (my term). Work in laboratory phonetics has shown that the effect of a pitch accents extends beyond the syllable associated with the accent (Turk and White, 1999; Turk and Sawusch, 1997). This increases duration in syllables to the left and right of the accented syllable, although more spillover is found to the right than to the left and it appears to be attentuated by word boundaries. Thus in addition to prominence factors a syllable is also marked if it is directly to the left or right of a pitch accent.

3.3 A Practical Prosodic Coding Strategy

In order to quantify the effect prosodic structure has on any acoustic correlates of articulatory care two questions must be resolved:

1. What factors in prosodic structure should be examined?
2. How should such factors be represented in a quantitative analysis?

Both practical and theoretical issues determine the response to these questions. From a practical point of view only factors that can be quantified reliably and (considering the amount of material required for any analysis of redundancy) with relative efficiency, can be included in this analysis. From a theoretical point of view, as this work is not attempting to promote or undermine any particular prosodic theory, only factors with which there is reasonable consensus will be included.

In general research has shown that, apart from segmental identity and certain segmental context effects, it is prominence and the boundaries of constituents that have the strongest effect on speech acoustics (see van Bergem, 1988; Price *et al.*, 1991; Beckman and Edwards, 1990). Thus the coding strategy I used puts a clear emphasis on describing prominence and boundary features.

In order to simplify a large scale statistical analysis a primitive will be adopted as the standard data point. For example, in corpus linguistics such a primitive is often the word, in phonetics the segment or phoneme. In this work, for reasons detailed below, syllables will form the basic primitives that coding is applied to.

3.3.1 Issues in Coding Constituents

3.3.1.1 The Syllable

Every data point in my analysis represents an individual syllable with prosodic, redundancy and care of articulation information associated with it.

The syllable was chosen as the primitive because it was the smallest easily usable constituent. The HCRC corpus is word segmented and as 70% of the words are monosyllabic most of the syllabic durations have been measured by hand. Thus, although autosegmentation was used to segment syllables in polysyllabic words, most of the data analysed was unaffected by inaccuracies caused by automatic techniques. It was unrealistic, for this thesis, to hand-segment 15 hours of spontaneous speech into smaller constituents such as phonemes and autosegmentation of constituents this small was regarded as too unreliable. Thus the syllable offered a compromise between size and the amount of duration measurement error caused by autosegmentation. In addition both redundancy and care of articulation metrics could be applied at the syllable level (see chapters 2 and 5).

The syllable has formed part of most modern theories of prosody either explicitly (Nespor and Vogel, 1986; Hayes, 1989; Selkirk, 1984; Couper-Kuhlen, 1993) or implicitly as in the AM approach (Beckman and Pierrehumbert, 1986; Ladd, 1996).

Attempts to define syllables have fallen into two main areas:

1. Phonological Definitions: There are restrictions in what sounds may be grouped together (for example /ng/ is not a permitted sequence of sounds in the same syllable in English). Permissible and non-permissible relationships between the constituents can be used to define the syllable (see Couper-Kuhlen, 1986, chapter I section 3.22 for a review).

2. Phonetic Definitions: Syllables represent peaks in sonority. The sonority of a sound is measured by comparing the acoustic intensity of a sound when spoken with similar stress, duration and pitch. Vowels are more sonorous that consonants and therefore become the nuclei of syllables (Ladefoged, 1982).

In this work a phonological definition, as implemented in the CELEX dictionary (Baayen *et al.*, 1995), is used to define each syllable. In general the syllabification present in an isolated word is preserved when the word is articulated. However there are examples of words which have different possible syllabifications (For example 'predatory' as 'pre-da-tO-ry/pre-dA-try' or 'city' as 'ci-ty/cit-y'. See (Ladefoged, 1982, p220) for a discussion). In spontaneous speech when a large amount of reduction occurs syllables can sometimes become squashed together or completely removed. Different speakers can sometimes pronounce words with different syllabic structure. In the material this work considers, 70% of all syllables appear in monosyllabic words, so although serious difficulties exist in defining the notion of a syllable, in this case, syllabification was mostly carried out as part of word segmentation. Where automatic syllabification was required a dictionary based on the CELEX database (Baayen *et al.*, 1995) was used to decide syllable boundaries. This syllabification was based on the primary pronunciation as specified in Gimson (1977) and used the maximal onset principle (Clements and Keyser, 1983).

The relationship between each syllable and other constituents was then coded as a number of prosodic factors (see section 3.3.2 for a complete list).

3.3.1.2 Other Constituents: Break Index Coding

From an experimental perspective, Price *et al.* (1991) take a pragmatic approach to the problem of defining boundaries and constituents by assigning a break index which represents the boundary strength between two words. By coding for boundaries directly the concept of a prosodic hierarchy is accepted but without the need to characterise the complex composition of the domains themselves. Pause length, out breath and phrase lengthening can be directly related to the break index (Wightman *et al.*, 1992).

Using break index to represent boundaries and implicitly higher level constituents is the approach I have used. I have not coded feet, prosodic words and clitic groups explicitly, not because I am denying these may be part of the phonological structure but because I am investigating the surface structure rather than theoretical differences in metrical phonology. Boundaries represented by break indices are coded on the basis of the word segmentation already carried out. In general each word is separated by a break index of 1. However in some cases words are heavily run together, for example 'do you have' might become 'dyuv'. When no sensible word boundary could be assigned these run together words were treated as a single word. In these cases a hand edited additional dictionary was used to assign appropriate syllabification (see section 3.4.5.3) and the missing word boundaries were regarded as a break index of 0. Break indexes 2-4 are used to represent the boundaries of intonational phrases where 2-3 are used to mark intermediate intonational phrases. This second level of intonational domain, the intermediate intonational phrase (IIP) were proposed by Beckman and Pierrehumbert (1986) to account for data in English and Japanese.

The overall result is a combination of prosodic facts associated with syllables (for example whether within a monosyllabic word or not) and an impressionistic boundary strength coded using the ToBI system (Beckman and Ayers, 1993). Section 3.4 will give a detailed description of the methodology and the exact coding carried out.

3.3.2 Summary

Prosodic structure that this work will investigate is as follows:

1. Constituents
   - Syllables form the basic building block for the analysis.
   - Higher level constituents are coded as boundaries occurring after a particular syllable.
   - Word boundaries and cliticisation are coded as break index 0-1
   - Intermediate intonational phrases are coded as break index 2-3
   - Full intonational phrases are coded as break index 4
2. Prominence
   - Nuclear accents are coded for presence or absence.
   - Pitch accents are coded for presence or absence.
   - Lexical Stress is coded for presence or absence.
   - Whether the vowel in a syllable is full or reduced.
   - Spillover: whether the syllable is directly to the left or right of an accented syllable.

3.4 Prosodic Coding: Methodology

3.4.1 Introduction

3190 words making up 679 full intonational phrases from the HCRC Map Corpus (Anderson *et al.*, 1991) were coded using GlaToBI (Mayo *et al.*, 1997), a variant of the ToBI tone and break index coding system which was adapted for the Glaswegian accent. Automatic techniques were then used to label nuclear accent placement on these materials (see section 3.4.5) as well as syllabic structure, lexical stress, phrase boundaries and word class for all materials in the corpus (approximately 200,000 syllables).

3.4.2 ToBI

A ToBI variant was used for coding for the following reasons:

1. ToBI is a well understood standard for prosodic coding which is used widely by the speech community.

2. The variant used for the Glaswegian accent has already been defined and evaluated.

3. The segmental style of coding made it easy to link features in the prosodic coding with individual syllables.

The ToBI prosodic coding system was developed by a team of academics in the U.S. in order to produce a common standard for coding intonation and prosodic structure for large corpora of speech held in digitised form on computer (Silverman *et al.*, 1992; Pitrelli *et al.*, 1994). The system itself was a compromise between those researchers focusing in intonation/prominence and those focusing on prosodic constituent structure. This led to a two tier coding system.

1. The tone tier codes changes in pitch which are associated with accents and phrase boundaries. The system is based on an Autosegmental/Metrical (AM) view of intonation (see Ladd, 1996, chapter 3 for a review). Phenomena such as accents are made of up of strings of tone symbols combining to produce accents and boundary tones. This tone level is heavily influenced by Pierrehumbert's intonational analysis of English (Beckman and Pierrehumbert, 1986).

2. The break index tier codes the strength of boundaries between lexical items. Break index values can also be combined to represent higher level structures such as intonational phrases. This tier is based on work by Price and her colleagues (Price *et al.*, 1991).

In addition to ToBI coding other prosodic features were calculated automatically for the whole corpus (about 200,000 syllables). Some of these prosodic features, such as the notions of words, syllables and lexical stress are implicitly part of a ToBI analysis.

3.4.3 GlaToBI

GlaToBI was developed at the University of Edinburgh by Matthew Aylett, Jacqueline Kowtko, Bob Ladd and Paul Taylor in order to produce a ToBI like coding system that could be applied to the HCRC Map Corpus. In this corpus most speakers spoke with a Glaswegian accent and a number of clear differences in intonation between this accent and a standard British accent meant changes were required to in order to use the ToBI system. Once an agreed system was in place it was evaluated by Catherine Mayo. Details of the GlaToBI system and its evaluation are presented in Mayo *et al.* (1997).

3.4.4 Method

Although GlaToBI was used for prosodic coding, only 2 items of information were retained from this coding, the presence or non-presence of an accent and the break indexes. Modifications adopted by GlaToBI affect accent and boundary type rather than accent and boundary presence. Consequently none of these modifications have any direct effect on the results in this thesis. However the GlaToBI evaluation is important in that the coder who coded all the materials used in this study was evaluated and he was found to be as competent as the other two expert coders.

In all 3190 words making up 679 full intonational phrases were coded using GlaToBI. The phrases were coded using Entropic's Xwaves software. The coder was able to listen to sections and parts of the speech as many times as required. The speech had already been word segmented by phoneticians at the Centre for Speech Technology Research at the University of Edinburgh. The coding was carried out over a period of several months. Earlier coding was systematically checked to ensure consistency was maintained. For polysyllabic words syllable boundaries were determined automatically using autosegmentation and an online dictionary (Baayen *et al.*, 1995) using a syllabification based on the primary pronunciation as specified in Gimson (1977) (see section 3.3.1.1 for details). The output of this coding was a set of syllables marked for accentedness (yes/no) and break index (0-4). Syllables marked as having a disfluent break index were ignored.

The materials were taken from all 64 speakers in the map task (34 male, 30 female). Some speakers were represented more than others (e.g. 33 phrases were the maximum for a speaker, 1 phrase was the minimum - mean 8, standard deviation 6).

This hand coded prosodic information, together with the automatic measurements described below, were the prosodic factors used in the comparison between redundancy and care of articulation.

3.4.5 Automatic Coding

3.4.5.1 Nuclear Accent Placement

Nuclear accents are regarded as the most prominent accent in an intonational phrase. ToBI does not code nuclear accent placement explicitly but rather implicitly in that, in English, the last accent before a phrase boundary is regarded as the nuclear accent. Thus given a ToBI coded phrase it is possible to determine nuclear accent placement by examining accent and boundary markers. All GlaToBI coded materials were automatically marked in this way.

3.4.5.2 Boundaries

The entire HCRC map task is word segmented and transcribed. In addition words were tagged for word class. Syllabification and phrase boundaries were not however explicitly marked.

Automatic phrase boundaries were placed after a stream of phonation when a pause or non-phonated noise (such as an in-breath) occurred. These automatic phrase boundaries were inferior to hand coded break indexes in that they would posit a phrase boundary at locations of disfluency when one may not exist and miss phrase boundaries marked with pitch change but no pause. The differences between the hand segmented break indices and automatically determined break indices are discussed in the results chapter (chapter 6). The advantage of these measures is that they could be deduced for the whole corpus rather than the small subset determined by the prosodic coding carried out by hand.

Syllable boundaries (for polysyllabic words) were determined using autosegmentation. This involved consulting an online dictionary containing a canonical phonemic representation for each word in order to establish the probably segmental contents of each syllable. A hidden markov model (HMM) speech recogniser (Young *et al.*, 1996) with a model for each segment already trained from previous speech was used to posit the likely boundaries of each phoneme. The syllabification as present in the dictionary lookup (see section 3.3.1.1) was then used to determine likely syllable boundaries. Although these syllable boundaries were not as accurate as the hand measured word boundaries the error was generally within 30-40ms (see chapter 5). However errors could also be introduced if a word was re-syllabified or a large percentage of the word was elided. It was decided that given the small number of polysyllabic words in the corpus and, of these, the small number of words that can have alternative syllabifications such errors would be rare and could be ignored.

3.4.5.3 Lexical Stress

Implicit to ToBI coding is a word segmented speech stream and the notion of lexical stress. However within ToBI:

'The orthographic tier is arguably not part of any core prosodic analysis, except inasmuch as the labels on this tier can be used to interface the transcription to dictionary entries which do indicate such things as which syllable is likely to be more stressed in each word, prosodic information which is otherwise not included in the ToBI system.' (Beckman and Ayers, 1993, section 1.1).

Consequently, as with syllabification, a dictionary was used to determine the location of lexical stress. The dictionary was hand modified so that compounds such as 'dyou' and 'dyouhava' were assigned appropriate lexical stress (dYOU and dyouHAVa rather than DO-YOU and DO-YOU-HAVE-A). Secondary stress was also marked and these values were associated with each syllable in the speech stream. This process was carried out on all the syllables in the corpus rather than just those which had been prosodically coded.

3.4.5.4 Word Class

Although not strictly a prosodic element, monosyllabic closed class words which have a structural role in language, such as articles and auxiliaries, show different prosodic behaviour. Often they do not carry pitch accents and often lose their lexical stress in connected speech (for example 'the' in 'go and see the doctor' would probably be realised as /ðə/ rather than /ði/. See (Cruttenden, 1986) 2.3). For this reason the word class for each word in the corpus was extracted from a hand modified automatic syntactic parse and associated with the word. Adjectives, non-auxiliary verbs, common and proper nouns were marked as open class. All other words were regarded as closed class.

3.4.5.5 Automatic Marking Evaluation

Because all the automatic marking was carried out on the whole corpus a lot more material was coded in this way allowing two analyses of the effects of prosodic structure on redundancy and care of articulation. The first analysis was carried out over the small set of hand coded materials augmented with some automatic coding (e.g lexical stress, syllabification). The second analysis was carried out over the whole corpus with only automatic coding and approximations to hand coded factors.

3.5 Summary

As mentioned earlier, when dealing with a large corpus practical considerations are vital in any coding strategy. The strategy described here is the result of a number of compromises:

1. The compromise between a phonetic/descriptive coding and a phonological/interpretive coding.

2. The compromise between time consuming hand coding and fast but less accurate automatic coding.

3. The compromise between complex detailed coding systems which are difficult to quantify and to code reliably but represent much depth and complexity against simple coding systems where agreement is greater between coders but does not capture much of the complexity we know exists.

The approach taken here has a number of advantages and disadvantages. The advantage of the prosodic coding carried out here is that it balances hand coding with automatic coding, uses simple prosodic features common to most modern theories of prosodic structure and has been carried out on spontaneous connected speech taken from a very large corpus. Some potential disadvantages include the possibility of inaccuracies in hand coded and automatic coded materials and the use of spontaneous materials which are only a subset of speech styles and speakers.

A summary of the output of this coding is as follows:

3638 syllables hand coded for:

- accent placement
- break index
- nuclear accent placement.
- spillover.

169464 syllables (including the materials above) automatically coded for:

- Syllabic position and total number of syllables in the overall word.
- Phrase initial or phrase final : In terms of immediately preceding or immediately following a pause.
- Lexical stress: Whether lexically stressed or not determined from consulting a hand checked dictionary.
- Vowel Type: Whether the vowel is reduced or full.

Chapter 4

Care of Articulation: Literature Review

4.1 Introduction

So far I have presented a framework for relating prosodic structure, redundancy and care of articulation to each other (chapter 2), and considered basic terms in redundancy (chapter 2) and reviewed prosodic theory (chapter 3). The result of these discussions is to produce a practical solution to measuring language redundancy and to coding prosodic structure. In this chapter we will focus specifically on the acoustic characteristics of carefully articulated speech and, up until now, what evidence has been presented that such speech is associated with prosodic structure and redundancy. As we will see few studies have either taken redundancy into account when examining effects of prosody or prosodic structure into account when examining effects of redundancy. This thesis seeks to address this omission.

In this chapter I will define the terms used to describe differences in care of articulation and review the current literature that has investigated the effects of prosody and redundancy on care of articulation.

In doing so we will address the following questions:

- How do we define care of articulation?
- How does this definition relate to concepts of hyperspeech, 'clear speech' and intelligibility?
- What are the acoustic characteristics of carefully articulated speech in terms of spectral and durational changes in vowels and consonants?

- Can duration change alone be regarded as a strong correlate of carefully articulated speech?

- How do prosodic structure and redundancy relate to these acoustic characteristics?

4.2 Defining Care of Articulation

Before we can establish a link between prosody, redundancy, and care of articulation we need to define care of articulation and some related terms.

In order to explain phonetic variation Lindblom (1990) in his H&H (hyper- and hypospeech) theory presents the idea that differing degrees of articulatory effort are used in different circumstances. Lindblom argues that a speaker assesses the needs of a listener and balances the effort used in producing speech against the need for producing speech which is sufficiently discriminable. In doing so the speaker alters articulation in response to communicative and situational demands along a continuum of hyper- and hypospeech.

Hyperspeech is carefully articulated speech. Sounds produced in hyperspeech are easier to ascribe to individual phonemes, the variance within the production of speech sounds is less and the effect of coarticulation and reduction are minimised. Hypospeech in contrast is 'sloppy' speech with more variance in the speech sounds and greater coarticulation and phonetic reduction.

A number of laboratory studies have investigated the acoustic and articulatory effects of hyperspeech and clear speech (Hanley and Steer, 1949; Freed, 1978; Ferguson, 1977; Moon and Lindblom, 1994; Picheny *et al.*, 1985, 1986; Uchanski *et al.*, 1996; Bond and Moore, 1994; Bradlow *et al.*, 1996). These differences in articulation appear to systematically occur in conjunction with prosodic factors (Lehiste *et al.*, 1976; Wightman *et al.*, 1992; Price *et al.*, 1991; Beckman and Edwards, 1990; Cutler and Butterfield, 1990; Summers, 1987; de Jong, 1995; van Bergem, 1988; Turk and White, 1999) but also in conjunction with changes in predictability. Effects ascribed to predictability vary from those caused by word frequency, word structure, and word context (Lieberman, 1963; Hunnicut, 1985; Balota *et al.*, 1989; Luce, 1986; Goldinger and Summers, 1989; Wright, 1997) and also at 'higher levels' involving semantic and syntactic redundancy such as the use of referring expressions (Fowler and Housum, 1987; Fowler, 1988; Fowler *et al.*, 1997; Hawkins and Warren, 1994; Bard *et al.*, 1995; Samual and Troicki, 1998;

Shields and Balota, 1991). In sections 4.3, 4.4, 4.5 I will review this literature in detail. First I will define some of the terms relevant to this work.

In this work my definition of care of articulation is similar to the definition of hyperspeech:

Carefully articulated speech is speech which is articulated with more articulatory effort than usual in order to produce speech sounds that are more discriminable than usual.

To further elaborate on this definition I will begin by looking at the studies which have specifically attempted to elicit hyperspeech and examined the acoustic and articulatory effects of this carefully articulated speech. I will then examine work that has associated these acoustic and articulatory effects with prosodic factors and finally look at work which has looked at the relationship between predictability and care of articulation.

4.3 The Acoustic and Articulatory Correlates of Carefully Articulated Speech

Before looking at individual work we need to examine the differences between the term hyperspeech and the terms 'clear speech' and 'intelligibility'.

4.3.1 Clear Speech

Clear speech is a type of speech that has been hyper-articulated. In order to elicit clear speech Moon and Lindblom (1994) asked subjects to read a list of words as clearly as they could. In order to maintain this effect the subjects were periodically interrupted by the experimenter who pretended the word had not been understood and should be repeated. They then looked at the acoustic differences between vowel sounds in a normally spoken control utterance in contrast to vowels in 'clear speech' (see below for more details). Previous work in the acoustic differences between normal speech and types of clear speech have included looking at speech production in noisy environments (e.g. Hanley and Steer, 1949), the way people adopt a style of speech sometimes referred to as "Foreignese" when speaking to non-native speakers with limited comprehension skills (Freed, 1978), and the "Simplified Register" or "Motherese" that mothers use in communicating with infants (Ferguson, 1977).

Work on clear speech falls into two categories:

1. Work that concentrates on how clear speech differs from normal speech. This work is primarily interested in how to speak clearly and looks at the direct relationship between differences in articulation and intelligibility. Researchers in this area are often interested how one may speak clearly for the hard of hearing and how hearing aid technology could be improved to make speech clearer and more intelligible.

2. Work which uses clear speech as an example of hyper-articulated speech. This work doesn't just look at the acoustic factors which characterise clear speech but generalises these factors as characteristic of all hyper-articulated speech. The assumption is that articulation varies along a scale of hypo/hyper speech and that the clear speech style is, in general, more hyper-articulated. Thus the same section of speech with the same prosodic context will be measurably more hyper articulated in clear speech. However articulation also varies within an utterance. Once the acoustic factors that characterise clear speech are ascertained these factors can then be used to measure differences in care of articulation within as well as across speech styles. Such factors may be differences in amplitude, spectral characteristics and timing.

   For example if the sentence 'The cat sat on the mat.' is elicited as clear speech we would find that the /æ/ vowels tend to be longer than in the same sentence spoken in normal spontaneous speech because, in general, extended length indicates more carefully articulated speech (see section 4.3.5). We can also use this measurement within the utterance and compare differences in vowel length between /kæt/ and /sæt/ and say whether one word or the other has been more carefully articulated. Unfortunately there are problems when comparing phonemes within an utterance in this way. We need to know what length effects are purely due to phonemic context and normalise for this. However, in principle, providing such normalisation is carried out, we can use any factor that characterises clear speech as a potential metric for care of articulation within an utterance.

Although work in 'clear speech' concentrates on articulation, implicit to the term clarity is that there is a listener who finds clear speech easier to understand than unclear speech, that clear speech would, in general, be more intelligible. Work by Payton *et al.* (1994) and (Picheny *et al.*, 1985) confirm this. Because of the relationship between intelligibility and clear speech, and the fact that significant laboratory studies have looked at the phonetic characteristics of intelligibility, it is important to consider work on intelligibility with reference to care of articulation. In the next section I will discuss the meaning of intelligibility, review work in this area and discuss how results from intelligibility studies relate to care of articulation.

4.3.2 Intelligibility

Intelligibility as a measurement has been used by, amongst others, Bard *et al.* (1995) in the investigation of givenness, by Fowler and Housum (1987) also in the investigation of givenness, and by Bradlow *et al.* (1996) in looking for sources of its variability between speakers. The measurement of intelligibility is also used in studies of hearing disability and in human factors (e.g. Moore *et al.*, 1994; Payne *et al.*, 1994).

If something is easy for someone to recognise it is regarded as being intelligible, if it is impossible to recognise it is unintelligible. Intelligibility is a measure of this continuum including these two extremes.

Different experiments have used different methods to measure intelligibility. Fowler and Housum (1987) excerpted words of interest from their context and played them to the subjects at the rate of one every five seconds. The subjects wrote down what they thought the word to be and also how confident (between 1 and 5) they felt concerning the choice they made. Bradlow *et al.* (1996) asked subjects to transcribe whole sentences and chose five key words from each sentence. The sentence was scored as correct if all five key words were transcribed correctly. Bard *et al.* (1995) also excerpted words in the same way as Fowler and Housum but also added noise to the recording to make the words less easy to recognise.

In all three methods the transcriptions for the same utterances were pooled. The accuracy of transcription over all subjects is then regarded as a measure of intelligibility.

Clear speech and intelligible speech are related. For example: One of the conclusions reached by Bradlow *et al.* (1995, p201) is "...female speakers, who tend to have more precise articulations, also have higher overall intelligibility scores than males." This term 'more precise articulation' is close to the concept of 'clear speech'. Fowler and Housum (1987, p489) also make reference to the acoustics of intelligibility. "...talkers aim to provide an acoustic signal for a word that is sufficiently informative for listeners to identify the word." Implicitly it is not the choice of word or choice of sentence structure that is being examined here but the information in the acoustic signal resulting in differences in intelligibility. Intelligibility variation is being regarded as articulatory variation. Thus, when measurements of intelligibility are made within lexical item and within speaker, intelligibility variation reflects differences in articulation and the resulting acoustic change in the word. The interest in this change is spurred by the fact it appears non-random and related to discourse structure. This acoustic variation appears to be there for a purpose.

Because of the close relationship between hyper-articulated clear speech and speech which is more intelligible results from studies examining intelligibility have a bearing on work presented here. I will therefore include descriptions of some of this work below when looking at the acoustic properties of carefully articulated speech. However, when relating results from intelligibility studies to acoustic and articulatory studies of clear speech care is required. Often noise is added to the token that is used in an intelligibility experiment. The effect of this noise might well interact with the speech acoustics. For example if a fixed level of noise is used phonemes with higher amplitudes such as low vowels will be less affected. In contrast if noise is added dependent on the amplitude in the signal consonants such as plosives will retain more of their characteristic structure. Another problem in interpreting intelligibility results is a 'ceiling' and 'floor' problem in the measurement. A word can only get so intelligible that all subjects recognise it or so unintelligible that no one can recognise it. Acoustic and articulatory measurements in contrast may continue to change even when a word's intelligibility falls outside these bounds.

4.3.3 Carefully Articulated Vowels in Clear Speech and Intelligible Speech

In 1963 Lindblom put forward a target undershoot model of vowel articulation (Lindblom, 1963) (for more detail on the actual modelling process used see chapter 5 section 5.4.2.3). In this study Lindblom suggested that each vowel had a set of spectral targets that the articulators attempted to produce. If the duration of the vowel was reduced it became impossible for the tongue to reach the correct position in time and the spectral target was undershot. Further studies produced conflicting results with regards to this model. Some studies found no undershoot (e.g. Fourakis, 1991), others found that undershoot appeared to be speaker dependent (Flege, 1988) while yet more confirmed Lindblom's basic theory while presenting a more complex model of the undershoot phenomena (For example Broad and Clermont, 1987; van Son, 1993).

In response, Lindblom (1990), in his H&H theory, suggests that speakers can use different degrees of articulatory effort when producing speech. An explanation for inconsistencies in the results cited above could be that such differences in articulatory effort were not controlled for. When speakers were asked to produce tokens in a laboratory environment it is difficult to establish how much effort they made in trying to produce 'good' tokens.

In order to examine this problem Moon and Lindblom (Moon and Lindblom, 1994) specifically elicited speech which was hyper-articulated. In contrast to normal read speech this 'clear speech' was the result of the experimenter asking speakers to repeat a token because it was not understood. They used contexts for each vowel that would intensify formant transitions (see below for more detail) and they found evidence that:

- Vowels in clear speech were longer and displayed less average undershoot.
- There were clear differences between the amount of undershoot exhibited for different speakers.
- Tense vowels showed less duration independent undershoot than lax vowels.

The general effect of undershoot across a speech sample is for the spectral characteristics of the first two formants of the vowels to exhibit reduction. For example for the vowel /i/ the F2 value (averaged over speakers and different word lengths) was 223 Hz lower in citation speech in contrast with clear speech (Moon and Lindblom, 1994) in a /wVl/ context. Over five speakers and four front vowels this reduction was significant in all but two cases out of twenty. Because F1 and F2 values will generally have less extreme values in these contexts, if a two dimensional space described by F1/F2 is plotted, they group more strongly in the central area. This tendency to move towards the centre of the vowel space is termed centralisation and the tendency for less extreme F1/F2 values for a vowel is termed spectral reduction. Moon and Lindblom also found differences in vowel duration between clear and citation speech varying from 9 ms to 109 ms (from 6% to 40% reduction) depending on speaker and vowel type.

Moon and Lindblom's results for clear speech reinforce results from Picheny *et al.* (1986) and Bond and Moore (1994) who examined the acoustic characteristics of clear and conversational speech. In this study, amongst other effects (see below), vowels exhibited more spectral reduction (F1 ≈60Hz, F2 ≈200Hz) and shorter segmental durations (≈10-100ms, 10%-60% reduction) in conversational speech.

Bradlow *et al.* (1996) also report that in more intelligible speech the vowel space is more spread out than in less intelligible speech. This implies that less vowel reduction occurs in intelligible speech. As demonstrated by Moon and Lindblom (1994) this can either be a result of longer segmental duration or of increased articulatory effort or both.

All the studies reported above used speech read in a 'normal conversational' manner as a contrast to the clear speech. Very little work has been carried out on genuine spontaneous speech. Sotillo (1997) when examining intelligibility of spontaneous speech tokens found that differences in vowel duration significantly related to intelligibility of spontaneous tokens and were significantly longer in carefully produced citation forms.

To summarise, for clear speech, vowels generally have more distinct spectral characteristics and are longer (Picheny *et al.*, 1986; Bond and Moore, 1994; Bradlow *et al.*, 1996; Moon and Lindblom, 1994; Sotillo, 1997). However what this means for particular instances of vowels is less clear.

Vowel identity as noted by Moon and Lindblom (1994) has a measurable effect on changes in duration and spectral characteristics between clear speech and normal citation speech. Furthermore as I will discuss later (in section 4.4) prosodic factors also have a very important effect on both duration and spectral characteristics.

4.3.4 Consonants in Carefully Articulated Speech and Intelligible Speech.

Research in the acoustics and articulation of consonants is complicated by the sheer variety of acoustic cues and articulatory mechanisms for producing non-vocalic sounds. For example the spectral structure and the variation over time in the speech signal is completely different between an /s/ and a /b/. Therefore work investigating the effects of hyperspeech on consonants has tended to concentrate on particular cues or particular examples of articulation.

Recent work has examined differences in the release of obstruents (Picheny *et al.*, 1986; Bond and Moore, 1994; Bradlow *et al.*, 1996; Sotillo, 1997), differences in voice onset times in obstruents (Bond and Moore, 1994; Bradlow *et al.*, 1996), durational differences in the duration of interword /s/ (Bradlow *et al.*, 1996) and place assimilation of word final nasals (Sotillo, 1997).

The most widely cited work which investigated the acoustics of clear speech was the study carried out by Picheny, Durlach, and Braida (1986). They looked at obstruents in clear and conversational speech and found that word final stop bursts were released more often and the RMS energy[1] of obstruents is greater for clear speech. Bond and Moore (1994) also found obstruents in more intelligible speech tended to be released more often and had a more distinctive voice onset time (VOT).

Looking at differences between spontaneous speech and carefully read speech Sotillo (1997) found that stops were more likely to be deleted in spontaneous speech. An examination of place of articulation change in nasals was more problematic. The difficulties of measuring differences in place of articulation using the acoustics proved difficult and this part of the study remained inconclusive.

Because of the large number of speech cues involved in consonant recognition and many different factors in consonant production the work described above leaves many questions with regards to specific cues unexplored. It is also uncertain how cues described above might be combined to give an overall measure of the care of articulation in a whole word. These problems as well as the relative ease of measuring duration has encouraged the use of durational measurements as indications of clear or unclear speech.

4.3.5 Duration Differences in Carefully Articulated Speech and Intelligible Speech.

Segments and words tend to be longer in clear speech than other speech styles (Picheny *et al.*, 1985, 1986; Uchanski *et al.*, 1996; Moon and Lindblom, 1994; Sotillo, 1997; Bond and Moore, 1994; Cutler and Butterfield, 1990).

This increase in duration has been noted on vowel durations (Picheny *et al.*, 1985, 1986; Uchanski *et al.*, 1996; Moon and Lindblom, 1994; Sotillo, 1997; Bond and Moore, 1994). This effect differs substantially between lexically stressed and unstressed vowels in percentage terms and also appears to be dependent on vowel type (Moon and Lindblom, 1994; Sotillo, 1997, see section 4.4). Consonants have also exhibited lengthening. For example Picheny *et al.* (1986) show an increase in the length of /s/ in 'pass' in the context of the sentence "His quick world must pass in a flag" when spoken as clear speech.

[1]RMS energy is root mean squared energy. For speech which oscilates around 0 this represents the average variance in the signal and thus amplitude over time. Thus obstruents with a high RMS energy will have louder bursts.

This duration increase is also noted on (non-phrase final) word final syllables (Cutler and Butterfield, 1990) and over whole word durations (Picheny *et al.*, 1985, 1986; Uchanski *et al.*, 1996; Sotillo, 1997; Bond and Moore, 1994) when words are spoken in a clear speech style.

However examples exist of words which are longer and less intelligible. In data examined by Bard *et al.* (1995) where intelligibility was measured between clearly spoken citation forms and 2nd mentions of the same words in spontaneous speech 14% of the words which were shorter in the 2nd mention condition were actually more intelligible than their citation controls.

Similarly duration change is not a necessary result of clear speech at the segmental level. Lindblom specifically argues that distinctiveness is the primary characteristic of hyper-articulated speech (Lindblom, 1990). Although lengthening tends to occur as a side effect of more carefully articulated speech it can also occur when care is not being taken. For example Flege (1988) showed that vowel undershoot, although related to vowel duration, could be controlled differently by different speakers. Some speakers can and do articulate carefully as well as quickly.

Also Bradlow *et al.* (1996) also showed that increasing the duration (110ms-180ms) of a word initial /s/ in 'seems' in the context of the sentence "The play seems dull and quite stupid" led to more mis-recognitions of 'play' (it being recognised as 'place') and that careful articulators exerted more control on segmental timing making it shorter in this context.

I will return to issues in using duration as a care of articulation measurement in the following chapter. Despite the points raised above the general consistency of duration reflecting a general increase in care of articulation make it an attractive care of articulation metric.

4.3.6 Summary

I have given a brief description of the key terms used in research that has examined clear speech, intelligible and hyperspeech followed by an overview of current work which has examined the acoustic differences between clear speech and other speech styles. A great deal of consensus exists on general characteristics but less so when using these characteristics predictively. In general clear speech has longer duration and more spectrally distinct segments. Lindblom (1990) argues that control of reduction is oriented to the listeners' needs and that it is sufficient distinctiveness that drives the extent speech is hyper or hypo-articulated.

Before addressing the question of how predictability appears to affect care of articulation in line with some of Lindblom's predictions we will examine the direct relationship between prosodic factors and the acoustic characteristics of clear speech.

4.4 The Acoustic and Articulatory Effects of Prosodic Structure

Howell and Bonnett (1997) point out:

"All the factors discussed by Picheny, Durlach, and Braida (1986) show that prosody differs between clear and unclear speech. Stress appears to be particularly important in interpreting the results of Picheny, Durlach, and Braida (1986), as a word that receives high stress is usually found to have a higher pitch, its syllables are lengthened, it is likely to be louder and will probably be a content word rather than a function word. An unstressed word, on the other hand, will have reduced vowels, and often final plosives are not released. Thus, all the differences between clear and unclear speech that are discussed by Picheny, Durlach, and Braida (1986) are associated with differences in stress." (Howell and Bonnett, 1997, p96)

There is indeed much evidence to suggest that differences in the articulation and acoustics of clear speech can be attributed to prosodic structure (Lehiste *et al.*, 1976; Wightman *et al.*, 1992; Price *et al.*, 1991; Beckman and Edwards, 1990; Cutler and Butterfield, 1990; Summers, 1987; de Jong, 1995; van Bergem, 1988).

It should be noted that in the studies described in section 4.3 a detailed analysis of the prosodic structure in terms of accenting and boundary tones was not carried out. Prosodic structure was instead controlled implicitly through careful choice of word identity and carrier phrase. The only exception to this was the work by Cutler and Butterfield (1990, 1991). Here relationships between strong and weak vowels as well as effects of f0 caused by accenting are considered, however, the prosodic analysis did not extend to prosodic boundaries. This leaves open the fundamental question:

Is prosodic structure the means with which we change the clarity of speech?

As we will see there is persuasive evidence that this may be the case. In the following section I will review current literature which investigates the effect of prosodic structure on duration change and segmental spectral characteristics. For an overview of the terminology used in the following discussion on prosodic structure and for a review of the broader issues within prosody see chapter 3.

4.4.1 Prosodic Boundaries

As mentioned in section 4.3.1 clear speech generally exhibits lengthening compared to normally articulated speech. Clear speech also contains more and longer pauses (Picheny *et al.*, 1986; Cutler and Butterfield, 1990). Both lengthening and pauses are also associated with prosodic boundaries (Lehiste *et al.*, 1976, Shattuck-Hufnagel and Turk, 1996 for a review, Wightman *et al.*, 1992, Price *et al.*, 1991 among others).

Preboundary lengthening has been shown to occur at the end of an intonational phrase (Lehiste *et al.*, 1976; Wightman *et al.*, 1992; Price *et al.*, 1991; Beckman and Edwards, 1990), at the end of intermediate intonational phrases (Wightman *et al.*, 1992; Price *et al.*, 1991; Beckman and Edwards, 1990) as well as on word final syllables in polysyllabic words (Wightman *et al.*, 1992; Price *et al.*, 1991; Cutler and Butterfield, 1990; Beckman and Edwards, 1990).

The work by Wightman *et al.* (1992) deserves a more detailed description here as it serves as the most commonly cited piece of work (together with Price *et al.*, 1991) supporting the notion of prosodic hierarchy based on boundary-related lengthening. It is also of direct relevance to attempts to normalise duration measurements which will be discussed in chapter 5.

The work was based on a corpus of read speech developed by Price *et al.* (1991). The corpus consisted of 35 pairs of phonetically similar but syntactically ambiguous sentences. The sentences were read by four professional news announcers. These were then autosegmented and coded using seven levels of break index. The first five of these levels (0-4) correspond to break indexes described in chapter 3. The remaining two, *level 5* delimited a group of intonational phrases found in long sentences and *level 6* was reserved for marking sentence boundaries. The durations of all phones were normalised both for segment identity and speaking rate over the sentence (see Wightman and Ostendorf, 1991, for details).

The results showed that preboundary syllables were longer than similar syllables in different contexts. This lengthening appeared to be limited to the rhyme of the syllable and vowel length in particular showed significant lengthening between break indexes 1 to 4. A significant effect was only noted between breakindex 0 and 1 for stressed syllables and the authors suggested that differences between break indexes 4 to 6 could be marked by pause rather than extended lengthening.

To what extent gross changes of duration in clear speech may be directly caused by changes in prosodic boundaries remains unclear. No direct comparison has been carried out between the prosodic structure in clear speech as opposed to citation speech that I'm aware of. (Although studies of effects of speech rate on prosodic structure have been carried out see Shattuck-Hufnagel and Turk, 1996 for a review and also Caspers, 1994).

4.4.2 Prominence

Prominence also has a direct effect on acoustic factors linked with clear speech and careful articulation. Prominent syllables are longer (de Jong, 1995; Summers, 1987; van Bergem, 1988) and the vowels are less spectrally reduced (van Bergem, 1988; de Jong, 1995). Prominence is also associated with less spectral tilt (Campbell and Beckman, 1997), f0 transitions (e.g. Cruttenden, 1986; Ladd, 1996) and increased amplitude (de Jong, 1995).

Articulatory studies have also associated prominence with changes in articulation. The duration, velocity and spatial extensiveness of jaw opening is increased (Summers, 1987; de Jong, 1995) and the openness of the vocal tract increases (Beckman *et al.*, 1992). This results in increased acoustic power and more extreme spectral features in vowels (de Jong, 1995). de Jong argues that this shift in spectral features is made in order to increase perceptual clarity and is better regarded as hyper-articulation than a simple increase in amplitude. This view supports the acoustic findings of van Bergem (1988).

van Bergem (1988) carried out a detailed study on the effects of sentence accent (phrasal stress) and word stress (lexical stress) on vowel reduction. The study investigated 3465 vowels read by 15 male speakers. Both stress conditions had a significant effect on the steady state formant frequencies (F1,F2) of the vowels as well as on the vowel durations. He reports that lexical stress had a stronger effect than phrasal stress. However this analysis deserves some explanation. van Bergem treats phrasal stress as affecting the whole word thus allowing the context -[lexical stress] together with +[word stress]. Usually phrasal stress is regarded as being associated only with a lexically stressed syllable (Cruttenden, 1986; Ladefoged, 1982). In this more traditional view what van Bergem regards as -[lexical stress] and +[word stress] can also be regarded as spillover over from the accented syllable onto a neighbouring unstressed syllable (see Chapter 3 and also Turk and White, 1999 as noted below).

Overall, prominence meant that the vowels were less reduced and longer. This is the same effect reported by Moon and Lindblom (1994) for lax vowels. Ladefoged (1982) argues that the reduced/full vowel distinction is a level of prominence (see chapter 3 section 3.2.2). Tense vowels in unstressed syllables are often reduced vowels. Thus the differences reported by Moon and Lindblom between tense and lax vowels could be regarded as related to the effects of prominence at this reduced/full level rather than as an effect of vowel identity.

Prominence and Boundary effects can't, in fact be viewed in isolation. Turk and White (1999) show that the domain of accentual lengthening is affected by word boundaries. An accent placed on a syllable affects the syllable with which it is associated but also has an effect on unstressed syllables within a word. This is the similar to the effect reported by van Bergem (1988) (see above). This effect is much stronger in a rightwards direction unless attenuated by a word boundary where the effect is much smaller (see also Turk and Sawusch, 1997).

There is much evidence to demonstrate that prosodic boundaries and prominence both affect the duration of syllables and segments. Both also occur with f0 changes. Given this one may ask how we can tell the difference between a boundary effect and a prominence effect. It is possible that other acoustic factors such as reduction behave differently in a prominence as opposed to a boundary context although this is yet to be established. Work carried out by Fougeron and Keating (1997) suggest that boundary effects also include articulatory factors beyond increased duration which might suggest more careful articulation and could possibly lead to more spectrally distinct segments. However in their study a small amount of reiterant speech was used (e.g. 'nono no' instead of 'ninety nine'). This use of reiterant speech as well as the use of numerical sentences (e.g. 'ninety nine times ninety nine times ninety nine equals a lot') may have affected their results.

Overall given the wealth of prosodic factors and the large effect they appear to have on the same acoustic factors examined in clear speech and care of articulation there is a possibility that prosodic structure could account for these changes. However as I will discuss below redundancy also has a strong effect on these factors. To what extent can redundancy alone explain variation in care of articulation?

4.5 The Acoustic and Articulatory Effects of Redundancy

As we have seen in chapter 2 patterns of redundancy and predictability in natural language are complex. However, even given this, a number of studies have persuasively shown that more predictable sections of speech exhibit the same acoustic reduction and shortening that is common in hypospeech and avoided in hyperspeech. A criticism of this work, especially the much cited Lieberman (1963) is the lack of prosodic controls.

4.5.1 An Informal Observation

Bolinger (1963) points out:

"The more redundant something is, the shorter it tends to be, and conversely: 'the factor of novelty is relevant to the prolongation'[2]. I note two manifestations: the familiarity of a particular form or phrase, and the familiarity of a particular combination. An example of the first is the fusion of polymorphemic words. The relatively infrequent *sugar loaf* in my speech tends to be longer than the frequent *sugar lump*. For me, the relatively new and unfamiliar *robot* is slower and more disjointed at the syllable boundary than in *rowboat*, despite the fact that *rowboat* contains two morphs and *robot* one. The fusion of highly frequent individual verb-adverb phrases illustrates the same thing:..." (Bolinger, 1963, p7).

In this work no formal model of redundancy is appealed to and no formal phonetic laboratory study is carried out to establish the patterns of lengthening Bolinger describes. However considerable evidence from laboratory studies does indeed support Bolinger's observation. One much cited study is that of Lieberman (1963). Here Lieberman establishes an index of redundancy for a number of words in a different contexts by asking subjects to predict them from these contexts and then explicitly excerpts the words and plays them to subjects to see how intelligible they are.

4.5.2 Lieberman, Hunnicut and related studies

Lieberman (1963) used different contexts such as "A stitch in time saves ..." and "The number you will hear is ..." to elicit redundant and non-redundant tokens

[2](Sharp, 1960, p131)

of a word (In the above example *nine*). The prosodic context was not explicitly controlled for and in some examples could well have confounded the results. For example, the prosodic context of the word *budget* in the two sentences 'A wise and balanced budget is the core of good government' and 'Robert Budget is in jail' could be significantly different. In the first it could be marked with a nuclear accent and be followed by phrase boundary, in the second it is probably marked with a non-nuclear accent and is less likely to be followed by a phrase boundary.

Lieberman used 60 native U.S. speakers to guess the word from the contexts (in the *nine* example 85% guessed the word in the first context and 10% in the second). Listeners were played these words and asked to write down what they heard (in the above example 50% recognised the non-redundant token while only 33% recognised the redundant token).

The duration and peak amplitude of the words were also measured. Out of nineteen pairs 10 were longer when less redundant, 6 remained the same length and only 3 were longer in the redundant context. In terms of peak amplitude 9 were louder, 5 had the same amplitude and 5 were quieter when less redundant. In all, 15 tokens out of 19 were easier to recognise when excerpted from less redundant contexts.

Lieberman uses subjects to assess redundancy rather than calculating the redundancy on the basis of corpus statistics but he nevertheless appeals to two alternative models. The first, where only left context is given to the subject in order for them to guess the word, and the second, where both left and right context are given. In this work only the full context consistently reflected redundancy in the materials. For 9 out of 14 contexts, left context did not provide sufficient information for any of the 30 redundancy checkers exposed to this context to guess the word. This highlights a problem with a psycholinguistic approach to measuring redundancy. In many cases the chances of predicting a word from any context is small because the lexicon is large. For example the chances of any of the 30 listeners to guess that 'neither a' is followed by the word 'borrower' is low (In Lieberman's study none guessed the word given this context). However that does not mean the redundancy from a left context is 0; it just means that it is probably less than 0.03. A very large number of subjects would be required to give results for small probabilities (sometimes in the tens of thousands!). This also raises another objection to Lieberman's study. Only four sentences were used to establish the most redundant contexts and of these two were well known adages ("A stitch in time..." and "Neither a borrower nor a lender be.") and a third a fairly unusual compound ("witch hunts"). Given the highly redundant nature of these contexts it is difficult to be sure they are representative of redundancy effects in general.

In response to these criticisms Hunnicut (1985) followed up this study by looking at a larger set of Swedish sentences (80 were used in the analysis) in wider contexts (21 pairs were adages and 19 were text-type sentences). In this study care was taken to match sentence structure between examples of high and low redundancy contexts. This had the effect of producing similar prosodic contexts for both words. However the sentences were quite long and no prosodic analysis was carried out on the speech produced to establish that the prosodic context was produced as assumed. Overall the results support the notion of high redundancy, low intelligibility and thus implied poor articulation although only in the text-type context.

Her conclusion was:

"The results of the current study indicate that the relationship of intelligibility to redundancy is not clear. There may be dependency in certain conditions but not others. The question that has been asked in this study, and also in the Lieberman's study, concerns the intelligibility of a word in isolation and its dependency upon factors of redundancy in context. That is, redundancy is defined as the percentage of essential information present in a sentence without the test word. Then we can say that in the non-idiomatic, non-metaphorical sentences of a reader, these results indicate that there is a clear intelligibility advantage for words in lower-redundancy contexts." (Hunnicut, 1985, p53)

The problems in using intelligibility as a metric are highlighted in this paper. Blanket pink (speech like) noise was used to make the words harder to recognise. It was found that the signal to noise ratio fell significantly towards the end of the sentence. Thus the effect of added noise might have a much greater effect depending on sentence position and confound intelligibility results.

The final results of these studies are interesting but inconclusive. This is partly due to a number of non-trivial problems in this methodology:

1. Intelligibility measurements are noisy. Even with appropriate controls (which were not used in these studies) ceiling effects and the need to mask words with noise introduce serious fluctuations into individual results.

2. All materials are read. It is quite possible that tokens in spontaneous speech reflect redundancy differently. For example they may already be too reduced

to be affected any further.

3. The informal models of redundancy, although attractive because they can systematically reflect subjects' complete language knowledge also are strongly affected by low probabilities and unusual contexts.

4. No control was carried out for prosodic structure, which, as I have outlined above, is well known to affect articulation. Variations in prosody could confound some of these results.

This thesis specifically attempts to address these problems. In the next chapter a number of automatic measurements are developed to try and produce a consistent approach to measuring care of articulation across a large corpus of speech. The speech itself is spontaneous, running speech taken from a relatively normal dialogue situation and redundancy is explicitly and formally coded. Finally prosodic structure is considered in detail and the relationship between its effects and those of redundancy on care of articulation considered closely.

4.5.3 Given and New: Repetition Studies

These studies examine the effect of discourse structure on the way a word is articulated. Fowler and Housum (1987) suggest that this variation in articulation is used by speakers to signal differences between 'New' and 'Old' or 'Given' information. For example in the HCRC Map Corpus the following type of conversation often occurs:

> Do you have a *disused monastery*?
>
> No.
>
> Well you need to turn left under the *disused monastery* and then go south.

The first mention of *disused monastery* is an introductory mention. The speaker has not mentioned this landmark earlier in the dialogue and so it is 'New' information. The second mention, in contrast, is referring back to something the speaker has already talked about and so this is 'Given' information. Given and New can be regarded as examples of redundancy at a 'higher level' than, for example, word frequency in that the redundancy in this case is also dependent on semantic, syntactic and discourse knowledge.

In this situation according to Lindblom's H&H theory we might expect an acoustically reduced form of the second mention because it is more readily inferable by the listener. This is the general finding; in spontaneous speech in the form of monologue (Fowler and Housum, 1987; Fowler *et al.*, 1997), in dialogue (Hawkins and Warren, 1994; Bard *et al.*, 1995; Bard and Aylett, 1999), read speech (Fowler, 1988), (Samual and Troicki, 1998)[3] and when spoken from memory (Shields and Balota, 1991). When no context existed to help predict the second mention such as in a list context (Fowler, 1988) then no reduction was observed. Again prosody was not controlled for in a majority of these studies. When it was (Bard and Aylett, 1999; Hawkins and Warren, 1994) conflicting results were obtained (see section 4.5.6).

In contrast to the studies described above 'Given' versus 'New' offers an easily determined difference in redundancy together with a useful control. If the same word is spoken by the same speaker differences caused by idiosyncratic articulation and word identity are controlled. In general, when spontaneous speech was examined this decrease in care of articulation for 'Given' mentions is more marked than in read speech (Fowler and Housum, 1987).

I will now turn to studies which have looked at redundancy and articulation with regards to formal probabilistic models in terms of both the lexicon and also in terms of phonemes.

4.5.4 Redundancy Caused by the Lexicon

Central to Lindblom's H&H theory is the idea that language is produced with a listener in mind and that speech should be sufficiently discriminable. Part of the task of any listener is to decide what words make up an utterance. The examples of redundancy discussed above considered the context surrounding the word. However structure within the lexicon can also increase the redundancy of individual words.

For example if you were told to guess a three letter word that had been randomly found in a book the word 'the' would be a sensible guess. 'the' is the most predictable word given this information and thus the most redundant. In general more frequently used words are shorter and undergo more severe articulatory reduction when produced (Balota *et al.*, 1989). The most common words used in English are function words such as 'the', 'and', 'to'. In spontaneous speech these three examples, rather than produced /ði/, /and/, /tu/ are often produced as /ð/,/n/,/t/.

[3]Reduction was only found for children and adults who had good control of the production situation.

The structure within a word, given the lexicon, also leads to redundancy. For example, if you were told to guess a three letter word ending in 'at' and beginning with 'c' or 'g' you would guess 'cat' because the word 'gat' does not exist in the normal lexicon. In contrast if the word ended 'ap' you could choose 'cap' or 'gap'. In the first example the c/g distinction is redundant in the second it is not. Measuring this redundancy is non-trivial. As discussed in chapter 2 redundancy is only meaningful with regards to a model. In terms of the lexicon there is considerable debate concerning models of word recognition and the different importance of different cues within a word. A detailed review of word recognition literature is beyond the scope of this work however a number of important findings with regards to care of articulation will be discussed.

Pisoni *et al.* (1985) found that a word's intelligibility was affected by the the *neighbourhood density*: the number of phonologically similar words in the lexicon and the *relative frequency*: the word's frequency compared to its nearest phonological neighbour. Words which had more competitors, in other words words with less redundant phonemic distinctions were more intelligible and thus more carefully articulated. Words which were relatively less frequent and thus less predictable than words they could be compared with were also more intelligible.

More direct articulatory measurements reinforce this result. Goldinger and Summers (1989) carried out a study looking at differences in VOT between voiced/voiceless minimal pairs. They asked subjects to read minimal pairs chosen from sparse and from dense lexical neighborhoods. Each subject read each pair four times. They found that the VOT difference between voiced/voiceless pairs was greater for pairs taken from dense neighborhoods than from sparse neighborhoods. However as Wright (1997) points out the study was flawed because the use of minimal pairs made the subjects aware of the distinction being studied and could cause them to exaggerate the contrast. Wright (1997) looked instead at vowel undershoot in sparse and dense lexical neighborhoods. He took monosyllabic CVC words of equal familiarity but varying in the density of their lexical neighborhoods. He measured the F1 and F2 values in the central region of each vowel in Bark and measured the Euclidean distance of each from the centre of the speakers' vowel space. He found a significant centralisation for the vowels from words taken from sparse lexical neighborhoods ($F(1, 480) = 130.92, p < .0001$).

However in contrast Sotillo (1997) found, in a clear contrast with predictions made by Lindblom's H&H theory, that: *"The degree of hypo-articulation... is independent of any kind of assessment of potential lexical competition."* (Sotillo, 1997, p270) when examining nasal assimilation. Sotillo (1997) carried out a perceptual experiment to measure perceived nasal assimilation in tokens taken from spontaneous task oriented dialogue. Sotillo found no significant effect of a close competitor set (a similar notion to a dense lexical neighborhood) on the degree of assimilation perceived. However the materials Sotillo used were spontaneous speech tokens excerpted from dialogue. This contrasts with the read words used in both the Wright (1997) and the Goldinger and Summers (1989) studies. It is unclear the extent such differences in materials lead to these different results. It is possible that normal spontaneous speech is already maximally reduced in many contexts thus making tendencies observed in read speech difficult to detect.

Sotillo does, however, present evidence that hypo-articulation (d-deletion and reduced vowel duration) is more prevalent in word offsets (which are more redundant) than word onsets but with an important caveat. Different parts of the word are more perceptually salient than others and hypo-articulation within the word is not just dependent on internal word structure and redundancy within the lexicon but also the acoustic identity of items within the word.

Even taking these complexities into account, assimilation per se does not necessitate poor intelligibility. Shillcock *et al.* (1994) argue than many types of assimilation, far from making words harder to recognise actually reduce the size of the neighborhood density surrounding the word. In other words the assimilation actually increases the amount of information in the word rather than reducing it. For example the labial assimilation of 't' to 'p' in *batman* results in smaller competitor sets if the word is represented in the mental lexicon as 'bapman'.

A degree of caution is required however when dealing with such assimilations with regards to a model of lexical access. It is unclear whether in actual spontaneous speech the 't' becomes a 'p' in the above example or whether it becomes some sort of stop which could be characteristic of both a 't' **and** a 'p'. The effect on redundancy is quite different in these two cases.

In all there is compelling evidence that regularities within the lexicon contribute to redundancy and that these differences in redundancy affect care of articulation. However there are several potential models of word recognition and it is therefore difficult to characterise the precise nature of the redundancy that occurs within a word. This, together with the difficulties comparing read speech studies with studies carried out on spontaneous speech, mean that we are far from a clear understanding of the precise relationships involved.

4.5.5 Summary

Research in articulation and in intelligibility has consistently suggested that redundancy affects care of articulation. In general if something is predictable from context or from the lexicon then care of articulation is reduced. A number of problems have been outlined in some of this work. In particular the lack of work that looks at spontaneous speech in a normal communicative environment and the lack of prosodic controls in the work discussed above. In the next section I will consider some research that has specifically looked at whether prosodic structure can account for the redundancy effects noted in some of the work described above.

4.5.6 Prosody, Intelligibility and Redundancy

Ladd states:

"...it is well known that accent tends not to be placed on elements that are repeated or 'given' in discourse..." (Ladd, 1996, p175)

This naturally raises the question of whether the intelligibility differences noted in given/new studies are a direct consequence of accenting differences. Hawkins and Warren (1994) and Eefting (1991) present evidence that this is indeed the case. In contrast, Bard and Aylett (1999) show that accent change alone does not explain intelligibility differences between given/new tokens as there is still a significant intelligibility reduction between accented first and accented second mentions. The differences in these results can be attributed to differences in the materials examined. The result obtained by Bard and Aylett was that accent change certainly did alter intelligibility but that in normal spontaneous dialogue deaccenting doesn't happen very often. 75% of the materials had no change in accentedness. Differences in these results could be attributed to very different sample sizes and means of eliciting the material. The studies described by Bard and Aylett examined the intelligibility differences of 408 pairs of repeated mentions produced by 64 speakers in task oriented dialogue. In contrast Eefting (1991) used 16 target words read by a single experienced newsreader and Hawkins and Warren (1994) examined 19 words produced by three subjects in a picture description exercise.

Differences in styles of production (whether read speech or spontaneous, whether monologue or dialogue), communicative setting and speaker differences make comparisons between studies difficult. Intelligibility studies are very resource intensive as are studies which involve the hand measurement of acoustic cues associated with hypo-articulation. Thus in much of the work described above small amounts of controlled material were necessarily used. To my knowledge no large scale study examining differences in care of articulation and relating them directly to different factors in prosodic structure and differences in redundancy has been carried out. This work seeks to address this.

4.6 Summary

There is considerable variation in the care with which sections of speech are articulated. Different acoustic measurements and differences in intelligibility have been directly associated with these differences in care of articulation. By using these measurements it has been shown that this variation is non-random and systematically associated with both prosodic structure and differences in the predictability of language. The extent to which these two factors are independent of each other remains unclear. This thesis will address this question. Firstly by suggesting a framework to explain a prosodic structure/redundancy relationship (chapter 2), secondly by coding and measuring prosodic structure, redundancy and care of articulation over a large corpus of spontaneous speech (chapters 2,3,5) and finally by carrying out a large scale quantitative analysis of these materials (chapter 6).

Chapter 5

Care of Articulation: Measurement

5.1 Introduction

In the previous chapters I have presented a compelling case for exploring the relationship between prosodic structure, redundancy and care of articulation. Extensive evidence has been presented that both redundancy and prosody affect care of articulation:

1. Sections of speech which are difficult to predict are generally articulated more carefully than redundant sections of speech.

2. Prosodically prominent sections of speech are generally articulated more carefully. Speech at prosodic boundaries tends to undergo lengthening which is associated with careful articulation (chapter 4 section 4.3.5).

However the main question, the extent prosodic structure implicitly represents the effects of redundancy and the degree redundancy exerts an effect independent of prosodic structure remains unanswered. In order to address this question using a quantitative framework we need to examine a considerable amount of speech. The factors we need to consider are the prosodic codes, detailed in chapter 3, the redundancy measurements, detailed in chapter 2 and finally the care of articulation measurements which are detailed in this chapter. Measuring care of articulation for almost every syllable in 15 hours of speech is a significant research task in itself. This task is addressed here.

5.2 The Options

A good care of articulation measurement should conform to the following criteria:

1. There should be extensive laboratory work that associates the measurement with careful or hyper-articulated speech.

2. In this work it should be possible to apply the measurement at the syllabic level and to as many syllables as possible. This is because, as already described in chapter 3 section 3.3.1.1, each data point in this work represents a syllable.

3. For practical purposes the measurement needs to be largely automated. The large amount of data considered here precludes any complex hand coding.

Two acoustic properties were measured, syllabic duration and vowel quality. For each two different metrics were used:

Syllabic Duration. Raw syllabic duration is the first duration measurement. However, as outlined below, syllabic content and context exerts a strong influence on a syllable's raw duration. Thus in addition to raw syllabic duration a duration measurement based on a combined log segmental distribution model was also used. This second measurement is more complex and tries to normalise duration to take into account the internal structure of individual syllables (see section 5.3.2).

Vowel Quality. A measurement of vowel centralisation is the first vowel quality measurement. This measurement reflected the distance of a vowel instance from the centre of a speaker's vowel space. The second measurement of vowel quality was a target measurement. In contrast to the centralisation measurement this metric measured how much a vowel instance undershoots a vowel target in a vowel space generated from clearly articulated speech (see section 5.4).

Syllabic duration and vowel quality were chosen because they both have consistently been shown to be associated with carefully articulated speech (see section 4.3.5 and section 4.3.3 in chapter 4) and because the measurements complement each other (see section 5.3.1 below). Duration is simple to measure but less directly connected with care of articulation. Vowel quality is more difficult to measure but more directly connected to care of articulation.

The measurements offered potentially good coverage of the data as most syllables have a vowel nucleus and all have a duration thus the measurements could be applied to most of the syllables in this study.

5.3 Measuring Care of Articulation using Syllabic Duration

5.3.1 Does Longer equal More Careful?

In general more carefully articulated speech or 'clear speech' is longer (see chapter 4 section 4.3.5). Word duration is greater in 'clear speech' than when the same word is spoken in spontaneous or citation speech (Uchanski *et al.*, 1996). At the phonemic level this increase in word duration can be ascribed both to:

1. Lengthening of individual phonemes. For example vowels, when taken from the same contexts, are generally longer in clear speech than in citation and spontaneous speech (Moon and Lindblom, 1994).

2. Less deletion and reduction. Segments such as word final /d/ in 'poisoned stream' have less tendency to be deleted (Bradlow *et al.*, 1995).

Thus, in general, syllables with longer durations are more like 'clear speech' and thus are more likely to be articulated more carefully.

There are, however, two major problems with using syllabic duration as a metric for care of articulation:

- Although lengthening tends to occur as a side effect of more carefully articulated speech it can also occur when care is not being taken. For example (Flege, 1988) showed that vowel undershoot, although related to vowel duration, could be controlled differently by different speakers. Some speakers can and do articulate carefully as well as quickly. In other words, speakers can mumble slowly; they just tend not to. Similarly an increase in duration does not *necessitate* an increase in care of articulation. Un-accented phrase final syllables may well be examples of longer but not more carefully articulated speech (see chapter 6 section 6.4.4.5)

  For this reason it was felt that syllabic duration, taken by itself, was a potentially unreliable measure of care of articulation, but, if analysed with another measure of care of articulation that did specifically address the issue of distinctiveness, it offered a simple and effective global measurement. In section 5.4 the alternative measurement of care of articulation, care of vowel articulation, is discussed. By analysing data with both measurements some of the weaknesses in a duration measurement are offset.

- Comparing duration change between different syllables, or even the same syllable in a different context, is hard. A normalisation process is required (see section 5.3.2). Otherwise it is difficult to argue that a specific syllable is longer in a particular context than you might expect or that a different syllable in the same context is lengthened in a similar fashion. If we cannot predict syllabic duration change in this way then we cannot compare the effect that prosodic structure and redundancy measurements have on this metric.

  A number of normalisation techniques have been explored (Campbell and Isard, 1991; Aylett and Bull, 1998). The technique used in this work (described below) is a compromise between simplicity and accuracy. There is certainly potential to improve this normalised duration measurement but it was felt that such work was a significant research task in itself and beyond the scope of this thesis.

5.3.2 Comparing Syllabic Duration between Different Syllables in Different Contexts

In order to compare duration change in different syllables it is necessary to generate a model of a syllable's duration and then calculate the extent the actual duration deviates from it. Because prosodic factors are explicitly part of the analysis carried out here they do not need to be (in fact must not be) included in the model. This leaves the following factors that might be accounted for in any normalisation procedure:

- Number of segments. For example, if a syllable has five segments in it rather than three we would expect this to significantly affect the duration.
- Phonemic identity. As the inherent duration of particular phonemes varies (e.g. Klatt, 1976) we might expect that the different phonemes present in a syllable would significantly affect the duration of the syllable.

Aylett and Bull (1998) present a number of different duration models that can be used to normalise raw duration scores based on work by Campbell and Isard

(1991). The basic model used a combined log distribution model of each phonemic segment, and assumed that a change in the duration of a word is divided equally among the segments of that word in terms of z-scores for each segment's duration. Therefore, the change between a word's predicted duration and actual duration could be measured in terms of a single z-score calculated for all of a word's segments. This value, called here the 'k-score', was used as a measure of how much a word had been 'stretched' or 'compressed' from a citation form.

The predicted duration, d, of any word may be expressed as:

$$d = \sum_{i=1}^{n} \exp^{(\mu(i)+k\sigma(i))} M \tag{5.1}$$

where:
n = the number of phonemes in a word,
k = a constant function of average segment length,
μ = the mean log duration of a segment,
σ = the standard deviation of the log distribution of a segment's duration
M = an optional multiplier which defaults to 1.

Aylett and Bull (1998) found that phonemic content, the fact 'beach' is made up of /b,i,ʧ/, was not as important as syllabic context (see below) when normalising duration. In Aylett and Bull (1998) it was found that syllabic factors such as:

- Whether the syllable was lexically stressed in the word.
- The position of the syllable in the word for example if it was initial, middle, or word final.
- Whether the word was monosyllabic.
- The number of segments in the syllable.

were more effective at predicting prominence, when used to normalise duration, than the actual phonemic contents of the syllable. This was particularly true for long words where segments are significantly reduced. The syllabic context was a fundamental factor in this reduction. More surprising was that combining phonemic content and syllabic context information produced only a minor improvement in results and appeared to be worse at generalising duration change across speakers and unseen data. This maybe because phonemic content is not independent of syllabic context. For example the phoneme ð occurs mostly as "th" in the word "the". Because of this the distribution calculated from a large numbers of observations of ð will underestimate the duration of this significantly in a stressed open class context e.g. the “th” in “mother”. This lack of independence between phonemic contents and syllabic structure is widespread. Taking the consonants **s,k** we find a marked difference in the frequency that syllables containing them are of a particular segmental length. 53% of syllables containing **s** are 2 or 3 segments in length whereas 73% of **k** syllables are this length and an enormous 94% of ð are 2 or 3 segments long. Because of syllabic structure, vowel and consonant distributions are also markedly different. For example 74% of syllables containing the diphthong **aI** (The 'i' in 'bite") are 3 segment syllables. This lack of independence between phonemic content and syllabic structure (in this case the number of segments in a syllable), together with the fundamental importance of syllabic structure in word duration, means that generalising durational effects on the basis of syllabic context is more effective than generalising durational effects on the basis of phonemic contents.

Phonemic identity could and, in the long term, should be used in such a duration model. However gathering data not confounded by these other factors is difficult and remains an avenue for further research. In this model each phoneme was regarded as being identical with the same log distribution (μ=-2.7478 (64ms) σ=0.5702 (-1 sd=36ms, +1 sd=113ms)) representing its characteristic duration.

The multiplier M depended on the number of syllables, whether the syllable was lexically stressed and the number of segments in the syllable. In order to use this model in this work it was necessary to ignore number and position of syllables in a word and lexical stress so that the resulting k score would not confound further analysis of prosodic factors which included this information. Therefore the multiplier was restricted to modifying overall syllabic duration based only on the number of segments in the syllable. The multiplier, calculated on the basis of data collected from the ATR database (Campbell, 1993) by Campbell (1992), regards three segments as the default and expects segments in longer syllables to be reduced while in shorter syllables to be extended (See table 5.1).

Multipliers					
Number of Segments	1	2	3	4	5+
M	1.60	1.14	1.00	0.93	0.87

Table 5.1: Multipliers for different number of segments in a syllable. For example if a segment is in a three segment syllable the multiplier is 1.00, if it is in a four segment syllable the multiplier is 0.93 (see equation 5.1). These multipliers are derived from duration results presented by Campbell (1992).

5.3.3 Summary

In this chapter I have put forward arguments for using a syllabic duration score as a metric of care of articulation and described a means of calculating a normalised score for each syllable in the corpus. Using such a metric raises two major difficulties:

1. In some cases longer duration clearly would not indicate more carefully articulated speech.
2. Current duration models are rough approximations and will introduce noise caused by errors in the model.

We address the first problem by using this metric together with a metric specifically designed to measure care of vowel articulation (see sections 5.4.4.1.1,5.4.4.1.2). By using two different approaches, vowel quality and duration, to measure care of articulation some confidence can be ascribed to results that are returned by both, and interest to results that differ.

The second problem is more difficult to address. In order to control for noise introduced by the model a simple raw syllabic score (from here on termed DUR1) was also used. By looking at the differences between this raw score and the normalised score (from here on termed DUR2) in the final analysis we can get a feeling for the extent the normalisation helps reduce or add noise. For example, if prosodic factors are much better at predicting variation in DUR1 than in DUR2 we would suspect that the normalisation process was not working very well. Again, where both measurements predict the same behaviour, it is possible to be confident concerning the direction and type of relationship.

5.4 Measuring Care of Articulation using Vowel Quality

5.4.1 Introduction

Vowel quality relates to the spectral characteristics of a vowel. There is evidence that, in carefully articulated speech, the quality of vowels is measurably different from the quality of the same vowel spoken in a less careful context (such as in a spontaneous speech style). See section 4.3.3 in chapter 4 for review.

In order to assess vowel quality we need to decide:

1. What spectral characteristics we will use to characterise vowels.

2. Where or how, in the constantly varying spectral characteristics of normal speech we will measure these characteristics.

3. What we will use as a reference point to compare these measurements to.

The approach taken here is as follows:

- Speech is pre-processed using:

    1. An LPC based formant tracker.

    2. A frequency to Bark transformation. This scale better reflects differences in the perception of frequency. Although the model described here is a production model there is an implicit assumption that care of articulation is connected to discrimination. This makes a perceptually based scale more appropriate.

    3. A parametric curve fitting algorithm to calculate the achieved F1 and F2 targets of each vowel.

- These values are normalised and compared to a normalised model of a speaker's vowel space based on the speaker's citation speech. This comparison produces two values:

    1. A measure of how centralised each vowel is (see section 5.4.4.1.1).

    2. A measure of how likely a vowel was produced as a carefully articulated vowel (see section 5.4.4.1.2).

Much of this section is a fuller account of work already presented in previously published papers. Two specifically concentrated on presenting the modelling technique (Aylett, 1996, 1998), another looked at the modelling technique from an information theory approach (Aylett, 1999), and (Aylett and Turk, 1998) presented the evaluation of the care of articulation measurement based on the model. What follows here is the most up-to-date account of this work.

Crucial to the approach used here is the acoustic model of each speaker's vowel production. Before looking at the methodology used to generate this model I will first present a brief review of recent work carried out in this area.

5.4.2 Acoustic Models of Vowel Production

The leading model in research on vowel production is that of target undershoot in articulation (e.g. Lindblom, 1963; Broad and Clermont, 1987; van Son, 1993). This approach takes formant values as the main means with which to describe the spectral characteristics of vowels. The attraction of formant models is that the first two formants can be related directly to the articulatory movement of the tongue as it produces vowels. This allows the use of acoustic data to generate articulatory models. Before discussing these undershoot models it is useful to review the use of formants to characterise vowels. I will do so firstly in terms of the resulting *vowel space* which is created in the two formant approach and secondly in terms of *formant transitions*.

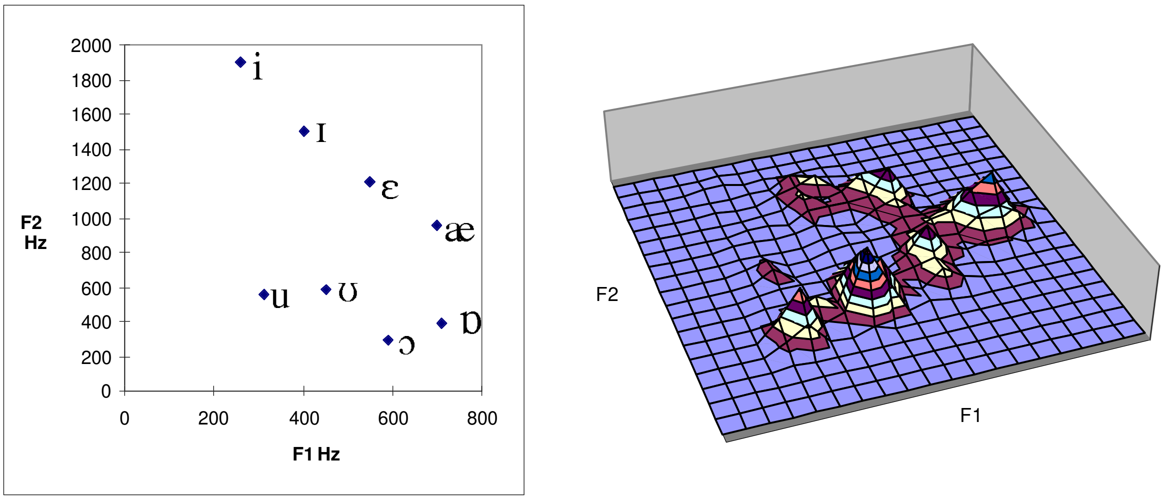
5.4.2.1 The Vowel Space.

Different vowels have different characteristic spectral qualities. Areas within the spectrum of a vowel with relatively high energy frequency components (i.e areas around these peaks) are termed formants. (For a more detailed definition see Ladefoged, 1962). In vowels the frequency of formants, generally the first and second formant (F1, F2), can be used to categorise vowels.

"For vowel sounds generally, and this is true of the English system, a significant part of the information listeners use in distinguishing the sounds is carried by the disposition of F1 and F2" (Fry, 1979, p78).

The higher the tongue in the mouth when producing the vowel the lower F1. The further forward the tongue in the mouth when producing the vowel the higher F2. So, for example, /i/ (in heed) which is a high front vowel (i.e. the tongue is high and to the front when producing this vowel) has a high F2 and a low F1 while /ɒ/ (in hod) which is a low back vowel (i.e. the tongue is low and to the back when producing this vowel) has high F1 and a low F2. It is possible to plot the F1 value against the F2 value of different vowels (See Figure 5.1a).

This two dimensional space can be referred to as the vowel space. The triangular shape made by the three vowels /i, u, ɒ/ (heed, who'd, hod) is often referred to as the vowel triangle. A scatter plot of F1/F2 values from vowels in citation speech show how actual values produced relate to the vowel space. If the density of the scatter is plotted as a third dimension, a 3D plot of the vowel space is produced (figure 5.1b). In this plot the hills show locations of high density. In general the values of F1 and F2 making up each hill will correspond to an example of a



(a) (b)

Figure 5.1: **(a)** The 'vowel space'. A formant chart showing the frequencies of the first and second formant for eight American English vowels. heed /i/, hid /ɪ/, head /ɛ/, had /æ/, hod /ɒ/, hawed /ɔ/, hood /ʊ/ and who'd /u/. **(b)** A three dimensional view of citation speech. A scatter plot of F1/F2 values from vowels in citation speech show how actual values produced relate to the vowel space. If the density of the scatter is plotted as a third dimension a 3d plot of the vowel space is produced. No scale is marked as data is first Bark transformed and then normalised.

particular vowel.

5.4.2.2 Formant Transitions

Vowels are traditionally described as having potentially both steady state and transition regions. Formants do not remain at a static value within a vowel but instead change value at the edge of the vowel and in the case of diphthongs within the vowel. The transitions at the edge of a vowel reflect the articulation of the surrounding phonemes. In fact these transitions play an important role in consonant recognition. For an example of formant transitions see Figure 5.2.

A target model of vowel production assumes that the formant is moving towards and away from an ideal value that describes this vowel. Thus the ideal target value may not be reached depending on such factors as phonetic context, vowel duration and care of articulation. If the ideal value is not reached then the formant is said to undershoot the target (see section 5.4.2.3 for more detail on target-undershoot models).

The effects of care of articulation on vowel quality described in chapter 4 section

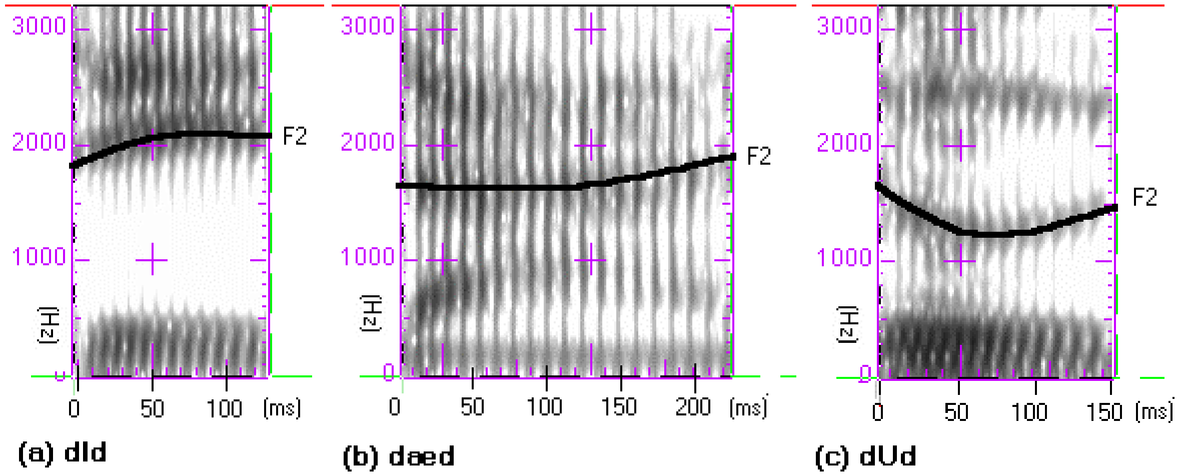


Figure 5.2: Formant transitions for (a) dId, (b) daed, (c) dUd. The figures each show the spectrogram of the vowel. The second formant (F2) is marked by hand with a black line to show the transition of the formant at the edges of the vowel.

4.3.3 can be explained by undershoot. In the studies that examined F1/F2 values in carefully and less carefully articulated vowels it was found that the formants in the central region of the vowel tended to be less extreme in less carefully articulated speech and closer the centre of the vowel triangle. This *centralisation* could be caused by the formant not reaching the extreme vowel target that it would in carefully articulated speech. This occurs because the speaker makes less effort to move the articulators to the extremes required to produce these ideal values.

In order to find these representative F1/F2 values of vowels a method is required to model the transitions described above. The method used in this work involves fitting a parametric curve to the formant values and is described in detail in section 5.4.3.2. This approach is based on target-undershoot models of vowel production (e.g. Lindblom, 1963; Broad and Clermont, 1987; van Son, 1993) described below.

5.4.2.3 Target-Undershoot Models of Vowel Production

The ideal F1/F2 values for a vowel can be described operationally as the F1/F2 values reached when a vowel is articulated slowly, clearly and in a context which has little effect on the formants in the initial part of the vowel such as the vowel (V) in /hVd/. The extent such ideal targets are speaker independent is complicated by factors such as age, sex, f0 range, native language and accent. Individual speaker characteristics aside, undershoot is also related to phonemic context, vowel duration and, crucially for this work, care of articulation.

Let us first consider vowel duration effects. Lindblom (1963) showed that, in general, the shorter the realisation of the vowel the greater the undershoot. Lindblom used several sentence frames to generate eight different Swedish lax vowels in a /b-b/, /d-d/, and /g-g/ context of between 80 to 300ms. In order to elicit these different durations he produced materials with the same vowel context in different stress conditions. In order to control for these stress differences in the carrier sentence Lindblom also used supplementary speech data of vowels spoken in the same sentence at different speech rates.

Lindblom was able to model just under half the variance in the original speech materials with the following equation:

$$F_{no} = k(F_{ni} - F_{nt})e^{-ad} + F_{nt} \tag{5.2}$$

where:

> F_{no}=frequency of formant n at the formant's maxima or minima. This is where the formant's rate of change is 0 and normally corresponds to the value at around the centre of the vowel.
>
> F_{ni}=frequency of formant n at the beginning of the vowel. This value depends upon the surrounding consonants and each vowel.
>
> F_{nt}=ideal vowel target for formant n. This target will be reached if the vowel is long enough.
>
> d=duration of the vowel in milliseconds.
>
> k, a=constants which depend upon the surrounding consonants.

This equation was inspired by a damped mass-spring analogy (see Lindblom, 1983) where the effort required to move the articulators increases in order to reach a target in less time. Figure 5.3 demonstrates the values predicted for achieved vowel targets (the maxima or minima of the formant track) for F2 for different vowel durations for three vowels /I,æ,ʊ/. It is interesting to compare these predictions with the values of F2 in figure 5.2. Despite the fact that this speech is from a different, non-Swedish, male speaker the predictions for /I/ and /æ/ are accurate within 50Hz. The predictions for /ʊ/ are poorer possibly due to a difference in the Swedish and English /ʊ/.

The damped mass-spring analogy was extended to produce more complex equations to model the actual path of the formants in more complex and less symmet-

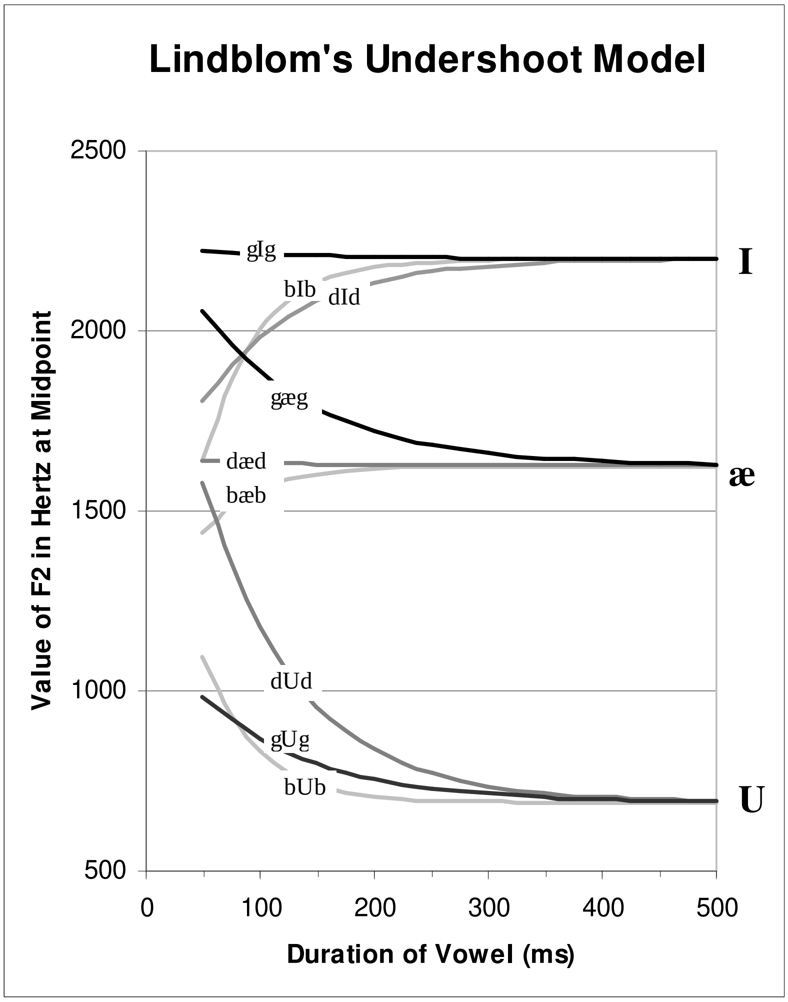


Figure 5.3: Graph showing achieved formant target as modelled by Lindblom (1963). As the vowel duration increases the achieved target gets closer to the ideal target.

rical phonemic contexts by Broad and Clermont (1987). However as there is no evidence of a linear relationship between the movement of the articulators and the formant transitions, as van Son (1993) points out, the actual choice of the function to model the formant transitions is one of convenience. Any function that fits the data effectively could be used. For example van Son (1993) uses instead Legendre polynomials to model the formant tracks. These functions are able to model the flat topped hill shape often seen in formants as well as the more complex curves seen in diphthongs.

The theoretical implications of these models are less clear. It has been shown that local duration is certainly not the only factor to influence vowel undershoot. Speaking style (e.g. Moon and Lindblom, 1994), speaking rate (e.g. Flege, 1988)) and individual speaker strategy (Flege, 1988) also affect undershoot. This raises the question of how intentional undershoot is. van Son concludes:

"...that the amount of vowel formant-undershoot is planned by the speaker." (van Son, 1993, p129)

He makes this claim on the basis that increased speaking rate, and therefore duration alone, did not influence the vowel formant undershoot or time-normalised track shape in his data. Results for individual speaker strategies and for speaking style also suggest that undershoot is a choice rather than a by-product of durational constraints.

However if undershoot is planned then this leads to a rather confusing use of the terms undershoot and target. After all, generally a target is something you intentionally try to hit. If undershoot is intentional then the speaker is not *trying* to hit this ideal vowel target at all. Instead speakers are trying, and succeeding, in hitting a vowel target which is more reduced. This is important in terms of the work presented here because the model I present for measuring vowel undershoot is based solely on such achieved targets. No *ideal* targets are used in this model. The method used is purely observational. I first build a statistical model of what achieved vowel targets look like in clearly articulated speech. I then compare these achieved targets with achieved targets in spontaneous speech. The more alike they are the more clearly articulated the vowels in spontaneous speech are assumed to be. In fact not only are *ideal* targets ignored but even the vowel identity is ignored. This is because, in spontaneous speech, we don't really know what vowel the speaker was trying to produce. Accent differences, use of schwa and idiosyncratic pronunciation mean that the vowel produced by a speaker may not even have meant to be the vowel suggested by a canonical pronunciation retrieved from an online dictionary.

In a general sense the method used here to measure vowel quality follows the same approach as that used in coding prosodic structure and measuring redundancy. It is as simplistic as possible while taking into account generally accepted findings in the literature. A summary of the results that underpin my approach are as follows:

1. Care of vowel articulation is related directly to undershoot (Picheny *et al.*, 1986; Bond and Moore, 1994; Bradlow *et al.*, 1996; Moon and Lindblom, 1994). We already have a general duration metric but undershoot has been shown, in some cases, to be independent of duration. A measurement of

undershoot offers a potential metric for care of articulation for the whole syllable which will complement the duration measurements described in section 5.3.1.

2. By using undershoot we are accepting a target based account of vowel articulation. However the extent we also assume *ideal* vowel targets underlying such an account depends on how we measure the undershoot. For example, a simple centralisation measurement does not assume such ideal targets. An alternative metric based on target change between speech styles will be presented in section 5.4.4.1.2. As with a simple centralisation measurement this also avoids the problem of deciding what an *ideal* vowel target might be.

3. *Achieved targets* will be defined, as by Lindblom (1963), as the minima or maxima of a formant track. However ascertaining these *achieved targets* in spontaneous speech is hard. I will describe the method for doing so in section 5.4.3.2.

5.4.3 Measuring Care of Vowel Articulation: Methodology

The method for calculating the target undershoot and centralisation measurements can be split into four stages:

1. Pre-processing to extract the first two formants for each vowel in a clear speech style.

2. Using curve fitting to estimate the achieved target for each of these vowels.

3. Building a model of the speaker's clear speech from these achieved targets.

4. Comparing vowels from running speech with this model and producing a numerical magnitude which reflects care of vowel articulation.

Intra-speaker differences were avoided by building a different model for each speaker. In the HCRC Map Corpus (Anderson *et al.*, 1991) every speaker reads out a list of all the landmarks on his/her map after completing all the dialogues. The subjects are asked to read this list twice, slowly and clearly. This citation list forms the basis of the clear speech models.

5.4.3.1 Pre-processing

The speech was recorded on separate channels for each speaker and digitised at 20 Khz (Anderson *et al.*, 1991). The speech was then processed using:

Entropic's LPC formant tracker. The output of the tracker is a value for F1 and F2 for every 10ms frame of speech.

Entropic's F0 tracker. The output of the F0 tracker is the probability of voicing for each 10ms frame of speech. This, together with autosegmentation, is used to establish the location of vowels.

The Cambridge HTK toolkit. The Cambridge HTK toolkit was used to autosegment each hand segmented word into a set of phones dictated by a hand-modified online dictionary (see chapter 3 section 3.3.1.1). The output of the autosegmenter was used to find the approximate location of each vowel in the syllable. The output from the autosegmenter was also used to generate approximate syllable boundaries in polysyllabic words.

Conversion to Barks. The F1/F2 values output from the formant tracker were converted to the Bark scale. The transformation used to convert frequency into Barks is an approximation suggested by Zwicker and Terhardt (1980). It is a mixture of two arctan curves as follows:

$$z = 13 \arctan\left(\frac{f}{1000} 0.76\right) + 3.5 \arctan\left(\frac{f}{7500}\right) \tag{5.3}$$

Where z is a value on the Bark scale and f is the frequency in Hz.

The Bark scale represents the ability of the human ear to distinguish different tones at different frequencies (Zwicker, 1961; Zwicker and Terhardt, 1980). For example the human ear is more sensitive to tonal differences between 1000Hz and 2000Hz than between 4000Hz and 5000Hz. The use of the Bark scale has the effect of stretching the vowel space where the human ear is most sensitive and contracting the space where tonal differences are difficult for the ear to perceive. The Bark transformation was chosen over the Mel, Koenig and ERB-rate scales simply because a simple mathematical approximation was readily available. In fact all these perceptual scales are fairly similar (see Rosner and Pickering, 1994, p16 for a review).

Normalisation. The F1 and F2 values were normalised for each speaker to have a mean of 0 and a standard deviation of 1. This has the effect of stretching and squashing the F1/F2 dimensions so that nearly all the data falls within a square of size -2.5 standard deviations to 2.5 standard deviations. This made it easier to inspect and compare pre-processed output.

The voicing information, together with the segmentation information was used to constrain which formant data to examine. Only formant tracks in voiced speech (in this case with a probability of voicing of 0.99) and within an expected vowel segment were used as input to the next phase of finding achieved targets. This helped offset errors caused by the autosegmentation and ensured that only formant data for reliably voiced speech was considered.

A problem encountered and not readily solved with the tools I had available was that the formant tracker is based on an all pole LPC model and therefore had difficulty in finding correct formant tracks in nasalised vowels. A second problem was that the temporal positioning of the 10ms frames had a small but significant effect on the formant values produced. Thus an identical section of speech with the 10ms frames offset by say 5ms would not produce quite the same formant tracks. A different formant tracker might well produce more consistent results but one was not available for the work carried out here. As we will see, these pre-processing problems contributed to some errors in the achieved targets calculated by curve fitting (See the evaluation in section 5.4.3.3.1).

5.4.3.2 Finding Achieved Targets

A variety of mathematical functions can be used to model formant transitions.[1] The most well known approach is that used by Lindblom (1963). Lindblom, in proposing a target model for vowel production modelled the formant transitions using exponential functions based on the mathematics of a damped spring (see section 5.4.2.3). van Son (1993) gives a detailed review of the target/undershoot model and its variations since 1963. He also discusses the use of Legendre polynomials to model formant tracks (van Son, 1993, chapter 4).

In this work a simple parametric curve of the form $y = ax^2 + bx + c$ is used to model formant tracks. The curve is fitted to the data on the basis of mean squared error and the maximum or minimum of the curve is used as the vowel

[1] My implementation of this technique is based on a talk given by Steve Isard to the Phonetics and Phonology group at Edinburgh university.

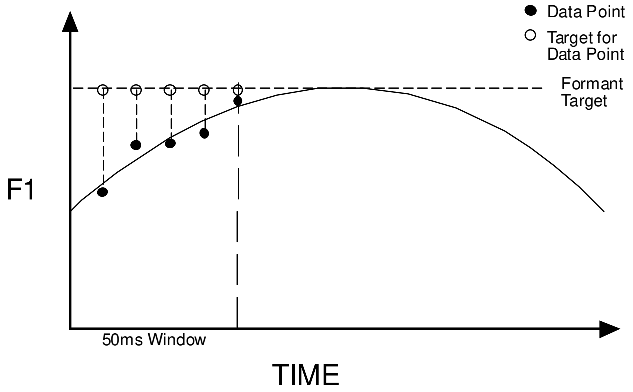


Figure 5.4: Using a parametric curve to calculate the achieved spectral target of a formant.

target or approximation to the steady state central formant value in the vowel (see appendix C for details).

This method can be used to estimate the mid vowel formant targets by fitting the best parametric curve to a number of formant values over a time window. The maximum or the minimum of the curve can be regarded as the achieved spectral target that this formant is heading towards or away from (see Figure 5.4).

A major problem in applying this technique to unsegmented speech is to decide on how many points to use (or the window size) and whether such a window should overlap. If windows do overlap or a number of window sizes are used, it is necessary to choose between different values predicted by different curves for the same point in time. See Figure 5.5 for an example of this effect using different sized windows.

The method selected each target depending on how well the overall curve it belonged to fitted the data by averaging the fit error by the window size (see appendix C for details).

Figure 5.6 shows the result of applying this technique to the speech "you gotta map". The top part of the figure shows the results from the formant tracker for F1,F2,F3 and F4. The lower part shows the targets estimated using parametric curve fitting. The targets are normalised on the basis of the speaker's citation speech. Lack of voicing and poor autosegmenation have meant that targets were not found for most of the /ɒ/ in /gɒtə/. The /u/ in /ju/ has an unusually high F2 suggesting this was pronounced more as /jy/. This highlights the problem of using expected vowel identity in any modelling process. In the approach used here vowel identity is ignored.

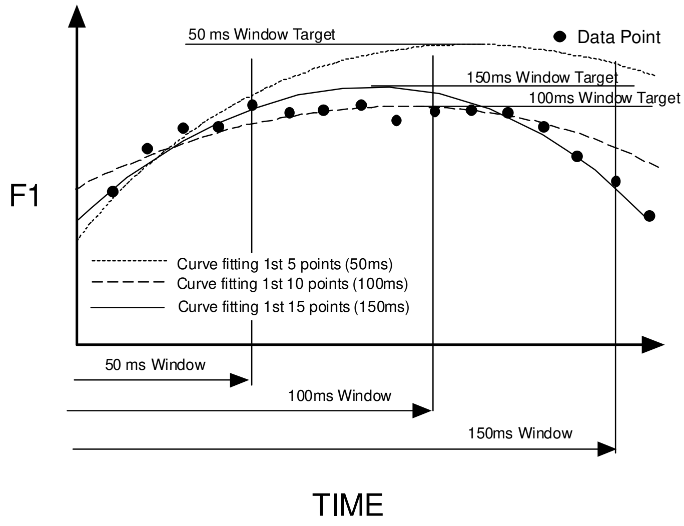


Figure 5.5: The effects of differing time windows on fitting parametric curves. An intended target is calculated for each frame on the basis of the parametric curve. As we can see different window sizes generate different curve fits. For each frame, the curve that fits best over the window (with a bias to longer curves) is selected.

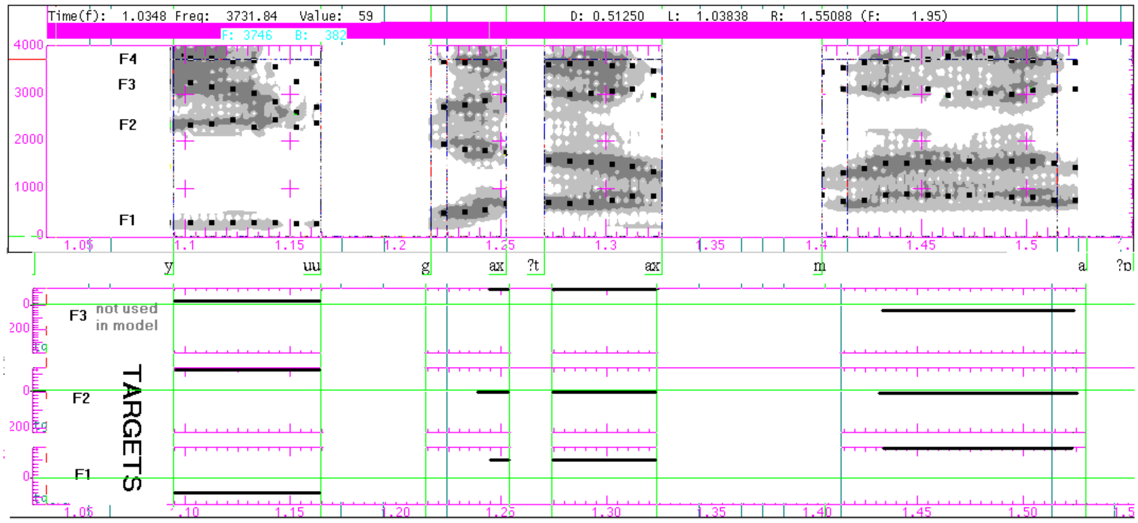


Figure 5.6: The result of applying the curve fitting technique to the speech "you gotta map". The top part of the figure shows the results from the formant tracker for F1,F2,F3 and F4. The lower part shows the targets estimated using parametric curve fitting for F1,F2 and F3 (F3 was not used in the modelling process). The targets are normalised on the basis of the speaker's citation speech.

During this process, clear citation speech, which is to be used to produce a model of clear achieved vowel targets, and spontaneous speech, which will be compared against this model, were treated a little differently. The data we wish to use to build the model needs to be as clean as possible, but, when calculating values for spontaneous speech a value is required for as many vowels as possible. To make the citation model as clean as possible only achieved target values that had remained within 1 Bark for a minimum of 4 frames (40ms) were accepted. An example of a vowel which wouldn't meet this criterion is the /ɒ/ in /gɒtə/ in figure 5.6. The vowel is just long enough but voicing and segmentation problems meant targets were only assigned for two thirds of the frames. In contrast, for the spontaneous speech any target, however transitory, was given a value.

Initially this process was carried out without any autosegmentation information. In these models all voiced speech was included. Figure 5.8 shows an example of the resulting density of targets in the vowel space for citation speech (a) and for spontaneous speech (b). In these examples the long thin hills to the left are not actually vowels but nasals. In contrast when the autosegmentation is used to filter out non-vowel voiced speech the result is closer to the classic vowel triangle. Figure 5.9 shows the result for citation speech (a) and spontaneous speech (b). Despite problems with noise the citation speech does produce a set of achieved targets which are clearly more distinct and more extreme than the spontaneous speech where many targets are centralised and the distinct hills representing individual vowels are merged one into the other.

Comparing figure 5.9a with figure 5.1b (The raw F1/F2 values from citation speech) we see that the vowel spaces produced seem very similar. This is partly due to the normalisation process. However if we look at individual vowels without normalising the F1/F2 Bark measurements (for example see figure 5.7 for the /i/ vowel) we find that although the peaks of the raw and fitted distribution are very similar there are differences in the spread of the data. The target fitting has made the data more granular in low probability target areas and has increased the concentration around the peak. This results in a lower standard deviation for the data on the F1 dimension. In addition the points tend to be less centralised (for /i/, low F1/high F2) suggesting the target fitting has indeed found the F1/F2 value that a vowel's formants are moving towards. This results in a higher mean for F2. Similar results are found for vowels located elsewhere in the vowel space, for example /o/ and /æ/, where slightly less variance was noted and the means where slightly shifted away from the central area of the vowel space. Over the entire vowel space this resulted in a reduction of entropy (randomness) by 4%

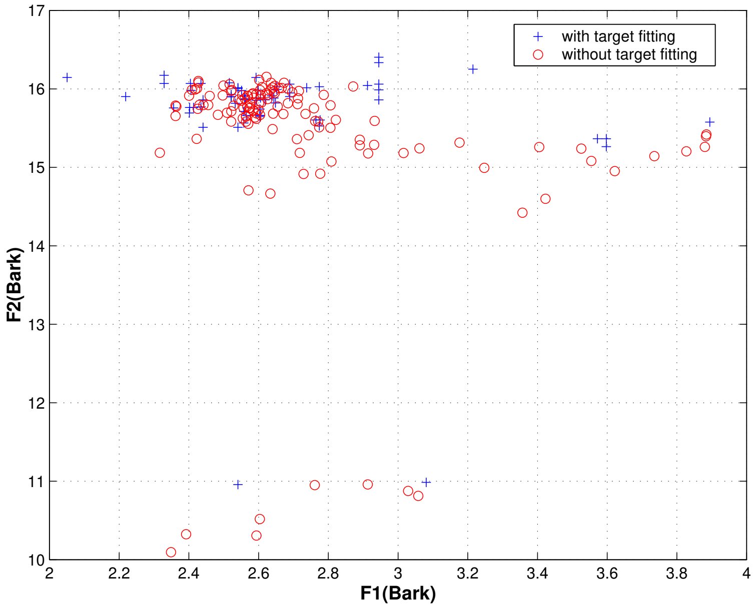


Figure 5.7: Comparison of F1/F2 values for the vowel /i/ with (n=125) and without (n=128) target fitting.

suggesting a less noisy and more defined data set.

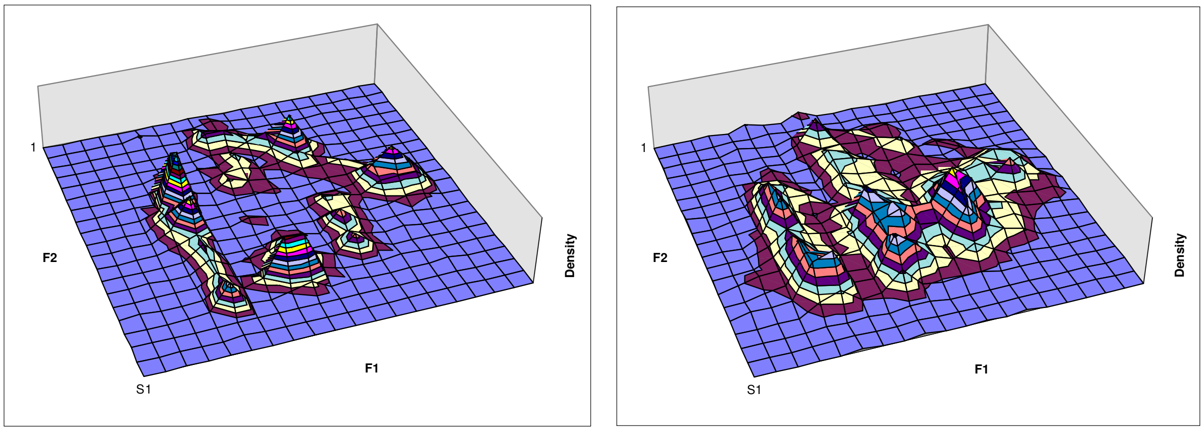
These results, together with the strong theoretical basis of an undershoot model, make the target fitted vowels from each speaker's citation speech useful for producing a model of a speaker's clear speech. In turn this model offers a potential means of measuring the care of vowel articulation, by a speaker, over an individual syllable in spontaneous speech (see section 5.4.4.1.2).

A disadvantage of the process is the loss of data for very short vowels (less than 40ms). For the relatively well articulated citation speech this is not a serious problem. For spontaneous speech this does raise some important issues which I address in detail in chapter 6, section 6.4.4.2 and chapter 7, section 7.2.1.

5.4.3.3 Evaluation of Achieved Target Calculation

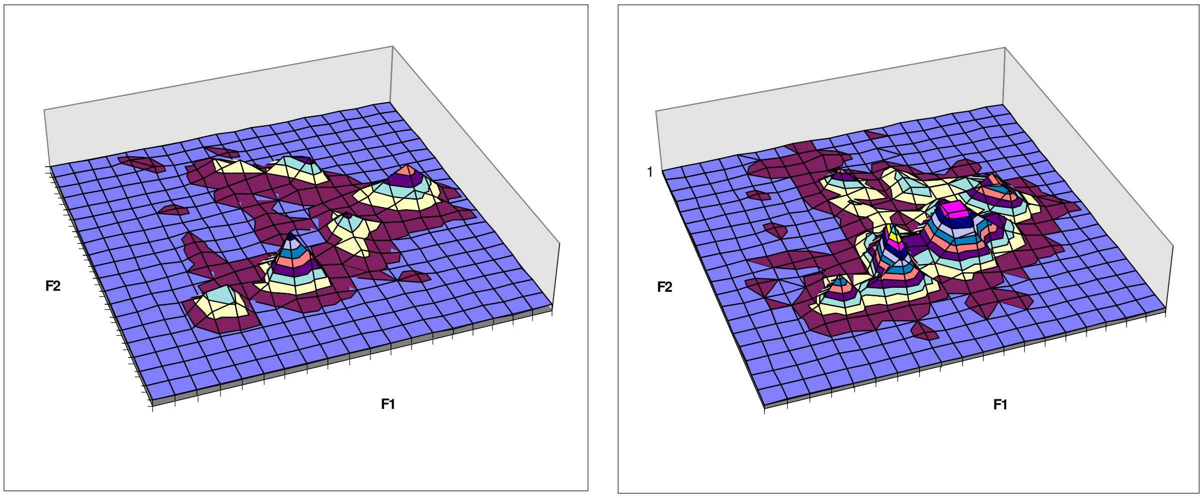
In order to ascertain the accuracy of the achieved targets calculated by the above technique, 37 vowels were examined by two phoneticians and 180 by a single phonetician. The 180 vowels consisted of 60 /i/, 60 /æ/ and 60 /u/. Half the vowels were citation speech and half were spontaneous speech. The vowels were taken from multiple speakers. The 37 cross labelled vowels represented a balanced sample of the 180 vowels.

For each vowel the start time and end time were marked using a wide band



(a) **(b)**

Figure 5.8: Achieved targets found by parametric curve fitting for (a) clear voiced citation speech and (b) spontaneous voiced speech.



(a) **(b)**

Figure 5.9: Achieved targets found by parametric curve fitting for (a) clear voiced citation speech excluding non-vowels and (b) spontaneous voiced speech excluding non-vowels.

spectrogram and by listening to the speech. The spectrogram was also used to assess the achieved target values for F1 and F2. The achieved target was regarded as the point in the spectrogram that the formant appeared to be heading towards or away from. Time constraints meant that a more detailed analysis of the spectral envelope was not carried out.

5.4.3.3.1 Accuracy of F1/F2 Achieved Targets. In order to compare values it was necessary to decide how to translate the series of achieved target values produced for each 10ms frame of the vowel into a single value in order to compare with the human judgements.

The values for each vowel were grouped into stable areas. By stable I mean that the intended targets had not changed by more than 1 Bark from one value to the next. Table 5.2 shows an example of such a grouping for a diphthong. The vowels considered in the evaluation were all monophthongs and thus only the largest group was used to represent the overall F1/F2 targets for comparison with the hand measured results. (This grouping process was also used to generate the COVA1 measurement, see section 5.4.4.1.1). The result of this grouping process is to use the mode to evaluate the values rather than the mean.

These values were compared using linear regression with hand coded values. Reported here is the percentage of the variation the automatic values predict in the hand coded values together with the regression coefficient. Complete agreement would give a percentage of 100% and a coefficient of 1. The advantage of this method is that it takes into account the different variance of F1 and F2.

1. Comparisons between automatic values and two coders (37 vowels)

	Citation Speech n=17		Spontaneous Speech n=20		Both n=37	
	coef.	agree	coef.	agree	coef.	agree
Auto/C1 F1	0.73	58%	0.85	92%	0.75	80%
Auto/C1 F2	0.71	67%	0.89	80%	0.78	72%
Auto/C2 F1	0.52	65%	0.85	94%	0.73	83%
Auto/C2 F2	0.73	68%	0.96	93%	0.82	78%
C1/C2 F1	1.09	91%	0.99	96%	1.01	94%
C1/C2 F2	0.98	98%	0.94	88%	0.96	93%

   As we can see the hand coders (C1/C2) agree with each other better than the automatic values.

2. All vowels compared with a single hand coder (180 vowels)

Time (ms)	F1 Target (Bark)	F2 Target (Bark)	Stable Groups
0	5.884604	13.375496	Group 1 - 30ms
10	5.884604	13.375496	
20	5.884604	13.375496	
30	5.884604	15.056014	Group 2 - 20ms
40	5.884604	15.056014	
50	3.936561	14.975552	Group 3 - 60ms
60	3.980757	14.975552	
70	3.980757	14.975552	
80	3.980757	14.975552	
90	3.980757	14.975552	
100	3.980757	14.962492	

Table 5.2: Grouping the vowel targets into stable groups. The values shown above are targets in Bark for F1 and F2 for the diphthong /ɑI/. The table shows each 10ms frame regarded as being within the vowel according to autosegmentation and voicing. The frames are grouped on the basis of target values remaining within 1 Bark. The middle group is probably noise, the two largest groups probably represent targets for the two parts of the diphthong. In the evaluation of the achieved target calculation (all monophthongs) the average targets of the largest group was used as a comparison with human judgements. For calculating COVA1 (see section 5.4.4.1.1 the targets of the two largest groups were used. (Data taken from Giver in dialogue Q4NC1 from 'right' at 51.637 seconds.)

	Citation Speech n=90		Spontaneous Speech n=90		Both n=180	
	coef.	agree	coef.	agree	coef.	agree
Auto/C1 F1	0.71	66%	0.65	50%	0.67	58%
Auto/C1 F2	0.87	85%	0.85	83%	1.12	84%

Although agreement for F2 is reasonable the overall agreement of 58% for F1 values is disappointing.

3. Raw error scores (Bark).

	Citation Speech		Spontaneous Speech		Both	
	mean	sd	mean	sd	mean	sd
F1 error	-0.516	1.129	-0.127	0.627	-0.306	0.901
F2 error	-0.208	1.979	-0.109	1.194	-0.037	1.587

The overall distribution of F1 has a mean of 5.073 and an sd of 1.688 Bark. For F2 the mean is 11.235 and the sd 2.748 Bark.

Looking at the raw error scores we can see that the automatic method tends to underestimate formant values from 0.127 to 0.516 Bark. The variance of the error rate F1 is high considering the low standard deviation of F1 (1.688 Bark)

There are two sources of error that can account for these results. The first are errors within the formant tracker (addressing errors at this level is beyond the scope of this thesis) and errors at the level of the parametric curve fitting. While hand coding the materials it appeared that the poor results for F1 were connected with the formant tracker mistakenly regarding female f0 as f1 (A high female voice may easily have an f0 of more than 200Hz). To investigate this I looked at the agreement and coefficient between C1 and only automatic measurements produced for male speakers. For F1 the results jumped to a coefficient of 0.82 and an agreement of 79% and for F2 a coefficient of 0.86 and an agreement of 90%.

The conclusion reached was that an improved formant tracker would have a significant effect on these results.

However overall these results do need to be put in perspective. All the techniques described here are automatic. Noise is an expected problem with such automatic techniques especially when applied to spontaneous speech. It must be borne in mind at every stage described here that, by using automatic techniques, I am able to include 170,000 vowels in this study. It would take an estimated 8,500 hours to hand code this number of vowels if it took 3 minutes to examine the F1/F2 of each and this does not include time for coding the citation speech.

5.4.3.3.2 Accuracy of Autosegmentation. Although only indirectly used to calculate vowel targets this was regarded as a good opportunity to check the accuracy of the autosegmentation carried out on the corpus. The results were as follows:

	Citation Speech		Spontaneous Speech		Both	
	mean	sd	mean	sd	mean	sd
Start time error	-25ms	15ms	-23ms	14ms	-24ms	14ms
End time error	-16ms	16ms	-10ms	19ms	-13ms	17ms

As we can see the autosegmentation has consistently placed the start and end of the vowel early. The high standard deviations show that the autosegmentation is only really accurate to within 30-40ms.

Given that the average vowel length in this set is 89ms, this is not a very good result. Fortunately for the work presented here the vowel segmentation is only used as a filter to remove non-vowel voiced segments and as a means of splitting polysyllabic words into syllables. In both situations, therefore, although poor segmentation is unwelcome, it isn't critical. The segmentation could certainly be improved if a more complex model was used (a unigram model was used in this work[2]) and if a substantial set of material was hand segmented in order to train the segmentation model.

Finally, although a 30-40ms error is poor for segmenting individual vowels, in terms of syllabic segmentation (the other main use of autosegmentation in this work) this error is reasonably acceptable (≈15% error for polysyllabic syllables as opposed to ≈50% error for the phonemes themselves).

5.4.3.4 Building Models of Carefully Articulated Vowels

The achieved targets, calculated for each speaker's citation speech from voiced speech excluding non-vowels (figure 5.9a), is then used to generate a model of clearly articulated vowels.

5.4.3.4.1 Vowel Centralisation Metric. The simplest model and the one used to measure centralisation is to regard the centre of the vowel space, or the mean F1 and F2, as the most centralised and thus most poorly articulated example of a vowel. The further away a vowel's targets are from this central region the more carefully articulated the vowel. In this work the mean F1 and

[2] For a detailed description of speech recognition methods for autosegmentation see the Cambridge HTK documentation (Young *et al.*, 1996)

F2 were calculated from vowel targets in the speaker's citation speech.

Looking at the vowel space generated by citation speech achieved targets (e.g. figure 5.1b) we can see a number of problems with this metric.

1. The vowels are not oriented in a circle around the mean. This means that some vowels will always tend to be clearer than others. For example, a well articulated /i/ will always be further from the centre of the vowel space in absolute terms than a well articulated /ɛ/ (see figure 5.1a).

2. The vowel space is a complex space. A simple centralisation measurement completely ignores this complexity. For example, looking at figure 5.1b we can see some areas in the vowel space are quite empty and others quite crowded. A simple centralisation metric ignores this structure.

However a centralisation measurement does have some advantages. It is simple, assumes less, and does give a rough idea of how much undershoot might have occurred. Also such a measurement can act as a control for the more complex target undershoot model.

5.4.3.4.2 Target Undershoot Metric. For this more complicated measurement we need to model the vowel space in much more detail. The basic idea for this measurement is that some areas in the vowel space are more distinct and preferred in clear speech. The more the achieved target of a vowel falls within these areas the clearer the vowel is and the less undershoot has occurred in the spontaneous speech.

One method of modelling the complex space is to fit a two dimensional histogram over the top of the citation speech's achieved targets. The more points that are in each bin the more preferred the region. However the disadvantage of this technique is that it is strongly affected by individual points especially in sparse areas in the vowel space. One way of avoiding this problem as well as producing a model which generalises well is to fit a continuous probability function onto the data. In effect, we fit a number of hills to the data. A probability density function (pdf) constructed from two dimensional Gaussian distributions can achieve this and the EM (expectation maximisation) algorithm can fit this pdf to the data.

5.4.4 The EM Algorithm

A two dimensional Gaussian curve resembles a hill. The north/south width of the hill is the variance of the Gaussian in one dimension and the east/west width is the variance in the second dimension. The location of the peak of the hill is the mean of the Gaussian. A number of these Gaussians can be added together to model a complex distribution. The expectation maximisation (EM) algorithm will, given a specified number of Gaussians, fit them to a distribution. I will not give a detailed account of the mathematical thinking behind the EM algorithm. This has been treated in some detail in other statistics and maths literature. For a clear and detailed account refer to (Bishop, 1995, chapter 2) or (Duda and Hart, 1973).

The algorithm works as follows:

1. Pick a number of Gaussians
2. Randomly place them on the distribution with random standard deviations, random probabilities of occurring and random means.
3. While the fit continues to improve take the points that 'belong' to each Gaussian and use them to recompute the means, standard deviations and probability of occurring for that Gaussian. The fit is calculated by summing the probability of the pdf producing every point in the data set.

The calculations that are required to run the algorithm are as follows.

Given a set of n points with vectors $\mathbf{x}$, $\mathbf{M}$ Gaussians, the initial probabilities of a jth Gaussian occurring $P(j)$, a covariance matrix Σ_j and a vector of means μ_j, recompute new $P(j)$, Σ_j and μ_j.

For the case where we allow no covariance between dimensions (in fact F1/F2 are fairly independent) the covariance matrix has only the variance for each dimension along the diagonal. To simplify the calculation this can be thought of as a vector of standard deviations σ_j.

The formulae to recompute the parameters are as follows:

To recompute the new means:

$$\mu_j^{new} = \frac{\sum_n P^{old}(j|\mathbf{x}^n)\mathbf{x}^n}{\sum_n P^{old}(j|\mathbf{x}^n)} \tag{5.4}$$

To recompute the new variances:

$$(\sigma_j^{new})^2 = \frac{\sum_n P^{old}(j|\mathbf{x}^n)(\mathbf{x}^n - \mu_j^{new})^2}{\sum_n P^{old}(j|\mathbf{x}^n)} \tag{5.5}$$

To recompute the new probabilities of a Gaussian occurring:

$$P(j)^{new} = \frac{1}{N}\sum_n P^{old}(j|\mathbf{x}^n) \tag{5.6}$$

Where:

$$P(x|j) = \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_j)^T \Sigma_j^{-1}(\mathbf{x} - \mu_j)\right\} \tag{5.7}$$

Taking Σ_j as the covariance matrix with σ_j^2 along the diagonals, this is the basic equation for a Gaussian.

And where:

$$P(x) = \sum_{j=1}^{M} P(\mathbf{x}|j)P(j) \tag{5.8}$$

And using Bayes theorem:

$$P(j|x) = \frac{P(x|j)P(j)}{P(x)} \tag{5.9}$$

The fit function being maximised is the average log likelihood of the data fitting the distribution:

$$Fit = \frac{1}{n}\sum \log(P(x)) \tag{5.10}$$

The EM algorithm is an iterative algorithm that will reach a maximum fit although the maximum fit it finds may only be a local maximum. This problem is general to all hill climbing algorithms such as the EM algorithm. The number of local maxima depends on many complex interactions in what is a multi-dimensional search space. The more local maxima the more sensitive the algorithm becomes to starting criteria and the more likely it will find not the best

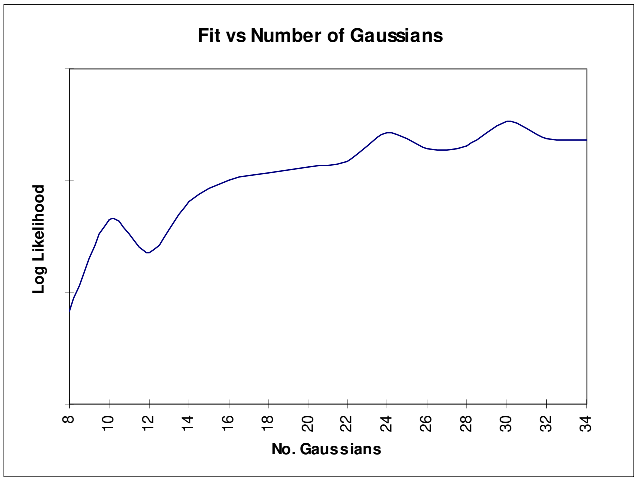


Figure 5.10: Fit of models for different numbers of Gaussians. Fit is poor for too few Gaussians but becomes more unstable and risks over fitting with too many. 20 Gaussians were chosen for the modelling process. (Fit is the probability of the model generating the data set.)

solution but a secondary solution. The EM algorithm will find a fit for a set of n Gaussians but in order to feel secure that this fit is a good fit it may be necessary to run the algorithm a number of times from different random starting positions.

The algorithm is unsupervised. It is only necessary to specify the number of Gaussians used in the model; it is not necessary to specify what the data points in the distribution represent.

There are, however, two disadvantages. Firstly it is necessary to choose the number of Gaussians in advance. On what basis do we choose this number? Secondly how can we ensure the algorithm does not get stuck in a local maximum? There is no theoretically bomb proof means of answering these questions. However a pragmatic approach to the problem can produce interesting results.

If we examine the final fit using different numbers of Gaussians we can see in figure 5.10 that improvement appears to level off and become more unstable (probably due to more local minima with models containing more Gaussians). This levelling off together with an inspection of the actual density distribution we wish to model can be used to estimate a good number of Gaussians. Models with a similar number of Gaussians behave in similar fashions so it is not necessary to be absolutely correct. The number I chose for my model was 20 partly because that seemed a sufficient number to model the data by inspection (figure 5.8a, figure 5.9a) and because (as can been seen in Figure 5.10) the improvement appears to both level off and become more unstable after about 20 Gaussians.

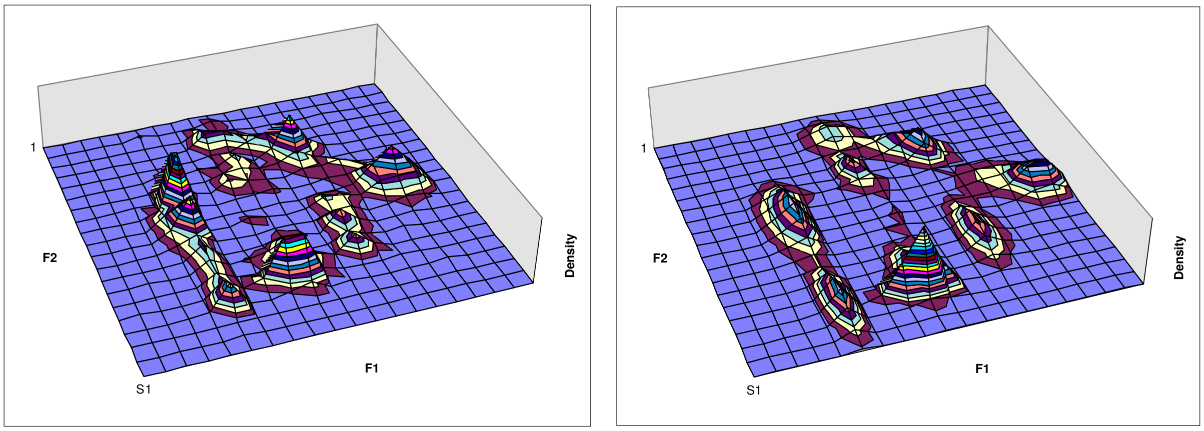
In order to avoid local maxima it is necessary to run the EM algorithm a number of times. The hope is that local maxima will generally be less stable than global maxima and thus it would be very unlucky, using random starting parameters, to find the same local maxima on several occasions. Over 10 trials the results from the model appeared generally stable.

Figure 5.11 and figure 5.12 show the result of applying the 20 Gaussian mixture model to the citation data with and without the non-vowel filter. Some care must be taken when comparing these figures. In order to produce them each are quantised over a 20x20 grid. If this grid is increased in the size, the detail in the original data will appear to increase while the detail in the mixture model will stay relatively unchanged. However, bearing this in mind, the mixture model has fitted the original data with some degree of success. Although some hills have been merged the advantage of the mixture model is that it both generalises and smoothes the data. This helps deal with data sparsity in low probability areas within the vowel space as well as allowing smooth transitions between high and low probability areas. The degree the model fits the citation speech will vary depending on random starting criteria and the type of structure present within the data. However, providing the model represents the broad structure within the citation speech, it can be used to assess care of articulation. This is because it is the differences between the model and the spontaneous speech (see figures 5.8b and 5.9b) which are important to the metric. Providing the model assigns high probability to the peripheral vowel target areas in the vowel space it can function as a clear speech model.

In general this is the case with these models. Even the rather poorly fitting model shown in figure 5.12 is constructed of Gaussians with means located at more extreme areas in the vowel space than the means of original citation targets of individual vowels. At worst the model acts as a simple centralisation metric. As we will see in chapter 6 a metric based on this Gaussian model appears to do better than such a centralisation metric suggesting the additional structure modelled by the Gaussians does help to give a better idea of where clear vowel targets are likely to be.

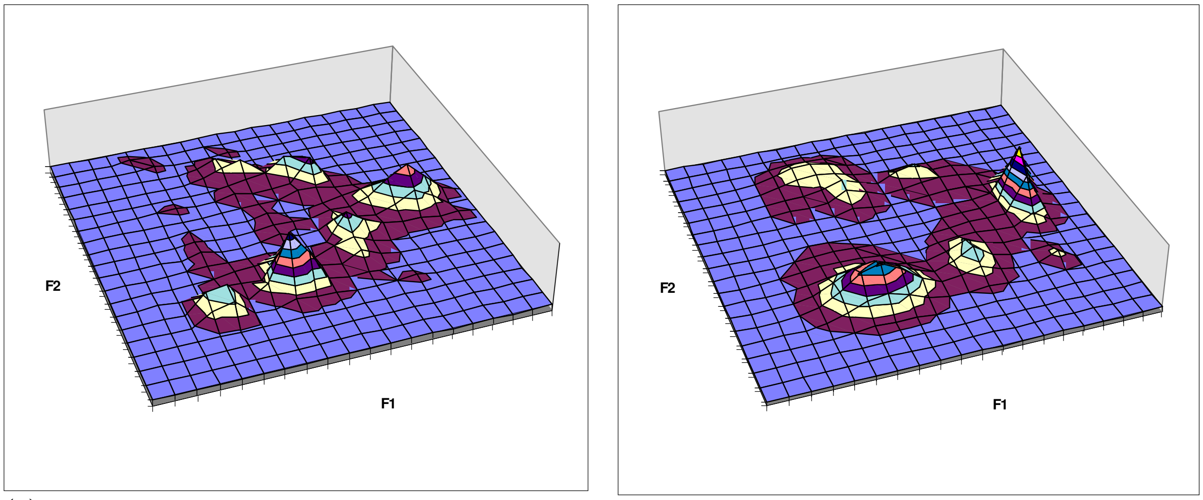
5.4.4.1 Calculating the Care of Vowel Articulation Metrics

5.4.4.1.1 Vowel Centralisation Metric. For convenience this metric will be referred to from here on as COVA1 (Care of Vowel Articulation 1). A major



(a) **(b)**

Figure 5.11: **(a)** Achieved targets found by parametric curve fitting for clear voiced citation speech (as in figure 5.8) and **(b)** A 20 Gaussian model built using EM.



(a) **(b)**

Figure 5.12: **(a)** Achieved targets found by parametric curve fitting for clear voiced citation speech excluding non-vowels (as in figure 5.9) and **(b)** A 20 Gaussian model built using EM.

problem with calculating this metric is dealing with outliers. Because this is a distance measurement, as with linear correlations, single outlying values will have a disproportionate effect on the final value. Such values can occur at formant transition points, for example in a diphthong when the target changes or at the edge of a vowel when other voiced speech has been incorrectly segmented as belonging to the vowel.

In order to deal with this problem the method used to group values in section 5.4.3.3.1 to evaluate the accuracy of achieved targets was used (see also table 5.2).

The two largest groups were selected to represent the overall vowel targets of the vowel. If the smaller of these two groups was at least 25% of the size of the larger group the vowel was regarded as possibly being a diphthong and both groups were retained. If not, only the larger group was retained. By averaging the values in these groups up to two values for F1 and F2 were produced.

These values are normalised with regards to the speaker's vowel space. The effect of this is to produce values which are z-scores for each F1/F2 measurement. The Euclidean distance of the F1/F2 pairs are then calculated and averaged.

The overall calculation is as follows:

$$\mathrm{COVA1} = \frac{\sum_{i=1}^{n} \sqrt{\left(\frac{f1(i)-f1\mu}{f1\sigma}\right)^2 + \left(\frac{f2(i)-f2\mu}{f2\sigma}\right)^2}}{n} \tag{5.11}$$

where $n = 1$ (monophthong) or $n = 2$ (diphthong), $f1(i)$, $f2(i)$ are the proposed grouped target(s) for the vowel, $f\mu, f\mu$ are the means of the speakers vowel space and $f\sigma, f\sigma$ are the standard deviation of the F1/F2 values in the vowel space.

5.4.4.1.2 Target Undershoot Metric. For convenience this metric will be referred to from here on as COVA2 (Care of Vowel Articulation 2). This measurement depends on the statistical model of a clear vowel space constructed from citation speech for each speaker described above. This model maps out areas of the vowel space which are desirable for clear vowels. In effect the clear vowel targets are expressed in probabilistic terms. Rather than a single point we have hills of probability which represent the achieved vowel targets in clear speech. In order to calculate undershoot for each vowel we can produce a value which is the probability of this clear vowel model producing those points. The more carefully articulated the vowel in spontaneous speech the more likely it is that the clear speech model could have produced it. Undershoot will have a tendency to pull the achieved targets of the vowel from spontaneous speech away from the hills and towards the middle and less likely areas of the model.

This metric does not suffer from the same problem as COVA1 with regards to outliers. This is because spurious values will produce low probabilities (approaching 0), and, providing they do not occur frequently, will have only a marginal effect on the overall score.

To calculate the probability that the vowel in spontaneous speech could have been produced by the clear speech model we calculate the average log likelihood for each 10ms frame in the vowel which has a valid achieved target.

$$\mathrm{COVA2} = \frac{1}{n}\sum_{i=1}^{n}\log(p(x_i|M)) \qquad (5.12)$$

By using this method we have avoided the need to take into account vowel identity or any set of idealised vowel targets. However there are some problems with this measurement.

1. If a vowel in the spontaneous speech was *more* clearly produced than in the clear citation speech the COVA2 value would be wrong and regard it as a bad example of a vowel.

2. If a vowel was misproduced, so for example a lousy /i/ was produced as a decent /e/ again the COVA2 value would be inaccurate.

3. No phonemic context is taken into account. This context could well mediate the extent increased care of articulation can prevent undershoot. For example in the syllable /dʊd/ it is harder for the tongue to achieve the /ʊ/ targets than in the syllable /bʊb/ because it has to attempt to reach a /d/ target on the alveolar ridge.

The first problem does not appear to be critical if the citation speech is reasonable quality. If you inspect the vowel spaces of the citation and spontaneous speech (figure 5.8, figure 5.9) there really aren't many examples of vowels which have achieved targets more extreme than in the citation speech.

The second problem is ignored. If a vowel has the acoustic properties of a good vowel then it is a good vowel. Without being able to read the mind of the speaker we can never be sure what vowel was intended only what was produced. For example in figure 5.6 it was noted that "you" was probably pronounced more like /yi/. In a Glaswegian accent this is perfectly possible. To say such an /i/ is a bad example of a /u/ is prescriptive and not the approach taken in this work.

The third problem, phonemic context, is indeed a potential weakness and one that it would be good to address in future work. In order to do so more citation speech would be required to build the speaker's model either so context effects could be normalised out of, or, in some way, included in, the model.

In fact the main problem encountered in this work was the quality, type and amount of citation speech available for each speaker. Some speakers did not produce very clear citation speech. The corpus was not collected for the purposes used here and although, in general, the citation speech is a lot more carefully articulated than the spontaneous speech this cannot be guaranteed.

These doubts concerning COVA2 are addressed in two ways. Firstly the measurement is used together with three other measurements of care of articulation, DUR1, DUR2 and COVA1. Secondly a perceptual evaluation of COVA2 was carried out to see how predictive it was from a psycholinguistic perspective.

5.5 Evaluating COVA2

Although COVA2 is based on a production model it is assumed that poorly articulated vowels sound unclear to listeners. If COVA2 is measuring care of vowel articulation then you might expect these measurements to agree with human listeners when asked to judge vowel quality. To test this assumption a perceptual experiment was carried out.

5.5.1 Method

32 subjects (23 British English native speakers of which 12 had a Southern British accent, 7 were Northern British, 3 were Scottish and 1 Irish together with 4 North American English native speakers and 5 non-native speakers) were played 90 vowels excerpted from spontaneous speech together with 90 matched fillers taken from citation speech and asked to rate their 'goodness' using magnitude estimation. Magnitude estimation is a technique often used in psychophysics to validate and construct scales of physical sensations. The main advantage of magnitude estimation over more traditional rating scales or visual analogue scales is that the scale used to measure subjects' response does not affect the response. In magnitude estimation a subject decides on their own scale based on the first

stimulus and uses that first response as a yardstick to measure all others. In order to compare results between subjects the responses are log transformed.

For example, the first /i/ vowel is played. The subject decides this sounds like a good vowel and decides that a good vowel is scored at 10. The subject then hears the next /i/ vowel and decides it sounds twice as good as the previous vowel and scores it as 20. The third /i/ vowel is played and the subject decides it sound nearly but not as good as the first vowel and scores it with a 9. The only restriction on the scores is that they must be positive and non-zero. By allowing the subjects to decide on their own scale they can always score a vowel that sounds better or worse than the ones they have already heard. For a clear and concise introduction to magnitude estimation see Lodge (1981).

The vowels used all had durations between 90-110ms, had their amplitude normalised and were excerpted from the HCRC Map Corpus (Anderson *et al.*, 1991). Segmentation was achieved by combining word segmentation done by hand with phonemic auto-segmentation carried out using the HTK toolkit (Young *et al.*, 1996) and hand corrected entries from the CELEX online dictionary (Baayen *et al.*, 1995). The vowels represented 3 vowel types (one from each corner of the vowel triangle), 3 levels of COVA2 (high, medium, low) as calculated using the model described. Each cell of ten stimuli had a matching set of ten citation fillers with similar COVA2 scores, durations and speakers. The speakers who produced each of the ten stimuli in each cell were different and split equally between male and female speakers. Where possible the same speakers were used in each cell.

COVA2 groups were decided on the basis of the distribution of the COVA2 score of all 90-110ms vowels. The mean of the log likelihood COVA2 score of the vowels was -16.912. The COVA2 data was grouped by quartiles. Any vowels with a COVA2 of less than -16.75 (in quartiles 1 and 2) were regarded as low COVA2 items. Items above -16.5 (in quartiles 3 and 4) were divided into two further groups, those with a COVA2 between -16.5 and -15.5 (quartile 3) which were regarded as medium and those with a COVA2 of greater then -15.25 (quartile 4) which were regarded as high COVA2 items. The standard deviation of the COVA2 score was 2.154.

Each subject was first given a practise exercise in Magnitude Estimation training them to use this technique to judge line lengths. They then listened to some randomly selected sections of spontaneous speech produced by Glaswegian Speakers and to some example vowels excerpted from this speech. They then carried out a short practise session judging the vowel quality of 10 vowels before taking part in the main experiment. In the main experiment they were played 60 randomised examples of each vowel (/i/ as "ee" in "street", /o/ as "o" in "gold" and /æ/ as "a" in "cat"), they were given the word the vowel was taken from and asked to judge how good they thought the vowel sounded. The order of presentation of vowels was varied amongst subjects to control for an ordering effect.

Each vowel was presented twice with a 2 second gap between each presentation and a 4 second gap and a beep between each vowel. Vowels were blocked into groups of ten and data was captured using netscape and a web interface.

5.5.2 Results

There are two main questions that this evaluation hopes to answer:

1. If vowel quality is a good metric of care of articulation, and related to acoustic redundancy as argued in chapter 2, then we would expect subjects to be sensitive to the vowel quality differences in the materials. We can gauge how sensitive subjects are by the amount they agree with each other when judging vowel clarity. The cluster analysis of subjects responses in section 5.5.2.1 addresses this question.

2. If COVA2 is successfully measuring vowel quality then the average subject response of 'vowel goodness' should relate to the vowel quality goodness as dictated by COVA2. This raises two questions:

   - Is there a significant relationship between subjects regarding a vowel as good and the vowel being classed as good by COVA2. Do the other factors controlled for in the experiment (vowel type, speaker sex, subject nationality) affect this relationship? This question is addressed by the by-subjects and the by-materials ANOVA analysis in section 5.5.2.2.

   - How predictive is the COVA2 score. Can we use COVA2 to predict subject responses? This question is addressed by carrying out a linear correlation between COVA2 and pooled subject responses in section 5.5.2.3.

5.5.2.1 Cluster analysis of subjects responses

In order to investigate agreement between subjects a cluster analysis was carried out on subject's responses. The clustering was carried out using correlation

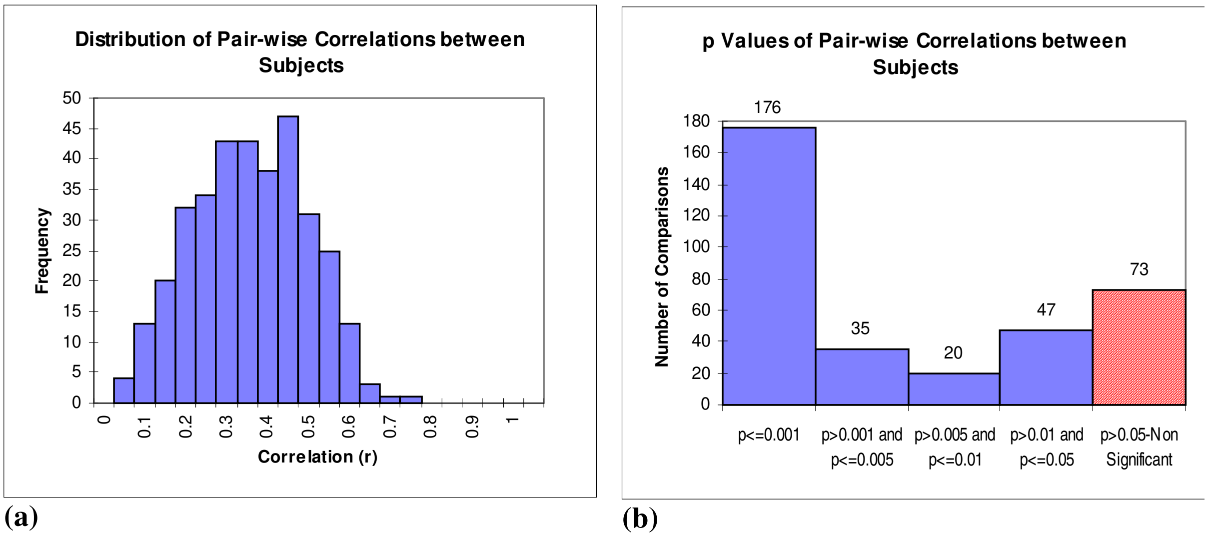


Figure 5.13: The results from all 27 subjects were compared with each other producing 351 pair wise comparisons. (a) shows the spread of the correlation co-efficient r over these comparisons (Average=0.33). (b) Shows the spread of significance of these correlations.

as a distance measurement and maximum similarity (minimum distance), single linkage to combine clusters (Hartigan, 1975). No grouping effect was apparent. Agreement between subjects varied considerably. The average correlation between any two subjects is quite low ($r = 0.33$) but the significance of the agreement between subjects is generally high (79% with a $p <= 0.05$) between all pairwise comparisons (see figure 5.13).

Subjects are sensitive to vowel quality differences in the materials but not strongly so.

5.5.2.2 By-Subjects and by-materials ANOVA.

The by-subjects ANOVA used subject linguistic background (Native English, Native North American, Non-Native) as a grouping variable with vowel (i, o, a) and COVA2 as calculated by the model (high, medium, low) as crossed variables.

Surprisingly the linguistic background had no significant effect on the responses. Subjects from Germany and Poland rated vowels similarly to Native English speakers. As I will discuss later this probably has more to do with the basic difficulty of the task than some underlying similarity in vowel sensitivity.

Similarly vowel type alone had no significant effect on results although there was a vowel/COVA2 interaction ($F(4, 96) = 4.15, p < 0.005$). However COVA2 group

($F(2,48) = 20.75, p < 0.001$) did have a significant effect on the subjects' responses. The means of the responses for spontaneous speech within each COVA2 group were as follows:

By-Subjects Responses			
COVA2 Group	High	Med	Low
Geometric Mean	0.883	0.799	0.777

This supported the hypothesis that the COVA2 model was modelling subjects' response to some extent. Low, medium and high COVA2 groups as decided by the COVA2 model reflected low, medium and high responses from subjects.

Following the insignificant effect of subjects' linguistic background these responses were pooled. In the by-materials ANOVA, sex of speaker, vowel type and COVA2 group were used as grouping variables.

The COVA2 group result persisted in the by-materials analysis ($F(2,72) = 3.71, p < 0.05$). Again the pattern of means supported the hypothesis:

By-Materials Responses			
COVA2 Group	High	Med	Low
Geometric Mean	0.69	0.625	0.582

The difference in significance between by-subject and by-materials analyses suggests there is too much variance unaccounted for in the materials. This suggests that COVA2 is a noisy measurement. From the evaluation of the method for calculating F1 and F2 achieved targets (section 5.4.3.3) we know that a proportion of the noise resides here. If the achieved targets calculated for a vowel are very unusual due to such noise they will produce very low COVA2 values. These low values correspond to very unusual and thus low probability locations in the vowel space (i.e. nowhere near the distribution of the speakers vowels). Thus very low COVA2 scores (more than 2 standard deviations from the mean) should be treated with suspicion.

5.5.2.3 Linear correlation between COVA2 as assigned by the statistical model and pooled subject responses.

Before carrying out a linear correlation between pooled subjects response and raw COVA2 score it was decided to remove low valued outliers (that is with a value lower than 2 standard deviations from the mean), firstly because of suspicions concerning their validity and secondly because of the large effect outliers can have on linear correlation tests. This removed 7 data points from the 90 vowels taken from spontaneous speech. The result was a weak but significant correlation

($r = 0.313, p < 0.005$).

The model appears to predict only about 10% of the subjects responses.

However bearing in mind the difficulty faced by subjects when carrying out the task of rating vowel goodness (average agreement $r = 0.33$) the statistical model performs comparatively well ($r = 0.313, p < 0.005$).

5.5.3 Summary

Can subjects reliably judge the clarity of vowels excerpted from spontaneous speech without duration cues? The answer is yes but it's hard. They reliably agree with each other about 10% of the time. Can the COVA2 score reliably predict the subjects' response to such vowels? Again the answer appears to be yes but, again, it's quite hard only predicting about 10% of the subjects' responses. Basically the COVA2 score is roughly as good – or bad – a predictor of any one listener's judgement as any other listener's judgement.

Vowel quality in spontaneous speech does contribute to subjects' perception of vowel 'goodness'. However the failure of subjects to agree on individual vowels suggests that this contribution is not a strong one. Duration is likely to be a primary factor. Of the 170,000 vowels segmented in the HCRC Map task nearly 100,000 are either too short to measure the spectral target reliably (less than 40ms) or were unvoiced. The materials we used in our perceptual experiment did not reflect these short vowels or devoiced vowels. In contrast to materials generated in 'clear speech' experiments, where the scale of vowel articulation varies from clear to very clear, in spontaneous speech the spectral quality of vowels often varies from poor to very poor. Perhaps in these conditions the difficulty in relying on spectral cues alone to perceive vowel quality leads to more reliance on segmental duration. However, in order to establish this, further experiments varying the duration of the segments used would be required.

Finally a clear problem with the approach taken in the modelling strategy is the fact that phonetic context is not taken into account. Rather than the model assigning a COVA2 score based solely on the F1/F2 targets of the vowel it might be more productive to assign this score on these values given the pre and/or post segmental context. However modelling these factors effectively using the statistical approach described here would require substantial quantities of controlled citation data from each speaker. It is also important to bear in mind that other acoustic factors such as spectral tilt, f0 and amplitude might also make an important contribution to any judgement of a vowel's 'goodness' in spontaneous speech. Although the model could be altered to take such factors into account it is not entirely clear how such factors should be automatically measured and incorporated.

5.6 General Summary

The measurements of care of articulation, especially COVA1 and COVA2, described in this chapter are noisier than I would like. For COVA1 and COVA2 noise is from the following sources:

- The formant tracker introduces errors due to nasalisation and by miscategorising female f0 as F1. The 10ms frame method used also causes variation in results. The start and end frame of a token will generally contain data from other tokens. In short vowels these transition frames have more influence than in long vowels.
- The autosegmentation is unreliable. Although this problem is mitigated by also using voiced speech to determine where vowels are it means that more data is lost than I would like. Again this is more of a problem for short tokens.
- Phonemic context is ignored and is known to play an important role in formant transitions.
- The parametric curve is only an approximation to the formant transitions and will not model the transition perfectly.
- The quality of the citation speech used to produce speaker models was variable.
- More information than F1 and F2 may be required to model vowel quality (for example amplitude, f0, spectral tilt, F3 etc.).

Despite room for improvement (some possible approaches to improving COVA2 are discussed in chapter 7) COVA2 does reflect human responses to the question of "how good is a vowel?". The achieved targets calculated from parametric curve fitting also reflect human judgements of F1 and F2 vowel targets. The methods are also well grounded on results from laboratory phonetics. Thus COVA1 and COVA2, together with DUR1 and DUR2, offer a practical solution to the problem set by this work, the quantitative assessment of the relationships between prosodic structure, redundancy and care of articulation over a large amount of spontaneous speech. As we will see in 6, even noisy measurements, when applied to a lot of data (200,000 syllables) produce interesting results.

Chapter 6

Results

6.1 Introduction

In chapter 2 I formalised the two central hypotheses that underly the work carried out here:

The Smooth Redundancy Hypothesis: Strong Version

Prosodic structure smoothes signal redundancy by controlling care of articulation

The Smooth Redundancy Hypothesis: Weak Version

Prosodic structure smoothes signal redundancy by controlling care of articulation except when it acts as a checking signal

To recap:

The strong hypothesis claims that, firstly, there is an inverse relationship between language redundancy (such as word frequency, trigram frequency) and acoustic redundancy (how clearly a sound is produced) and, secondly, that prosodic structure is responsible for effecting this relationship. Implicitly it suggests that this redundancy relationship can explain most of the effects that prosodic structure has on care of articulation and therefore the main reason prosodic structure exists in English.

The second hypothesis takes a weaker stance and accepts that another major factor, checking, modifies the relationship between prosodic structure, care of articulation and language redundancy. In chapter 2 it was argued that checking offered an alternative strategy to smoothing signal redundancy in order to produce robust communication at the signal level.

Figure 6.2 and figure 6.1 (taken from chapter 2) show the difference between these two hypotheses.

6.2 Testing the Hypotheses

The first step in testing these hypotheses is to confirm that prosodic structure does indeed relate to care of articulation. To do this we need to carry out a multiple linear regression using care of articulation (COA) metrics as the dependent variables and the prosodic structure factors as predictive variables.

On the basis of work reviewed in chapter 4 we would expect that the more prominent a syllable the greater the COA. For prosodic boundaries, although we would certainly expect prosodic boundaries to be associated with lengthening, it is less clear whether such lengthening increases care of articulation in terms of distinctiveness. This is an interesting question in itself, but, with regards the hypotheses it is peripheral. Providing we can show that part of prosodic structure significantly relates to COA then we can argue that prosodic structure could control COA.

This issue of *control* is central to both hypotheses and presents some difficulties. A strong correlation between factors, although supportive evidence of a causal relationship, does not necessitate one. For example, there is a strong, significant relationship between the number of radios purchased by year between 1940 and 1970 in the United States and the number of suicides. This does not mean that listening to DJs necessarily causes people to end their lives. It is only in the light of a theoretical prediction of causality that a correlation can be regarded as evidence of such causality.

However, if we look at figure 6.3 we can see that a traditional view of prosodic structure does imply such a causal connection. It suggests that prosodic structure controls the way speech sounds are realised. An alternative view might be to regard prosodic structure as simply emergent from phonetic structure. In the same way shadows are produced by an interaction between light and solid objects perhaps prosodic structure is produced by an interaction between phonetics and language. However evidence from both psycholinguistics and phonetics (see chapter 3) does suggest that prosodic structure exists, and that it does affect phonetics, and thus the acoustic realisation of speech, and thus, in the work reported here, potentially control care of articulation.

The second step in testing the hypotheses is to confirm that language redundancy

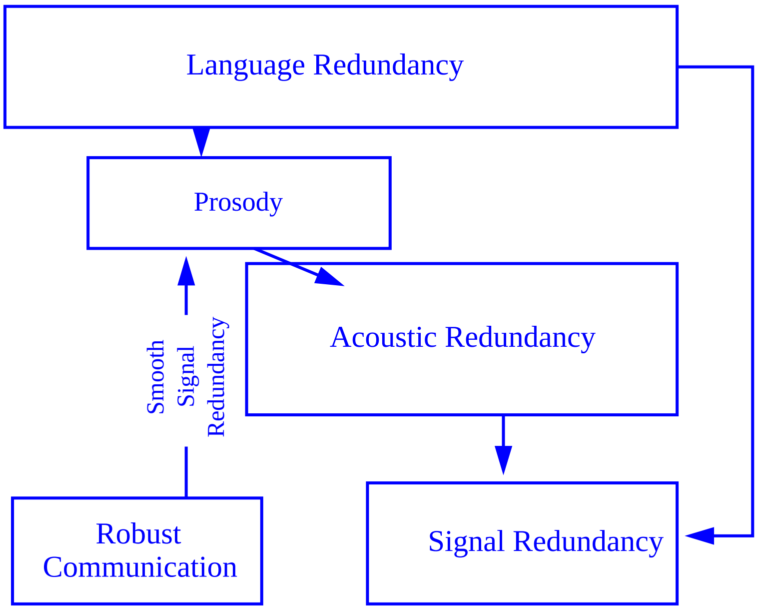


Figure 6.1: Strong smooth redundancy hypothesis.

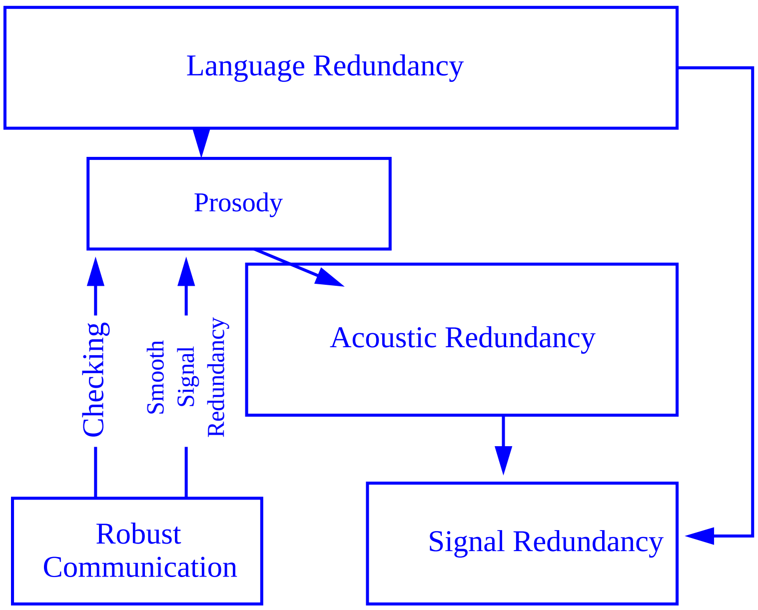


Figure 6.2: Weak smooth redundancy hypothesis.

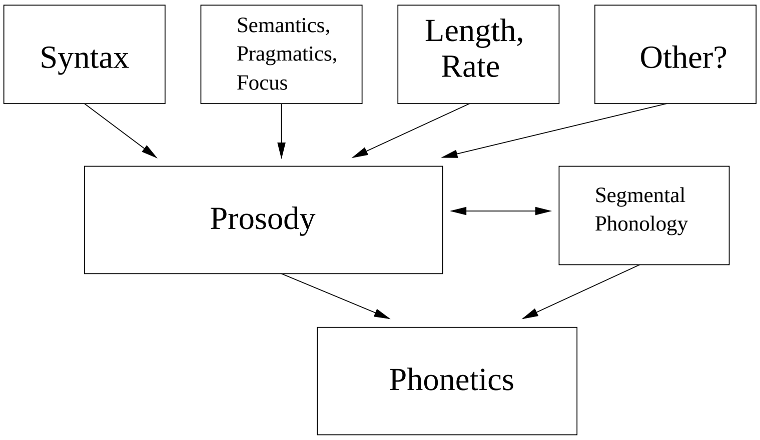


Figure 6.3: One view of the role of the prosodic component of the grammar (taken from Shattuck-Hufnagel and Turk, 1996, p237).

has an inverse relationship with COA metrics. This would show that signal redundancy is indeed smoothed by changes in COA. The stronger this relationship the more we can argue that smoothing is the main result of changes in COA and evidence for the strong hypotheses described above. In contrast, if such a strong relationship exists only in contexts where a checking signal is unlikely to be present, this is evidence which supports the weaker hypothesis.

If we do see strong, significant correlations between prosodic factors and COA and an inverse relationship between language redundancy and COA the third step is to examine how independent prosody and redundancy are with regards to COA.

Using maximum likelihood we can determine the extent the predictive power of a prosodic model and a redundancy model are shared. The less predictive power is shared, the more independent the models are of each other. Both hypotheses predict a strong shared contribution. In order for prosodic structure to implement smoothing it must alter COA in the **same way** as language redundancy. Both hypotheses also predict that the independent contribution from the redundancy factors is small. If redundancy makes a large contribution independent of prosody then prosodic structure is not implementing much of the significant effects of redundancy. If this is the case then smoothing is being carried out either by direct reference to redundancy factors or by other means.

In contrast, if prosodic factors show a strong independent contribution in addition to a large shared contribution then this undermines the strong smooth redundancy hypothesis. This is because it suggests prosodic structure is altering COA in a ways which **do not** smooth signal redundancy.

However, if this independent prosodic effect is limited to likely checking locations, and absent when checking is unlikely to occur, we can argue that the occasions when prosody does not smooth signal redundancy are the occasions when prosody is producing a checking signal. This would support the weak smooth redundancy hypothesis.

6.2.1 Summary

To summarise we are looking for the following to support the hypotheses:

- Prosodic factors show a strong significant correlation with COA metrics.
- Language redundancy factors show a strong, significant and inverse correlation with COA metrics.
- The shared contributions of the redundancy and prosodic models is high.
- The independent contribution of the redundancy model is low.
- That, for the strong hypothesis, the independent contribution of the prosodic model is low, **or**, for the weak hypothesis the independent contribution of the prosodic model is low in contexts which exclude potential checking locations.

6.3 Establishing Confidence in the Coding and Care of Articulation Metrics

While testing the hypotheses as described above we can also establish confidence in our coding and measurements. This is achieved by examining the direct relationship between prosodic structure and the care of articulation dependent variables (DUR1, DUR2, COVA1, COVA2) and also between the redundancy factors and these variables. If our measurement are effective we would expect these relationships to support previous findings in laboratory phonetics.

These can be summarised as follows:

Prosody:

- Syllables are lengthened before prosodic boundaries. The stronger the boundary the greater the lengthening.

- Syllables are lengthened by prominence. The greater the level of prominence the greater the lengthening.

- Vowels in syllables are more clearly articulated the greater the level of prominence. This leads to more extreme vowel targets and decreased centralisation.

Redundancy:

- The more predictable a syllable, as in greater word frequency, increased trigram likelihood and/or more mentioned, the less carefully the syllable is articulated resulting in shorter syllables and less carefully articulated vowels.

Once I have established how effectively the different factors and measurements model these expected relationships, and shown the nature of the direct relationships between prosody and COA as well as redundancy and COA, I will then examine the extent, and the contexts, prosodic factors and redundancy factors are independent of each other. In doing so we can test whether redundancy is implicitly represented in prosodic structure and the extent prosody may also be used as a checking signal.

6.4 Methodology

The procedure for analysis is as follows:

1. Carry out a multiple linear regression with appropriate factors to examine the degree and significance of these factors as separate prosody and redundancy models in predicting each of the four dependent variables.

2. For both the redundancy and prosodic models compare a set of reduced models, each with a factor removed. Use these comparisons to calculate the independent contribution and significance of each factor in each model using maximum likelihood (also termed the likelihood ratio test) (Neter *et al.*, 1990).

3. Graph significant results (and, where appropriate, non-significant results) to give a clear impression of the size and direction of these effects. In all following results if the direction of the effect is not discussed it was

in the expected direction. For example more redundant = less carefully articulated, prominent = more carefully articulated etc.

4. Finally, use maximum likelihood to calculate the independent contribution of prosodic and redundancy models in predicting care of articulation in spontaneous speech.

6.4.1 Materials and Coding: Review

This work is based on a large corpus of spontaneous task oriented dialogue collected by the HCRC at the University of Edinburgh - the HCRC Map Corpus (Anderson *et al.*, 1991). The corpus is comprised of about 15 hours of spontaneous speech, 64 speakers and around 200,000 syllables.

As explained in chapter 3 each data point in this analysis is a syllable. The syllables are coded with prosodic, redundancy and care of articulation factors. Not all syllables have the same coding:

- Any syllables which were not coded for redundancy factors (such as syllables forming words unknown to the BNC corpus) were ignored. This removed just under 10% of the data (18225 syllables).

- a proportion have been prosodically hand coded so in addition to the lexical and automatic factors these syllables are also coded for accent and break index.

- A proportion of syllables, those that are within references to landmarks on the maps used in the dialogues, have also been coded for mention.

The overlap between these groups and the entire data set is shown in figure 6.4.

Two further factors affect the total number of syllables examined in each analysis:

1. When examining the DUR1 and DUR2 measure of care of articulation (raw syllabic duration and normalised syllabic duration respectively) all these materials are examined because every syllable has a duration. However for COVA1 and COVA2 (vowel centralisation and vowel quality respectively) some of these syllables remain uncoded because, in order to measure care of vowel articulation the syllable must have a vowel nucleus as well as at least 40ms of voiced speech in order to fit a parametric curve to formant values.

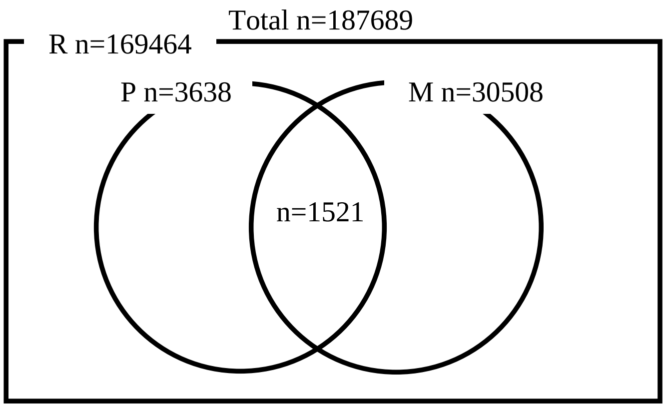


Figure 6.4: Materials examined in the analysis. R: The number of syllables with valid word frequency and trigram information. P: The number of syllables with hand coded prosodic factors. M: The number of syllables coded for mention.

2. The *weak* redundancy hypothesis (see chapter 2) claims that redundancy factors are only important when no boundary checking signal is present. To test this all syllables with a pause following them or that were part of polysyllabic words are removed from the analyses.

Table 6.1 shows the number of materials for all possible conditions: DUR, COVA, DUR+Weak, COVA+Weak.

	Total Coded	Mention Coded	Prosodic Coded	Mention + Prosodic
DUR Coded	169464	30508	3638	1521
COVA Coded	71747	13366	1482	707
DUR Coded + Weak	89532	12295	1186	205
COVA Coded + Weak	32213	4654	438	122

Table 6.1: Number of syllables in each condition.

The analyses carried out can be grouped as follows:

1. Prosodic factors: A test of the relationships between the different prosodic factors and the dependent care of articulation variables.

2. Redundancy Factors: A test of the relationships between the different redundancy factors and the dependent care of articulation variables.

3. Independence of Redundancy and Prosody: A test of the extent prosodic factors implicitly account for redundancy effects and the extent redundancy

factors offer an independent contribution to predicting the dependent variables.

6.4.2 Summary of Variables and Factors for each Coding Set

Before considering the results from these analyses I will first give a brief summary of the variables and coding used (see chapters 2, 3, 5 for details).

Independent Variables: Prosody

- Prosodic Boundaries: Binary variables

  wboun: Word boundary. This corresponds to a ToBI break index of 1.

  iip2boun: Intermediate Intonational Phrase with a ToBI break index of 2.

  iip3boun: Intermediate Intonational Phrase with a ToBI break index of 3.

  ipboun: Full Intonational Phrase Boundary. This corresponds to a ToBI break index of 4.

  Aipboun: Automatically coded Full Intonational Phrase Boundary. For materials not hand coded, if the syllable was followed by a pause it was regarded as having a high likelihood of being followed by a full intonational phrase boundary.

- Prominence: Binary variables

  vtype: Vowel type. Whether the vowel is full or reduced (where reduced equals unstressed /I,ə/). This corresponds to the first level of prominence described by Ladefoged.

  lexstr: Lexical stress. Whether the syllable is lexically stressed. This corresponds to the second level of prominence described by Ladefoged and the first level of prominence as described by Cruttenden. (lexstr is not strictly a binary variable as although primary lexical stress is coded as a 1, secondary stress is also coded as 0.5.)

  acc: Phrasal Accent. Whether a phrasal accent has been marked using ToBI. This corresponds to the second level of prominence as described by Cruttenden.

  Aacc: Automatically coded Phrasal Accent. For materials not hand coded, if the syllable was lexically stressed **and** open class, it was marked as having a high likelihood of having a phrasal accent.

pps: Primary Phrasal Accent. The last accent before an intermediate or full intonational phrase boundary (as coded using ToBI) is marked as having primary phrasal stress. This corresponds to the third level of prominence described by Ladefoged and the third level of prominence as described by Cruttenden. Automatic coding of primary phrasal stress was considered too unreliable.

- Spillover

spill: This factor is based on work by Turk and White (Turk and White, 1999) and is used in hand coded prosodic data only. It represents the amount durational effects of prominence spill over from an accent. This is mostly in a rightward direction (20%), leftwards by much less (5%), when no word boundary blocks the effect. When a word boundary is present only a spill of 4% is reported in a rightward direction.

Independent Variables: Redundancy

wf: Word Frequency. The log of the COBUILD word frequency of the word containing the syllable.

trigram: Trigram Probability. The log probability of guessing the syllable correctly based on the two syllables preceding it.

men: How many times a particular landmark has been referred to in the dialogue up until this point. Only references to landmarks in the HCRC Map Corpus are coded in this way.

Dependent Variables: Care of Articulation

DUR1. Raw syllabic duration in milliseconds.

DUR2. Syllabic duration normalised for number of segments and based on chained log normal distributions. Measured in k which are a combined z score for the chained distributions (see chapter 5 section 5.3.2).

COVA1. Centralisation. How close to the centre of a speaker's vowel space the vowel targets were realised. Measured in distance in Bark normalised across F1 and F2.

COVA2. Clear Speech Target. How likely a model of a speaker's clear speech would have generated the vowel target. Measured in average log probability of the clear speech model producing the target values.

6.4.3 The Problem with Using Total Number of Syllables as a Prosodic Factor

It has been shown that the same syllable in a polysyllabic word tends to be more reduced the greater the total number of syllables (e.g. Campbell, 1992). Thus the total number of syllables would appear to be an important prosodic factor in predicting care of articulation at least in terms of duration change.

However it was found that if this factor is included together with lexical stress and word boundary information, not only is the independent contribution of this factor to predicting duration change very small (0.05%) but, although significant, it predicts a greater rather than a reduced duration.

Two reasons account for this result:

1. When examined alone number of syllables does behave as expected although predicting less variance than word boundary and lexical stress factors. This may be because over 88% of all syllables in the HCRC Map Corpus are within words with either one or two syllables (with 75% being monosyllabic) giving little scope for number of syllables to act as an accurate predictor in most cases.

2. The effect is in the unexpected direction when these factors are included because number of syllables correlates strongly and negatively with word frequency ($r = -0.55$ $p < 0.001$). Overall in the HCRC Map Corpus, once lexical stress and word boundary are taken into account, the number of syllables in the word no longer predicts shorter syllables but instead predicts less frequent words which in turn predict longer syllables.

For these reasons the number of syllable factor was removed from this analysis in favour of lexical stress and word boundary information. As a final check analyses were carried with number of syllables as a controlling factor. This was accomplished by carrying out linear regressions separately over syllables in monosyllabic, bisyllabic and trisyllabic words. There was no indication that total number of syllables was a confounding factor for either redundancy or prosodic factors when controlled for in this way. Thus in all further analyses reported here total number of syllables is ignored.

6.4.4 Do Results from these Materials and Measurements Support Results Obtained in Laboratory Phonetics: Prosody and Care of Articulation

6.4.4.1 DUR1/DUR2

The r, r^2 for the hand coded prosodic model as well as the independent contributions from each factor are shown for DUR1 and DUR2 in tables 6.2, 6.3.

Figures 6.5, 6.6 on page 132 show the average values for each factor. In general results do confirm previous laboratory results apart from leftwards within word spillover as reported by Turk and White (1999). The differences between DUR1 and DUR2 are not great. Raw syllabic duration appears to be a surprisingly good durational measure of care of articulation. However DUR2 is more sensitive to prosodic change showing greater differences within prominence and boundary effects as well as a neater linear relation between strength of prominence and average DUR2. However it is possible that including number of segments within a syllable in the prosodic analysis and using raw duration may produce better results than using number of segments as a normalising factor in DUR2. Overall the hand coded and lexical prosodic factors together account for nearly 60% of the variation in syllabic duration.

6.4.4.2 COVA1/COVA2

The r, r^2 for the hand coded prosodic model as well as the independent contributions from each factor are shown for COVA1 and COVA2 in tables 6.4, 6.5. As expected COVA1 as a raw centralisation measure was very sensitive to vowel type. In general results are mostly insignificant.

Work reviewed in chapter 4 suggests that prominence should have a strong effect on vowel articulation. Although the low r value suggests that these measurements are very noisy the results are still disturbing in that a number of the significant results in COVA1 are contrary to the theoretical predictions. Although phrasal accents and break index 2 are significant the direction of the relationship is the reverse to what we would expect. The results for these materials suggest that accented syllables and syllables with a break index of 2 are articulated less clearly rather than more clearly.

For COVA2 only the nuclear/non-nuclear accent distinction is significant. This is interesting, especially as this distinction appeared to have no impact on duration

DUR1: Raw Syllabic Duration			
	Regression Results	$r = 0.7710$	$r^2 = 0.5944$
Prosodic Factor	Independent Contrib. to r^2	F(1,3638)	p value
vtype	00.09%	8.59	0.01
lexstr	01.17%	105.52	0.001
acc	03.30%	296.41	0.001
spill	01.21%	109.12	0.001
pps	00.01%	0.90	NS
wboun	06.97%	625.73	0.001
iip2boun	00.72%	64.74	0.001
iip3boun	00.04%	3.98	0.05
ipboun	00.15%	14.02	0.001

Table 6.2: Regression analysis of hand coded prosodic factors against raw syllabic duration. See section 6.4.2 for details of factors.

DUR2: Normalised Syllabic Duration			
	Regression Results	$r = 0.7238$	$r^2 = 0.5238$
Prosodic Factor	Independent Contrib. to r^2	F(1,3637)	p value
vtype	00.45%	34.85	0.001
lexstr	04.54%	346.95	0.001
acc	02.70%	206.79	0.001
spill	02.23%	170.39	0.001
pps	00.00%	0.37	NS
wboun	07.46%	569.88	0.001
iip2boun	00.74%	56.92	0.001
iip3boun	00.00%	0.11	NS
ipboun	00.07%	5.85	0.05

Table 6.3: Regression analysis of of hand coded prosodic factors against normalised syllabic duration. See section 6.4.2 for details of factors.

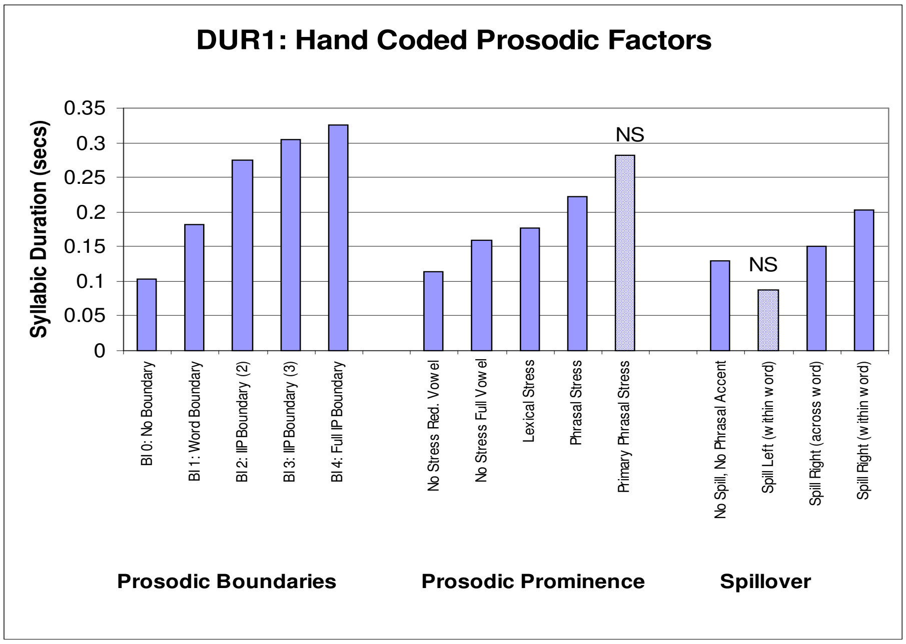


Figure 6.5: Prosodic boundaries, as they increase in strength (*BI*=Break Index), are associated with longer mean syllable duration. Similarly as prominence increases mean syllabic duration increases. Spillover in a rightwards direction confirms results reported by Turk and White (Turk and White, 1999).

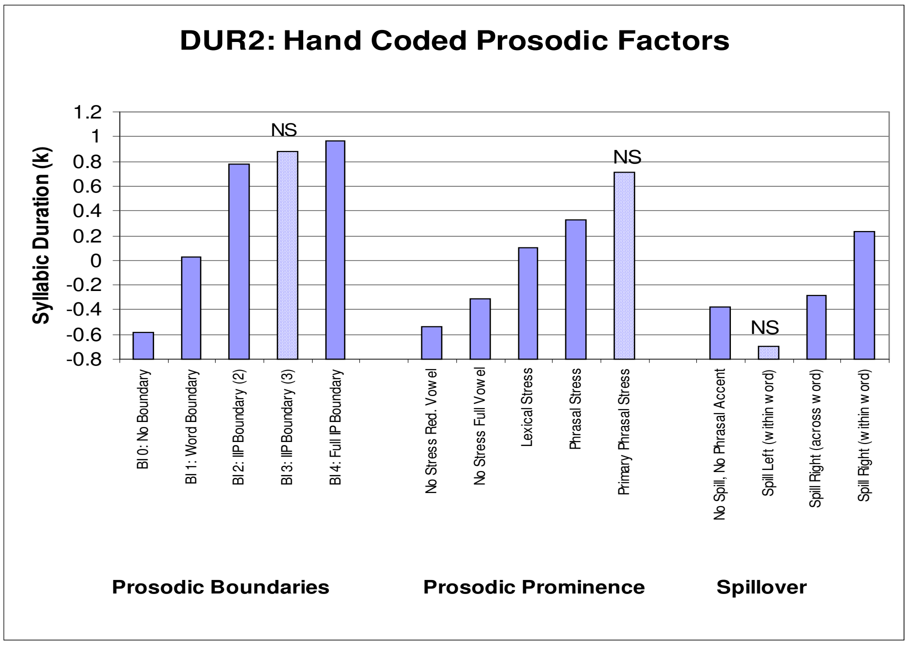


Figure 6.6: Results are very similar to those shown for DUR1 (figure 6.5) except that the differences between Break Index 2-4 are less marked and in the case of Break Index 3 non significant.

COVA1: Vowel Centralisation			
	Regression Results	$r = 0.1336$	$r^2 = 0.0178$
Prosodic Factor	Independent Contrib. to r^2	F(1,1481)	p value
vtype	00.55%	8.38	0.01
lexstr	00.23%	3.60	NS
acc	00.45%	6.85	0.01
spill	00.17%	2.64	NS
pps	00.20%	3.13	NS
wboun	00.11%	1.75	NS
iip2boun	00.44%	6.71	0.01
iip3boun	00.00%	0.02	NS
ipboun.	00.25%	3.86	0.05

Table 6.4: Regression analysis of hand coded prosodic factors against vowel centralisation. See section 6.4.2 page 127 for definitions of factors.

COVA2: Vowel Targets			
	Regression Results	$r = 0.1042$	$r^2 = 0.0109$
Prosodic Factor	Independent Contrib. to r^2	F(1,1481)	p value
vtype	00.18%	2.63	NS
lexstr	00.14%	2.00	NS
acc	00.01%	0.08	NS
spill	00.05%	0.68	NS
pps	00.31%	4.52	0.05
wboun	00.04%	0.61	NS
iip2boun	00.09%	1.26	NS
iip3boun	00.24%	3.55	NS
ipboun.	00.06%	0.86	NS

Table 6.5: Regression analysis of hand coded prosodic factors against vowel targets. See section 6.4.2 page 127 for definitions of factors.

scores. It is possible that at this high level of prominence duration cannot be further extended but care of vowel articulation can be increased. However, given the poor results overall, it is difficult to have much confidence in this observation.

One direct cause of these poor results is that not only are the COVA measurements noisy, but that they are not representative of the data as a whole. As described in chapter 5 COVA measurements were only taken for vowels which remained voiced long enough to analyse the formant tracks using conventional target-undershoot techniques to assess achieved targets. For 60% of the syllables in the spontaneous speech in the corpus the vowels were either unvoiced or too short (less than 40ms) to be measured in this way. If we examine the proportion of syllables that could not be measured for COVA1 and COVA2 we see a strong relationship with regards to prosodic factors(table 6.6). For example, 80% of syllables that were marked as having no prominence could not be measured for COVA1 and COVA2 while in contrast only 25% of syllables marked as carrying a primary phrasal accent could not be measured. Thus prosodic category is an important conditioning factor on whether we have a COVA1/2 measurement to consider and therefore COVA1/2 naturally produces a rather unrepresentative data set.

By doing so, much of DUR1/2 variance predicted by prosodic factors is removed. In table 6.7 we can see that the standard deviation for DUR1/2 for these unmeasured syllables is similar to those measured. Thus about half of the duration variation which forms the basis of the results detailed for DUR1 and DUR2 are within these unmeasured syllables

Prosodic Prominence	% measured by COVA	% too short for COVA	Prosodic Boundary	% measured by COVA	% too short for COVA
none	19.5	80.5	none	30.4	69.6
+vtype	37.6	62.4	wboun	38.0	62.0
+lexstr	44.7	55.3	iip2boun	66.3	33.7
+acc	65.7	34.3	iip3boun	73.3	26.7
+pps	74.0	26.0	ipboun	69.5	30.5

Table 6.6: The proportions of prosodic types coded by COVA1/2.

This raises a number of possibilities for the failure of the COVA measurements to reflect major findings with regards to spontaneous speech in the hand coded prosodic data set:

- Differences in care of articulation in spontaneous speech and how they reflect prosodic structure are below the range measurable with these care of vowel

	% too short for COVA	% measured by COVA
DUR1		
mean	0.134	0.237
sd	0.086	0.122
DUR2		
mean	-4.13	0.498
sd	0.842	0.697

Table 6.7: DUR1/2 mean and standard deviations of materials that could and could not be measured by COVA1/2.

articulation techniques.

- The generalisation of undershoot and centralisation models to compare different vowels in different phonetic contexts is inadequate. In chapter 7 I will discuss possible improvements to the modelling approach used here and discuss how phonemic context and identity could be included in the model.

- The COVA measurements are only representative of the data in longer stressed syllables. Due to noise, and the much smaller set of materials hand coded for prosodic factors, no significant results are obtained.

- The interaction between which materials could be measured and the prosodic factors confound the results.

Despite the disappointing performance of the COVA metrics we will return to them when analysing the whole corpus with redundancy factors and the prosodic factors available for the whole data set. This is because of the much larger size of the full data set (200000 vs 3000). This huge number of tokens may help counter problems of noise in the COVA metrics.

6.4.4.3 Examining Prosodic Effects over the Whole Corpus

Although the corpus as a whole consists of nearly 200,000 syllables due to time constraints only about 3,500 could be prosodically hand coded for break index and accent.

However a number of the prosodic factors can be applied to this larger set, namely, vowel type, lexical stress, word and syllable boundary. In addition estimations of full IP boundaries and phrasal accents are also considered. These estimations are assigned IP boundaries on the basis of a syllable being followed by a pause and for phrasal accents by examining the lexical class of the word (see sections 6.4.4.3.2, 6.4.4.3.1 below).

6.4.4.3.1 Guessing Accented Syllables: In general an accent will only occur on a stressed syllable. In general accents occur much more frequently in open class content words such as 'beach' than in closed class function words such as 'the'.

If, on the basis of this, we automatically assign phrase accents to stressed syllables in open class words and then compare the results with the hand coded data (table 6.8) we find that just over 60% of accents are correctly coded with a false alarm rate of just under 16% (The number of unaccented syllables incorrectly coded). This is sufficiently accurate to give an idea of potential accentedness in the whole corpus.

	-Aaac	+Aaac
-acc	**2105 (84.1%)**	398 (15.9%)
+acc	435 (38.0%)	**710 (62.0%)**

Table 6.8: The number of accurately guessed phrasal accents in hand coded materials.

6.4.4.3.2 Guessing IP boundaries: I make no attempt to guess Intermediate IP boundaries. Such boundaries are less common than full intonational phrase boundaries. Also, as we saw in section 6.4.4, Break Index 2 and 3 do not have as strong an effect on care of articulation (in terms of DUR1/2) as Break Index 0, 1 and 4. On this basis Intermediate IP boundaries are ignored in the automatic analysis.

By regarding every syllable before a pause as being at the edge of an IP we guess 90% of coded IPs (table 6.9). Less than 2% of word or syllable boundaries are incorrectly coded as IP boundaries. IIP boundaries (Break Index 2/3) account for 6% of all boundaries mistakenly coded as a full IP boundary.

The automatic coding (these two automatic coding together with vowel type, lexical stress, word boundary and syllable boundary data) compares well with fully hand coded factors over the hand coded materials. Automatic coding predicts 45% ($r = 0.6714$) of the variance of DUR1 and 35% ($r = 0.5943$) of the variance

	-Aipboun	+Aipboun
none	**1403 (99.1%)**	13 (0.9%)
wboun	**1534 (98.7%)**	20 (1.3%)
iip2boun	*66 (69.5%)*	*29 (30.5%)*
iip3boun	*66 (62.9%)*	*39 (37.1%)*
ipboun	48 (10.0%)	**430 (90.0%)**

Table 6.9: The number of accurately guessed phrase boundaries in hand coded materials.

of DUR2 whereas hand coding predicts 59% ($r = 0.7710$) of the variance of DUR1 and 52% ($r = 0.7238$) of the variance of DUR2.

6.4.4.4 DUR1/DUR2: Whole Corpus with Automatic Prosodic Coding

Examining the automatic coding and all materials with DUR1/DUR2 measurements we see that all prosodic factors significantly and independently predict these variables (see tables 6.10, 6.11 on page 139).

Prosodic boundaries account for the majority of the effect, in particular the automatically tagged IP boundary. However prominence also makes a strong contribution. See figure 6.7 and figure 6.8 for the magnitude and directions of the automatic prosodic factors. Overall the results suggest that the DUR1/DUR2 measurements and automatic prosodic factors are behaving as we expect from literature reviewed in chapter 4.

6.4.4.5 COVA1/COVA2: Whole Corpus with Automatic Prosodic Coding

The large amount of data considered in this full analysis was sufficient to uncover significant prosodic effects for both COVA metrics. Given the very low r values and the reservations discussed in section 6.4.4.2 these results should be treated with caution. However as argued in chapter 5 one main function of the COVA metrics was to act as a comparison with the DUR measurements.

In both cases the prosodic factors predicted less than 1% of the variance of the COVA metrics. All factors were significant for COVA1.

Significance at these very low r values and very high population sizes deserves some discussion. Significance can be thought of as indicative of a tendency but not as predictive in these contexts. A good analogy is the significant effect of left

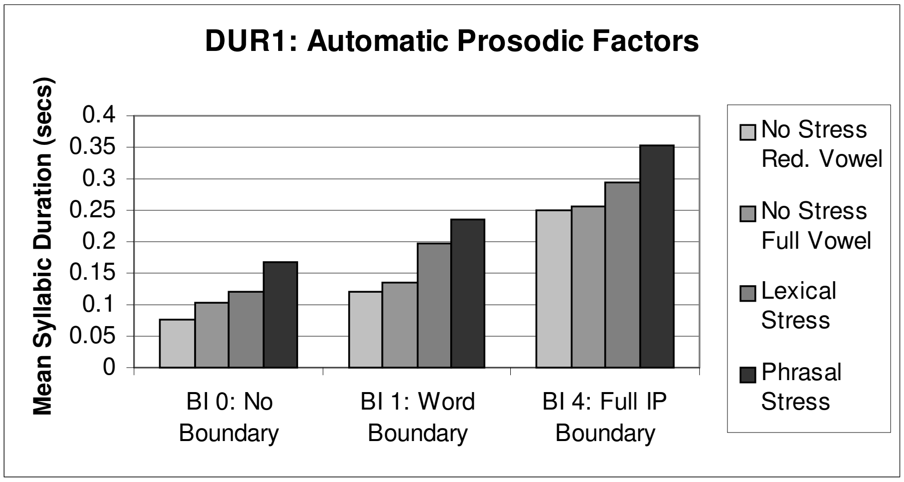


Figure 6.7: Effect of automatic prosodic factors (boundary and prominence) on DUR1.

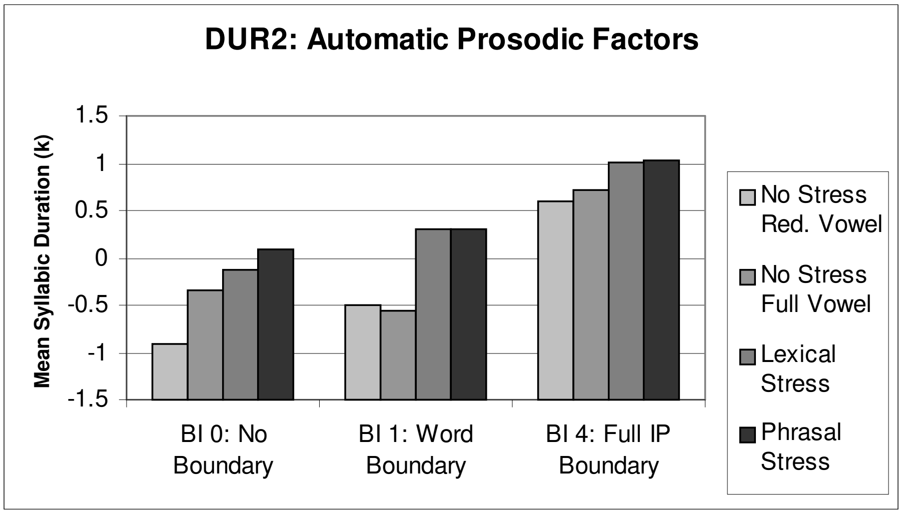


Figure 6.8: Effect of automatic prosodic factors (boundary and prominence) on DUR2.

DUR1: Raw Syllabic Duration			
	Regression Results	$r = 0.6473$	$r^2 = 0.4190$
Auto Prosodic Factor	Independent Contrib. to r^2	F(1,169461)	p value
vtype	01.08%	3139.49	0.001
str	00.83%	2421.31	0.001
apacc	01.49%	4335.15	0.001
awboun	03.62%	10561.72	0.001
aip.	19.72%	57523.91	0.001

Table 6.10: Regression analysis of automatic prosodic analysis with raw syllabic duration. See section 6.4.2 page 127 for definitions of factors.

DUR2: Normalised Syllabic Duration			
	Regression Results	$r = 0.6077$	$r^2 = 0.3693$
Auto Prosodic Factor	Independent Contrib. to r^2	F(1,169461)	p value
vtype	00.37%	997.12	0.001
str	03.31%	8901.64	0.001
apacc	00.03%	79.00	0.001
awboun	01.46%	3926.77	0.001
aip.	13.10%	35208.99	0.001

Table 6.11: Regression analysis of automatic prosodic analysis with normalised syllabic duration. See section 6.4.2 page 127 for definitions of factors.

handedness on life expectancy[1]. Left handedness is a highly significant factor in life expectancy however the amount of variance it explains is very small (a shorter life of about two weeks). However the fact it is significant is important and is indicative of an underlying cause. In the same way the significant results obtained for both prosodic factors and redundancy with COVA metrics do indicate an underlying relationship. However the weakness of the relationship means, that in this work they form a basis for discussion and speculation but are not used to justify any hypotheses.

Firstly looking at the results for COVA1 (table 6.12) we see that the results are very strongly affected by whether the vowel is full or reduced. This is not unexpected considering that COVA1 attempts to measure centralisation. More surprising is that lexically stressed vowels appear more centralised than unstressed

[1]Thanks to Paddy O'Donnell at the Psychology Department, University of Glasgow for explaining this analogy to me.

full vowels (see figure 6.9). However, given the very small predictive power of the regression and the comparatively strong effect of vowel type this result should be treated with caution. In contrast with the results from DUR1/DUR2 neither prosodic boundary factor increased COVA1 instead predicting less centralisation when absent rather than more. I will return to the lack of boundary effects after looking at COVA2 results. In general phrasal stress does contribute a tiny positive effect but again the impact of vowel reduction may be undermining this result.

The COVA2 results are more interesting (table 6.12). The strong effect of vowel reduction is absent. This supports the idea that COVA2 is better at measuring undershoot in individual vowels than COVA1. Like COVA1 the measurement is noisy. However phrasal stress does seem to have a highly significant effect whereas all other prominence factors do not. Of the pitifully tiny 0.5% predictive power of COVA2 that the model achieves most of this is from the automatically guessed phrasal accent factor. Looking at figure 6.10 we see this difference whatever the prosodic boundary context.

Ignoring the vowel type effect in COVA1 and taking COVA1/COVA2 results together it seems that these vowel articulation measurements are more sensitive to phrasal stress than prosodic boundaries. Unfortunately the noisy nature of the metrics make this observation far from conclusive. The results hint at the following:

- Care of articulation in terms of spectral quality is only within speakers' control in already relatively prominent syllables. This would explain the lack of a COVA1/COVA2 effect between lexically stressed, and unstressed syllables, when no pitch accent was present. The majority of the vowel studies which showed spectral quality differences related to lexical stress only (van Bergem, 1988, for example) were on citation speech. In fast, running spontaneous speech where 60% of the vowels are less than 40ms these effects seem to disappear.

- Although prosodic boundaries have a strong effect on duration they do not appear to have a strong effect on vowel articulation. In accented syllables such vowel effects appear almost independent of lengthening due to prosodic boundaries. This could be used by the human language system to differentiate duration change signalling a boundary (the checking signal discussed in chapter 2) and careful articulation used to smooth the overall signal redundancy.

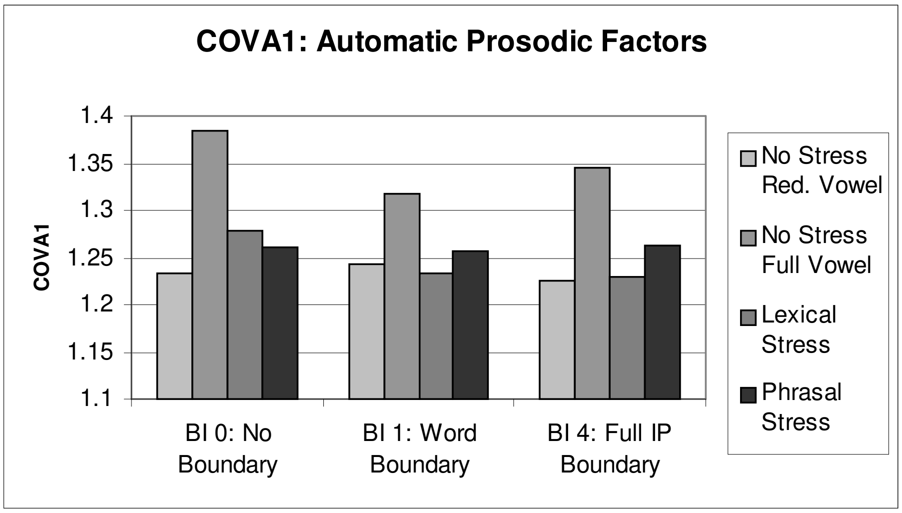


Figure 6.9: Effect of automatic prosodic factors (boundary and prominence) on COVA1.

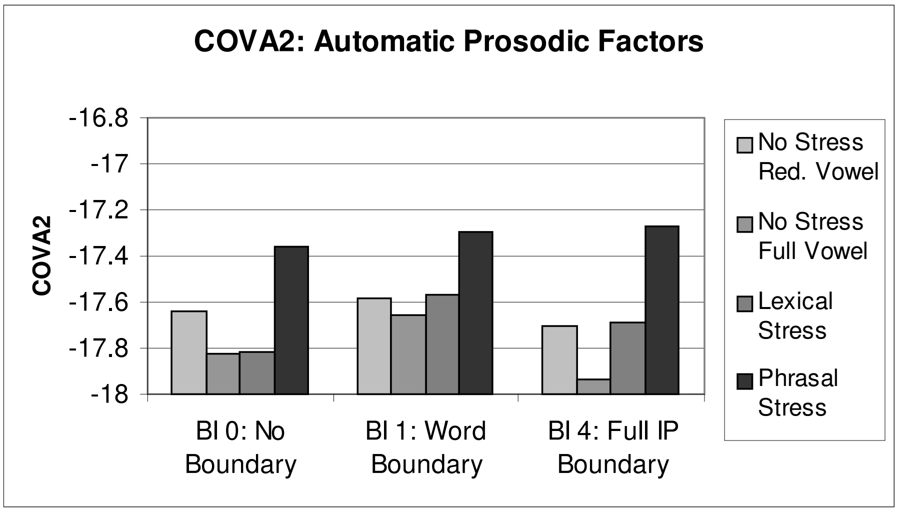


Figure 6.10: Effect of automatic prosodic factors (boundary and prominence) on COVA2.

COVA1: Vowel Centralisation			
	Regression Results	$r = 0.0579$	$r^2 = 0.0033$
Auto Prosodic Factor	Independent Contrib. to r^2	F(1,80577)	p value
vtype	00.29%	234.37	0.001
str	00.21%	171.23	0.001
apacc	00.01%	8.00	0.01
awboun	00.01%	7.95	0.01
aip.	00.01%	10.09	0.01

Table 6.12: Regression analysis of automatic prosodic analysis with vowel centralisation. See section 6.4.2 page 127 for definitions of factors.

COVA2: Vowel Targets			
	Regression Results	$r = 0.0729$	$r^2 = 0.0053$
Auto Prosodic Factor	Independent Contrib. to r^2	F(1,80577)	p value
vtype	00.00%	1.13	NS
str	00.00%	0.09	NS
apacc	00.37%	304.87	0.001
awboun	00.04%	30.23	0.001
aip.	00.00%	3.78	NS

Table 6.13: Regression analysis of automatic prosodic analysis with vowel targets. See section 6.4.2 page 127 for definitions of factors.

However as emphasised earlier this is really just speculation. A much more robust measure of vowel undershoot would be required to test these ideas in spontaneous speech.

6.4.5 Do Results from these Materials and Measurements Support Results Obtained in Laboratory Phonetics: Redundancy and Care of Articulation

Generally if things are predictable we would expect them to be shorter and less carefully articulated. We have three redundancy measurements, log of COBUILD word frequency, the log of the syllabic trigram prediction and the number of times the referent has already been mentioned in the dialogue.

We would firstly expect these factors to have a significant effect on the dependent variables and we would expect to see these dependent care of articulation variables fall in value as these factors increase in value.

6.4.5.1 DUR1/DUR2

Results from the regression and maximum likelihood analysis of these factors are complicated by the different data sets we have to consider.

I will first consider word frequency effects and trigram probability on DUR1 and DUR2 over the entire corpus. Then I will examine these factors together with mention over only those materials with mention coding and finally both these analyses again but only looking at syllables in a monosyllabic context with no intonational phrase boundary following them. This final context controls for any possible checking effect (see chapter 2) and will be used to support the weak smoothing redundancy hypothesis.

6.4.5.1.1 Word frequency effects and trigram probability effects on DUR1/DUR2 over the entire corpus: If we look at tables 6.14 and 6.15 we see that these factors predict about 15% of the variation in DUR1 and 9% of the variation of DUR2. Both factors are highly significant in both cases. Looking at figures 6.11 and 6.12 we can see that as expected the more redundant the syllable the shorter it tends to be.

6.4.5.1.2 Mention effects on DUR1/DUR2 over mention coded part of corpus: If we look at tables 6.16 and 6.17 we see that a significant mention effect is present although the independent contribution it makes to the model is less than 1%. Again looking at figures 6.11 and 6.12 we can see that as with the factors over the entire corpus the more a syllable in a referent has been mentioned the shorter it becomes.

Also of interest is that over these mention coded materials the redundancy models as a whole are more predictive (31% of the variation in DUR1 and 27% of the variation in DUR2) than for all materials. This is probably due to the more homogeneous nature of this reference coded material. There will be few verbs, open class words will tend to be adjectives and two syllable nouns (such as "white mountain"), and the function words will mostly consist of pronominals, deictics and determiners.

DUR1: Raw Syllabic Duration: All			
	Regression Results	$r = 0.3811$	$r^2 = 0.1452$
Redundancy Factor	Independent Contrib. to r^2	F(1,169464)	p value
wf	03.20%	6357.96	0.001
trigram.	06.28%	12454.40	0.001

Table 6.14: Regression analysis of redundancy factors applicable to the entire corpus with raw syllabic duration. See section 6.4.2 page 127 for definitions of factors.

DUR2: Normalised Syllabic Duration: All			
	Regression Results	$r = 0.2976$	$r^2 = 0.0886$
Redundancy Factor	Independent Contrib. to r^2	F(1,169464)	p value
wf	02.77%	5138.14	0.001
trigram.	02.95%	5470.34	0.001

Table 6.15: Regression analysis of redundancy factors applicable to the entire corpus with normalised syllabic duration. See section 6.4.2 page 127 for definitions of factors.

DUR1: Raw Syllabic Duration: M			
	Regression Results	$r = 0.5594$	$r^2 = 0.3130$
Redundancy Factor	Independent Contrib. to r^2	F(1,30507)	p value
wf	03.78%	1675.82	0.001
trigram	12.48%	5540.36	0.001
men.	00.50%	221.30	0.001

Table 6.16: Regression analysis of redundancy factors applicable to the mention coded part of the corpus with raw syllabic duration. See section 6.4.2 page 127 for definitions of factors.

DUR2: Normalised Syllabic Duration: M			
	Regression Results	$r = 0.5203$	$r^2 = 0.2707$
Redundancy Factor	Independent Contrib. to r^2	F(1,30507)	p value
wf	06.65%	2781.59	0.001
trigram	06.37%	2665.76	0.001
men.	00.93%	386.51	0.001

Table 6.17: Regression analysis of redundancy factors applicable to the mention coded part of the corpus with normalised syllabic duration. See section 6.4.2 page 127 for definitions of factors.

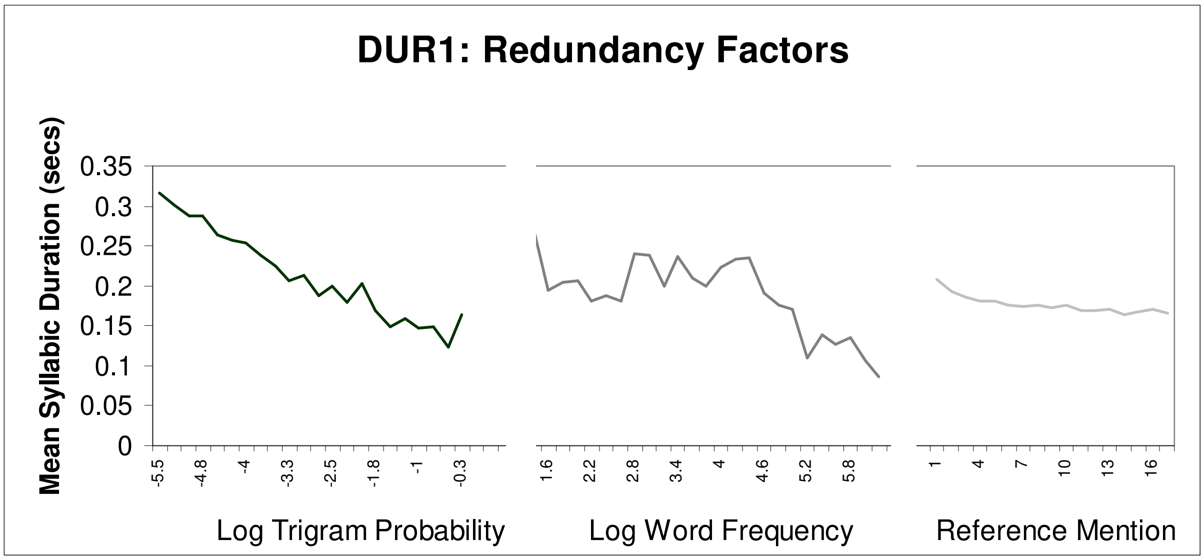


Figure 6.11: The relationship between redundancy factors and DUR1. Redundancy increases left to right. Trigram and word frequency factors are calculated over the entire corpus whereas mention is calculated over only mention coded materials.

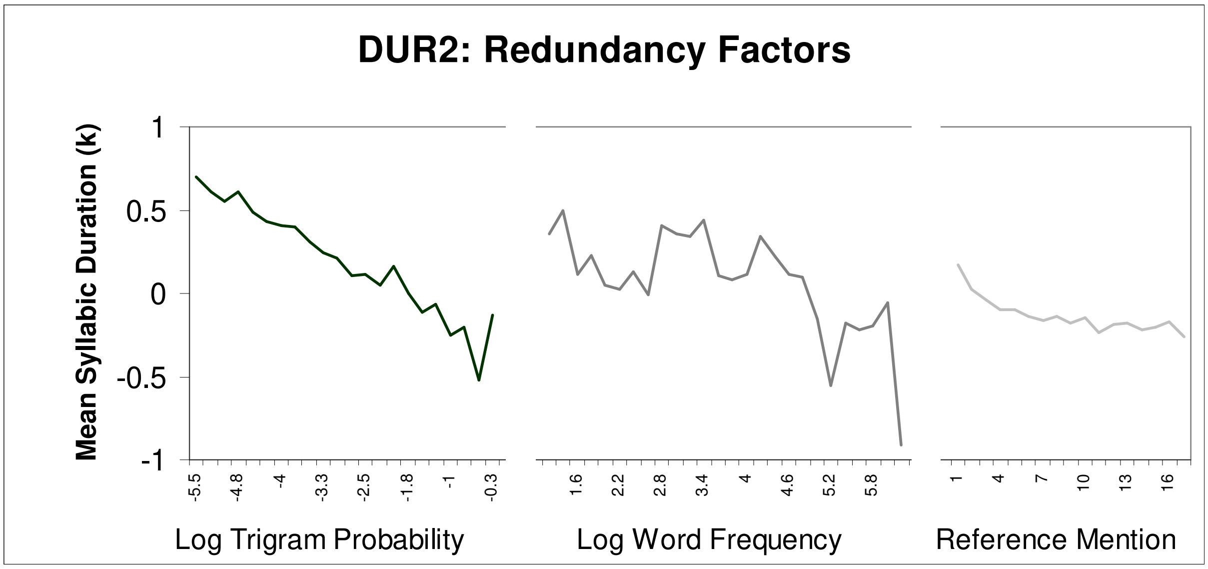


Figure 6.12: The relationship between redundancy factors and DUR2. Redundancy increases left to right. Trigram and word frequency factors are calculated over the entire corpus whereas mention is calculated over only mention coded materials.

DUR1: Raw Syllabic Duration: All: Weak			
	Regression Results	$r = 0.6081$	$r^2 = 0.3698$
Redundancy Factor	Independent Contrib. to r^2	F(1,89531)	p value
wf	10.11%	14361.29	0.001
trigram.	01.93%	2736.84	0.001

Table 6.18: Regression analysis of redundancy factors applicable to entire corpus with raw syllabic duration (weak model). See section 6.4.2 page 127 for definitions of factors.

DUR2: Normalised Syllabic Duration: All: Weak			
	Regression Results	$r = 0.4250$	$r^2 = 0.1806$
Redundancy Factor	Independent Contrib. to r^2	F(1,89531)	p value
wf	04.44%	4850.10	0.001
trigram.	01.21%	1322.12	0.001

Table 6.19: Regression analysis of redundancy factors applicable to the entire corpus with normalised syllabic duration (weak model). See section 6.4.2 page 127 for definitions of factors.

DUR1: Raw Syllabic Duration: M: Weak			
	Regression Results	$r = 0.8085$	$r^2 = 0.6536$
Redundancy Factor	Independent Contrib. to r^2	F(1,12294)	p value
wf	06.06%	2150.52	0.001
trigram	00.74%	263.66	0.001
men.	00.33%	116.28	0.001

Table 6.20: Regression analysis of redundancy factors applicable to the mention coded part of the corpus with raw syllabic duration (weak model). See section 6.4.2 page 127 for definitions of factors.

DUR2: Normalised Syllabic Duration: M: Weak			
	Regression Results	$r = 0.6603$	$r^2 = 0.4360$
Redundancy Factor	Independent Contrib. to r^2	F(1,12294)	p value
wf	05.45%	1187.79	0.001
trigram	00.11%	24.75	0.001
men.	00.90%	196.06	0.001

Table 6.21: Regression analysis of redundancy factors applicable to the mention coded part of the corpus with normalised syllabic duration (weak model). See section 6.4.2 page 127 for definitions of factors.

6.4.5.1.3 Word frequency effects and trigram probability effects on DUR1/DUR2 with no boundaries (weak model):

In order to explore the weak smoothing redundancy hypothesis we consider all four analyses described above but this time for only syllables in monosyllabic words without a following intonational phrase boundary. If it is true that a checking signal is confounding our redundancy results we would expect redundancy to predict much more of the variation in these materials (tables 6.18, 6.19, 6.20, 6.21).

The main effect of the weak model is to increase the predictive power of the redundancy factors substantially and to reduce the independent contribution made by the trigram probability measurements. These differences are summarised in table 6.22.

Summary of Weak/Strong model differences for DUR1/DUR2				
	Strong		Weak	
	trigram	r^2	trigram	r^2
DUR1 entire corpus	06.28%	0.1452	01.93%	0.3698
DUR2 entire corpus	02.95%	0.0886	01.21%	0.1806
DUR1 with mention	12.48%	0.3130	00.74%	0.6536
DUR2 with mention	06.37%	0.2707	00.11%	0.4360

Table 6.22: Differences between the independent contribution of the trigram factor and the overall r^2 for strong (all materials) and weak (syllables in monosyllabic words with no subsequent IP boundary) materials.

Overall the redundancy factors perform in a way predicted from the literature reviewed in chapter 4 in that redundant equals shorter. The importance of boundaries confounding this result with regards to duration measurements is also supported. When considering the independent contributions of redundancy factors outwith prosodic factors these weak models must be also taken into account.

6.4.5.2 COVA1/COVA2

Our experience with COVA1/2 in previous sections suggests we are unlikely to achieve robust correlations with redundancy factors. However as with prosodic factors we do find the similar low r but highly significant effects throughout these regressions (see tables 6.23 and 6.24 and for the +mention model tables 6.25 and 6.26). The trigram probability factor is insignificant for COVA1 when viewed over the whole corpus although when used in the model together with mention this changes. In contrast with COVA2, it is the mention effect that is insignificant. If we look at the effect these factors have on the magnitude of COVA1 and COVA2 (figures 6.13 and 6.14) we can see quite clearly what the effect of noise is on these metrics. Although the significant factors do have a perceivable down drift the random variation is much more intense than in the DUR1/DUR2 examples (figures 6.11 and 6.12).

Because of the low r values of the COVA regressions it is unwise to compare r^2 values and different contributions across materials with very different populations. Unfortunately this means that the comparison between weak and strong models carried out for DUR1/DUR2 cannot be meaningfully made for COVA1/COVA2.

Thus although it looks as if care of vowel articulation is not as strongly affected as duration metrics by prosodic boundaries it is impossible to establish this without improving the COVA1/2 measurements significantly by addressing some of the sources of noise mentioned in chapter 5.

6.4.5.3 Summary

Overall DUR1 and DUR2 behave as expected for both prosodic and redundancy factors. COVA1 and COVA2 although succeeding in acting as an interesting contrast to DUR1/DUR2 appear too noisy to act as reliable variables on their own. A number of possible explanations are put forward for the failure of these measurements to act as a robust control for DUR1/DUR2 in section 6.4.4.2. COVA1/2 have highlighted a potential problem with duration measurements reflecting care of articulation at strong prosodic boundaries (section 6.4.4.2). It is possible that the lengthening we see caused by boundaries may be independent of careful articulation and that these syllables may be longer but not more acoustically distinct. However the noise in COVA1/2 and low r values mean that it is not possible to use these measurements to make strong comparisons. For this reason no comparisons for weak and strong models were carried out for COVA1 and COVA2.

COVA1: Vowel Centralisation: All			
	Regression Results	$r = 0.0467$	$r^2 = 0.0022$
Redundancy Factor	Independent Contrib. to r^2	F(1,71747)	p value
wf	00.17%	122.06	0.001
trigram.	00.00%	1.62	NS

Table 6.23: Regression analysis of redundancy factors applicable to the entire corpus with vowel centralisation. See section 6.4.2 page 127 for definitions of factors.

COVA2: Vowel Targets: All			
	Regression Results	$r = 0.0684$	$r^2 = 0.0047$
Redundancy Factor	Independent Contrib. to r^2	F(1,71747)	p value
wf	00.24%	172.20	0.001
trigram.	00.07%	48.40	0.001

Table 6.24: Regression analysis of redundancy factors applicable to the entire corpus with vowel targets. See section 6.4.2 page 127 for definitions of factors.

COVA1: Vowel Centralisation: M			
	Regression Results	$r = 0.0927$	$r^2 = 0.0086$
Redundancy Factor	Independent Contrib. to r^2	F(1,13365)	p value
wf	00.14%	19.36	0.001
trigram	00.20%	27.19	0.001
men.	00.22%	29.26	0.001

Table 6.25: Regression analysis of redundancy factors applicable to the mention coded part of the corpus with vowel centralisation. See section 6.4.2 page 127 for definitions of factors.

COVA2: Vowel Targets: M			
	Regression Results	$r = 0.1172$	$r^2 = 0.0137$
Redundancy Factor	Independent Contrib. to r^2	F(1,13365)	p value
wf	00.12%	16.71	0.001
trigram	00.71%	96.06	0.001
men.	00.00%	0.29	NS

Table 6.26: Regression analysis of redundancy factors applicable to the mention coded part of the corpus with vowel targets. See section 6.4.2 page 127 for definitions of factors.

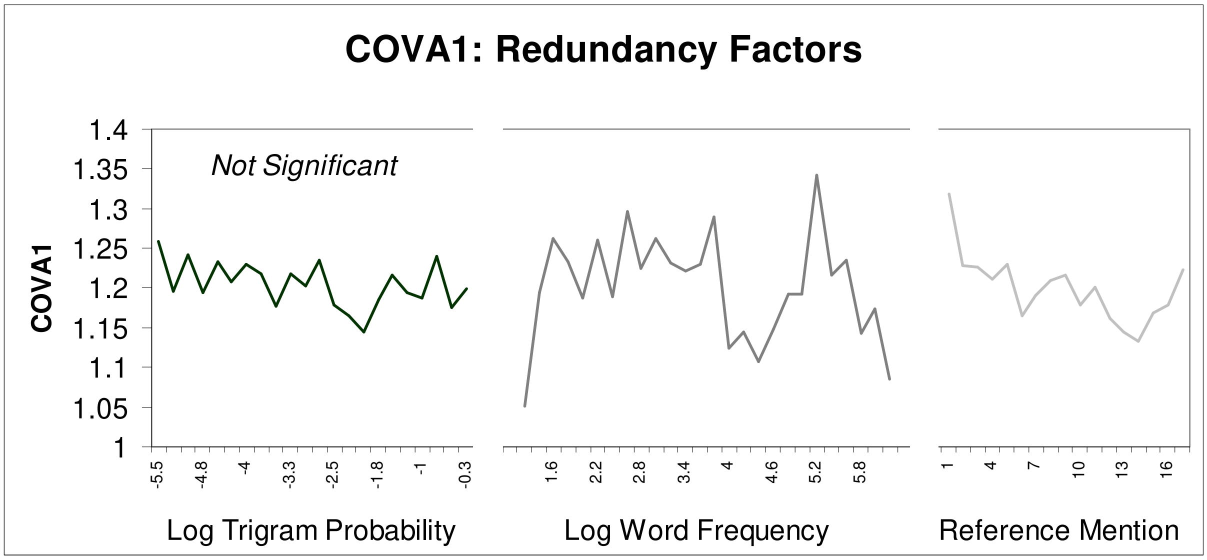


Figure 6.13: The relationship between redundancy factors and COVA1. Redundancy increases left to right. Trigram and word frequency factors are calculated over the entire corpus whereas mention is calculated over only mention coded materials.

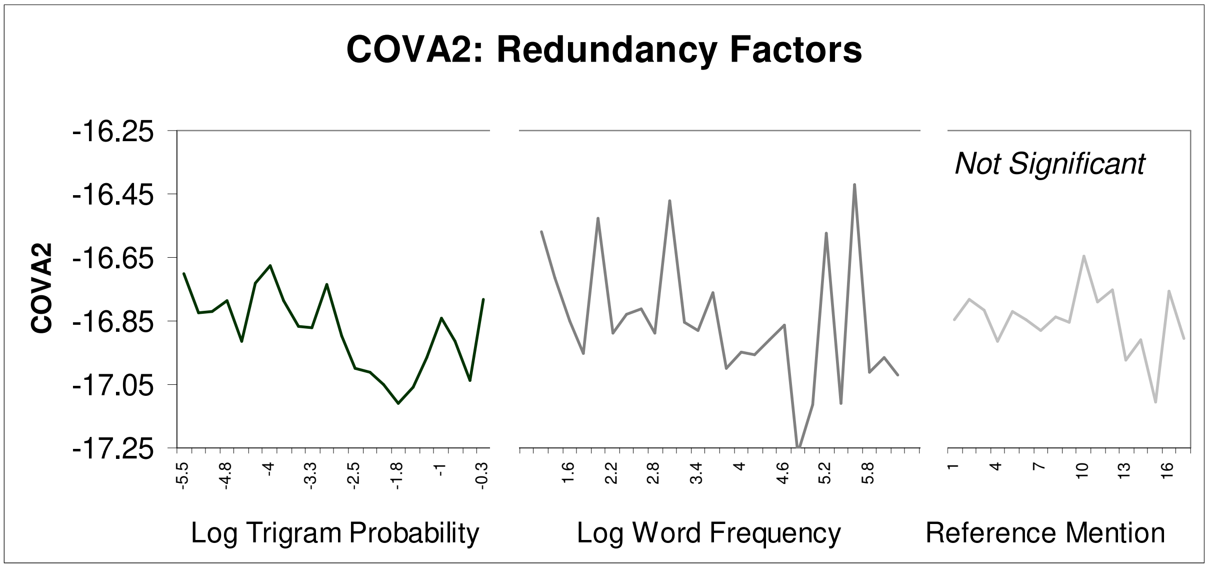


Figure 6.14: The relationship between redundancy factors and COVA2. Redundancy increases left to right. Trigram and word frequency factors are calculated over the entire corpus whereas mention is calculated over only mention coded materials.

6.4.6 The Independent Contribution of Redundancy to Care of Articulation Change.

With these results in mind it is now possible to return to our two main hypotheses once more.

The Smooth Redundancy Hypothesis: Strong Version

Prosodic structure smoothes signal redundancy by controlling care of articulation

and

The Smooth Redundancy Hypothesis: Weak Version

Prosodic structure smoothes signal redundancy by controlling care of articulation except when it acts as a checking signal

To test these hypotheses we need to compare the independent and shared predictive power of the redundancy model with the prosodic model with regards to the care of articulation measurements. The more factors outside of language redundancy that prosody is representing the greater the independent contribution of the prosodic model. This can be related to the dashed boxes in figure 6.15 (This figure was originally shown in chapter 2. Also see this chapter for more detail on the hypotheses mentioned here).

Specifically, if the strong hypothesis is correct we would expect to find the following relationships:

Redundancy is inversely related to COA: That language redundancy factors influence care of articulation and that the more predictable a syllable the less carefully articulated it is. This establishes that changes in care of articulation do indeed smooth signal redundancy as argued in chapter 2.

Prosody relates to COA: That prosodic factors do influence care of articulation. This establishes that prosodic structure *can* control care of articulation.

Redundancy is implicitly expressed by prosody: That language redundancy offers only a small independent contribution to a joint prosody/redundancy model. This establishes that the variation that smoothes signal redundancy is implicitly and only expressed in terms of prosodic structure.

Prosody only smoothes redundancy: That prosodic structure offers only a

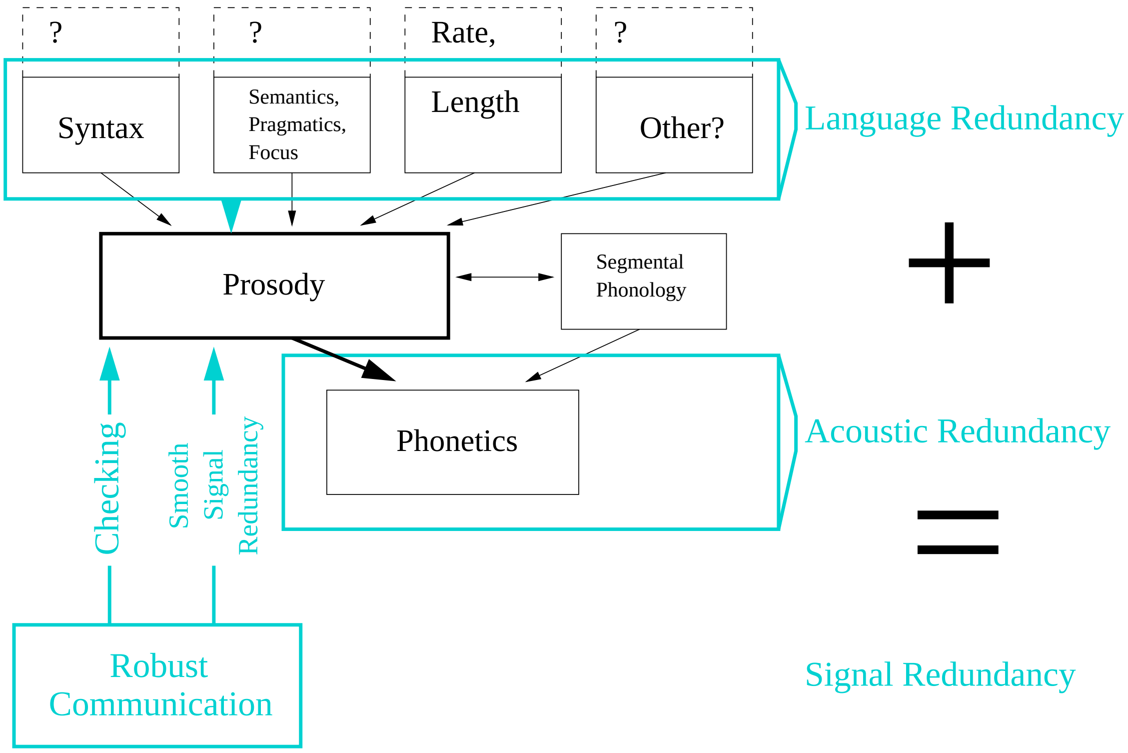


Figure 6.15: How the weak smoothing signal redundancy model could be amalgamated with more traditional views of prosody. (based on the figure Shattuck-Hufnagel and Turk, 1996, p237)

small independent contribution to a joint prosody/redundancy model. This establishes that there are not conditions where prosodic structure alters care of articulation in a way that *does not* smooth signal redundancy. This is a fairly strict restriction. It is possible that an independent contribution from prosody may just reflect inadequacies in the current language redundancy model. However a large independent contribution certainly suggests that prosody is acting strongly outside the predictions of language redundancy and on that basis the strong hypothesis could not be accepted.

For the weak hypothesis we would also expect to find the above but:

Prosody only smoothes redundancy when not checking: Smoothing only occurs for materials which had boundary conditions controlled.

Prosody does not only smooth redundancy at checking locations: That for materials that do not have boundary conditions controlled that prosodic structure did show a substantial independent contribution to a joint model above that modelled by redundancy.

'Redundancy is inversely related to COA' and **'prosody relates to COA'** are strongly supported for DUR1 and DUR2 in section 6.4.4.1 and section 6.4.5.1. They are also more weakly supported for COVA1 and COVA2 in section in section 6.4.4.2 and section 6.4.5.2.

A problem testing the other relationships is deciding exactly what a "small" as opposed to a "substantial" independent contribution is in numerical terms. In table 6.27 we see the independent contributions of the prosodic and redundancy models to the overall r^2 and also the shared contribution.

DUR1/2:Independent and Shared Contribution								
	Strong				Weak			
Materials	r^2	Pros.	Red.	Shared	r^2	Pros.	Red.	Shared
DUR1: P	63.11%	39.52%	3.67%	19.92%	53.17%	12.10%	7.97%	33.10%
DUR1: P∪M	68.35%	38.52%	2.29%	27.54%	53.29%	4.73%	11.27%	37.29%
DUR2: P	53.80%	38.27%	1.42%	14.11%	33.25%	21.13%	1.80%	10.32%
DUR2: P∪M	61.91%	37.18%	1.67%	23.06%	31.44%	16.69%	$2.31\%^{ns}$	12.44%
DUR1: All	49.01%	34.49%	7.11%	7.41%	41.44%	4.46%	9.70%	27.28%
DUR1: M	61.06%	29.76%	9.64%	21.66%	67.83%	2.47%	3.62%	61.74%
DUR2: All	40.22%	31.36%	3.29%	5.57%	29.68%	11.62%	2.84%	15.22%
DUR2: M	51.90%	24.83%	9.12%	17.95%	55.01%	11.41%	1.16%	42.44%
Average	56.17%	34.24%	4.78%	17.15%	45.64%	10.58%	5.08%	29.98%
% of Explained Variance		60.96%	8.5%	30.53%		23.18%	11.14%	66.69%

Table 6.27: Independent contributions of redundancy - *Red.* and prosodic models - *Pros.* in predicting variance of DUR1 and DUR2 over all materials. The non-independent, shared contribution is shown under *Shared.* P: Materials hand coded for prosody. M: Materials with mention coding. All: Materials with automatic prosodic coding and a trigram/word frequency redundancy model. P∪M: Materials both hand coded for prosody and for mention. All results are significant except for the redundancy model's contribution to hand coded and mentioned coded material with respect to DUR2.

Values are not shown for COVA1/2 because r values are too small for such a comparison to be meaningful. Thus we have failed to produce sufficient evidence to support either hypothesis for COVA1 and COVA2 and for these measures of care of articulation they must be rejected.

For DUR1 and DUR2 we can make meaningful comparisons. Firstly taking **'redundancy is implicitly expressed by prosody'** we can see that in general the redundancy model made an independent contribution of just under 4% to predicting these variables. This represents around 10% of the total variance predicted in both a strong and weak context. This is arguably a small independent contribution and thus the relationship, **'redundancy is implicitly expressed**

by prosody', can be accepted.

In contrast over all materials (strong) the prosodic model independently accounts for about 60% of the predictive power of the overall model. This is a large contribution and over all materials (without attempting to control for a checking effect) prosodic structure does affect care of articulation in a way *not* predicted by the redundancy model. Thus the relationship **Prosody only smoothes redundancy** and thus the strong hypothesis must be rejected.

However if we compare the independent contribution of the prosodic model between strong (all materials) and weak (materials without boundaries) we see this contribution is reduced from 60% to 23% while the shared proportion rises from 30% to 66%. That is 66% of of predictive power of the prosodic model is directly related to smoothing redundancy when no checking signal is likely to occur. The remaining 23% must either be accounted for by the unknown factors in the dotted boxes shown in figure 6.15 or to inadequacies of the redundancy model. Overall, given the simplicity of the redundancy model this result supports the weak hypothesis (and both the relationship, **'prosody only smoothes redundancy when not checking'**, and, **'prosody does not only smooth redundancy at checking locations'**).

Looking more closely at table 6.27 we can see that for some materials this conclusion is stronger than for others, in particular for DUR1 (see also figure 6.16). This is because word frequency is a strong predictor of overall word length and thus a good predictor of the number of segments in a monosyllabic word. For DUR2, where this information is normalised out of the metric, word frequency is not such a good predictor making the prosodic model relatively stronger.

Secondly we see that the hand coded prosodic materials show a generally stronger independent contribution from prosody (see also figure 6.16). The main difference between the hand coded and the automatic coded models in the weak materials is that the automatic factors are exclusively factors within the lexicon (vowel type, lexical stress, open/close class) whereas in the hand coded materials phrasal accents are not guessed on the basis of these lexical features but were assigned by inspection. Thus the hand coded model is stronger, again raising the general independent contribution of these factors.

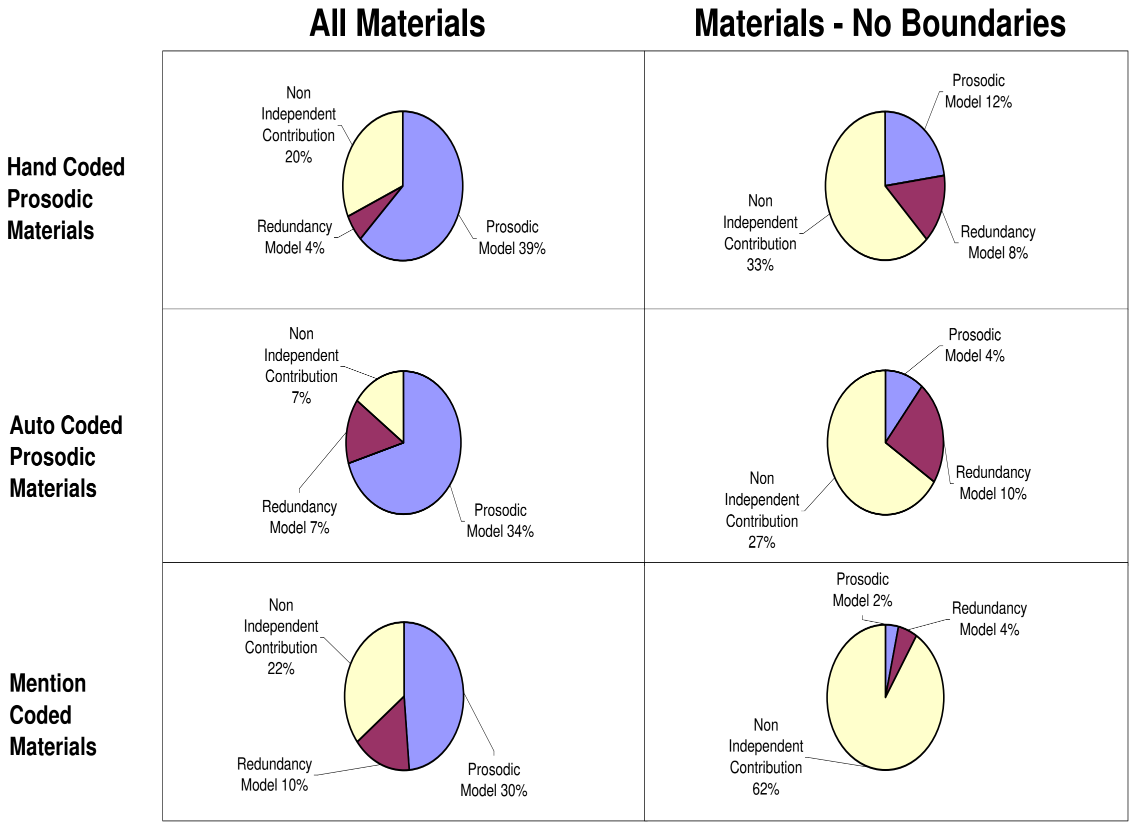


Figure 6.16: Pie charts showing the shared and independent contribution of prosodic and redundancy models in predicting raw syllabic duration (DUR1). Weak materials (no boundaries) show a much higher shared contribution.

6.5 Summary of Results

The results obtained from this work can be summarised as follows:

- Both prosodic factors and redundancy factors have a significant effect on care of articulation in a large corpus of spontaneous running speech.

- In terms of duration change prosodic factors predicts up to 59% of raw syllabic duration in hand coded materials. Results for automatic coding based on lexical information and pauses predicted 42% of the variation.

- Also in terms of duration change, redundancy factors, looking at a subset of landmark referents with controlled prosodic boundaries, predicted 65% of raw syllabic duration change. Results for other materials varied. For all materials, trigram and word frequency factors predicted 14% of the variation. The more predictable a syllable in terms of low level factors such as word frequency and syllabic trigram and a higher level factor, givenness/mention, the less carefully articulated a syllable is.

- Results for care of articulation measured in terms of the spectral quality of vowels were disappointing. It was found that the metrics were very noisy and although significant relationships were found they predicted very small amounts of variance.

- However the results from these metrics suggested that such articulatory care may be independent of lengthening due to prosodic boundaries.

- Comparing the independent contribution of redundancy factors and prosodic factors to predicting duration it was found that (see figure 6.16):

  1. Most of the contribution made by redundancy factors is implicitly represented by prosodic factors. However a significant but small percentage (2-4%) predicted even by these very simple redundancy metrics was not represented by prosodic factors.
  2. Prosodic factors, especially hand coded factors, made a large independent contribution to predicting duration change above that representing redundancy (about 35% compared to a shared prosodic/redundancy contribution of 17% over all sets of materials).
  3. This independent contribution was much smaller for syllables where prosodic boundaries were controlled for (11% over all sets of materials). This suggests a major role of prosodic structure, outwith boundaries, is to smooth signal redundancy by controlling care of articulation in a way which implicitly mirrors language redundancy factors. This led to tentatively acceptance of the weak smoothing redundancy hypothesis.

In the next chapter I will discuss the implications of these results more fully, consider future directions this work may take and speculate on the role of redundancy in spoken language.

Chapter 7

Discussion

7.1 Introduction

In the previous chapters we have explored the relationship between prosodic structure, redundancy and care of articulation. Results have strongly supported previous findings on the effect of these factors on duration. Prominent syllables and syllables next to prosodic boundaries are lengthened. Redundant, easy to predict syllables, are reduced. Results for vowel quality, although hampered by noise, show similar effects, with vowels becoming more centralised and less spectrally distinct when redundant and less centralised and more spectrally distinct when prominent.

When comparing the independent contribution of prosody and redundancy to predicting these changes it was found that, in general, prosodic structure implicitly shared a majority of the predictive power of redundancy.

In this chapter I will discuss the importance of these findings and what further work should be addressed. I will deal with the following issues:

1. Measuring Care of Articulation: The advantages and disadvantages of automatic techniques. How do we improve a metric of care of articulation?

2. Smoothing Redundancy versus Checking: In chapter 2 I considered why redundancy might affect care of articulation. How do my results support the smooth signal redundancy hypothesis?

3. Stochastic Suprasegmentals: Prosodic structure implicitly represents much redundancy information. How, or should, this be incorporated into prosodic theory?

7.2 Measuring Care of Articulation

In chapter 6 I wrote that measuring care of articulation for almost every syllable in 15 hours of speech was a significant research task in itself. In this section I would like to consider how successfully this task has been addressed.

The strong, significant results for all care of articulation variables supporting previous work in prosodic theory and redundancy are encouraging. The differences between the DUR and COVA metrics with regards to prosodic structure are also interesting. However the COVA metrics, were in the end, too noisy to carry out the complete analysis. With such poor r values it was not possible to compare the independent contributions of the redundancy and prosodic models.

7.2.1 COVA1/COVA2

As explained in chapter 6 some noise in COVA1/2 is attributable to the formant tracker used. Some of this noise was caused by the failure to track F1 for women with high f0. However even where this was not the case the automatic techniques did not do as well as the human coders. If the analysis was carried out for male speakers only, although this did improve results, the r values for COVA1/2 and the prosodic and redundancy models were still very low (for the combined model $r = 0.1058$). Different formant trackers are available and a more effective tracker might well exist for female and male spontaneous speech. In the same way better autosegmentation techniques would have also improved results slightly. For a 100ms vowel, autosegmentation error, on average, caused 20% of the data to be outside the vowel and lost 12% of the data within the vowel. Despite voicing being used in addition to autosegmentation to filter out non-vocalic data, for shorter vowels this error is sufficient to undermine F1/F2 assessments. Considering how clearly visible many boundaries in the speech signal are, current frame based autosegmentation techniques seem to do a poor job. There is room for improvement here.

Another source of noise was caused by considering the F1/F2 space in isolation. By ignoring other acoustic factors in the vowels in order to simplify the model I also discarded information that can be used to identify vowels and stress such as spectral tilt, amplitude and f0 transitions. Although the simplified model was very advantageous when used to test and view output from the system I believe more information should be retained. Local amplitude variation, especially, seems an obvious candidate to consider in the model. To do so would require sophisticated normalisation both in terms of speaker differences and in terms of phonetic contents.

The decision to ignore phonetic context was another source of noise in these measurements. It would be advantageous to build models that took into account such contexts so that a vowel following a consonant that is known to reduce F2 could be treated differently from a vowel following a consonant that is known to increase F2. To a large extent, the restriction on the data available used to build the citation model of each speaker, precluded the use of such context.

This brings me to the final point with regards to noise in the system. Although the citation speech used to build the speaker models was certainly a lot more carefully articulated than most of the spontaneous speech it was not specifically collected for this purpose. In some cases the citation speech was not as carefully articulated as I would have liked and it was also not phonemically balanced.

Overall I do believe that this technique could potentially be improved to give a better, less noisy metric of care of vowel articulation. Such a measurement would be potentially useful in discriminating between duration change caused by prosodic boundaries as opposed to duration change caused by prominence.

Given the problem of considering spectral characteristics of phonemic segments when they have been so heavily reduced (over 60% of vowels in my corpus were less than 40ms) it is unclear how such a metric could effectively be combined with duration to produce a single care of articulation measurement. However if a less noisy COVA2 can be produced this may at least become a practicality. The main reasons for rejecting short and voiceless tokens were:

1. The formant tracker could not produce valid results for voiceless speech.
2. The curve fitting algorithm used to estimate vowel targets could not meaningfully be applied to less than four points (in this case 40ms) of speech.

However it is possible to assess formants within voiceless speech with techniques other than LPC analysis (Wrench, 1995). It would also be possible to modify the curve fitting algorithm to accept a simple average for very short stretches of speech.

This was not carried out in this work for two reasons:

1. Noise: A clear problem with the COVA1/2 measurements was noise. The severity of the noise problem increases the shorter the vowel, firstly, because

of autosegmentation error, and secondly, because a single spurious value has more effect on the overall result. One of the functions of the curve fitting algorithm was to reduce the effect of spurious values. Given the problem with noise for longer vowels it is unlikely that taking an average for shorter vowels will produce usable results. Autosegmentation error alone would mean that on average, of the data points in a vowel under 40ms, 2 out of the 3 data points would be outside the vowel. It is possible to address this problem by reducing the frame size and by improving the formant tracker and autosegmentation. However to do so would require significant re-engineering of the automatic processes used in this work.

2. Phonemic Context: The second major function of the curve fitting algorithm was to assess the achieved target of the vowel and to take into account some of the different coarticulatory effects caused by phonemic context. In vowels less than 40ms long a simple average, even if all data points are representative of the vowel, will give undue influence to the coarticulated beginning and end of the vowel and give a false impression of how carelessly a speaker may have tried to achieve the vowel target.

Although these are significant problems they could (and should) be addressed in future work (see above for a discussion on dealing with both noise and phonemic context). If these issues are addressed a simple average could then be used to produce COVA1/2 values for short and voiceless vowels. This would address the question of whether care of vowel articulation does reduce further in these short tokens or whether it is only duration that can be further attenuated in these contexts.

Finally I hope that some of the results reported in this work will encourage more controlled laboratory work on the way care of articulation is expressed in terms of acoustic factors. Without these careful studies it is difficult to proceed with a modelling approach in a considered fashion because, as discussed in chapter 2, statistical modelling benefits from being theoretically led as well as being observationally driven.

7.3 Smoothing Signal Redundancy versus Checking

In chapter 2 I argued that care of articulation was related to language redundancy because the result was to smooth signal redundancy. Smooth signal redundancy is good if a message is likely to be degraded by a noisy environment. The results reported in chapter 6 support this view. Predictable, redundant syllables are shorter and their vowels have less defined spectral characteristics. This relationship is strongest for syllables where prosodic boundaries have been controlled. The effect of reduction in duration and in spectral characteristics is to make syllables harder to guess based solely on their acoustic properties. Their redundancy with regards to an acoustic model is reduced. The signal redundancy, the combination of these two models, is thus smoother (see figure 2.1 in chapter 2).

However as pointed out in chapter 2 the fact that final phrase lengthening occurs confounds a simple smoothing signal redundancy hypothesis. The ends of phrases are more predictable than the beginning of phrases yet we see an increase in syllabic duration of around 20%. Over all materials, the fact that a full intonational phrase boundary followed a syllable predicted about half of all variation in syllabic duration.

This led to the weak hypothesis, that when a checking signal caused by prosodic boundaries was controlled then the smooth signal redundancy hypothesis would be fulfilled. This does indeed appear to be the case with redundancy predicting much more duration variation outside boundary contexts. The independent contribution of prosody in these conditions, and thus its potential for reflecting factors beyond language redundancy, is much lower. We conclude that the smooth redundancy hypothesis in these conditions must be tentatively accepted.

To accept the weak smooth redundancy hypothesis with more confidence we need to explain what the independent prosodic contribution in these conditions is representing. Can such an independent contribution be ascribed to phonological constraints for example? Would it disappear if the redundancy model was more sophisticated, for example taking into account more semantic and syntactic information? Is it because of inaccuracies in the method used to control for prosodic boundaries?

The key to unravelling this problem and the major problem of deciding whether duration change is being caused by boundaries or prominence is to build and test a sophisticated checking model. One way of approaching this is stochastically.

Work looking at the statistical patterns in language suggests that prosodic boundaries can be identified by the high trigram probability of the section of speech before the boundary as opposed to the low trigram probability of the section of speech after the boundary. However care is required. Final phrase lengthening appears to occur primarily in the rhyme of the syllable. This suggests the larger syllabic domain used in my work is not ideal for addressing this problem. There is also a question of differences between monologues, where a listener is not supposed to interrupt, and more collaborative dialogues where a great deal of back and forth is expected. In the HCRC Map Corpus much is known concerning the interaction between speakers in terms of intervals between speakers and discourse structure. None of this knowledge is considered in the work presented here but could form the basis of a more sophisticated model of checking.

Finally the rather frustrating result from the COVA variables suggests that care of articulation is a combination of a number of factors and that changes caused by a checking signal might be different from changes caused by smoothing signal redundancy. As reported in chapter 6 section 6.4.4.5 I found that boundaries did not appear to affect COVA2 in the same way as prominence. Vowel quality did not seem to increase when a boundary was present, only when prominence was expected. The noisy nature of the COVA variables leaves this result as inconclusive. Before trying to build a checking model these metrics need to be re-examined and improved. The results linking COVA with redundancy in a broad sense are encouraging as is the relationship between them and human subjects' perception of what makes a good vowel. If the problem of noise can be addressed this approach may lead to more conclusive results.

7.4 Stochastic Suprasegmentals

The extent redundancy factors predicted care of articulation change, in terms of duration, varied across the different materials we examined in chapter 6. Redundancy factors were most successful at predicting raw syllabic duration in syllables occurring in references to landmarks when they were controlled for prosodic boundaries. In this case the redundancy model predicted 65% of the variation ($r = 0.8085$). Let's put this in perspective. For over 12,000 syllables without knowing anything about their phonetic contents, redundancy factors predicted over half the raw duration. Even taking into account the restricted contexts of these syllables (monosyllabic words in references to landmarks with a low probability of being followed by a major prosodic boundary) this predictive power seems high. Especially when you consider it is based on three very simple redundancy measurements, word frequency, syllabic trigram probability and how many times the landmark is mentioned.

For the same materials prosodic structure also accounted for about the same amount of variance (64%, $r = 0.8013$). Over these automatically coded materials, with prosodic boundaries factored out, all of the prosodic information coded is lexical in nature, as in whether the syllable has a full or reduced vowel, lexical stress and whether the word is open or closed class.

When we consider a joint model of redundancy factors and these lexical prosodic factors we find that an enormous 62% of the predictive power is shared.

Although not as extreme, results over all the other materials supported the extent prosodic factors embodied these redundancy factors. When boundaries were considered the shared predictive power fell to about a third of the variance predicted, when boundaries were controlled this rose to two thirds of the variance predicted.

These results are not accidental and I believe go some way to answering the question of not **what** prosody is but **why** prosody is.

If we take a critical look at the more traditional view of prosody embodied in figure 7.1 (previously shown in chapter 2) and compare it with the combined model shown in figure 7.2 (also previously shown in chapter 2) we can make some interesting observations.

Firstly the traditional view does not offer a theoretical framework for why some things affect prosody and others do not. Each area, syntax, semantics, discourse structure are treated independently in this traditional view. The reasons some syntactic factors affect prosody and some do not are not related to the reasons some semantic factors affect prosody and some do not. By looking at language in terms of redundancy we can relate these different factors to each other. Concepts as diverse as focus, syntactic structure, word class, length of utterance and word frequency can be looked at in terms of a predictive model and thus in terms of language redundancy (see figure 7.2). In addition, the reason language redundancy should affect care of articulation and thus be expressed in terms of prosodic structure follows persuasively from the requirements of getting information from A to B within a noisy environment. This does not mean that other factors outside redundancy do not affect prosody, for example psycholinguistic or phonological constraints, however it does shed some light on why we have prominence and with checking, why we have prosodic boundaries.

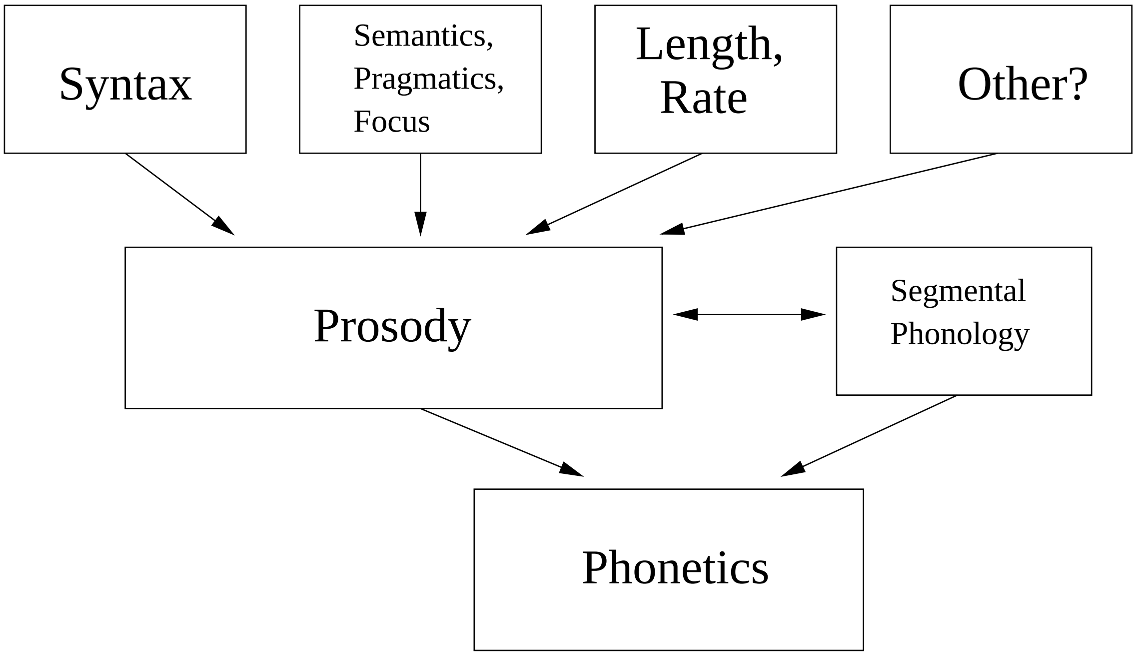


Figure 7.1: One view of the role of the prosodic component of the grammar (taken from Shattuck-Hufnagel and Turk, 1996, page 237).

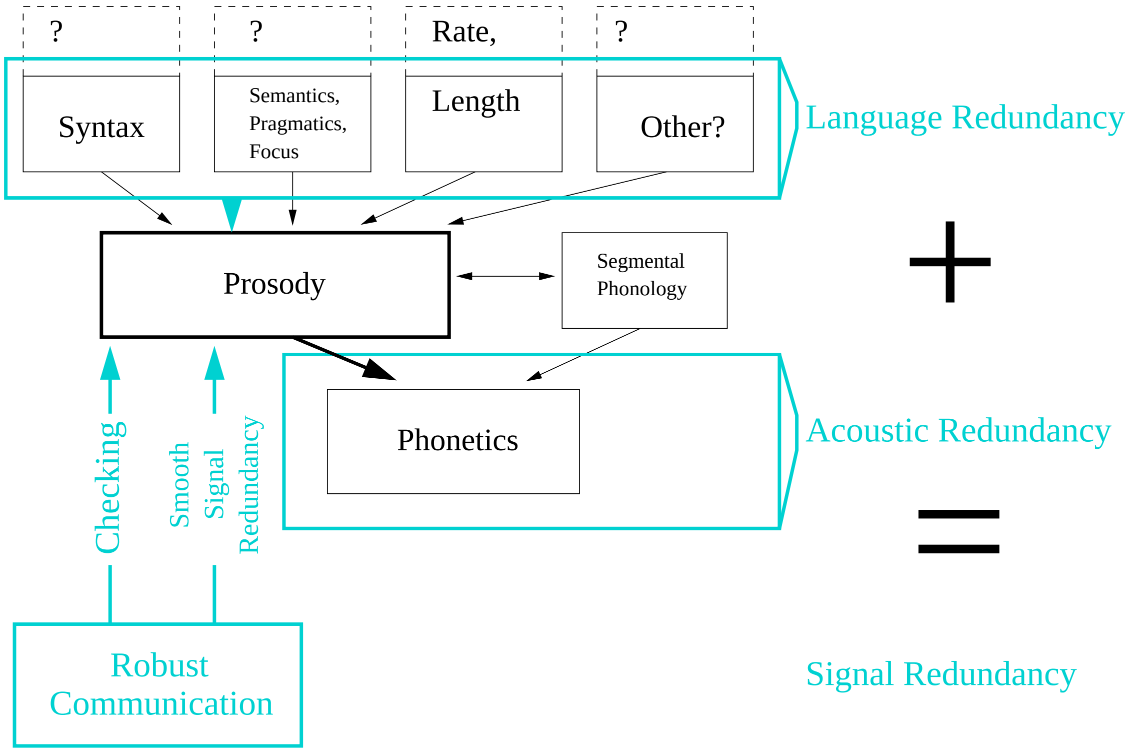


Figure 7.2: How the weak smoothing signal redundancy model could be amalgamated with more traditional views of prosody (based on the figure Shattuck-Hufnagel and Turk, 1996, p237).

This redundancy framework could in principle be applied to global questions concerning prosodic structure. For example the traditional approach of regarding languages as stress timed or syllable timed could be related instead to smoothing signal redundancy and checking.

So called stress timed languages may require (or just have) more smoothing. The natural tendency of word initial syllables to be unpredictable would then cause them to be lengthened and the rest of the word (however long) to be reduced. This would lead to a tendency for isochrony between lexical stresses. In contrast so called syllable timed languages may require (or just have) less smoothing and stronger checking making each syllable a self contained checked piece of information. In this case syllables would tend to be produced more regularly.

In addition languages with looser word ordering could use position as a means of smoothing redundancy rather than prominence. Thus, rather than attenuating repeated mentions, they could be placed further to the front of the phrase where they were less predictable in terms of context. Such differences in the constraints of a language would be reflected in differences in prosodic structure given the smooth signal redundancy hypothesis.

I do not begin to address these questions here. My point is that redundancy and stochastic modelling offers a common framework within which these questions can be meaningfully asked.

I would argue that the results from my work suggest that prosody acts as a interface between the compositional structure of language and the constraints of producing a robust and effective signal. This role of prosody in smoothing signal redundancy and checking is crucial to why prosodic structure is as it is and why it works as it does.

Given this crucial relationship between predictability in language and prosody it seems surprising that the role of redundancy has remained so unspecified in prosodic theory. It emerges occasionally in concepts of 'semantic weight', and in arguments linking prosody with both language perception and acquisition, but formal, stochastic, representations of redundancy have been generally ignored.

In my opinion redundancy information should be incorporated into prosodic theory in the same way that other lexical factors, such as stress and number of syllables within words are incorporated. The need for this is exemplified by the results on givenness found by Bard and Aylett (1999). In this work there is clear evidence that subsequent mentions of the same word are produced less carefully even when the traditional prosodic structure is identical. By including redundancy information into prosodic theory we could go some way to addressing this issue. For example, lexical stress could be modified to take redundancy into account so that syllables in unpredictable words were regarded as having stronger lexical stress than stressed syllables in predictable words. Perhaps these stronger stressed syllables could be regarded as more desirable sites for phrase accent placement than their more common neighbours. The probability of accentedness could then be directly related to acoustic parameters rather than the categorical +/-phrase accent from traditional prosodic phonology. In this way suprasegmentals could be connected to stochastic information and be used to produce the redundancy effects we have observed. In fact, even if we ignore the practical concerns of modelling data effectively using prosodic structure, we still need to include redundancy information more explicitly in prosodic theory because:

1. My results are consistent with the view that the requirement for robust transmission in a noisy environment drives prosody. This relationship should be formalised.

2. If prosodic theory is the means with which redundancy smoothing and checking are implemented then the small but consistently significant contribution from the redundancy model should be modelled by prosodic theory. It doesn't make sense to have such a small contribution represented as an independent factor when so much of the predictive power is shared.

3. Such stochastic information can be used as a useful interface between phonetic variation such as in duration, amplitude and pitch and a categorical phonological view. As with modern speech recognition such a statistical model allows the calculation of the most likely string of prosodic phonology given such phonetic observations without necessitating any particular sequence. This could potentially be used to produce more natural sounding synthesised speech.

Another major advantage of including stochastic information directly in this way is that, as suggested above, it offers a potential framework for comparing prosodic structure across languages. If it is true that a major role of prosody is to smooth signal redundancy and act as a checking signal then this should be the case cross linguistically. The same requirements for a robust an effective signal exist whatever language you are speaking in. Thus by including redundancy in our prosodic theory we could potentially go some way to making prosodic theory less language dependent.

A great deal of work would be required to pursue this approach. It is possible that if other languages are considered we may find that such a smoothing redundancy/checking role is not cross linguistic at all. This remains an open research question. We may also find that with more effective metrics of care of articulation and more complex models of redundancy that the view presented here must be modified. Perhaps, more crucially, the question of exactly how a checking signal is produced by prosody and how it interacts with redundancy will undermine any simple redundancy articulation relationship. Perhaps we will find that the independent contribution made by prosodic structure in my results is actually reflecting the music and rhythm of language and some sort of stately dance between speakers cooperating in a dialogue, rather than the mundane problem of making sure that bits of information have been effectively received.

However 62% is 62%. How can so much predictive power be shared by factors as different as lexical stress and word frequency and not demand being addressed by a combined theory?

7.5 Conclusion

I hope that this work has shown how important redundancy and ideas of redundancy are in the study of spoken language. As with much research the questions answered pose further more challenging questions. Current access to large corpora of digitised speech have made the work here possible and opened up an approach that could be described as corpus phonetics. Speech technology does not just benefit from the findings in phonetic research but phonetic research can itself benefit from applying such technology in order to explore relationships within large multi-speaker corpora. In conclusion I reiterate what I regard as the main findings in this work:

- Redundancy in language has a strong association with care of articulation. This association is implicitly represented by much formal prosodic theory.
- This is because spoken language needs to have a smooth signal redundancy and prosodic structure offers a linguistic means for effecting this.

References

Anderson, A. H., M. Bader, E. G. Bard, E. Boyle, G. M. Doherty-Sneddon, S. Garrod, S. Isard, J. C. Kowtko, J. M. McAllister, J. E. Miller, C. F. Sotillo, H. S. Thompson, and R. Weinert (1991) The HCRC Map Task Corpus. *Language and Speech*, **34**, 4, 351–366.

Aylett, M. (1996) Using statistics to model the vowel space. In *Proceedings of the Edinburgh Linguistics Department Conference*, pp. 7–17.

Aylett, M. (1998) Building a statistical model of the vowel space for phoneticians. In *SST98 ICSLP98*.

Aylett, M. (1999) Modelling clarity change in spontaneous speech. In R. J. Baddeley, P. J. B. Hancock, and P. Foldiak (eds.), *Information Theory and the Brain*. New York: Cambridge University Press.

Aylett, M. and M. Bull (1998) The automatic marking of prominence in spontaneous speech using duration and part of speech information. In *Proceedings of ICSLP-98.*, pp. 2123–6.

Aylett, M. and A. Turk (1998) Vowel quality in spontaneous speech: What makes a good vowel?. In *ESCA Workshop: Sound Patterns of Spontaneous Speech*.

Baayen, R. H., R. Piepenbrock, and L. Gulikers (1995) *The CELEX Lexical Database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania. Version 2.5.

Balota, D., J. Boland, and L. Shields (1989) Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language*, **28**, 14–36.

Bard, E. G. and M. Aylett (1999) The dissociation of deaccenting, givenness and syntactic role in spontaneous speech. In *ICPhs99*.

Bard, E. G., C. F. Sotillo, A. H. Anderson, G. M. Doherty-Sneddon, and A. Newlands (1995) The control of intelligibility in running speech. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 4, pp. 188–191.

Beckman, M., J. Edwards, and J. Fletcher (1992) Prosodic structure and tempo in a sonority model of articulatory dynamics. In *Papers in Laboratory Phonology II: Segment, Gesture and Tone.* Cambridge: Cambridge University Press.

Beckman, M. E. and G. M. Ayers (1993) *Guideline for ToBI Labelling.* 1st edn.

Beckman, M. E. and J. Edwards (1990) Lengthening and shortenings and the nature of prosodic constituency. In J. Kingston and M. E. Beckman (eds.), *Papers in Laboratory Phonology I.* Cambridge: Cambridge University Press.

Beckman, M. E. and J. Pierrehumbert (1986) Intonational structure in japanese and english. *Phonology Yearbook*, **3**, 255–309.

van Bergem, D. R. (1988) Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, **12**, 1–23.

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press.

Bolinger, D. (1963) Length, vowel, juncture. *Linguistics*, **1**, 5–29.

Bond, Z. and T. Moore (1994) A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication*, **14**, 325–37.

Bradlow, A. R., G. M. Torretta, and D. B. Pisoni (1995) Some sources of variability in speech intelligibility. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 1, pp. 198–201.

Bradlow, A. R., G. M. Torretta, and D. B. Pisoni (1996) Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, **20**, 255–72.

Broad, D. and F. Clermont (1987) A methodology for modelling vowel formant contours in CVC context. *The Journal of the Acoustical Society of America*, **81**, 1572–1582.

Brown, G. and G. Yule (1983) *Discourse Analysis.* Cambridge: Cambridge University Press.

Campbell, N. and M. Beckman (1997) Stress, prominence and spectral tilt. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis (eds.), *Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications.*, pp. 67–70. ESCA and The University of Athens.

Campbell, W. N. (1992) *Multi-level timing in speech.* Ph.D. thesis, Sussex University.

Campbell, W. N. (1993) Multi-level timing in speech. *Advanced Tellecommunications Research Institute Technical Report.*

Campbell, W. N. and S. D. Isard (1991) Segment durations in a syllable frame. *Journal of Phonetics*, **19**, 37–47.

Caspers, J. (1994) *Pitch Movements under Time Pressure: Effects of Speech Rate on the Melodic Marking of Accents and Boundaries in Dutch.* Ph.D. thesis, Leiden.

Chafe, W. (1974) Language and consciousness. *Language*, **50**, 111–133.

Charniak, E. (1993) *Statistical Language Learning.* Cambridge, Mass.: The MIT Press.

Clark, H. and E. Clark (1977) *Psychology and Language.* New York: HBJ.

Clarkson, P. and R. Rosenfeld (1997) Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of Eurospeech 97*, pp. 2707–10.

Clements, G. and S. Keyser (1983) *CV Phonology: A Generative Theory of the Syllable.* Cambridge, Mass.: The MIT Press.

Couper-Kuhlen, E. (1986) *An Introduction to English Prosody.* London: Edward Arnold.

Couper-Kuhlen, E. (1993) *English Speech Rhythm: Form and Function in Everyday Verbal Interaction.* Amsterdam: John Benjamins Publishing Company.

Cruttenden, A. (1986) *Intonation.* Cambridge: Cambridge University Press.

Cutler, A. and S. Butterfield (1990) Durational cues to word bounderies in clear speech. *Speech Communication*, **9**, 485–95.

Cutler, A. and S. Butterfield (1991) Word boundaries in clear speech. *Speech Communication*, **10**, 335–353.

Duda, R. O. and P. E. Hart (1973) *Pattern Classification and Scene Analysis.* New York: Wiley.

Eefting, W. (1991) The effect of "Information Value" and "Accentuation" on the duration of Dutch words, syllables, and segments. *The Journal of the Acoustical Society of America*, **89**, 412–424.

Ferguson, C. (1977) Baby talk as a simplified register. In C. Snow and C. Ferguson (eds.), *Talking to Children.* Cambridge: Cambridge University Press.

Flege, J. (1988) Effects of speaking rate on tongue position and velocity of movement in vowel production. *The Journal of the Acoustical Society of America*, **84**, 901–916.

Fougeron, C. and P. Keating (1997) Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, **101**, 6, 3728–40.

Fourakis, M. (1991) Tempo, stress, and vowel reduction in american english. *The Journal of the Acoustical Society of America*, **90**, 1816–27.

Fowler, C. A. (1988) Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech*, **31**, 307–19.

Fowler, C. A. and J. Housum (1987) Talkers' signalling of "New" and "Old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, **26**, 489–504.

Fowler, C. A., E. T. Levy, and J. M. Brown (1997) Reduction of spoken words in certain discourse contexts. *Journal of Memory and Language*, **37**, 24–40.

Freed, B. (1978) *Foreign talk: A study of Speech Adjustments made by Native Speakers of English when in Conversation with Non-native Speakers.* Ph.D. thesis, University of Pennsylvania.

Fry, D. (1958) Experiments in the perception of stress. *Language and Speech*, **1**, 126–52.

Fry, D. B. (1979) *The Physics of Speech.* Cambridge: Cambridge University Press.

Gimson, A. C. (1977) *EVERYMAN'S English pronouncing dictionary : containing over 59,000 words in international phonetic transcription (originally compiled by Daniel Jones; extensively revised and edited by A. C. Gimson).* Everyman's reference library.

Goldinger, S. and W. Summers (1989) Lexical neighborhoods in speech production: A first report. *Progress Report, Indiana University*, **15**, 331–342.

Halliday, M. (1967) *Intonation and Grammar in British English.* Mouton.

Hanley, T. and M. Steer (1949) Effect of distracting noise upon speaking rate, duration and intensity. *Journal Speech and Langauge Disorders*, **14**, 363–68.

Hartigan, J. A. (1975) *Clustering Algorithms.* New York: Wiley.

Hawkins, S. and P. Warren (1994) Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics*, **22**, 493–511.

Hayes, B. (1989) The prosodic hierarchy in meter. In P. Kiparsky and G. Youmans (eds.), *Phonetics and Phonology, Vol 1: Rhythm and Meter.*, pp. 201–260. San Diego: Academic Press.

Hogg, R. and C. McCully (1987) *Metrical Phonology: A Coursebook.* Cambridge: Cambridge University Press.

Howell, P. and C. Bonnett (1997) Speaking clearly for the hearing impaired: Intelligibility differences between clear and less clear speakers. *European Journal of Disorders of Communication*, **23**, 89–97.

Hunnicut, S. (1985) Intelligibility versus redundancy – conditions of dependency. *Language and Speech*, **28**, 45–56.

de Jong, K. J. (1995) The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, **97**, 491–504.

Klatt, D. (1976) Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, **59**, 1208–20.

Ladd, D. R. (1996) *Intonational Phonology.* Cambridge: Cambridge University Press.

Ladefoged, P. (1962) *Elements of Acoustic Phonetics.* Chicago: University of Chicago Press.

Ladefoged, P. (1982) *A Course in Phonetics.* New York: Harcourt, Brace, Jovanovich, 2nd edn.

Lehiste, I., J. Olive, and L.A.Streeter (1976) The role of duration in disambiguating syntactically ambiguous sentences. *The Journal of the Acoustical Society of America*, **60**, 1199–1202.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, **6**, 172–187.

Lindblom, B. (1963) Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, **35**, 1773–81.

Lindblom, B. (1983) Economy of speech gestures. In P. F. MacNeilage (ed.), *The Production of Speech*, pp. 217–245. Heidelberg: Springer-Verlag.

Lindblom, B. (1990) Explaining phonetic variation: a sketch of the H & H theory. In W. J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*, pp. 403–439. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Lodge, M. (1981) *Magnitude Scaling: Quantitative Measurement of Opinions.* Beverly Hills, California: Sage Publications.

Luce, P. A. (1986) A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics*, **39**, 155–158.

Mayo, C., M. Aylett, and D. Ladd (1997) Prosodic transcription of Glasgow English: An evaluation study of glatobi. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis (eds.), *Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications.*, pp. 231–234. ESCA and The University of Athens.

McEnery, T. and A. Wilson (1996) *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McGill, W. J. (1954) Multivariate information transmission. *Psychometrika*, **19**, 97–116.

Miller, G. A. and F. C. Frick (1949) Statistical behavioristics and sequences of responses. *Psychological Review*, **56**, 311–324.

Moon, S.-J. and B. Lindblom (1994) Interaction between duration, context and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, **96**, 40–55.

Moore, B., B. Glasberg, and D. Vickers (1994) Simulation of the effects of loudness recruitment on the intelligibility of speech in noise. *British Journal of Audiology*, **29**, 131–143.

Nespor, M. and I. Vogel (1986) *Prosodic Phonology.* Dordrecht: Foris Publications.

Neter, J., W. Wasserman, and M. H. Kutner (1990) *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Design.* Boston: Irwin, 3rd edn.

Nooteboom, S. G. (1997) The prosody of speech: Melody and rhythm. In W. Hardcastle and J. Laver (eds.), *The Handbook of Phonetic Sciences.*, pp. 641–673. Oxford: Blackwell.

Payne, D., L. Peters, D. Birkmire, M. Bonto, J. Anastasi, and M. Wenger (1994) Effects of speech intelligibility level on concurrent visual task-performance. *Human Factors*, **36**, 441–475.

Payton, K. L., R. M. Uchanski, and L. D. Braida (1994) Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, **95**, 1581–1592.

Picheny, M., N. Durlach, and L. Braida (1985) Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research*, **28**, 96–103.

Picheny, M., N. Durlach, and L. Braida (1986) Speaking clearly for the hard of hearing ii: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research*, **29**, 434–445.

Pierce, J. R. (1961) *Symbols, Signols and Noise: The Nature and Process of Communication.* New York: Harper.

Pisoni, D., H. Nusbaum, P. Luce, and L. Slowiaczek (1985) Speech perception, word recognition, and the structure of the lexicon. *Speech Communication*, **4**, 75–95.

Pitrelli, J. F., M. E. Beckman, and J. Hirschberg (1994) Evaluation of prosodic transcription labeling reliability in the tobi framework. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 123–6.

Price, P., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong (1991) The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, **90**, 2956–70.

Rosner, B. and J. Pickering (1994) *Vowel Perception and Production.* Oxford: Oxford Science Publications.

Samual, A. G. and M. Troicki (1998) Articulation control is inversely related to redundancy when children or adults have verbal control. *Journal of Memory and Language*, **39**, 175–194.

Selkirk, E. (1978) On prosodic structure and its relation to syntactic structure. In T. Fretheim (ed.), *Nordic Prosody II.* Trondheim: TAPIR.

Selkirk, E. (1984) *Phonology and Syntax: The relation between sound and structure.* Cambridge, Mass.: The MIT Press.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379–423;623–656.

Sharp, A. E. (1960) The analysis of stress and juncture in English. *Transactions of the Philological Society*, pp. 104–135.

Shattuck-Hufnagel, S. and A. E. Turk (1996) A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, **25**, 193–247.

Shields, L. W. and D. A. Balota (1991) Repetition and associative context effects in speech production. *Language and Speech*, **34**, 47–55.

Shillcock, R., J. Hicks, P. Cairns, N. Chater, and J. Levy (1994) Phonological reduction, assimilation, intra-word information structure, and the evolution of the lexicon of english. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society.*

Silverman, K. E., E. Blaauw, J. Spitz, and J. F. Pitrelli (1992) A prosodic comparison of spontaneous speech and read speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP).*

Sluijter, A. (1995) *Phonetic Correlates of Stress and Accent.* The Hague: Holland Academic Graphics.

van Son, R. (1993) *Spectro-Temporal Features of Vowel Segments.* Ph.D. thesis, University of Amsterdam.

Sotillo, C. F. (1997) *Phonological Reduction and Intelligibility in Task-Oriented Dialogue.* Ph.D. thesis, University of Edinburgh.

Summers, W. V. (1987) Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *The Journal of the Acoustical Society of America*, **82**, 847–863.

Turk, A. and L. White (1999) Structural influences on accentual lengthening in English. *Journal of Phonetics*, **27**, 171–206.

Turk, A. E. and J. R. Sawusch (1997) The domain of accentual lengthening in American English. *Journal of Phonetics*, **25**, 25–41.

Uchanski, R. M., S. Choi, L. Braida, C. M. Reed, and N. I. Durlach (1996) Speaking clearly for the hard of hearing IV: Further studies in the role of speaking rate. *Journal of Speech and Hearing Research*, **39**, 494–509.

Wightman, C. and M. Ostendorf (1991) Automatic recognition of prosodic phrases. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing.*, pp. 321–324.

Wightman, C., S. Shattuck-Hufnagel, M. Ostendorf, and P. Price (1992) Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, **91**, 1707–17.

Wrench, A. A. (1995) Analysis of fricatives using multiple centres of gravity. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, vol. 4, pp. 460–463.

Wright, R. (1997) Lexical competition and reduction in speech: A preliminary report. *Progress Report, Indiana University*, **21**, 471–485.

Young, S., J. Jansen, J. Odell, D. Ollason, and P. Woodland (1996) *The HTK Book.* Entropic. Version 2.00.

Zipf, G. K. (1949) *Human Behavior and the Principle of Least Effort.* Reading, Mass.: Addison-Wesley.

Zwicker, E. (1961) Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, **33**, 248–249.

Zwicker, E. and E. Terhardt (1980) Analytical expressions for critical bandwidths as a function of frequency. *The Journal of the Acoustical Society of America*, **68**, 1523–1525.

Appendix A

DISC and other phonetic codes used in CELEX

IPA	example	SAM-PA	CELEX	CPA	DISC
p	**p**at	p	p	p	p
b	**b**ad	b	b	b	b
t	**t**ack	t	t	t	t
d	**d**ad	d	d	d	d
k	**c**ad	k	k	k	k
g	**g**ame	g	g	g	g
ŋ	ba**ng**	N	N	N	N
m	**m**ad	m	m	m	m
n	**n**at	n	n	n	n
l	**l**ad	l	l	l	l
r	**r**at	r	r	r	r
f	**f**at	f	f	f	f
v	**v**at	v	v	v	v
θ	**th**in	T	T	T	T
ð	**th**en	D	D	D	D
s	**s**ap	s	s	s	s
z	**z**ap	z	z	z	z
ʃ	**sh**eep	S	S	S	S
ʒ	measure	Z	Z	Z	Z
j	**y**ank	j	j	j	j
x	lo**ch**	x	x	x	x
h	**h**ad	h	h	h	h
w	**wh**y	w	w	w	w
ʧ	**ch**eap	tS	tS	T/	J
ʤ	**j**eep	dZ	dZ	J/	_
ŋ̩	baco**n**	N,	N,	N,	C
m̩	idealis**m**	m,	m,	m,	F
n̩	burde**n**	n,	n,	n,	H
l̩	dangle	l,	l,	l,	P
*	fathe**r**	r*	r*	r*	R
(possible linking 'r')					

Figure A.1: Computer phonetic codes for English consonants. (Taken from the CELEX manual p4-25 Baayen *et al.*, 1995)

IPA	example	SAM-PA	CELEX	CPA	DISC
ɪ	p**i**t	I	I	I	I
ɛ	p**e**t	E	E	E	E
æ	p**a**t	{	&	^/	{
ʌ	p**u**tt	V	V	^	V
ɒ	p**o**t	Q	O	O	Q
ʊ	p**u**t	U	U	U	U
ə	**a**noth**er**	@	@	@	@
iː	b**ea**n	i:	i:	i:	i
ɑː	b**ar**n	A:	A:	A:	#
ɔː	b**or**n	O:	O:	O:	$
uː	b**oo**n	u:	u:	u:	u
ɜː	b**ur**n	3:	3:	@:	3
eɪ	b**ay**	eI	eI	e/	1
aɪ	b**uy**	aI	aI	a/	2
ɔɪ	b**oy**	OI	OI	o/	4
əʊ	n**o**	@U	@U	O/	5
aʊ	br**ow**	aU	aU	A/	6
ɪə	p**ee**r	I@	I@	I/	7
ɛə	p**air**	E@	E@	E/	8
ʊə	p**oo**r	U@	U@	U/	9
æ	t**im**bre	{~	&~	^/~	c
ɑ̃ː	dét**en**te	A~:	A~:	A~:	q
æ̃ː	l**in**gerie	{~:	&~:	^/~:	0
ɒ̃ː	bouill**on**	O~:	O~:	O~:	~

Figure A.2: Computer phonetic codes for English vowels and diphthongs. (Taken from the CELEX manual p4-26 Baayen *et al.*, 1995)

Appendix B

An example Dialogue from the HCRC Map Task (Q3NC8)

B.1 Instructions For Subjects

Map instructions

(once they're sitting down and have been given their copy of the map)

to the speaker

You and your partner have both got a map of the same place.

Your map has got a route on it; your partner's map does not.

Your job is to describe the route to your partner so that s/he can draw it on her/his map.

Your path is known to be the only reliable route through and around all the various obstacles.

You must try to describe your route carefully so that your partner can avoid the obstacles and hazards on the way.

It is important to avoid these obstacles, rather than to make the routes identical to the last millimetre!

As you do this, keep in mind that the maps have been drawn by different explorers and might not be quite the same.

then to the hearer

You and your partner have both got a map of the same place.

Your partner's map has got a route on it, which s/he's going to describe to you.

Your job is to draw the route on your map.

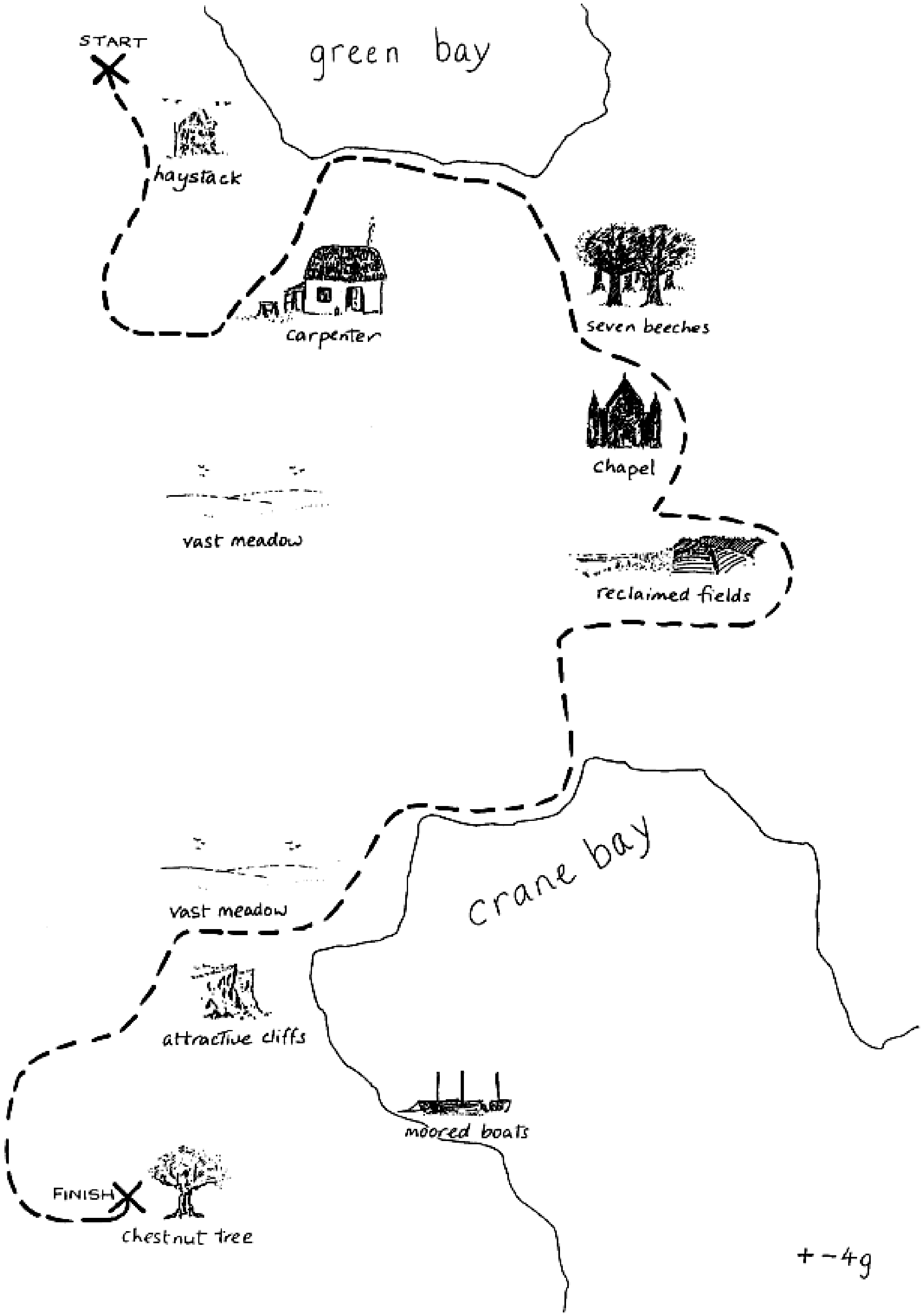
You must draw the route with care, because it's the only route known to avoid the various obstacles you may encounter.

Listen carefully to what your partner says, and ask questions if there's anything you're not sure about.

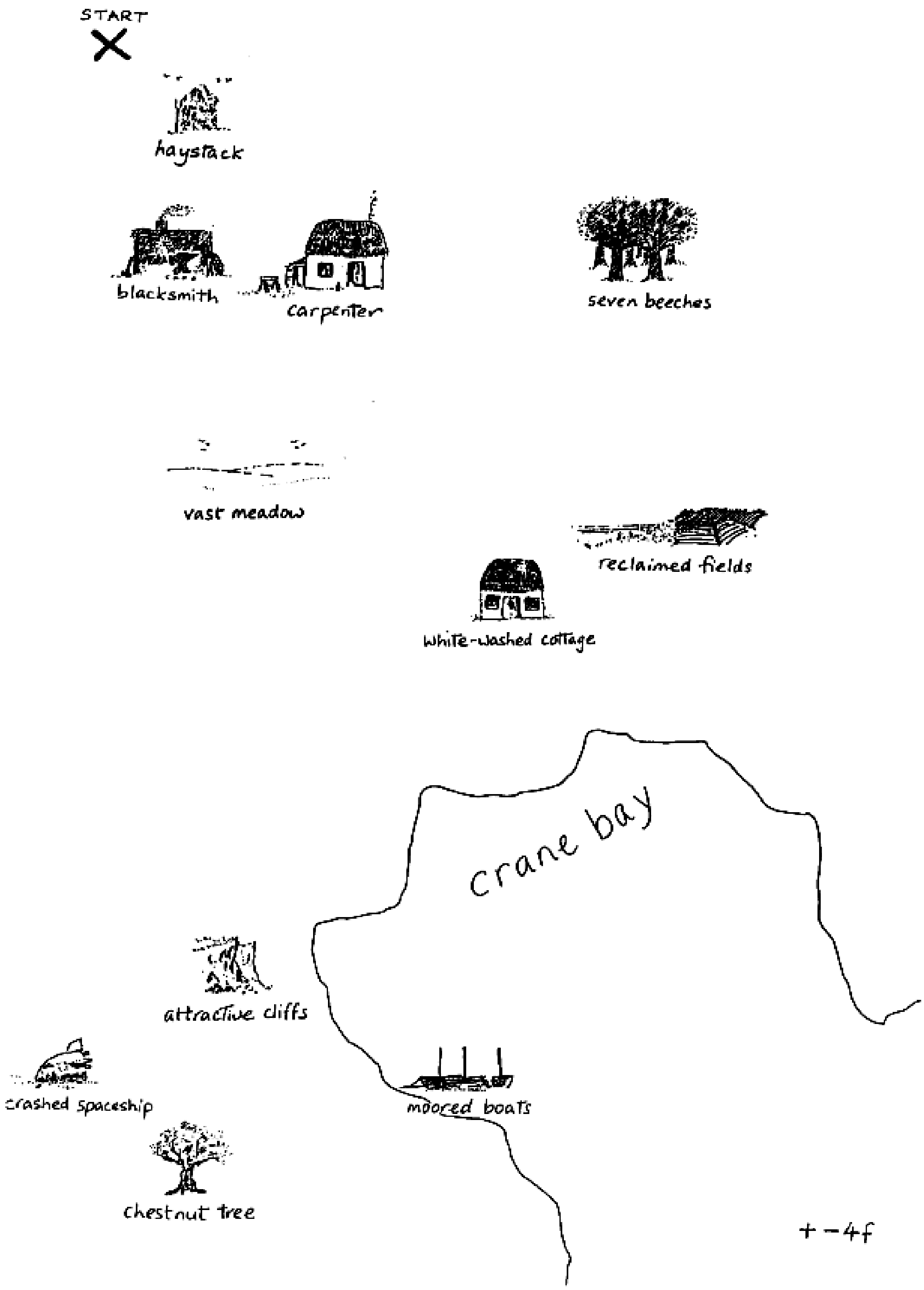
As you do this, keep in mind that the maps have been drawn by different explorers and might not be quite the same.

Do you understand what you're supposed to do?

B.2 Givers Map



B.3 Followers Map



B.4 Transcription of Dialogue

GIVER	FOLLOWER
M1(READY): Right	
M2(QUERY-YN): Have you got the start above the haystack	
	M3(REPLY-Y): Yeah
M4(ACKNOWLEDGE): Right	
M5(INSTRUCT): If you want to sort_of......... ehm...... head down towards the haystack,... right,... pass it by on its left-hand side	
	M6(ACKNOWLEDGE): Yeah, 'kay
M7(ALIGN): Right	
M8(INSTRUCT): And head... more_or_less straight down maybe curving slightly towards your left	
	M9(QUERY-YN): Have you got a blacksmith marked
M10(REPLY-N): No,... I don't	
	M11(ACKNOWLEDGE): No
M13(INSTRUCT): Ehm, head down for about,... eh two inches from the	

haystack,......... two and a half inches

M14(ACKNOWLEDGE): Two inches okay

M15(ALIGN): Right

M16(INSTRUCT): Ehm,... go to your right towards the carpenter's house

M17(READY): All right

M18(EXPLAIN): Well I'll need to go below. Got a blacksmith marked

M19(ACKNOWLEDGE): Right,... well you do that

M20(CHECK): Do you want it_to go below the carpenter

M21(REPLY-N): No

M22(REPLY-W): I want you to go up the left-hand side of it towards...... green bay and make it a slightly diagonal line,... towards,... ehm... sloping to the right

M23(ACKNOWLEDGE): Okay

M25(CHECK): So you want me to go above the carpenter

M26(REPLY-Y): Uh-huh

M27(ACKNOWLEDGE): Right

M28(CLARIFY): Towards... the bay

M29(QUERY-W): The bay

M30(QUERY-YN): Have you got the bay, no

M31(CHECK): What crane bay

M32(CLARIFY): Green bay

M33(REPLY-N): No, I don't have a crane bay

M34(ACKNOWLEDGE): Right

M35(INSTRUCT): Okay well head up above the carpenter's house for about, ehm,... it should be about... an inch above it

M36(ACKNOWLEDGE): Alright

M38(ACKNOWLEDGE): Okay

M39(INSTRUCT): And head... slope slightly... down the way... for about two inches

M40(ALIGN): Right

M41(QUERY-YN): Do you have the seven beeches

M42(ACKNOWLEDGE): Well okay

M44(REPLY-Y): Nope.... Oh, yes I do sorry

M45(ACKNOWLEDGE): You do

M46(READY): Right

M47(INSTRUCT): Go down past them on their left-hand side

M48(ACKNOWLEDGE): Okay

M49(INSTRUCT): And stop when you get to where it says seven beeches

M50(ACKNOWLEDGE): Okay

M51(INSTRUCT): Now you're going to go underneath that bit. You're going to make a slight curve, ehm... to the right,... right? While it's still going down the way

M52(ACKNOWLEDGE): Right

M53(EXPLAIN): Because you're avoiding a chapel which I don't think you've got

M54(ACKNOWLEDGE): No

M55(CHECK): So I'm going right

M56(REPLY-Y): Uh-huh, you're going right

M57(CLARIFY): Make it a curve sort_of

M58(CHECK): Down the way

M59(REPLY-Y): Down the way,... uh-huh

M60(CLARIFY): Out towards the... the right-hand side of your paper

M61(ACKNOWLEDGE): Okay

M62(QUERY-W): How far out towards right-hand side

M63(UNCODED): Right

M65(CLARIFY): Not too far, just... like you were drawing a circle but not quite

M66(ACKNOWLEDGE): Okay

M68(ACKNOWLEDGE): Right

M69(ALIGN): Ehm, now you're... slightly less than an inch below the chapel

M70(EXPLAIN): You haven't got the chapel. Ha ha ha

M71(ACKNOWLEDGE): No

M72(ALIGN): You're about... two and a half inches below the seven beeches.... Right you're above the

M73(CHECK): Turn to the right of them

M74(CLARIFY): Ehm, not really, you're underneath them

M75(REPLY-N): Oh. Oh no I'm not

M76(EXPLAIN): Well you should be

M77(ACKNOWLEDGE): Right

M78(EXPLAIN): You should be just above the reclaimed fields

M79(ACKNOWLEDGE): Right, okay

M80(EXPLAIN): I can go down there

M81(ACKNOWLEDGE): Right... ehm

M82(INSTRUCT): And, I want you to come... above them and round to the right-hand side of them and underneath them

M83(ACKNOWLEDGE): Right

M84(ALIGN): Right?... Dri--

M85(CHECK): Down to the right-hand side

M86(CLARIFY): Round the right-hand side and come right along underneath them

M87(ACKNOWLEDGE): Oh, right

M89(ACKNOWLEDGE): Okay

M90(INSTRUCT): Right and stop when you get to... the line where they stop

M91(ACKNOWLEDGE): Right, okay

M92(INSTRUCT): And I'd like you to come... straight down the way... towards crane bay

M93(ACKNOWLEDGE): Okay

M94(INSTRUCT): And when you get to that curve of crane bay stick closely to it

M95(ACKNOWLEDGE): Okay

M96(INSTRUCT): For... until you get to that corner

M97(QUERY-W): Which corner

M98(CLARIFY): You see where the... just opposite... the "c"... of crane bay,...... diagonally opposite

M99(CHECK): Towards the north of it, or to the

M100(REPLY-Y): Uh-huh

M101(ACKNOWLEDGE): Right

M102(READY): Right

M103(ALIGN): Now you're heading towards vast meadow and attractive cliffs

M104(REPLY-Y): Well... yes okay

M105(ACKNOWLEDGE): Right

M106(UNCLASSIFIABLE): Now I don't want you to stick to the coast,... just opposite vast meadow

M107(INSTRUCT): Right, you've got to come down in_between vast meadow and the attractive cliffs

M108(ACKNOWLEDGE): Okay

M109(UNCLASSIFIABLE): In a straight line between them once you've come down,... ehm...... at a southwesterly angle... towards them and then in

M110(UNCLASSIFIABLE): Straight in_between them

M111(ALIGN): Right

M112(REPLY-Y): Right

M113(CHECK): So I'm down near the

attractive cliffs

M114(REPLY-Y): Uh-huh, you're in_between vast meadow

M115(CLARIFY): And attractive cliffs

M116(ACKNOWLEDGE): Okay

M117(INSTRUCT): And then you come down in a southwesterly angle again,... down the left-hand side of the attractive cliffs

M118(ACKNOWLEDGE): Okay

M119(INSTRUCT): Stop when you get to the bottom of them

M120(ACKNOWLEDGE): Okay

M121(QUERY-YN): Have you got crashed spaceship marked

M122(REPLY-N): No

M124(EXPLAIN): Oh right... well I'm quite close to the edge

M123(INSTRUCT): Ehm,... I'd like you to head... m-- more_or_less westwards curving slightly down the way... towards... the left-hand side of_the page... very very close to the edge

M125(QUERY-W): I mean how far down

do you want me to go

M126(CLARIFY): I want you to well you're heading towards the chestnut tree but you're not,... ehm... going diagonally towards it

M127(ALIGN): Right

M128(CLARIFY): Just come down the side of the page for about an inch and a half

M129(ACKNOWLEDGE): Okay

M130(CHECK): Then head towards chestnut tree

M131(REPLY-Y): Uh-huh

M132(CLARIFY): Towards the finish

M133(ACKNOWLEDGE): Okay

M134(QUERY-W): Where's the finish

M135(REPLY-W): At the chestnut tree

M136(ACKNOWLEDGE): Right

M137(CHECK): North of it

M138(REPLY-N): No

M139(REPLY-W): Just by the side of it, at the the left-hand side of it

M140(UNCLASSIFIABLE): Left-hand side

M142(ACKNOWLEDGE): Okay

M143(EXPLAIN): That’s you.... I hope

Appendix C

Finding Formant Targets with Parametric Curves

C.1 Algorithm

The algorithm used in this work was as follows:

1. Find sections of voiced speech using the entropics pitch tracker.

2. Calculate F1 and F2 formant values over the area using the Entropic's formant tracker.

3. Move left to right across the voiced speech a 10ms frame at a time. Fit a simple parametric curve to the data using mean squared error (MSE) (see below) from that point over windows varying from 40-100ms.

4. For each frame retain the target value estimated by the parametric curve with the lowest average MSE per frame (Total MSE/window size).

C.2 Fitting Parametric Curves Using Mean Squared Error

The calculations required to fit a simple parametric curve using mean squared error are as follows:

For each point $[x(i), y(i)]$ over a window of n points, the mean squared error E for a curve $y = ax^2 + bx + c$ is:

$$E = \sum_{i=0}^{n} (y(i) - (ax(i)^2 + bx(i) + c))^2$$

To find the minimum error differentiate in parts with respect to a, b, c and giving:

$$\frac{da}{dE} = \sum_{i=0}^{n} 2ax(i)^4 + 2bx(i)^3 + 2cx(i)^2 - 2x(i)^2y(i)$$

$$\frac{db}{dE} = \sum_{i=0}^{n} 2ax(i)^3 + 2bx(i)^2 + 2cx(i) - 2x(i)y(i)$$

$$\frac{dc}{dE} = \sum_{i=0}^{n} 2ax(i)^2 + 2bx(i) + 2c - 2y(i)$$

Set all three equations to 0 for the minimum error and substitute to find a, b, c in terms of $x(i)$ and $y(i)$. To simplify for calculation in a subroutine in a computer program the resulting equations can be rearranged using temporary variables p, q, r, s, t as follows:

$$p = n \sum x(i)^3 - \sum x(i)^2 \sum x(i)$$

$$q = n \sum x(i)^2 - \sum x(i) \sum x(i)$$

$$r = n \sum x(i)^4 - \sum x(i)^2 \sum x(i)^2$$

$$s = n \sum x(i)y(i) - \sum x(i) \sum y(i)$$

$$t = n \sum x(i)^2y(i) - \sum x(i)^2 \sum y(i)$$

and a, b, c are calculated as follows:

$$a = \frac{ps - tq}{p^2 - rq}$$

$$b = \frac{pt - rs}{p^2 - rq}$$

$$c = \frac{\sum y(i) - a \sum x(i)^2 - b \sum x(i)}{n}$$