Summary of Ph.D., accepted January 1996

A Large-Scale Analysis Of The Acoustic-Phonetic Markers Of Speaker Sex

Gavin John Dempster

The research for this thesis lies within the field of speaker characterisation through the acoustic-phonetic analysis of speech. The thesis consists of two parts:

1. An investigation of the acoustic-phonetic differences between the speech of women and men;
2. An examination of the practicalities of automating the investigation to analyse a large speech database.

The acoustic-phonetic markers of speaker sex examined here are the fundamental frequency, the formant frequencies, and the relative amplitude of the first harmonic. The aims of the investigation were, firstly, to establish to what extent these markers differentiate between the sexes, and secondly, to examine the extent of between- and within-speaker deviation from the female and male norms, or average values for each sex.

These points were investigated by an automated acoustic-phonetic analysis of the TIMIT database, involving a data set of almost 16,000 segments of speech. An automated method was developed to enable the signal processing and statistical analysis of a data set of this size. The problems to be encountered in the analysis of a highly variable data source (i.e. the acoustic speech waveform) are addressed.

A Large-Scale Analysis Of The Acoustic-Phonetic Markers Of Speaker Sex

Gavin John Dempster
Departments of Computer Science and Psychology

Submitted for the Degree of Doctor of Philosophy (Ph.D.) February 1995
Accepted January 1996

Abstract

The research for this thesis lies within the field of speaker characterisation through the acoustic-phonetic analysis of speech. The thesis consists of two parts:

1. An investigation of the acoustic-phonetic differences between the speech of women and men;

2. An examination of the practicalities of automating the investigation to analyse a large speech database.

The acoustic-phonetic markers of speaker sex examined here are the fundamental frequency, the formant frequencies, and the relative amplitude of the first harmonic. The aims of the investigation were, firstly, to establish to what extent these markers differentiate between the sexes, and secondly, to examine the extent of between- and within-speaker deviation from the female and male norms, or average values for each sex.

These points were investigated by an automated acoustic-phonetic analysis of the TIMIT database, involving a data set of almost 16,000 segments of speech. An automated method was developed to enable the signal processing and statistical analysis of a data set of this size. The problems to be encountered in the analysis of a highly variable data source (i.e. the acoustic speech waveform) are addressed.

Contents

1 Introduction **4**

1.1 Some issues in speaker characterisation 4

1.1.1 The integration of speech technology and speaker characterisation . 4

1.1.2 Speaker characterisation and large-scale data analysis 6

1.1.3 Characterising speaker sex 7

1.1.4 The limitations of analysing variable data sources 7

1.2 Outline of the research 8

1.2.1 The development of the research plan 8

1.2.2 The investigation of the acoustic-phonetic sex markers 9

1.2.3 The design of a method for the analysis of speech databases 10

1.3 Outline of the thesis 11

2 An Introduction to Speaker Variability **15**

2.1 Introduction 15

2.1.1 A unified cognitive model of speech communication 15

2.1.2 Some concepts in the characterisation of speakers 16

2.1.3 Deciphering the linguistic message in the speech signal 18

2.2 A review of the sources of speaker variability 21

2.2.1 Describing speaker variability 21

2.2.2 Between-speaker and within-speaker variability 21

2.2.3 The consequences of speaker variability 25

2.2.4 Sources of speaker variability 28

3 The Variability in the Voice due to Speaker Sex **34**

3.1 Variation at the acoustic-phonetic level 36

3.1.1 The voice source and the fundamental frequency 37

3.1.2 The voice source and the relative amplitude of the first harmonic . . 45

3.1.3 The vocal tract and the formant frequencies 55
3.1.4 The vocal apparatus and the effects of aging, height and weight . . . 60
3.2 The perception of speaker sex 70
3.2.1 Pre-pubescent speaker sex recognition 70
3.2.2 Post-pubescent speaker sex recognition 78
3.3 Conclusions 84

4 An Analysis of the Acoustic-Phonetic Markers of Speaker Sex 87
4.1 Method 89
4.1.1 Discussion of the speech data used in the analysis 90
4.1.2 Techniques for measuring the acoustic-phonetic parameters 95
4.1.3 The TIMIT database 101
4.2 Results 111
4.2.1 Fundamental frequency 112
4.2.2 Relative amplitude of the first harmonic 132
4.2.3 Formant frequency 155
4.2.4 A Summary of the results 162
4.3 Discussion Of Results 167
4.4 Conclusions 176
4.4.1 Remarks on speaker characterisation 176
4.4.2 Remarks on the acoustic-phonetic characterisation of speaker sex . . 176
4.4.3 Remarks on the automated analysis of speech databases 178

A Analysis of the TIMIT Database using UNIX 182
A.1 A description of the UNIX tools used in the analysis 182
A.2 Using UNIX to analyse the TIMIT database 186

B The Database Analysis Procedure 188
B.1 The naming protocols 190
B.2 Setting up the structures for the analysis 197
B.3 Extraction of the slices from the TIMIT CD-ROM 201
B.4 Signal processing of the slices 203
B.5 Performing the slice statistics 212
B.6 Checking the analysis data 214
B.7 Statistical analysis of the slices and speakers 238

Chapter 1

Introduction

The research for this thesis lies within the field of speaker characterisation through the acoustic-phonetic analysis of speech. The thesis consists of two parts:

1. An investigation of the acoustic-phonetic differences between the speech of women and men;
2. An examination of the practicalities of automating the investigation to analyse a large speech database.

The acoustic-phonetic markers of speaker sex examined here are the fundamental frequency, the formant frequencies, and the relative amplitude of the first harmonic. The aims of the investigation were, firstly, to establish to what extent these markers differentiate between the sexes, and secondly, to examine the extent of between- and within-speaker deviation from the female and male norms, or average values for each sex.

These points were investigated by an automated acoustic-phonetic analysis of the TIMIT database, involving a data set of almost 16,000 segments of speech. An automated method was developed to enable the signal processing and statistical analysis of a data set of this size. The problems to be encountered in the analysis of a highly variable data source (i.e. the acoustic speech waveform) are addressed.

In the rest of this introduction, the first part, Section 1.1, looks at some of the issues in the field of speaker characterisation that are relevant to the research conducted for this thesis, in particular the need for large-scale analysis of speech in order to fully characterise different speech forms and speaker types. Section 1.2 describes the development of the research plan, and outlines the research reported in the thesis. Finally, the chapters of the thesis itself are summarised in Section 1.3.

1.1 Some issues in speaker characterisation

1.1.1 The integration of speech technology and speaker characterisation

At present, speech technology lacks the capability to deal efficiently with the diverse nature of speaker characteristics, from the differences between the speech of individuals, to the differences which signal membership of groups. One of the causes of this is the failure to address the inherent variability of speech. The variability that speakers exhibit at all levels of speech behaviour has typically been viewed as 'noise' to be filtered out,

or ignored if possible, in order to reach the linguistic message, or 'true' speech signal (Nolan 1983:3). It has often seemed to be the case that researchers have assumed that if the linguistic side of speech was mastered, then the variability associated with different speakers would somehow cease to be a problem. In effect it is assumed that speaker variability has a negligible effect on the acoustic speech signal. Research into automatic speech recognition (ASR) has concentrated on improving the recognition technology as it stands, the emphasis being placed upon deciphering the segmental and suprasegmental interactions in speech. By ignoring the poorly-understood acoustic differences between speakers, the development of ASR systems into a truly speaker-independent technology has been seriously hampered. Speech synthesis on the other hand has tended to concentrate on improving the understandability of the linguistic content of the synthesiser output; the synthesis of more naturalistic speech has been considered to be far less important. As a result, speech output systems are generallv incapable of adopting voice-types which could signal identity, personality or intent.

In the future, speech technology will benefit from increasingly powerful computers, able to integrate many more of the speech signal's layers of complexity. It will be possible to accommodate models of particular speaker characteristics to improve the performance of the technology. For instance, ASR systems will be able to identify characteristics from the input speech signal and use them to aid the deciphering of the linguistic message, for example by adjusting the expectations of the values of the fundamental frequency a speaker is likely to attain. Speech output systems will be capable of producing speech with a more naturalistic voice, overlaying extralinguistic and paralinguistic information onto the linguistic message. In effect, speaker characterisation could be the fine-tuning required to enable speech technology to be more flexible and responsive, and for ASR to be truly speaker-independent. However, much work needs to be carried out in order to understand how speaker characteristics are conveyed in the acoustic speech signal, and how we perceive them.

There are many sources of variability that create recognisable speaker characteristics in the acoustic speech signal. From a casual observation of any group of people involved in a conversation, it soon becomes clear that there is immense variation in their acoustic speech output. This is true at the acoustic-phonetic level, which considers the physical properties of individual speech sounds, and at higher levels, involving such factors as the intonation contour, speech rate and stress timing. For as well as the literal meaning of the words contained within it, the acoustic speech signal carries such information about the speaker as her or his sex, social and regional background, personality and emotional state. A person's voice identifies them both as an individual and as a member of particular groups. A more detailed observation of the conversants would reveal that as well as there being considerable variation between different the speakers, there is much within-speaker variability, to the extent that: "No two repetitions by a single speaker of a given phrase spoken at the same tempo are ever fully identical, at the level of articulatory and temporal microstructure" (Laver 1988:93). This capacity for variation in the acoustic speech signal, coupled with the ability of human beings to discount the variation in order to comprehend the linguistic message, and to perceive the sources of the variation in order to describe speakers according to voice type, suggests that one way to improve the performance of speech input and output systems would be to exploit this ability. By being able to characterise speech according to speaker type, an ASR system could pre-process the speech signal to remove the artifacts of variability, while retaining those components of the signal necessary for comprehension of the linguistic message. In speech synthesis, by describing the subtle cues that signal the membership of a particular group or the emotional state of a speaker, it would be possible to create more naturalistic artificial

voices.

1.1.2 Speaker characterisation and large-scale data analysis

The issue of speaker variability has only begun to influence the wider speech research community in recent years with the rise in the interest in speaker characterisation. Speaker characterisation attempts to identify the differences between the speech of individuals or of definable groups, for example by discovering the acoustic features which differentiate speakers of different ages, angry voices from happy voices, and working class speakers from middle class speakers. Much of the previous work in this area has been done in the field of sociolinguistics.

However, studies investigating the acoustic-phonetic markers of speaker characteristics have typically involved small numbers of speakers under particular research conditions. At the same time, researchers have often implied, or even explicitly stated, that their small samples of speakers are representative of the population as a whole without taking into account the influences of between- and within-speaker variability. This has had the effect of masking the diversity inherent in the acoustic speech signal. One reason for this is that research is generally carried out on homogeneous groups. While this is to some extent necessary in order to control for sources of variability that are not under investigation, there is often little acknowledgement that other voice types may deviate significantly from the feature values established for the sample group. Considerations of time and money have meant the sample populations are drawn from academia, and thus much research into real speech data has centred around white, middle class, often male speakers. A second reason is that speakers within any definable group will exhibit great variation in whatever feature is being measured, both in terms of their own mean values relative to the group mean for that feature, and in the range of values the speaker attains in different speaking conditions.

This indicates the need for large-scale studies to provide a sound statistical basis for description of the markers of speaker characteristics. The studies should involve sufficient speakers to describe the extent of between-speaker variability for a particular parameter, and sufficient speech from individuals to gauge how the parameter varies during the course of a person's normal speech. One of the reasons for the relative lack of interest in this area has been the difficulties in acquiring and analysing a sufficient quantity of data. There are a number of explanations lying behind this difficulty:

- The collection of speech data is a very expensive task. For instance, considerable effort must be put into designing the stimuli for the subjects to utter to ensure a sufficient sampling of their normal speaking characteristics. In addition, a comprehensive database will involve many speakers and many recording sessions.
- The amount of data contained in a large speech database will be immense, creating problems of storage and organisation. Speech must be recorded digitally at very high sampling rates to ensure the signal is not noticeably degraded. A typical sampling rate is 16,000 samples per second; in other words, 16,000 items of data for every second of speech on the database.
- The data must be organised into a consistent and computer-readable format to speed up access to and processing of the data. For example, phonetic transcriptions of the speech must be transparent to allow the targetting of specific types of phone. Furthermore, for a large-scale analysis, any small delay in a particular sub-process will be multiplied many times over in the full analysis.

- The processing power of the computers must be sufficient to handle the huge volumes of speech data, and the analysis of that data. For example, most signal processing of speech data involves some form of frequency analysis. This requires performing Fourier transforms on the speech, a method which is computationally expensive.

- The storage power of computers must also be able to cope with the amount of data and the output of its analysis. Because computer memory can be a scarce resource in research facilities, many databases are now available on CD-ROM. The memory saved by using a CD-ROM can be used to store the inevitably vast quantity of data generated by the analysis.

- Methods must be designed to handle the data analysis in an efficient and accurate manner. The correct segments of speech must be retrieved from the database, which must then be passed through the signal processing and statistical analysis algorithms, and finally the output must be coherently stored to enable further data retrieval and analysis.

1.1.3 Characterising speaker sex

Of the potential sources of variation in speech signal arising from the vocal characteristics of different speakers, the particular one examined in this thesis is that of speaker sex. Although the situation is changing, this has been a much-neglected area of research. Speech science, like most science, has traditionally been a male-dominated and androcentric area of research, dominated by male researchers and centred around investigating the male vocal apparatus. Most of our knowledge of speech is actually knowledge of male speech characteristics, and is thus sadly lacking in depth. A number of studies have reported that, when faced with a woman's voice, some speech recognisers give an inferior performance (see Doddington & Schalk 1981, Noyes & Frankish 1989, Waterworth 1984). Most previous attempts at dealing with variation due to speaker sex have involved simple formant and fundamental frequency scaling strategies. But the "analysis ... of female ... speech involve[s] more than a mere scaling of fundamental frequency" (Titze 1989a:1699). Rather, it is clear that some aspects of the output of the female speech production apparatus are radically different to that of the male, and should be treated as such.

1.1.4 The limitations of analysing variable data sources

Probably the most problematic aspect of research into speech, and in some respects one of the most overlooked, is the inherent variability of the acoustic speech waveform. There are an immense number of interrelated factors imposing their own influence and character on the speech signal, resulting in a highly variable data source. Thus, whilst in Section 1.1.2 above the need for a sound statistical basis for the description of the parameters of the voice has been acknowledged, the thoroughness with which the parameters can be described will almost inevitably be compromised by the inability to account for all speaker types in all speaking conditions. The results of any study into speech are inevitably constrained by the data upon which the analysis was based.

Thus, one of the most important criteria in the investigation of speaker characteristics is a recognition of the constitution of the analysed speech data. To illustrate this point, let us consider the research reported upon in this thesis. The speech data used in the analysis came from the TIMIT database, and this data imposes two restrictions on the results of this characterisation of speaker sex. The method of investigation adopted for this study

imposes one further restriction. Firstly, the speakers represented on this database are in the main white, middle class, university-educated U.S. citizens aged between 20 and 39 years (the attributes of the speakers are discussed in more detail in Section 4.1.3). This research is therefore an investigation of the acoustic-phonetic characteristics of speaker types, and it is insufficient to claim that the results of this study are representative of the population as a whole. Secondly, the database consists solely of read sentences, which is less dynamic than spontaneous speech. The data is therefore not entirely representative of a person's normal speaking patterns. Thirdly, the investigation reported upon in this thesis considered only certain vowels. It is possible that important sex-discriminating information is carried in other types of speech sounds, and in domains larger than single sounds.

It is therefore important for researchers to describe their data in sufficient detail to allow their work to be compared and contrasted with other investigations. Then, rather than assuming that, for instance, the characteristics of all female speakers are the same and that studies reporting very different results are in some way in error, a more complete and richer description of the influences on the acoustic speech signal can emerge.

1.2 Outline of the research

The research reported upon in this thesis consists of two distinct themes. The first of these is an investigation into the acoustic-phonetic correlates (or markers) of speaker sex. The second is a description of a method for the exploitation and analysis of a large-scale speech database. These themes will be explored in more detail below, but first it will be useful to consider the development of the research plan.

1.2.1 The development of the research plan

The original intention of this research was to model the characteristics of speaker sex in the acoustic-phonetic domain. However it became clear that the development of such a model was beyond the scope of this thesis. There were three principal reasons for this:

1. The variability imbued in the speech signal as a result of a speaker's sex exists in too many dimensions for it to be modelled successfully at present. In other words there are too many factors (both biological and psychological) involved in creating the percept of 'femaleness' or 'maleness' in the acoustic speech signal.

2. The academic literature provides insufficient data on the acoustic correlates of speaker sex from which a model may be constructed. The implications of the small numbers of speakers investigated in research studies have already been discussed. When research studies are compared, they are often inconclusive or contradictory, or throw light only on particular types of speakers. In addition, the data is usually reported in insufficient detail to gauge the variability of a particular parameter.

3. Many of the signal processing and parameter measurement algorithms required to produce the raw data for analysis are not reliable enough for a multi-speaker investigation. Algorithms are often developed using a limited range of speakers, and are optimised for a set range of values. Thus when they are faced with an anonymous database, containing many more speakers, who exhibit very different behaviour in the parameter to be measured, they can fail to produce accurate results. For example, formant frequency estimators are notoriously unreliable when faced with female speech. As a result, extensive evaluation tasks must be embarked upon to validate an algorithm's performance.

The scope of this research was thus restricted to three of the acoustic-phonetic variables which were cited most often in the literature, and which appeared to have especial importance in the signalling of speaker sex. These were the fundamental frequency (F_0), the amplitude of the first harmonic relative to the second (H_1-H_2), and the formant frequencies (F_1, F_2, F_3). The desire was to study these three variables in depth, encompassing as many speakers and as much speech as possible, to provide a fuller picture of (acoustic-phonetic) speaker sex characteristics and variability. To this end a computer-readable database of speech was selected to make available a sufficiently large amount of easily-accessible data. In addition, the signal processing of the speech to produce the data on the acoustic-phonetic variables was automated, this being the only way to realistically cope with this quantity of data. However, the techniques for the processing of such a large set of data were not readily-available, or had not been developed for speech research. In consequence, a regrettably large portion of the research time was spent in developing such techniques, including the evaluation of the performances of the signal processing

algorithms, which left less time for the statistical analysis of the acoustic-phonetic parameters under investigation. The statistical survey is therefore less comprehensive than was originally intended.

1.2.2 The investigation of the acoustic-phonetic sex markers

In order to set the research goals, and to have something to compare the results of this study with, a review of the literature on the acoustic-phonetic differences between female and male speakers was undertaken (see Chapter 3). From the review, a number of points become apparent:

1. The identification of a person's sex from their speech is perceptually a very easy task.

2. The acoustic-phonetic characteristics of women's speech are, in general, different from those of men's.

3. A speaker's sex is signalled not by a single parameter, such as the fundamental frequency usually cited, but by a number of parameters. More importantly, with regards to how sex is signalled in the voice, it is this combination of parameters that generates the percept of a speaker's sex.

4. The notion of an 'average' or 'idealised' speaker typifying each of the sexes is at best inadequate. While it is possible to compute average parameter values for particular categories of speakers, and while a group average may indicate a sex-dependent difference between the groups (for example, the average formant frequencies of /aa/ for women and men may be sufficiently different to indicate the presence of a sex marker), the members of the groups exhibit great variation both relative to the group's average, and within their own speech.

This research has set out to provide substance to the final three points listed here. This involved finding out in what ways female speakers as a group differ from male speakers, through an investigation of the acoustic-phonetic parameters of the fundamental frequency, the relative strength of the first harmonic, and the first three formant frequencies. Secondly, it involved an examination of the extent of between- and within-speaker variation in the measured parameters, to assess the parameter behaviour both within each sex group and the overlap between the sex groups.

1.2.3 The design of a method for the analysis of speech databases

Four main stages were identified to enable the automatic analysis of speech data, which can be summarised as follows:

1. Preparation of a database of input speech;

2. Establishment of structures to control the analysis of the data;

3. Formation of a database of analysed speech;

4. Statistical analysis of the database of analysed speech.

The TIMIT database satisfied the requirement of a computer-readable data source, although the format of the database rendered the speech data less amenable to analysis by commercial database software. This necessitated the development of software capable of extracting data from the CD-ROM. Structures to control the input, analysis and output of the data had to be designed to enable the complex cross-referencing required in the statistical analysis of the output. Algorithms were developed and comprehensively tested to measure the fundamental frequency, relative amplitude of the first harmonic and formant frequencies of the input vowel phones. The intention to investigate the extent of between- and within-speaker variability required the establishment of two output databases of signal-processed speech: one containing the results of the processing of individual speech segments, and one containing a summary of the results for individual speakers. The software was designed to be flexible, so that it was possible to target any type of phone for analysis, and to perform any type of signal processing on the data.

Finally, the problems associated with automating an analysis of a highly variable data type such as speech are addressed, particularly the signal processing analysis of speech, which can be fraught with difficulties. The speech behaviour of the population as a whole can be remarkably varied, and so any automated analysis of a large number of speakers must combine flexibility and reliability for there to be any confidence in the results it produces. While the algorithms designed to measure fundamental frequency and relative amplitude of the first harmonic were shown to be very robust, problems were encountered in ensuring the accuracy of the output of the formant frequency estimator. The extent of between- and within-speaker variability, and the added difficulties in measuring female formant frequencies (see Section 3.1.3 for further discussion of the problems associated with the measurement of female formants), made it very difficult to define a common search space to capture the three target formants for all speakers.

1.3 Outline of the thesis

The thesis is divided into four main parts: Chapter 2 consists of an introduction to the field of speaker characterisation; Chapter 3 consists of a review of the literature on the variability in the acoustic-phonetic parameters of voice due to speaker sex; and Chapter 4 presents the analytic research carried out for this thesis, while Appendix B describes the procedure developed to enable the data analysis. The thesis is described in more detail below.

Chapter 2 presents an overview of speaker variability, or speaker characterisation. It is limited in the main to the acoustic-phonetic level, but will also take in some higher level effects where necessary (such as the influence of emotional state on prosodic features). It emphasises the inherent variability of speech and the problems this poses for speech technology research: given that apparently distinct acoustic events can be heard as the 'same' sound, it indicates that extraction of speaker-specific information from the speech signal could be used to improve an automatic speech recogniser's performance; and by fully describing aspects of speaker characteristics, it indicates that the output of speech synthesis systems can be improved.

The chapter begins with an introductory section (Section 2.1) that places speaker variability within the wider context of the whole speech communication process and illustrates the many layers of information present in the speech signal. The speech communication process consists of a chain of integrated cognitive, biological and physiological sub-processes linking the speaker's and listener's minds. This is briefly illustrated in Section 2.1.1. Section 2.1.2 deals with ways of describing the information carried by the speech signal, and how this can be used to characterise both individuals and groups of speakers. This consists of the conceptual multidimensional space occupied by speakers, and the linguistic, paralinguistic and extralinguistic information carried in the acoustic speech waveform. Section 2.1.3 discusses some of the issues in the deciphering of the speech signal's linguistic message. Section 2.2 reviews the sources of speaker variability. It begins with a look at the categorisation and description of speaker variability sources (Section 2.2.1), and continues with a look at how they can be used to classify the between-speaker differences across groups and individuals, and the differences inherent in every person's speech (Section 2.2.2). Some of the consequences of speaker variability for the speech signal are discussed in Section 2.2.3). Finally, the sources of the variability are examined in Section 2.2.4.

Having considered the wider issue of speaker characterisation, Chapter 3 looks at the specific variability which arises from the sex of the speaker. In particular there is an attempt to discover the acoustic correlates of this variability, of special importance for the machine recognition and synthesis of speech. The anatomy and physiology of the vocal apparatuses of women and men are the source of the major sex-dependent acoustic contrasts between equivalent utterances, as one would expect from the generally smaller female frame. However, the differences in acoustic-phonetic output between women and men cannot be modelled by a simple scaling of the acoustic speech signal; they instead reflect the nonuniform differences in growth patterns between the sexes, and also the consequences of growing up in a world which places fundamental importance in the adoption and maintenance of sex roles. While socially-conditioned learned behaviour (acculturation) is more obviously evident in the suprasegmental characteristics such as intonation and in the use of language, there is evidence to suggest that acoustic-phonetic factors such as the formant frequencies are subject to sex role-determined behaviour. The chapter begins by stating that, as a result of the speech research community being male-dominated, relatively little research has been carried out into women's speech characteristics. Thus the variability due to speaker sex has been largely neglected in previous research (a statement

which is borne out in the performance of speech recognisers when dealing with women's voices. In fact, the history of the research into speech has been that of male speech, resulting in analysis techniques and theoretical models based on assumptions derived from the male voice.

Section 3.1 looks at some of the sex-specific factors that influence the variability at the acoustic-phonetic level. The emphasis is on the anatomical and physiological differences between women and men in the vocal apparatus, and how they mould the acoustic speech waveform into recognisably female and male voices. The acoustic-phonetic parameters examined are the fundamental frequency, the relative amplitude of the first harmonic, and the formant frequencies (and which are the three parameters investigated in the research carried out for this thesis – see Chapter 4). These are generally considered in the literature to be the most important acoustic-phonetic cues to speaker sex. The general differences in the biology of the vocal apparatus of women and men, and the effects on the values of the three acoustic-phonetic parameters, are examined in Sections 3.1.1, 3.1.2 and 3.1.3. Looked at another way, Sections 3.1.1 and 3.1.2 look at the influence of the laryngeal vocal tract on speaker sex variability, Section 3.1.3 the influence of the supralaryngeal vocal tract. The influence of age, height and weight on the acoustic-phonetic parameters is looked at in Section 3.1.4. Finally, Section 3.2 considers the perception of speaker sex for further clues to the sources of variation. It is shown that the perception of speaker sex is a relatively easy task. It is also shown that the perceptual cues are present in all parts of the acoustic speech signal, with the result that the percept of speaker sex is extremely robust. It is therefore suggested that in order to affect the satisfactory characterisation of female and male speech this must be taken into account.

The research carried out for this thesis is presented in Appendix B, a description of a method for the acoustic-phonetic analysis of large speech databases, and in Chapter 4, the results of a large-scale study into the acoustic-phonetic markers of speaker sex. Briefly, the speech data and the signal processing techniques used to analyse it are described in Section 4.1, the procedure developed to carry out the investigation is described in Appendix B, the results of the analysis are presented in Section 4.2 and are discussed in Section 4.3, and the conclusions are presented in Section 4.4.

Section 4.1 describes the method used in the analysis of the data on the TIMIT database. The segments of speech used as input data to the analysis are described in Section 4.1.1. The data comprised all instances of the vowel phones /aa/, /ae/, /ao/, /iy/, /uw/ and /ux/ contained on the TIMIT database, a total of nearly 16,000 speech segments. This section discusses the reasons why these particular vowels were chosen, the establishment of the core data set, and gives some statistics on the lengths of the speech segments and the number uttered per speaker. Section 4.1.2 describes the signal processing techniques used to analyse the input speech data and measure the values of the acoustic-phonetic parameters. Briefly, the fundamental frequency was measured using cepstral analysis; the relative amplitude of the first harmonic by locating the first two harmonics and comparing their amplitudes; and the formant frequencies using an algorithm developed by the Centre for Speech Technology Research (CSTR) at Edinburgh University. Finally, the source of the speech data, the TIMIT CD-ROM, is described in full in Section 4.1.3. This consists of an outline of the structure of the database held on the CD-ROM, in particular the directory structure it uses to organise the speech waveform files; a description of the TIMIT notation used to label the phones, and adopted in this thesis; and the information it provides about the extralinguistic attributes of the speakers (e.g. age, height). There is also a discussion of the limitations imposed on this study by the type of speech data available on the database.

The results of the study are presented in Section 4.2 for the fundamental frequency, the relative amplitude of the first harmonic and the formant frequencies. Where possible the results are compared with the relevant data from the literature. The results are reported in the following form: an analysis of the overall data, an analysis of the data by phone, and an analysis of the data by speaker variable (i.e. by age, dialect, etc.). Particular attention is paid to the distribution of the mean and the range of values produced by each speaker to facilitate the analysis of between- and within-speaker variability in Section 4.3. Also discussed in this section are the sex-differentiating potentials of the three acoustic-phonetic measures, and the effects of different speaker variables on the values of the measures. The conclusions are presented in Section 4.4, and includes remarks on speaker characterisation in general, on the characterisation of speaker sex, and on the automatic analysis of speech databases.

The use of UNIX in the analysis is described in Appendix A. Apart from the signal processing programs, which were written in C, nearly all of the software used in the extraction and analysis of the speech data was written in the form of UNIX shell scripts. Particular use was made of the pattern matching command **grep** and the pattern matching and processing language **awk**. This section describes the functions of the various UNIX commands, and outlines how they were incorporated into the shell scripts.

Finally, the automated analysis procedure is described in Appendix B. Four main stages in the analysis procedure were identified: the preparation of a database of input speech (the requirements for which were fulfilled by the TIMIT database); the establishment of structures to control the analysis of the data (Sections B.1 to B.2); the formation of a database of analysed speech, including the extraction of data from the input database, and the signal processing of the data to measure its frequency characteristics (Sections B.3 to B.6); and the statistical analysis of the output database (Section B.7). A guide to each of the sections follows.

Section B.1 describes the naming protocols, i.e. the protocols for the naming of speech slices, files and directories and the definition of data structures. These ensure the consistent structuring and storage of the input and output data. Section B.2 describes the directory structure for the organisation of data input and output, and the establishment of the files of labels pointing to the speech slices on the TIMIT database. Section B.3 describes how the speech slices were extracted from the sentence files on the CD-ROM. Section B.4 describes the signal processing of the speech slices, or more specifically, the implementation of the signal processing algorithms. Section B.5 describes how the frame-by-frame results from the signal processing are passed through a simple statistical analysis to produce the data on the frequency characteristics of each slice. Section B.6 describes the exhaustive checking of the slice statistics, to ensure the signal processing programs performed as intended, and to seek out any unusual results (which may come from, for example, unusual articulations of the vowels). This amounts to an evaluation of the algorithms designed to measure the fundamental frequency, relative amplitude of the first harmonic and formant frequencies. Finally, Section B.7 describes the procedures for the full statistical analysis of the databases of slice and speaker frequency characteristics. The procedures allowed for analysis of within- and between speaker variability, and of the effects of speaker attributes and phonetic context.

Chapter 2

An Introduction to Speaker Variability

2.1 Introduction

Walk into any room, and, providing everyone is speaking your language, you will typically experience no problems understanding what they are saying. The human speech interpretation systems exhibit a remarkable ability to adapt to speakers with vastly different speech characteristics, enabling us to understand the utterances of thickly-accented Scousers, squeaky-voiced children and pipe-smoking drunkards alike - although some of these may prove more difficult than others. Speech communication is evidently an inherently variable process.

There are many contributing factors to the sources of this variation in the speech signal, encompassing aspects of each speaker's anatomy, physiology, culture, personality and the situation s/he is in. The variability may take the form of differences between groups of people, for example between accent groups such as natives of Scotland and Pakistan; between individuals, such as differences in the length of the vocal tract; and within a single person because of a change in mood, or simply because an amount of time has elapsed between one utterance and another. The remainder of this introduction will explore some of the cognitive issues involved in the variability of the speech communication process to place speaker variability in context.

2.1.1 A unified cognitive model of speech communication

It is perhaps useful to consider a hypothetical situation in which we require a population sample to relate the same message. In this way, we can control the linguistic content of the utterance and we do not have to deal with any grammatical or semantic differences between the speakers[1]. Thus we can focus on the issue of speaker variability without considering the use of language.

For our hypothetical group of speakers, let us assume that at some higher level the message is represented internally in the same way. Before the acoustic signal leaves each persons'

[1]Unfortunately, this analogy imposes the constraint of having this group of speakers use exactly the same words. In reality, an important source of variability between speakers (and often, as will be seen, within speakers) is the way dialect will cause some words, and indeed word structures, to change. However, this analogy is adequate for the purposes of introducing the topic.

lips, the utterance passes through a system of cognitive and physical processing, which at the various stages imposes a series of changes upon this representation. In other words. from the creation of the message within our brains to the signal that is sent from our lips. each speaker imparts a unique voice quality upon the acoustic speech waveform. Through this process, no two waveforms representing the 'same' utterance are ever truly the same.

Laver (1988) places this aspect of the speech communication process within the framework of a unified cognitive model, a system consisting of a 'chain' of cognitive, biological and physical sub-parts linking the speaker's and listener's minds. Once integrated, the sub-systems control "the overall process of spoken communication ... from the ideational creation of the message to be transmitted, through the neurolinguistic, neuromuscular and neurosensory mechanisms of the speaker, through the acoustic characteristics of the transmission phase, to the sensory, perceptual and interpretive mechanisms exploited by the listener to reach an understanding of the message transmitted" (p83-4). In other words, only by considering the system as a whole, by recognising the varied contributions of the different sub-systems, can speech behaviour be fully modelled.

2.1.2 Some concepts in the characterisation of speakers

The speech signal is not simply a carrier of the bald linguistic message uttered by a speaker. It also contains information about the speaker which s/he intends to convey, such as a feeling of anger, as well as information which is merely a consequence of the speech act - for instance, cues to the speaker's sex and geographical origins.

Lyons (1977:33) drew a distinction between two aspects of the speech signal, dependent upon the speaker's intent. A signal is *informative* "if (regardless of the intentions of the sender) it makes the receiver aware of something of which he was not previously aware"; while a signal is *communicative* "if it is intended by the sender to make the receiver aware of something of which he was not previously aware". Of course, it is not possible to determine whether a signal, or some part of it, is communicative or informative merely by inspection.

We can further distinguish the speech signal by identifying the types of information it carries (Laver & Trudgill 1979:6):

1. **Linguistic information.** This may be considered as the 'message' the speaker is sending in terms of the speech sounds s/he articulates. The form of the linguistic information is therefore the semantic and grammatical units which are structured to produce meaningful utterances. Thus it is solely communicative in nature.

2. **Paralinguistic information.** This is a psychological marker of the speaker's attitude or mood, and is generally known as the 'tone' of a voice (e.g. the use of whispered speech to convey a conspiratorial intent, or the use of smiling to reassure the listener). It tends to exploit features with a relatively long time-base (i.e. whole phrase or utterance). While not 'linguistic' in the sense that it is not constructed from sequential units, it does carry a communicative message in speech.

3. **Extralinguistic information.** Extralinguistic features in speech are those habitual qualities of the voice that serve as markers of the speaker's physical and social identity. They are solely informative. They may be contained in such long-term parameters as average fundamental frequency and formant bandwidths (Künzel 1989:117) caused by the physiological makeup of her/his vocal apparatus, and such short-term

Signalling function	informative		informative and communicative	
Relation to language	extralinguistic voice characteristics		paralinguistic 'tone of voice'	phonetic realizations of linguistic units
Temporal perspective	permanent	quasi-permanent	medium-term	short-term
Vocal variables	vocal features deriving from anatomical differences between individuals influencing both quality and dynamic aspects	voice settings, i.e. habitual muscular adjustments of the vocal apparatus, including voice quality settings and voice dynamic settings	'tone of voice' achieved by temporary use of voice settings, including paralinguistic quality settings and paralinguistic dynamic settings	momentary articulatory realizations of phonological units, including short-term manipulations of phonetic quality features and short-term manipulations of phonetic dynamic features
Marking function	physical markers	social and psychological markers		
Potential controllability	uncontrollable, therefore unlearnable	under potential muscular control, therefore learnable and imitatable		

Figure 2.1: The relationship between local variables and their marking functions. From Laver & Trudgill (1979:8).

parameters as the idiosyncratic and linguistically irrelevant inflections of a person's accent.

All three types of information are combined in the speech waveform, exploiting many of the same phonetic features, and differing mainly in their time-base (Laver 1988:87). Their functions and interrelations, as defined by Laver & Trudgill (1979), are summarised in Figure 2.1. The paralinguistic and extralinguistic factors have the effect of distorting the acoustic speech waveform containing the linguistic communication. For example, the paralinguistic act of smiling alters the waveform through the spreading of the lips; while a longer term, extralinguistic habitual nasalisation will introduce antiresonances into the formant patterns. To recover the linguistic message, the listener must perform some sort of perceptual manipulation of the waveform, at the same time acknowledging the speaker's mood, or sex or accent if relevant to the discourse taking place.

If we imagine a multidimensional space that encompasses all speakers and their attributes, and comprises each of the acoustic, perceptual and physiological domains, then we can consider each that speaker occupies their own subspace. While a person's anatomy will define the boundaries of the subspace an individual can in theory occupy, they will habitually occupy a certain portion of that subspace defined by their culture and personality. This is known as the speaker type. The current location of the speaker within their own potential volume is determined by speaker state (the emotional and physical condition of the speaker at that particular time) and speaker style (the societal code of convention which controls the use of careful (formal) and casual (informal) speech).

2.1.3 Deciphering the linguistic message in the speech signal

Despite the immense variability of speech and the fact that the acoustic signal contains much more non-linguistic information than is necessary to comprehend the meaning it carries, we generally do not have difficulty in recognising a word spoken to us, or in deciphering the meaning of the message in an utterance. The above classifications of speech highlight both this excess of information in terms of the linguistic content, and also its richness regarding its capacity for providing a description of the speaker. Fry (1979:129) considers that there are two factors mainly responsible for our ability to understand speech. The first is *redundancy*, the ability to more or less predict the course followed by an utterance from our knowledge of the language, the speaker and the situation. This enables the listener to discard much of the information irrelevant to the understanding of the speaker's message, and instead concentrate on those aspects pertinent to the discourse taking place. This also has consequences for the speaker, allowing her or him to reduce vowels to make them more like the neutral schwa, and even to delete them. The listener has a powerful store of phonological, grammatical, semantic and pragmatic knowledge upon which to draw in order to make up for these speaker-imposed deficiencies (Koopmans-van Beinum & van Bergem 1989:285). This can, for instance, allow the listener to replace words that are semantically illogical either because s/he has misheard them or because the speaker has mispronounced them.

The second factor is the ability to select certain *acoustic cues* or features from the mass of information contained within the acoustic speech signal. Thus, if we receive a telephone call from a person in, say, an agitated and excited state, essentially babbling a question at us, we are able firstly to extract the appropriate cues to recognise both the caller and the caller's emotional state, thereby setting the context in which the utterance is taking place, and then to concentrate our attention upon those cues required to understand what the speaker is trying to say.

Given this variability in the realisation of a particular sound, it is clear that the acoustic cues used to make decisions about the identity of sounds must depend upon the *relations* between physical quantities, and not their *absolute* values (Fry 1979:130)[2]. For example, consider the use of formant frequencies as a cue for vowel recognition: despite the fact that a child's formant frequencies are much higher than those of a man's (see Peterson & Barney 1952), we have no difficulty in categorising the vowel sounds as the 'same'. We have a substantial capacity for discarding a large proportion of this information, instead concentrating our perceptive mechanism on a few acoustic cues. And yet speech researchers have consistently failed to pinpoint the intricacies involved in human speech perception. Remez & Rubin (1990:313) describe two "critical factors about speech" which they say lie at the heart of this failure: there does not seem to be a "core set of acoustic cues" for the use of the perceptual system (see Liberman & Cooper 1972), and the inherent variability of the speech waveform "does not appear to indicate a normal set of acoustic events about which variation occurs" (see Bailey & Summerfield 1980).

The problem of coarticulation and the overlaying of cues

In identifying a particular sound, we will generally use several acoustic cues (Fry 1979:130). Given that /p/, /b/, /w/ and /m/ are all bilabial consonants, we would not be able to distinguish between all of them on the basis of a single piece of acoustic information.

[2] At the acoustic level, the sound waves reaching our ears contain frequencies from 30 to 10000Hz and have intensity variations over a range of 30 dB or more, with those frequency and intensity levels changing many times a second (Fry 1979:129).

Although, with the redundancy present in a message, a single acoustic cue may suffice to make this distinction. However, the particular cues, or the combinations in which they are used, may well be different in different situations; and different speakers may well use different cues. There is, says Fry (p130), "substantial evidence that English listeners, for instance, do not use all the same cues for a given distinction", although there is a lot of consistency between the speakers of a language with regard to the cues they do use. As Fry (p130) points out, "the only necessary qualification for an acoustic cue is that it should enable the listener to ... recognise correctly the word that has been spoken."

Research into the perception of speech has shown that individual speech sounds, and therefore the acoustic cues that point to the identity of those sounds, are 'squashed together' during the production of speech, a process Liberman *et al.* (1967) termed *encoding.* This has the effect of blurring the boundaries of individual segments so that they run into each other. For example, consider the production of the word 'bat'. The articulators and vocal tract are first positioned to produce the characteristic labial burst for /b/, but are then moved towards the positions that would be necessary to elicit an isolated, steady-state /ae/. This target is never reached in normal speech however, because the speaker has already begun to move towards the articulatory configuration for /t/. Thus, to use Lieberman & Blumstein's (1988:145) term, a 'composite' sound has been formed, into which the individual speech sounds comprising the /b/, /ae/ and /t/ have been merged. While it is possible to identify the parts of the acoustic signal relating to each phoneme through the use of acoustic cues, and even to isolate the segment corresponding to the vowel, it is impossible to separate the /b/ or the /t/ without including at least some portion of the vowel.

The consonants can in fact be shown to be modulations of the formant pattern of the vowel (Lieberman & Blumstein 1988:145), manifesting themselves as formant transitions. This can be seen in Figure 2.2, which shows synthetic spectrograms of voiced stops followed by various vowels (see Delattre *et al.* 1955). Both the frequency and dynamics of the second formant are entirely dependent upon the following vowel, particularly so for the /dV/ context. Similarly, Liberman (1970) experimented with two-formant approximations to the syllables /d iy/ and /d uw/. Heard in isolation, the F_2 transitions corresponding to the /d/ were reported by listeners to be a rising (for the transition spliced from /d iy/) or a falling (from /d uw/) frequency modulation.

The problem of inconsistency in the production of speech sounds

Considering the second of Remez & Rubins' 'critical factors', there appears to be a lack of consistency across speakers as to the articulatory strategies they use to produce phonetically the same results (Nolan 1983:55). This is because: "In speech, production of the individual phonemes which compose words requires the coordination of intricately-timed laryngeal and supralaryngeal gestures" (Monsen & Engebrettson 1977:981). A number of studies bear this out. Harshman *et al.* (1977) used factor analysis of vocal tract cross-sectional area over a set of vowels to show that different speakers use different proportions of the two principal 'movement' factors (possibly as a result of anatomical differences, they suggest). Perkell's (1979) study used direct palatography to show there is considerable variation across subjects in tongue-palate contact for particular vowels. Riordan (1977) prevented subjects from rounding their lips, but found that they nevertheless lowered their formants by lowering their larynxes. Delattre (1967) used x-ray and spectrographic evidence to show that some American English speakers achieve a retroflex voice quality, not by raising the tongue tip but by bunching the tongue. Perkell (1979:375) suggests that his results, along with those of others, show "that each individual does what is necessary

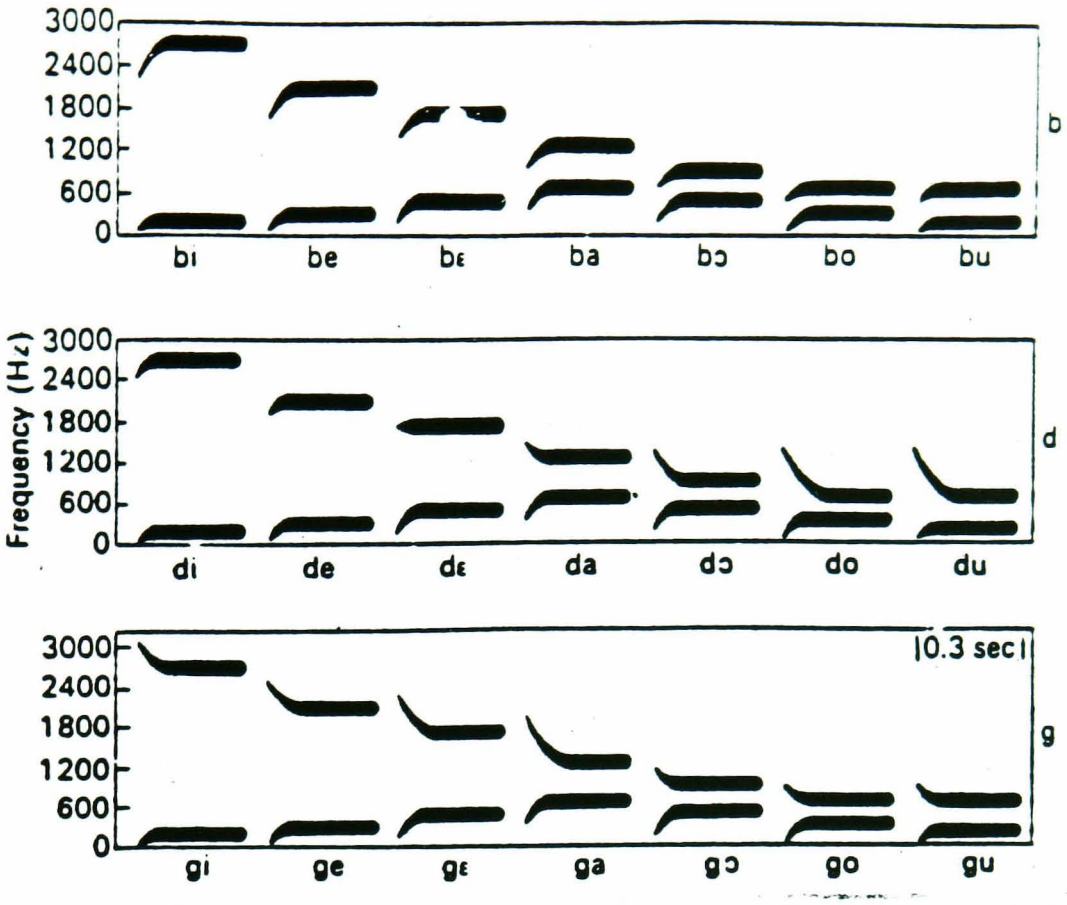


Figure 2.2: Synthetic spectrograms of voiced plosives followed by vowels using only F_1 and F_2 information. From Lieberman & Blumstein, 1988:144, after Delattre *et al.* (1955).

to produce an appropriate acoustic output."

As well as gross articulatory deviations, the movements of the articulators, particularly in fast speech, are never precisely time-locked, "even though their co-ordination has to satisfy some necessary and demanding acoustic and auditory temporal constraints" (Laver 1988:93). For example, in the production of a nasalised vowel, the timing of the movements of the soft palate with respect to the tongue can be observed to differ. Thus, in pursuit of an auditory goal, the speaker is free to implement the articulators as best s/he can.

2.2 A review of the sources of speaker variability

This section begins with a look at the categorisation and description of speaker variability sources (Section 2.2.1), and continues with a look at how they can be used to classify the between-speaker differences across groups and individuals. and the differences inherent in every person's speech (Section 2.2.2). Some of the consequences of speaker variability for the speech signal are discussed in Section 2.2.3. Finally, the sources of the variability are examined in Section 2.2.4.

2.2.1 Describing speaker variability

Scherer & Giles (1979:xii) coined the term 'speech markers' as a way of categorising the variability speakers exhibit. They intended this descriptor to be applied in a general sense to mean those extralinguistic, paralinguistic and linguistic "speech cues that potentially provide the receiver with information concerning the sender's biological, psychological and social characteristics." This allows us to assign particular speech cues as being of a particular type.

In the same publication, Laver & Trudgill (1979:3) report on a typology proposed by Abercrombie (1967:7-9) which aims to classify speech types. This typology proposes a hierarchy allowing people to be characterised according to differences between groups of speakers, between individual speakers, and within the speech of a single speaker. They posit that within speech there are: *group markers*, indicating membership of a group (e.g. a regional or social group); *individuating markers*, characterising the individual (e.g. a 'breathy' voice, or a lisp); and *affective markers*, characterising the changing states of the speaker. This final category was further defined by Lyons (1977:108) to include "any information in a ... signal which indicates to the receiver that the sender is in a particular state, whether this be an emotional state (fear, anger, sexual arousal or readiness, etc.), a state of health (suffering from laryngitis, etc.), a state of intoxication, or whatever."

Most other descriptions of speaker variability (e.g. Vaissière 1985:202-3, Laver 1988:107, Shevchenko 1989:131) tend to follow this categorisation, adding to their lists what are basically subcategories of the above. For the sake of clarity, it would seem appropriate to adopt the descriptors outlined by Laver & Trudgill: they allow us to classify speakers according to the between-speaker differences across groups and individuals, and to the differences inherent in every person's speech.

One further, useful descriptor of voice features is the use of the terms *intrinsic* versus *extrinsic*. They allow us to differentiate between those features outside of the speaker's control and those features within it, whether voluntary or not. Laver (1976:57) describes them as follows: "Intrinsic features ... derive solely from the invariant absolutely uncontrollable physical foundation of the speaker's vocal apparatus ... extrinsic features are made up of all aspects of vocal activity which are under the volitional control of the speaker, whether 'consciously' or not."

2.2.2 Between-speaker and within-speaker variability

The listener brings to bear a wealth of linguistic, paralinguistic and extralinguistic knowledge about the speech waveform to compensate for the relatively impoverished nature of the linguistic data it contains. "That distinct acoustic effects can be heard as the 'same sound' (e.g. the 'same vowel' spoken by a man and a woman) has long been a challenge to

perceptual physics" (Nolan 1983:198). However, by treating the production of speech as simply the result of highly trained gymnastic routines performed by the vocal apparatus, we obscure the way in which a speaker exploits her/his vocal apparatus for a variety of communication uses. Bolt *et al.* (1979:15) refer to "the lack of a mature scientific discipline for analysing speech in terms that characterise the speaker". Within-speaker differences, with few exceptions, have been seen as noise to be filtered out in order to obtain the 'true' speech data, i.e. the linguistic message.

Work on the theory and practice of recognising the differences between speakers is now increasing and is clearly of relevance in modelling human speech perception and production (see, for example, Hollien & McGlone 1976:39). That objectively different acoustic signals from two speakers can be normalised and perceived as the same by a listener, and that individuals may use different articulatory strategies to produce equivalent auditory effects, are just two of the phenomena which are worthy of more research. "The human perceptual system copes with ... inter-speaker and intra-speaker variability very well indeed: in the speech technology world, variation within a single speaker is itself a major part of the speech recognition task. Adding the acoustic variation that is found between speakers to the task of recognising spoken messages from continuous speech at least doubles the difficulty of the task, even when no change of accent-community is involved. It is almost certain that the problem of successful recognition of continuous speech from a single speaker using a large vocabulary will be solved well before the solution is found to the problem of adequate automatic adaptation to new speakers of different accents" (Laver 1988:100).

Sources of between-speaker variability

"The acoustical signal contains not only information about the linguistic content of the speech, but reflects also the unique properties of the speaker's vocal apparatus and conveys information about the speaker himself (his age, sex, identity)" (Vaissière 1985:202). Laver (1988:107) identifies four main types (or layers) of extra- and paralinguistic attributes present in the speech signal which contribute to the differences between speakers. These are:

- **Physical** - differences due to a person's sex, age and individual anatomy
- **Social** - socioeconomic, ethnic and regional contributions to accent
- **Psychological** - differences due the speaker's personality
- **Situational** - changes in the speech patterns due to variations in mood, attitude, etc.

The consequences of between-speaker differences manifest themselves in the same dimensional acoustic space in a complex interaction that affects the perceived speech signal in a multitude of ways. Thus any attempt to classify a group of speakers into particular representative groups - say along the lines of sex or accent - is fraught with difficulties. The complex interactions of the between-speaker variations can result in a significant deviation from, for example, the average range of formant frequencies associated with a group. These deviations from the group norm must "be taken into account if reference values are to allow optimal classification" (W Barry *et al.* 1989:356). For example, consider the use of vocal tract size to differentiate women, men and children. According to W Barry *et al.* (1989:356): "What is rarely considered in the light of such large *group* differences is

the range of vocal tract variation *within* the group." Thus the classification of a sample of speakers into groups "can ... be significantly affected unless within-group adjustment is undertaken." To complicate matters further, all the factors affecting between-speaker variability "exist simultaneously, and to an unknown degree in the signal, and adaptation to any one of these presupposes prior correction for the [others]" (W Barry *et al.* 1989:356).

The sources of between-speaker differences presented above are examined more fully in Section 2.2.4.

Sources of within-speaker variability

Speakers are unable to reproduce precisely the same speech sound, even when speaking in the same conditions. "[H]uman speech is inherently variable - a given speaker will never produce the same sound in exactly the same way" (Fallside 1985:49). In fact, the consistent perfect reproduction of speech sounds is not required of the speaker for the listener to comprehend what is being said.

Nolan (1983) demonstrated that speaker-dependent information is not encoded in the absolute values of particular parameters, but is instead incorporated into a range of values. "The way a speaker speaks on a given occasion is the result of a complex interaction between his communicative intent, the language mechanisms he controls and the context in which he is speaking" (Nolan 1983). In other words, the particular parameter values used by the speaker on a particular occasion are entirely dependent upon the situation, and different situations will yield different values.

There are a number of factors affecting the consistency of the acoustic speech waveform from a single speaker. Laver (1988:93-6) identifies four sources of variability:

1. The movements of the articulators relative to one another are never precisely time-locked. "[I]n the case of speech, fluent motor execution and a high event rate may depend on overlapping movements rather than synchronisation of movements ... Synchronous patterning may be a default principle that is overridden by phonetic and motor learning to yield the highly-overlapping patterns that characterise rapid, fluent motor execution" (Kent 1983:70). This may well be due to the complex mechanisms involved in their anatomy and manipulation.

2. The coarticulatory effect on a given phonological unit varies according to its phonological context. This is governed by biomechanical factors (inertia, elasticity, muscle-geometry) constraining the movements of the articulators and therefore the production of a particular sound in a particular phonological context; and phonetic factors such as accent governing the pronunciation of a given phonological prime in a particular sequence. For example, on some occasions a speaker will release a word final stop, and on others will not (Doddington & Schalk 1981:29).

3. Changes in speaking rate affect the speech patterns in the signal. For a speaking rate increase, this can result in such phonological changes as vowels and/or consonants being reduced or deleted. Speakers compensate when speaking faster by either decreasing articulator movements (and thereby undershooting phonological targets) or by increasing articulator velocity. Changes to the speaking rate cause nonconstant changes in the acoustic structure of the speech signal. For example, Miller *et al.* (1984 – cited Miller & Volaitis 1989:505) found that interviewees on a radio chat show "changed speaking rate frequently and substantially." 29 of the 30 speakers

tested changed the average duration of syllables by as much as 100ms, while 20 of the 30 changed it by as much as 300ms.

4. Choice of speaking style, i.e. formal or informal. This has implications at the phonological level (Laver 1988:96). The more formal the speech style or speaking situation, the fewer deletions, reductions and coarticulations occur due to the speaker increasing her/his articulatory effort (van Bergem & Koopmans-van Beinum 1989:285). As an illustration, English words have on average two or three reorganised forms for use in informal speech (or, for that matter, in speeded-up speech). For example, 'actually', spoken in a formal situation by an upper class English person, may be pronounced /ae k t y uw er l iy/. However, faced with increasing informality, the pronunciation may lead through /ae k ch uw er l iy/, /ae k ch uw l iy/, /ae k ch er l iy/, /ae k ch l iy/ and /ae k sh l iy/ to possibly the most informal derivative /ae sh l iy/ (Laver 1988:95).

In addition to these is the (possibly) more minor effect of *voice perturbation*, fluctuations in the speech waveform due to an inability to maintain a steady speech sound (see page 30 for a more thorough treatment). However, there is evidence that the magnitude of these perturbations in people without pathological vocal fold disorders is small (Milenkovic 1987:529).

An emotional state in the speaker will produce characteristics patterns of articulatory and respiratory movements which play a part in modifying the acoustic speech signal (see page 32 for more details). Temporary physiological effects such as colds and drunken speech will add to the variation imposed upon the speech signal. Other sources of variability include increased vocal effort, resulting in an increase in energy in the spectrum (Thomas *et al.* 1989), and idiosyncrasies of speech such as stuttering, lisping, slurring and speaker-generated noises such as lip smacks, heavy breath-intake, 'um's and 'er's (Pallett 1985:373).

Factors affecting pronunciation and their consequences for ASR systems

M Cooper (1987; see also reviews by Doddington & Schalk 1981, and Pallett 1985) identified some of the particular problems of within-speaker differences to automatic speech recogniser applications. This has obvious practical applications in that recognisers are intended for use in the workplace away from controlled laboratory conditions. Cooper notes that: "Current speech recognisers rely on consistent pronunciation by the user, [whereas] people differ considerably in their ability to speak in a consistent manner" (p85)[3]. In the real world, a number of factors combine to cause significant within-speaker variability in performance[4]:

- **Ambient noise.** This impairs the operation of the recogniser and affects the manner in which people speak. For example, half the people studied by Rollins (1985) spoke very differently when hearing loud industrial noise over headphones than under quiet conditions. She also found that if the application was for noisy and quiet environments, then training utterances obtained under noisy conditions were better for recognition.

- **Stress on speaker.** For example, Reed (1985) reported studies by Poock in which he found that 'moderate' stress applied to subjects in the laboratory caused recognition accuracy to drop by 23%. See also Hecker *et al.* (1968).

[3] In speech recogniser terms, people who are consistent in their speech and therefore cause the recogniser to perform well are known as 'sheep', while those whose speech is variable between utterance repetitions are known as 'goats' (Doddington & Schalk 1981).

[4] For more experimental results, see also Loven & Collins (1988).

- **Physical work.** If a person's workload is sufficient to affect breathing, then it is likely to change her/his manner of speaking (M Cooper 1987:84). For example, Visick *et al.* (1984) reported that the physical work of parcel sorting affected recognition accuracy. They suggested this could have been due to movement of the headset microphone as well as speech characteristics.

- **Temporal congruence.** This refers to the (in)ability of a speaker to reproduce the same utterance in the same fashion after a period of time has elapsed (Thomas *et al.* 1989:409). This has obvious consequences for the performance of speaker-dependent systems (see Nolan (1983:12) for a discussion of the change in speech patterns over different test sessions, and the consequences for machine recognition). Dew *et al.* (1986) found decreasing recognition rates during test sessions lasting an hour. Studies by Waterworth (1984) and Wilpon & Roberts (1986) found a longer term drifting effect present in speech. The maximum recognition score was obtained immediately after training the system, which declined with the passing of time. As well as considerations pertinent to the speaker, other factors which can alter the speech signal over time include changes in speaking environment (especially when noise is introduced) and microphone placement (Doddington & Schalk 1981:29).

2.2.3 The consequences of speaker variability

Intrinsic voice features exist due to the extralinguistic voice characteristics imbued by the anatomical differences between individuals. These anatomical influences on voice quality - due to such factors as vocal tract length, lip, jaw, tongue, nasal cavity and pharynx dimensions, dental characteristics and laryngeal geometry (Abercrombie 1967:92) - impose absolute limits (static constraints) on the range of the speaker's acoustic output. By their very nature they are permanent and uncontrollable. In addition, dynamic, physiological constraints are placed upon the articulators due to the inability, for example, of the tongue to curl back and touch the soft palate.

However, the anatomy of the vocal apparatus implies only physiologically determined maxima and minima for a speaker's voice features. Within these limitations, speakers will adopt certain voice 'settings', ranging from quasi-permanent extralinguistic features such as accent to more fleeting paralinguistic gestures associated with transitory emotions, such as anger or happiness. These settings are attained through a varied use of muscular action (Laver & Trudgill 1979:14).

The quasi-permanent settings are the result of learned patterns of behaviour, realised as long-term articulatory settings (W Barry *et al.* 1989:356), reflecting the speaker's life experience as a member of a locality, social class, culture and ethnic group. And as such these settings also act as markers, signalling to others the membership of these groupings. They are 'quasi-permanent' in the sense that they are subject to outside influences; for instance, a working class student at university may have her/his regional accent 'watered down' through mixing with the predominantly middle class, RP-speaking student population. Other long-term characteristics are those markers of personality or voice 'type' (e.g. harshness, nasality), the result of a configurational trend in the action of the vocal apparatus.

The widely accepted model of between-speaker variation quoted in the literature consists of two parts (Nolan 1983:26): organic (structural) differences - physiological differences in the vocal apparatus i.e. shape, size, dynamic limitations, etc.; and learned (functional) differences - differences in the manner of speaking from accent, dialect, etc., i.e. stored habit patterns. As Nolan (1983:27) says, this is excessively simplistic and belies the

impact of within-speaker variation. He introduces the notion of the *plasticity* of the vocal tract, the ability of speakers to alter their voice quality by adjusting the settings of the vocal apparatus - for example, raising or lowering the larynx, or applying a greater or lesser degree of tension on the vocal folds. The consequence of this is that the scope for variation in a given set of circumstances (e.g. stress, noise, illness) is considerable, only constrained by the physiological limits imposed by the dimensions of the vocal tract and the vocal folds. To say that a given speaker has an habitual and consistent way of speaking is to ignore "the vast core of knowledge the speaker has about the phonetics and phonology of his language" (Nolan 1983:28) and the ways in which this may be used to in some way alter her/his way of speaking when faced with a given situation. It also ignores the different levels of learning, and subsequently their susceptibility to alteration by the speaker. The complex interactions of the between-speaker variations, combined with those of within-speaker variations, can result in a significant deviation from the average range of resonance frequencies associated with a group. These deviations from the group norm must "be taken into account if reference values are to allow optimal classification" (W Barry *et al.* 1989:356). For example, consider the use of vocal tract size to differentiate women, men and children. According to W Barry *et al.*: "What is rarely considered in the light of such large *group* differences is the range of vocal tract variation *within* the group" (p356). To complicate matters further, all the factors affecting inter-speaker variability "exist simultaneously, and to an unknown degree in the signal, and adaptation to any one of these presupposes prior correction for the [others]" (p356). In other words, the whole issue of speaker variability, encompassing both between- and within-speaker differences, must be taken into account when generalising across groups of speakers.

The above anomaly is highlighted in sociolinguistic studies investigating the use of language in regional and social groups. Labov made the discovery that each speaker has control over a range of speech styles for use in different contexts, e.g. in a formal setting, or in casual conversation: "As far as we can see, there are no single-style speakers" (Labov 1972:208). For many sociolinguistic variables - e.g. the use of /n/ *vs.* /N/ in the *-ing* suffix in New York (Labov 1966) and Norwich (Trudgill 1974a) - a speaker in a given context produces neither 0% nor 100% of the variable. Similarly, there is a range of styles within regional accent groups, where a particular speaker's manifestation of a regional accent may lie anywhere between the full regional form and the standard national form (e.g. between 'Geordie' and RP), and may even contain certain elements of other regional accents. Thus the occurrence of the value of a variable in a speech sample cannot be used to reliably predict its occurrence throughout all of that person's speech in the same context. Labov also found that the sociolinguistic stratification (what social/geographic group you belong to) and stylistic variation (the use of different modes of speech in different contexts) noted above take place along the same dimensions (i.e. with the same variables). Thus "it may therefore be difficult to interpret any signal by itself - to distinguish, for example, a casual salesman from a careful pipefitter" (Labov 1972:240). This is illustrated schematically in Fig. 2.3 adapted from Labov (1972). Each line represents one socioeconomic class. The figure shows how speakers will adopt a higher percentage of a prestige form both because of membership of a 'higher' social grouping and because of an increase in the formality of the context. As a further example, consider Table 2.1, derived from a study by Labov (1972:103) concerning the social significance of post-vocalic r-colouring. In one instance, both speakers, from different social groups, pronounce the same number of *r*'s while talking in different styles.

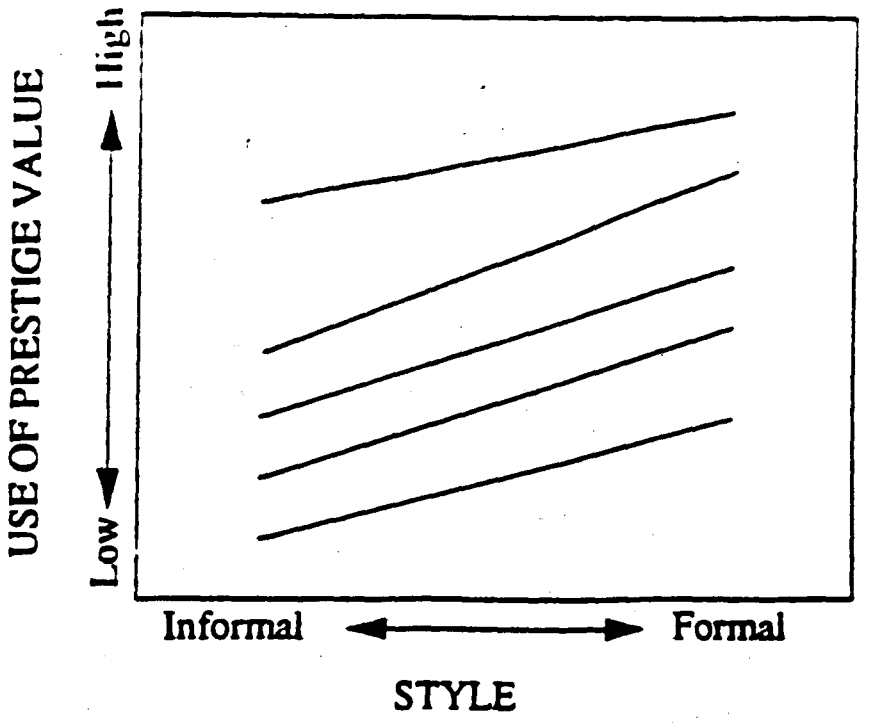


Figure 2.3: Schematic representation of the use of a prestige value of a sociolinguistic variable according to class and role. After Labov (1972).

Class	Casual speech	Careful speech
upper middle	32	47
lower middle	0	31

Table 2.1: Proportion of *-r* pronounced (post-vocalic r-colouring) by two women of different socioeconomic class, in informal (casual speech) and formal (careful speech) settings. After Labov (1972:103).

2.2.4 Sources of speaker variability

This section will aim to gather together the sources of variability exhibited by speakers. In keeping with the generality called for by Abercrombie (1967) for such a typology (see Section 2.2.1), the sources have been classified into five areas - namely accent markers, physical markers, personality markers, situational markers and emotional markers - while still retaining the overall structure defined by Laver & Trudgill (1979). Obviously, this can impose some unwelcome constraints. The physical marker of sex would more comfortably be considered as a grouping factor rather than an individuating one; and the milieu of regional, social, cultural and ethnic group contributions to accent constitute a very individual voice, or idiolect (Wells 1982:1). But considerations of clarity demand a structured approach, and so the few anomalies will be overlooked.

Group markers

Accent

A person's accent is the result of a learned experience (Vaissière 1985:203), embodying the complex interaction of a number of influences, such as area of upbringing, socioeconomic class, and ethnic, educational and cultural background. Diphthongs, glides, r-coloured sounds, tense vowels are particularly susceptible to these influences, especially across dialects (see Holbrook & Fairbanks 1962, Sambur 1975, Goldstein 1976). An important factor in the makeup of an accent is the history of that accent, specifically its development and interaction with other accents (Pallett 1985:373). This can incorporate aspects of parental accent, influences from peer groups, from different social classes, and, for people who have moved into a region, the influence of local accents and dialects.

Regional influences on accent can cause speakers from different localities to use different sound systems. This can mean differences in both the number of phonemes used by the speaker, and in their phonetic realisation (W Barry *et al.* 1989:355). Audibly, these sources of between-speaker variability manifest themselves as differences in phonetic quality, while acoustically they can be detected as differences in the time-frequency-energy patterns of functionally equivalent utterances. There will also be differences in the occurrence of the phonemes in the speaker's lexicon. An accent can cause wide-ranging deviations between accent systems (W Barry *et al.* 1989:355). For example, Southern British English speakers rhyme 'pass' with 'farce', while Northern British English speakers rhyme it with 'gas'; the first vowel in 'leisure' is rhymed with 'lesson' for British English speakers, and 'lesion' for American English speakers. Table 2.2 shows a number of pairs of words for which some accent groups make vowel distinctions. Thus accents do not simply result in a speaker pronouncing particular sounds differently, they also significantly alter the speaker's lexicon (Nolan 1987:476). These differences apply between all major English accents.

The influence of social class results in a range of accents being in existence in any given locality, from RP at the top of the social scale, down through a milder form of the regional accent in the middle classes, to a much broader form at the bottom, perhaps with more dialectal variation (Wells 1982:14). Thus: "A person's social position is reflected in the words and constructions he uses, as well as in the way he pronounces them" (Wells 1982:13). This is generally more applicable in England than elsewhere, although Labov (1966) found that variability in the speech of New Yorkers also depends upon the class of the speaker. The existence of social stratification in speech is amply demonstrated in the data reproduced in Table 2.3, from a study of speakers in Norwich by Trudgill (1974b), where the percentages in each column indicate average deviations from RP.

Word pair	SBE	NBE	Scot	NAE
trap-bath	+	–	–	–
foot-strut	+	–	+	+
foot-goose	+	+	–	+
lot-cloth	–	–	–	(+)
lot-thought	+	+	–	(+)
lot-palm	+	+	+	–
start-palm	–	–	+	+

Table 2.2: Vowel distinctions between gross regional accents (+ = distinction; – = no distinction). SBE = Southern British English, NBE = Northern British English, Scot = Scottish English, NAE = North American English. Note that the brackets in the NAE column indicate that some subgroups do not have such an opposition. From W Barry (1987:468), after Wells (1982).

Class	(ng)	(t)	(h)
Middle middle	31	41	6
Lower middle	42	62	14
Upper working	87	89	40
Middle working	95	92	59
Lower working	100	94	61

Table 2.3: Percentage deviations from RP of residents of Norwich for: (ng) = /n/ *vs.* /ng/ in *-ing*; (t) = /q/ *vs.* /t/, as in *butter* and *bet*; (h) = absence *vs.* presence of /h/, i.e. h-dropping. After Trudgill (1974b:48).

The ethnic background of speakers will also play a role in shaping their speech. For instance, Hudson & Holbrook (1981) showed that both female and male young black US Americans have on average a lower fundamental frequency than equivalent whites, even if the difference is relatively small (6.5Hz for males, 23.9Hz for females).

Individuating markers

Physical

The anatomy of a person's vocal apparatus, and the effects upon this of sex and age, all have characteristic effects on the speech signal.

The length of vocal tract is the main determinant for the dispersion of the resonance frequencies (Vaissière 1985:202): the shorter the vocal tract, the higher the resonance frequencies of the speaker (Peterson & Barney 1952). Formant frequencies for female vocal tracts are about 15% higher than for male vocal tracts, while those of children are higher still; while the formant bandwidths for women's vocal tracts are about 50% wider (Vaissière 1985:202). Adult males, adult females and children appear to have proportionately different pharyngeal and oral cavities (Fant 1966). Together with the wide range of sizes between speakers of the same sex, this results in variant distributions of the formant frequencies.

One source of variability often omitted is that of the nasal tract. The physical makeup of the nasal tract affects nasal resonances and thus affects the acoustic characteristics of nasal and nasalised sounds (Vaissière 1985:203).

One obvious difference between voices is the fundamental frequency of vibration of the vocal folds, perceived by listeners as pitch. According to Johnson (1990b:230): "It is well documented that hearers are influenced by F_0 when they make judgements about vowel quality (Miller 1953, Fujisaki & Kawashima 1968, Slawson 1968, Ainsworth 1975, Traunmüller 1981)". Therefore the physical makeup of the vocal folds will affect the perceived voice quality. According to the myoelastic aerodynamic theory of phonation, the frequency of vocal fold vibration (i.e. the fundamental frequency) is governed by the elasticity, tension and mass of the vocal folds[5] (Borden & Harris 1984). The more elastic the folds, the higher the fundamental because they will tend to bounce back to their steady-state quicker. A greater amount of tension in the folds will also raise their fundamental frequency. Increased tension is primarily achieved by stretching the folds through laryngeal muscular action. And the more massive (longer and thicker) the folds, the lower the fundamental. Titze (1989b:901) adds that F_0 is also controlled by the subglottal pressure[6].

There are other, more physiological, factors which influence the use of the vocal folds. The modes of vibration of the vocal folds are known as phonation types. They may be described as a deviation from the modal, or normal, type (see Nolan (1983:140) for a description of modal phonation). The different types may be achieved by adjusting the separation between the arytenoid cartilages (Klatt & Klatt 1990:821). There are two main phonation types which can affect voice quality, depending upon whether the vocal folds are being loosened (for breathy voice) or tightened (for pressed voice)[7]. With *breathy voice* the folds are not adducted (closed) fully, but are close enough to be vibrated causing continuously released air to accompany the sound wave. The glottal source spectrum becomes richer in low frequency energy, and noise may be created at the glottis during voiced sounds (Vaissière 1985:203). M Barry (1986:2) identifies three characteristics of breathy voice: increased spectral tilt, greater energy in the fundamental, and the presence of noise in the spectrum. *Laryngealised* (or *pressed*) *voice* is achieved by tightening the larynx (using the arytenoid cartilages) to permit only a small volume velocity pulse through the glottis. It is usually accompanied by a lowering of F_0 (Klatt & Klatt 1990:821). The glottal source spectrum becomes richer in high frequency energy (Vaissière 1985:203). An extreme form of laryngealisation is *creaky voice* (or *vocal fry*), comprised of extremely low vibrations of the vocal folds (Borden & Harris 1984).

It is an accepted fact that during voicing the vocal folds produce a waveform that is not entirely periodic. As far back as 1927, Simon concluded that "there are no tones of constant pitch in either vocal or instrumental sounds" (Simon 1927:83 – cited Orlikoff & Baken 1990:31); while Titze *et al.* (1987:252) say: "Even the most serious attempt by a speaker to produce steady phonation with constant pitch, loudness and quality results in perturbations in F_0, amplitude and waveshape of the speech signal". These perturbations are fluctuations in the speech waveform due to an inability of the speaker to maintain a steady speech sound. There are three main types of voice perturbation (Hillenbrand 1987:448, Titze *et al.* 1987:252): jitter (fundamental frequency perturbations) - rapid, and generally relatively small, cycle-to-cycle variations in F_0 (see Higgins & Saxman (1989) for a more thorough treatment); shimmer (amplitude perturbations) - analogous cycle-to-cycle

[5]Note that each of these factors will have the effect of working against each other. For example, the longer the folds are, the greater will be their mass, which will tend to lower F_0; conversely, the longer folds will also increase the tension across them, leading to a higher F_0.

[6]This is the primary variable in the control of vocal intensity. F_0 and intensity are not controlled independently of each other: a speaker will tend to raise the pitch of her/his voice when talking more loudly (Titze 1989b:901).

[7]Klatt & Klatt mention glottal stop as a further phonation type, which which may be considered as an even more extreme form of laryngealisation than creaky voice.

variations in amplitude; and perturbations due to additive noise - turbulent airflow through the glottis during phonation. Experiments in the synthesis of speech have shown that jitter seems to be an essential cue for natural voice quality (Lieberman 1967:36). When speech is synthesised using only a smooth glottal waveform, listeners report a "mechanical" quality to the speech (F Cooper *et al.* 1957, Kersta *et al.* 1960). A high degree of jitter, probably combined with shimmer, accounts for 'harsh' voice, "probably caused by extreme adductive tension and medial compression [in the vocal folds]" (Nolan 1983:141).

Pathological disorders of the vocal folds will also have an effect on the output of the glottal source, causing distortions to the frequency of vibration and affecting the quasi-periodicity of the glottal waveform. There may be irregularities along the edges of the folds due to contact ulcers (from too much tension focused in the larynx), nodules (from overuse of the voice) or swellings and irritation during colds with laryngitis. This can cause the voice to sound hoarse. Permanent damage to the vocal folds requires the listener to bring about compensatory action in order to produce vibrations. In the event of the paralysis of one of the folds (or the partial paralysis of both of them), one vocal fold can sometimes be trained to move more than halfway to meet the paralysed one. Part or all of the larynx may have been removed due to cancer, requiring the speaker to learn how to vibrate other tissues and muscle masses, such as scar tissue or the cricopharyngeous muscle (Borden & Harris 1984). Related to this is another potential source of speaker variability, that of the initiation of phonation. Good use of the voice requires that phonation be initiated from lightly adducted folds. However, some speakers begin voicing with a glottal attack. The folds are tightly adducted prior to vibration, followed by a plosive burst at the glottis. This is an affliction suffered by people making heavy use of their voices e.g. singers, teachers (Borden & Harris 1984). Also, by varying the timing of the onset of voicing relative to supraglottal articulatory movements, prevoiced, voiceless-unaspirated and voiceless-aspirated consonants can be achieved (Klatt & Klatt 1990:821). One of the effects of old age on the vocal folds is that the muscles atrophy, causing the glottis to have a bowed appearance (Luchsinger & Arnold 1985). An excess of mucus causing inefficient phonation combined with the necessity of the exertion of greater effort to bring the folds together results in a harsh, whispery voice (Laver & Trudgill 1979:10).

Long term health problems, such as respiratory diseases due to cigarette smoking, will have characteristic effects on the speech signal. For example, Gilbert & Weismer (1974) found significant differences between the fundamental frequencies of smokers (163.8Hz) and nonsmokers (182.8Hz). Subsequent laryngeal examination discovered vocal fold thickening in 87% of smokers as compared to 7% of nonsmokers. W Barry *et al.* (1989:356) comments on the reduced jaw opening exhibited by inveterate pipe smokers.

Personality

A person may have an emotional disposition which causes her/him to speak in a particular manner e.g. the quiet-voiced, shy person. This involves habitual settings of the vocal tract and larynx (Laver & Trudgill 1979:14). Other manifestations of this are the voice types such as 'nasalised' voice, where the soft palate is habitually lowered during speech, and palatalised voice, in which the tongue has a tendency to be raised towards the roof of the mouth (see Laver (1975) for an account of such voice settings).

Affective markers

Situational

A speaker will change the characteristics of her/his speech in accordance with the context s/he is in (or more precisely, the context the speaker *thinks* s/he is in) (Nolan 1983:36).

Increasing formality tends to cause the speaker (in many urban communities at least) to change the values of sociolinguistic variables in the direction of those people of higher status (Labov 1972, Shevchenko 1989). The more formal the speech style or speaking situation, the fewer deletions, reductions and coarticulations occur, due to the speaker increasing her/his articulatory effort (van Bergem & Koopmans-van Beinum 1989:285). For example, Shevchenko (1989) found that: "The degree of deviation from RP norms ... increases in spoken dialogue as compared to reading" (p132).

In other words, more care is taken with pronunciation in formal speech (note that this has obvious correlations with speeded-up speech). As an illustration, English words have on average two or three reorganised forms for use in informal speech (consider, for example, the different realisations of 'actually' on page 24). Shevchenko (1989:133) says that the context of increased formality leads to a higher maximum F_0, wider F_0 range, greater intensity, longer tone groups and shorter pauses. In addition to the above, a speaker may 'converge' or 'diverge' her/his speech to make it more or less like that of the addressee(s) depending upon her/his interpretation of the relative status of the addressee(s) (see Giles *et al.* 1979). "Increasingly it is being observed that sociolinguistic markers are not invariant, but depend on (the participants' interpretation of) social aspects of the interactional context (see e.g. Brown & Levinson 1979)" (Nolan 1983:36).

A number of studies have looked at the effects of temporary physical and working conditions on the variability of speech, with particular reference to the performance of automatic speech recognisers. These studies have shown the detrimental effect on recognition performance of both ambient noise and the imposition of a physical workload on the speaker (see page 24 for a more detailed analysis). Similarly, extreme fatigue can cause the mode of phonation to become inefficient, resulting in a whispery voice, or a weak, breathy voice (Laver & Trudgill 1979:13).

Scherer (1981a) noted that the effects of stress on speakers are increased F_0, a stronger energy concentration between 500 and 1000Hz, and possibly an increase in intensity. Testing the effects of stress, as induced by random electric shocks, on automatic speaker recognition, Hollien & Majewski (1977) found the machine identification rates dropped from 100% for normal speech to 92% under stress. In a similar experiment, Doherty & Hollien (1978) found an even more marked reduction in identification, from 100% to 72%. See also the effects of stress imposed on speakers on the performance of automatic speech recognition systems discussed on page 24.

A number of researchers have noted that: "Many speakers find it difficult to say the same word in a similar fashion after a period of time has elapsed" (Thomas *et al.* 1989:409). This lack of temporal congruence, the inability of a speaker to reproduce the same utterance precisely, is present even in the short term (e.g. Dew *et al.* 1986). Using an automatic speaker recognition system, Nolan (1983) found that from one set of test stimuli recorded at the same time as a reference, and one recorded three months previously, mean identification rates dropped dramatically from 53.6% to 34.6% for /l/ and from 61.1% to 36.3% for /r/. See also page 25.

Emotional

As well as specific vocalisations, a change in emotional state in the speaker in response to particular situations (e.g. anger, joy, fear, surprise) will produce definite patterns of articulatory and respiratory movements which modify the acoustic speech signal (for reviews see Scherer 1981b, Williams & Stevens 1981). Responses to emotional situations produce characteristic differences in the F_0 contour, average spectrum, temporal characteristics, precision of articulation and waveform irregularities in the glottal pulse train (see

Williams & Stevens (1972); see also Tartter (1980) for the effects of mood changes, particularly smiling). However, as Scherer (1981b:201) says: "It has been difficult for speech scientists and psychologists to study the effects of emotion on speech empirically ... In spite of the pervasiveness of human emotions, they are very difficult to capture by means of objective research ... [and it is] difficult to produce emotions in the laboratory since they are the emergency responses of the organism." Thus the number of studies on the acoustic correlates of emotional speech has been limited, often involving simulations by actors. However, what results there are show surprising consistency. For example, both happiness and anger will induce a higher pitch and larger pitch variation, greater loudness and faster tempo, while grief or sadness will have the opposite effects.

Chapter 3

The Variability in the Voice due to Speaker Sex

> "Women and children have been somewhat neglected groups in the history of speech analysis by machine." (Klatt & Klatt 1990:820)
>
> "Much of our knowledge about speech production comes from studies on male speakers. In the area of speech acoustics, a half-century of contributions by engineers (mostly male) has led to substantial theories about sound generation, propagation, and resonation in the vocal tract. Assumptions and simplifications in these theories were often based on speech samples derived from the investigator's own voice, or the voices of his colleagues. One wonders, for example, if the source-filter theory of speech production would have taken the same course of development if female voices had been the primary model early on." (Titze 1989a:1699)
>
> "The major part of phonetic studies has been based on research concerning male voices. Female voices came somewhat more into the picture rather recently." (Tielen 1989a:127)
>
> "It is a commonplace of speech synthesis to observe that the field is markedly biased towards the voice of the adult male, and that attempts to synthesise the female voice have met with little real success." (M Barry 1986:1)

As these citations illustrate, there has been a paucity of research into the female side of the speech analysis equation. In all aspects of speech research, most of the work has been based around the male vocal apparatus. Much of the pioneering work on the acoustic theory of speech production dealt exclusively with the male vocal apparatus (Laver 1988:99, Titze 1989a:1699).

This preoccupation with the vocal apparatus of (adult) males may prove to be a major block to the advancement of the speech production model. As Laver (1988:99) says, "a major problem that speech technology has inherited from this theoretical work [on the source-filter model of speech production] is that a model of speech acoustics has been worked out in great detail only for an idealised vocal apparatus", that of a 'typical' male speaker with a vocal tract approximately 17cm long. As Titze says above, it is interesting to speculate how speech science might have developed had the source-filter theory of speech production been modelled on female speakers.

Summing up, "the concentration on an idealised vocal apparatus has left us relatively ignorant about the acoustic differences between speakers which are due to anatomy" (Laver

1988:99). This chapter attempts to redress the balance by examining what literature there is on variation due to speaker sex, although principally at the acoustic-phonetic level. The first section, Section 3.1, concentrates on the variation at the acoustic-phonetic level due to the anatomy and physiology of the vocal apparatus, and reviews the published data on the fundamental frequencies, the relative amplitudes of the first harmonic and the formant frequencies of women and men. The perception of speaker sex is examined in Section 3.2. Important information about the cues to speaker sex perception can be gained from looking at speaker sex identification studies, and the high rates of identification using different experimental stimuli. Perception of the sex of children is examined in Section 3.2.1, and of adults in Section 3.2.2.

3.1 Variation at the acoustic-phonetic level

The major differences in the speech signal between the sexes are due to the anatomical and physiological structuring of the vocal apparatus. This is to be expected, firstly because of the readily-observable average size differences between women and men, and secondly because it is the vocal apparatus which imparts acoustic identity onto the speech signal. from the strength of airflow from the lungs, to the movements of the vocal folds, to the structure and positioning of the resonating chambers of the vocal tract. What follows is an review of the literature concerning these differences, with an attempt to relate the biological differences to their acoustic consequences.

The first three parts to this section (3.1.1, 3.1.2 and 3.1.3) examine the variability in the three acoustic-phonetic measures investigated in the analytic part of this thesis (see Chapter 4), namely the fundamental frequency, the relative amplitude of the first harmonic, and the formant frequencies. These parameters have been considered the most important acoustic-phonetic cues to speaker sex in the literature, and therefore offer a sound starting point from which to examine the extent of speaker sex variability. Each part examines the relevant anatomical and physiological aspects of the vocal apparatus (3.1.1 and 3.1.2 the laryngeal tract, and 3.1.3 the supralaryngeal tract), and then reviews the published data on the values of the parameters.

Section 3.1.1 considers the gross anatomy of the vocal folds and Section 3.1.2 considers vocal fold dynamics in more detail[1]: differing vocal fold vibration patterns, the acoustic waveform emanating from the glottis, and the acoustic correlates of breathy voice, in particular the relative amplitude of the first harmonic. Section 3.1.1 covers the female and male group averages for F_0, and more importantly, the range of F_0 for groups and individuals. Section 3.1.2 reports on those studies measuring the relative amplitude of the first harmonic in the frequency spectrum. Section 3.1.3 deals with the biology of the supralaryngeal vocal tract: its overall length, and the ratios between the pharynx and the oral cavity. It then covers the published data on the formant frequencies of women and men for different vowels. Finally, Section 3.1.3 also looks at the difficulties to be encountered in the measurement of female formant frequencies.

Section 3.1.4 looks at how the vocal apparatus, and the larynx in particular, is affected by aging and height and weight. The anatomical growth during childhood of the laryngeal and supralaryngeal tracts, the differences, if any, between girls and boys, and the effects on F_0 and the formant frequencies is dealt with in the first part. This is particularly important to the issues raised in the review of the child sex perception literature (see Section 3.2.1). The survey of F_0 studies covered in Section 3.1.1 is expanded upon in the second part to this section to consider how F_0 changes with age in adulthood. Finally, the few studies examining the relation between a person's height and weight and their fundamental frequency are looked at.

[1]Note that this is a simplified description of the voice source, but one which is nevertheless suitable for the purposes of this thesis.

3.1.1 The voice source and the fundamental frequency

The difference in average speaking fundamental frequency (SFF)[2] between women and men is often quoted as being about one octave (Vaissière 1985:202, Daniloff *et al.* 1980). Luchsinger & Arnold (1965:95,99) even talk about an octave phenomenon, implying some significant auditory/perceptual link between this sexual dimorphism and the pitch scale. However, "analysis and synthesis of female ... speech involves more than a mere scaling of [male] fundamental frequency" (Titze 1989a:1699). This is illustrated by studies of the average SFF of talkers, and will be investigated below.

The biology of the voice source

First, let us consider the anatomy of the vocal folds, and their relation to F_0. Luchsinger & Arnold (1965:99) consider it "an elementary fact that the dimensions of the larynx, particularly of the vocal folds, determine the vocal range of each individual." Thus a large larynx with long, broad vocal cords will allow a range of low tones, while a small larynx with short, thin vocal cords will allow a range of high tones. The vocal folds of adult females are smaller in both mass and length (Lieberman & Blumstein 1988:100) and are thinner in cross-section (Daniloff *et al.* 1980:203) than those of adult males. However, the female voice source is not simply a scaled-down version of the male. Reviewing the literature on the morphological differences between the female and male larynx, Titze (1989a) arrived at the following average size differences: The anterior-posterior dimension (roughly, the length) of the male thyroid cartilage is approximately 20% longer than the female; the membranous length of the male vocal folds is approximately 16mm, compared to 10mm for the female, a difference of 60%; the thickness of the vocal folds is approximately 20-30% greater than the female[3]. The subsequent effect of these differences on the vocal folds' vibratory properties causes significant differences in the acoustic output of the voice source (M Barry 1986:1)[4]. One of the results of this combination of differences is that the power of women's and men's voices is similar, in that while the more massive male voice source has a greater strength, the female source vibrates at higher frequencies (Titze 1989a:1706). This can be illustrated by drawing a comparison with woofers and tweeters in loudspeaker design: the woofer has a greater cone surface and so drives more air, but the tweeter radiates at higher frequencies, therefore their output power can be the same (Titze 1989a:1706). RO Coleman *et al.* (1977) found similar absolute ranges of sound pressure level in twelve female and ten male subjects.

However, the length and thickness of the vocal folds are not the only influences on a person's level of fundamental frequency at any particular moment: sub-glottal pressure and tension in the muscles of the larynx are also used. Graddol & Swann (1983:352) suggest that the use of laryngeal muscular tension may be even *more* important than the length and thickness of the vocal folds in determining the range of a person's SFF. As a general rule, an individual's potential fundamental frequency range is around two to three octaves. In measuring the vocal ranges of Germans, Schilling (1929 – cited Luchsinger & Arnold 1965:100) found 25% had a range of 1.7-2.0 octaves, 50% had ranges from 2.0-2.5 octaves, and 15% had a range greater than 2.5 octaves. RO Coleman *et al.* (1977:200)

[2]The term *speaking fundamental frequency* is used to emphasise the fact that we are referring to the range of F_0 a speaker uses in normal speech, rather than the range the speaker is physiologically capable of.

[3]The measurements were made on the excised larynges of human cadavers (see the studies by Kahane (1978) and Hirano (1983)).

[4]See Titze (1989a) for a more detailed analysis, some of which will be discussed in greater depth in Section 3.1.2.

Source	Lang. spoken	n	Age (years)	SFF (SD) (Hz)	Range (Hz)	Range (st)
Lass *et al.* (1976)	US Engl.	10	19-24	224 (29)	186-385	12.6
Hollien *et al.* (1982)	US Engl.	37	18-26	204 –	–	–
Graddol & Swann (1983)	RP	15	26-39	153 –	126-183	6.5
Deem *et al.* (1991)	US Engl.	30	18-30	223 –	185-247	5.0

Table 3.1: Survey of studies of female speakers reporting mean SFF values for isolated vowels and vowels in a CVC context. Note: (1) Range values are given in Hertz and semi-tones.

found a range of approximately three octaves for both female and male speakers, and they cite a study (Hollien, Dew & Phillips 1971) with almost identical results. Individual ranges for their twenty-two speakers went from 2.4-3.7 octaves for the males and 2.5-3.5 for the females. However, the usual range of fundamental frequency produced by a speaker lies more closely around their average SFF, and is located at the lower end of the range they are physically capable of. This is optimal and least fatiguing according to voice experts, and corresponds to the neutral rest position of the larynx (Luchsinger & Arnold 1965:100). Graddol & Swann (1983:352) positted three physical restrictions on a person's range of SFF, as opposed to the range they are physically capable of: certain frequencies may be difficult and uncomfortable to produce; some of the person's vocal mechanisms may be more efficient, and therefore less likely to suffer from long-term damage, than others; acoustically, certain frequencies will be more efficient than others due to the structure of the vocal cavities, and will therefore carry more effectively.

Mean SFF

The most obvious difference between women's and men's voice sources arising from the biological differences is that women have higher fundamental frequencies, as the heavier and longer male vocal folds vibrate more slowly (Daniloff *et al.* 1980:203). This is reflected in the results reported in the literature, and reproduced in Tables 3.1, 3.2, 3.3 and 3.4, from studies investigating the average SFF of female and male speakers. Considering particularly the larger sample studies (whose results are summarised in Tables 3.5 and 3.6), average SFF for women appears to be in the region 190-220Hz, and 110-130Hz for men, suggesting female average SFF is 50-100% greater. If this difference seems rather vague, then it merely reflects the wide variety of experimental procedures under which these studies were conducted, and, indirectly, the nonuniformity of individual F_0 performance within apparently homogeneous speaker groups.

There are a number of experimental factors which render comparisons between these studies difficult. These include the choice of stimulus, the number and type of subjects, and how the data is reported. This theme of disparate experimental procedures will be returned to in subsequent sections, but a number of points are worth making here to try to make sense of these data. The type of utterance spoken by subjects is obviously of much significance, as this will determine the range of F_0 used by each person, and can subsequently affect the average SFF. For this reason the data reported for each sex here has been split into data derived from vowels (see Table 3.1 for the women's vowel results, 3.3 for the men's) and from connected speech (see Table 3.2 for the women's connected speech results, 3.4 for the men's). The vowel data comes from either a CVC context, or from sustained productions of vowels in isolation. The types of isolated vowels range from

Source	Lang. spoken	n	Age (years)	SFF (SD) (Hz)	Range (Hz)	Range (st)
Snidecor (1951)	US Engl.	8	Adult	213 –	–	*12.5
		8	Adult	213 –	–	*10.0
McGlone & Hollien (1963)	US Engl.	10	65-69	196 –	–	*9.4
		10	80-95	199 –	–	*8.6
Hanley & Snidecor (1967)	US Engl.	8	21	204 –	–	2.7
Saxman & Burke (1967)	US Engl.	9	30-39	196 –	–	4.5
		9	40-49	189 –	–	3.7
Fitch & Holbrook (1970)	US Engl.	100	17-25	217 –	165-255	7.5
Linke (1973)	US Engl.	27	adult	200 –	–	*9.3
Gilbert & Weismer (1974)	US Engl.	15	30-54	183 –	–	4.7
		15	30-54	164 –	–	6.5
Hudson & Holbrook (1981)	US Engl.	100	18-29	193 (19)	139-266	11.1
Stoicheff (1981)	US Engl.	21	20-29	224 –	192-275	6.2
		18	30-39	213 –	181-241	4.9
		21	40-49	221 –	190-273	6.3
		17	50-59	199 –	176-241	5.4
		15	60-69	200 –	143-235	8.6
		19	70-83	202 –	170-249	6.6
de Pinto & Hollien (1982)	Australian	11	52-60	180 –	–	3.4
Graddol & Swann (1983)	RP	15	26-39	203 –	174-233	5.1
	RP	15	26-39	208 –	188-238	4.1
Krook (1988)	–	35	20-29	196 –	–	–
	–	100	30-39	195 –	–	–
	–	83	40-49	191 –	–	–
	Swedish	83	50-59	182 –	–	–
	–	85	60-69	181 –	–	–
	–	63	70-79	188 –	–	–
	–	11	80-89	188 –	–	–
Künzel (1989)	German	78	19-61	211 (17)	–	–
Henton (1990)	US Engl.	5	25-37	180 (25)	156-215	*7.8
		5	25-37	182 (24)	157-216	*7.2

Table 3.2: Survey of studies of female speakers reporting mean SFF values for connected speech. Note: (1) Range values are given in Hertz and semi-tones; (2) * indicates a 90% range.

Source	Lang. spoken	n	Age (years)	SFF (SD) (Hz)		Range (Hz)	Range (st)
Provonost (1942)	US Engl.	6	18	132	–	122-143	2.8
Mysak (1959)+	US Engl.	15	32-62	113	–	–	*9.5
		12	65-79	124	–	–	*10.0
		12	80-92	141	–	–	*11.2
Hollien & Jackson (1973)	US Engl.	157	adult	129	–	93-178	11.3
Lass *et al.* (1976)	US Engl.	10	19-24	111	(14)	88-146.	8.8
Wilcox & Horii (1980)	US Engl.	20	18-26	124	–	96-141	6.6
		20	60-80	122	–	90-135	7.0
Horii (1982)	US Engl.	20	22-37	127	–	–	–
Graddol & Swann (1983)	RP	12	26-39	86	–	68-111	8.5
Deem *et al.* (1991)	US Engl.	30	18-30	120	–	89-147	8.6

Table 3.3: Survey of studies of male speakers reporting mean SFF values for isolated vowels and vowels in a CVC context. Note: (1) Range values are given in Hertz and semi-tones; (2) * indicates a 90% range; (3) + indicates median scores of 110, 125, 142 respectively.

Source	Lang. spoken	n	Age (years)	SFF (SD) (Hz)		Range (Hz)	Range (st)
Mysak (1959)*	US Engl.	15	32-62	113	–	–	*9.4
		12	65-79	124	–	–	*9.6
		12	80-92	141	–	–	*11.4
Fitch & Holbrook (1970)	US Engl.	100	17-25	117	–	85-155	10.4
Hollien & Jackson (1973)+	US Engl.	157	adult	129	–	91-165	7.5
Hudson & Holbrook (1981)	US Engl.	100	18-29	110	(16)	82-159	11.4
Graddol & Swann (1983)	RP	12	26-39	114	–	98-136	5.7
		12	26-39	122	–	111-136	3.5
Künzel (1989)	German	105	19-61	116	(17)	–	–
Henton (1990)	US Engl.	5	25-37	113	(13)	93-125	*8.3
		5	25-37	116	(14)	96-132	*7.1

Table 3.4: Survey of studies of male speakers reporting mean SFF values for connected speech. Note: (1) Range values are given in Hertz and semi-tones; (2) * indicates a 90% range; (3) + indicates median scores of 110, 125, 142 respectively.

Source	Lang. spoken	n	Age (years)	SFF (SD) (Hz)	
Fitch & Holbrook (1970)	US Engl.	100	17-25	217	–
Hudson & Holbrook (1981)	US Engl.	100	18-29	193	(19)
Stoicheff (1981)	US Engl.	111	20-83	211	–
Krook (1988)	Swedish	460	20-89	188	–
Künzel (1989)	German	78	19-61	211	(17)

Table 3.5: Summary of large-sample ($\geq$ 50) studies of female speaker's mean SFF values for connected speech.

Source	Lang. spoken	n	Age (years)	SFF (SD) (Hz)
Fitch & Holbrook (1970)	US Engl.	100	17-25	117 –
Hollien & Jackson (1973)	US Engl.	157	adult	129 –
Hudson & Holbrook (1981)	US Engl.	100	18-29	110 (16)
Künzel (1989)	German	105	19-61	116 (17)

Table 3.6: Summary of large-sample ($\geq$ 50) studies of male speaker's mean SFF values for connected speech.

/aa/ phonated at the speaker's lowest possible pitch (Graddol & Swann 1983) to 5sec productions of eight different vowels (Horii 1982). The connected speech measurements were taken solely from read passages (as opposed to spontaneous speech). The passages read by the subjects were either taken from books or other printed media, or were specially constructed to force the subjects to use a range of phonemes and/or fundamental frequency. The effect of choice of test utterance on F_0 and F_0 range is immediately apparent: researchers generally ask their subjects to produce a vowel, especially an isolated vowel, with 'a level tone', i.e. at a steady F_0; a read passage can affect the F_0 attained by its length (longer sections of speech allow more scope for intonational dynamism) and its content (which might be solely declarative, or might contain, for example, question inflections). A number of the U.S. studies reported here (Benjamin 1986; Fitch & Holbrook 1970; Hudson & Holbrook 1981; Linke 1973; Saxman & Burke 1967; Stoicheff 1981) used the so-called Rainbow Passage, a phonetically-balanced declarative piece of prose (Fairbank 1960:127; see Nolan (1983:167) for a full transcription of the passage). However, even this source of between-study similarity is not as clear-cut as it seems: for example, Linke, Stoicheff and Saxman & Burke instructed their subjects to read the passage as though to an audience of 25 people, while Benjamin had her subjects read it in a conversational manner. A number of studies try to control their groups for particular factors, further adding to the heterogeneity of these studies. Stoicheff's study looked at the difference in F_0 across age groups of female non-smokers. Thus her group of subjects consisted of women who had never smoked or had not done so for 15 years, and were also from a variety of occupations (e.g. secretaries, housewives, nurses, students), although with no indication of socioeconomic class. Subjects were rejected if they had any formal voice training. About Linke's group we are told only that they are young adult students who were chosen to represent a range of effectiveness of voice usage. Similarly, Graddol & Swanns' (1983) group all worked in academia and were aged 26-39, but no more information was given.

It is tempting then to consider only those studies which directly compare female and male SFF from controlled groups, and preferably those which measured a large number of subjects to allow for a better sampling of the intra-sex range of the speakers' mean SFFs. Too many studies report the average SFF for the sex as if it applies to all women or all men, and in all speaking conditions. One of the most important conclusions to be drawn from a cross-study comparison, is that people are very varied in their own average SFF. More detailed information must be drawn from individual studies, especially those which seek a comparison across controlled groups which may allow us to draw conclusions about the effect of particular speaker attributes, e.g. Stoicheff's (1981) study of female non-smokers, Künzel's (1989) study of the relation between SFF and height and weight, Hudson & Holbrooks' (1981) study of young African Americans. The effects of age, height and weight will be discussed in more detail in Section 3.1.4.

Considering then the female-male direct-comparison studies, in particular the large sample

studies by Fitch & Holbrook (1970), Hudson & Holbrook (1981) and Künzel (1989), we can revise our estimate of the difference in average SFF between women and men: The average SFF of female speakers is between 193-217Hz, and the average SFF of male speakers is between 110-117Hz, suggesting the female average is 75-85% greater than the male.

Range of SFF

One oft-repeated claim is that women have greater SFF ranges than men. Henton (1990) challenges this claim on the grounds that such views "may have been based on incorrect interpretations of the experimental measurements" (p300). The claim is centred on the fact that measurements were made using the linear hertz scale. These do indeed show a wider *fundamental frequency* range for women. This, says Henton, is misleading when the ear judges F_0 range "not by measuring hertz, but by using a logarithmic, or non-linear scale" (p301). Pitch, the perceptual correlate of fundamental frequency, is perceived on a logarithmic scale related to the structure of the basilar membrane in the ear (Smith 1985:58). Such a view is also put forward by Graddol (1986), who argues: "Whenever intervals in pitch must be compared at different frequencies, a logarithmic scale is to be preferred" (p228). While absolute measurements of F_0 less than 1kHz tend to linearity (Traunmüller 1981), the important factor in characterising F_0 (and intonation) patterns, says Henton, is the F_0 interval, i.e. a rise, a fall, or a broad or narrow range. Thus she argues for converting the measurements of F_0 range on the linear hertz scale to "an *interval-preserving* F_0 scale using the logarithmic semitonal scale" (p302), using the relationship

$$semitones = 39.68 \log_{10} \frac{F_{high}}{F_{low}}$$

where F_{high} and F_{low} are the highest and lowest values (Hz) respectively in the range. Hudson & Holbrook (1981) took a similar stance in reporting the data for their study of young black female and male adults. They reported their F_0 ranges in octaves (where an octave consists of five tones and two diatonic semitones, i.e. 12 semitones[5]), and noted that: "The frequency difference expressed in hertz represents a linear progression whereas the ear responds to this information in a nonlinear fashion" (p199).

Following Henton (1990:303), the F_0 ranges reported in Tables 3.1, 3.2, 3.3 and 3.4 are shown using both the Hertz and semitone scales[6]. Nolan (1983) has indicated that at least 40 seconds of continuous speech are necessary to obtain reliable measurements of fundamental frequency parameters – i.e. to capture the full range of the speaker's fundamental – and so the results reported for the vowels in Tables 3.1 and 3.3 are unlikely to be

[5] A doubling of F_0 results in a perceived increase in vocal pitch of approximately one octave on the musical scale (Luchsinger & Arnold 1985 – cited Smith 1985:58).

[6] Note that it is sometimes unclear whether the ranges reported refer to the range of the average SFFs of individuals, the average range of SFF produced by individuals, or the range of all fundamental frequencies produced. Studies often do not explicitly state the source of their ranges, i.e. whether they are a range of means, a mean range, or the total range of F_0 produced by that sex. However, the range of average SFFs of individuals measured from the TIMIT data (which was also comprised of read speech) was 146-270Hz for the women and 82-183Hz for the men, while the mean range limits (i.e. the mean of the minimum and maximum F_0 produced by each speaker) were 151Hz and 273Hz for the women and 89Hz and 154Hz for the men, which are roughly comparable (see Section 4.2.1). They are also roughly comparable to the range values cited in the literature (and reproduced in Tables 3.2 and 3.4, for read speech). Thus whatever the source of their ranges, the figures serve as a rough guide to both the range of mean SFFs and the mean SFF range. The latter is certainly true for the studies of Henton (1990), Hudson & Holbrook (1981), Linke (1973) and Snidecor (1951). Furthermore, the lowest and highest F_0s measured from the TIMIT data were 63Hz and 384Hz for the women and 40Hz and 277Hz for the men, implying much greater ranges than those reported in the literature.

representative of a person's SFF range. This is, in general, borne out by the data, which show substantially larger ranges for the read passages.

It should be noted that a number of authors report the full F_0 range of their subjects. Henton (1990:304) argues that such data should be treated with caution as they fail to eliminate what she terms "rogue extreme values" which may not be representative of the population. For example, consider the difference reported by Linke (1973:179) between the total range of 23.3 semitones and the 90% range of 9.3 semitones. Also, consider the upper range of Lass *et al.*'s (1976) data for women. A fundamental frequency of 385Hz seems rather large, but as they published no individual figures then we cannot judge whether this reflects the true upper range of women's SFF, whether it was one speaker with a very high SFF, or even if it is a misprint. However, a number of the studies measured range from a 90% dispersal around the mean, and this is indicated by an asterisk in the final columns of the tables.

Henton (1990:304) points out: "The lack of accessible data for females and males from a well-controlled population is a familiar handicap in studying sex-linked phonetic behaviour." For instance, Peterson & Barneys' (1952) study involved a large spread of subjects from across the states. So making comparative judgements across heterogeneous groups is generally unavoidable. Bearing this in mind, the studies in which women and men were recorded under the same conditions, are obviously of the most interest. What the data show is that while the fundamental frequency range for women on the hertz scale is greater than that for men, using the semitonality scale it is, if anything, less. While not providing irrefutable evidence for such a statement, bearing in mind the variability between the groups of speakers, it does at least highlight a significant trend. And so it would appear that using a metric more closely related to our perceptions of the fundamental frequency, there is in fact little difference between the range of F_0 used by women and by men. Henton (1990:307) concludes, "it is possible to state that on the basis of this evidence females do not employ a greater pitch range in English." Considering all of the studies, it is also apparent that speakers in general employ a wide range of F_0, and that this is not confined to women.

There is an apparent inconsistency between the data reviewed by Henton (and repeated in Tables 3.1 to 3.4) and her own results, which involved speakers reading two passages. The semitonal ranges for the other studies are calculated from the range of the population sample, and so reflect the F_0 range used by the *entire sex*. On the other hand, the semitonal ranges for her data come from an average of the ranges of each subject. Is it then the case that on a person by person basis women *do* have greater F_0 ranges than men? Fortunately Henton (1990:311) gives us the data for each individual, and this is reproduced in Table 3.7. Referring to this table, what is apparent is the wide variety of F_0 ranges across both sexes with, on the basis of the results of this study at least[7], men having the greater 'range of ranges': from speaker JB's 3.91 semitone spread for his nasal passage to speaker OG's non-nasal 13.52.

The data also highlight precisely the point Henton made about 'outlier' extreme values. Speaker OG's range of 13.52 semitones for the non-nasal passage is vastly different from those of the other men in the same speaking condition. Excluding his result causes the mean of the ranges to fall from 8.30 semitones to 6.99. Similar, though less major, effects can be seen in the other sets of data in Table 3.7. Such anomalies are only to be expected in samples of this size, but at least the data do indicate that F_0 ranges, when converted

[7]Henton's experiment is lent credence by the careful selection of subjects for "homogeneity with regard to age, accent, race, stature, socio-economic and educational levels" (p305), something which cannot be guaranteed in the other studies.

Speaker	Sex	1st passage		2nd passage	
		Av. F_0 (Hz)	Range (st)	Av. F_0 (Hz)	Range (st)
CB	f	161.9	7.11	164.9	4.48
HC	f	173.4	6.59	174.4	5.36
AJ	f	155.5	9.45	157.4	7.00
GJ	f	215.1	8.88	216.3	11.19
IB	f	194.8	7.10	195.2	7.72
JB	m	121.4	5.58	122.6	3.91
OG	m	106.0	13.52	107.6	8.81
BG	m	93.2	6.77	95.7	6.22
MJ	m	125.1	7.77	131.7	9.03
KE	m	119.4	7.84	121.2	7.42

Table 3.7: Individual semitonal ranges for women and men reading two passages, calculated from four Standard Deviations around the mean. The first passage contained a high proportion of nasalised vowels, while the second passage contained no adjacent vowel and nasal segments. After Henton (1990).

to a tonality scale that approximates the human auditory system, are fairly well spread across the sexes, and as such support Henton's argument.

Considering now the average range of SFF in Hertz from the large sample studies, and being deliberately approximate, the average range of SFF lies between 150-265Hz for female speakers, and 80-160Hz for male speakers. Note that alot of the smaller sample studies report smaller ranges, indicating they failed to capture a sufficiently varied sample of speakers. More importantly, there is a distinct lack of overlap between the female and male averages, although some overlap is evident in the region 140-165Hz. This would indicate that an individual's average SFF is an extremely good speaker sex discriminater. Furthermore, from the study by Künzel (1989), there appears to be little overlap in the average SFFs of individuals (see Figure 3.1). However, Hudson & Holbrook (1981) state that their 100 male subjects produced F_0s ranging from 55 to 245Hz, and their 100 female subjects from 105 to 345Hz, indicating there is a potentially large overlap in the range of speaking fundamental frequency used by each sex as a group.

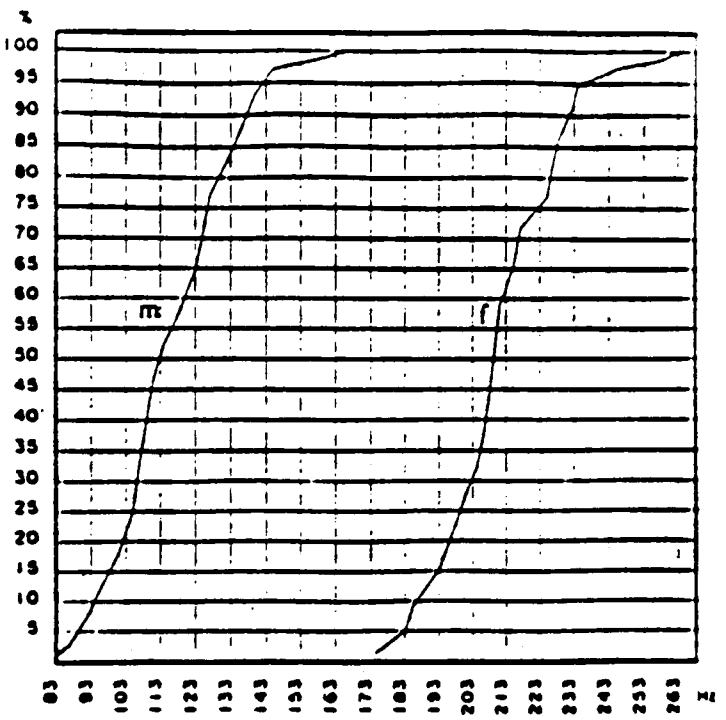


Figure 3.1: Cumulative distribution of average F_0 (Hz) for 78 female (f) and 105 male (m) subjects. From Künzel (1989:121).

3.1.2 The voice source and the relative amplitude of the first harmonic

Introduction

Three factors may be considered as being responsible for variations in the glottal waveform[8]: the speaker's sex, the linguistic context of the speech (Nittrouer *et al.* 1990:774), and the mode of phonation (including the use of a soft or loud voice) (Monsen & Engebretson 1977:991). However, individual speakers exhibit a wide range of variation, encompassing not only the fundamental frequency and RMS intensity, but also the appearance and shape of the glottal wave, the phase spectrum, and the intensity spectrum (Monsen & Engebretson 1977:984, see also p984-5). Finally, average values for speakers show great variation within each sex group (Nittrouer *et al.* 1990; Klatt & Klatt 1990).

Despite the individual variation, general sexual characteristics can be ascribed to the glottal waveform. Examination of the glottal waveforms for female and male speakers reveals a number of differences (see, for example, Monsen & Engebretson 1977:986 and Klatt & Klatt 1990:822-3). The female waveform tends to symmetricality and is quasi-sinusoidal in shape; in other words the opening and closing portions of each period are roughly equal. There is also a constant or nearly constant DC airflow through the glottis. The male waveform is more asymmetrical, with the frequent appearance of a prominent 'hump' in the opening portion, and a sharp corner at the end of the glottal pulse. The closing portion of the period occupies 20-40% of the total, i.e. there is a fairly rapid closure of vocal folds (Monsen & Engebretson 1977:986). Examples of female and male glottal waveforms obtained by Monsen & Engebretson are shown in Figure 3.2, and schematic waveforms are illustrated in the second row of Figure 3.3.

This differential behaviour of the airflow leaving the glottis is a result of the different vocal fold dynamics in the female and male larynx. Several studies have shown that incomplete vocal fold closure in the posterior parts of the glottis is common in women

[8]The glottal waveform, or the volume velocity wave from the glottis, is a representation of the airflow through the glottis. It is distinct from the speech waveform in that the speech waveform has the acoustics of the vocal tract overlaid on top of the glottal waveform.

Figure 3.2: Examples of typical glottal waveforms for normal phonation of /schwa/, from Monsen & Engebretson (1977:986). The waveforms for two male subjects are on the left, and for two female subjects on the right. The waveforms were obtained using a long reflectionless tube, which acts as a pseudoinfinite termination of the vocal tract, thereby significantly reducing the vocal tract's resonances (see p981-4 for a full description of the experimental details).

with normal speech (Biever & Bless (1989); Söderston & Lindestad 1990 – cited Söderston & Hammarberg 1992:23; Bless *et al.* (1986) – cited Söderston & Hammarberg 1992:23 and Klatt & Klatt 1990:826; Söderston & Hammarberg 1992). This was observed in 80% of the female subjects (and 20% of the males) in the Bless *et al.* study, and in all of the female subjects in the Söderston & Hammarberg study, even after formal voice training. The source of this incomplete closure, or glottal chink, are the arytenoid cartilages at the rear of the larynx, which hold open the vocal folds, ensuring the constant airflow. They are however sufficiently approximated to create the muscular tension required for phonation to occur. As a result, vocal fold closure during phonation is relatively gradual, from the front (anterior) to the rear (posterior) of the larynx, and the volume velocity wave from the glottis has a rounded corner at closure (Klatt & Klatt 1990:822).

In male speakers, typical laryngeal behaviour during voicing is for the vocal folds to be nearly approximated, producing the type of phonation known as modal or normal. As they are held more closely together at either end, the vocal folds snap shut during the closing phase of phonation, closing simultaneously along their length. This creates the sharp corner at the end of the glottal pulse in the volume velocity waveform. There is also a suggestion that laryngealisation is used as a marker of maleness (Klatt & Klatt 1990:823). For laryngealised, or 'pressed' voice, the arytenoid cartilages hold the vocal folds shut, allowing only a narrow glottal pulse and a small open quotient[9]. Creak, a more extreme form of laryngealisation, has also been cited as a male speech marker.

It is possible that this differential voice source behaviour is due to differences in anatomy between the female and male larynx. Preliminary data from Hicks (1989 – cited Söderston & Hammarberg 1992:26), from the measurement of the dimensions of cricoid cartilages in 47 female and 43 male larynges, suggested a difference in shape and attitude. The shape

[9]The open quotient refers to the ratio of time the glottis is open to the total duration of the period during phonation (Klatt & Klatt 1990:838).

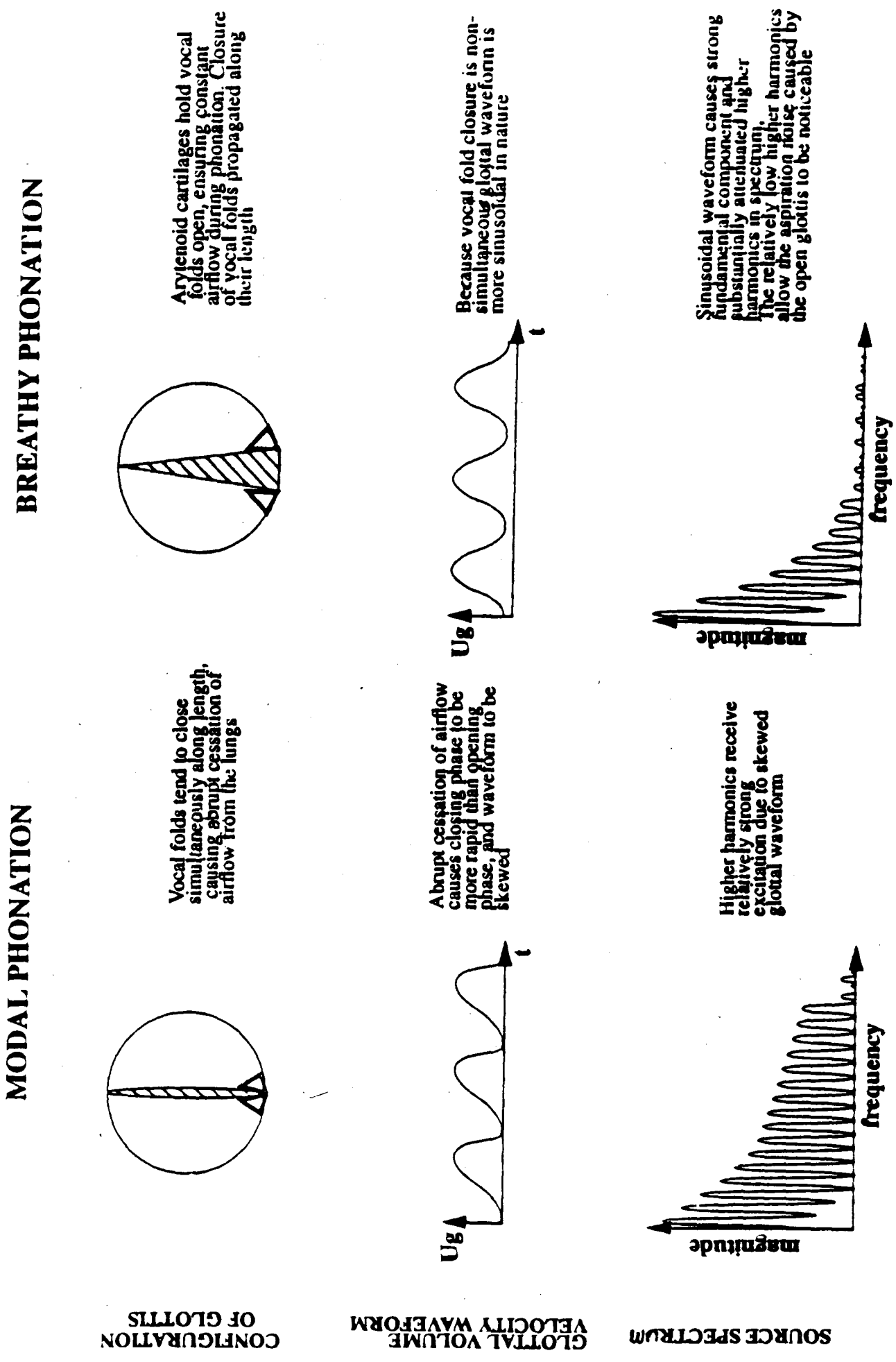


Figure 3.3: Schematic representation of the glottal configurations for modal and breathy phonation. The first row illustrates a cross-section of the larynx at the vocal folds, showing the articulatory configurations of the vocal folds (half circles), arytenoid cartilages (triangles) and glottis (shaded area). The second row shows the volume velocity waveforms for these voice source configurations. The third rows shows the frequency domain representations of the waveforms. Adapted from Klatt & Klatt (1990:822).

of the cricoid cartilage is considered to have an impact on the interarytenoid distance. The male cartilages tended to have an oval shape from front-to-back of the larynx, while the female cartilages tended to be rounded or were oval-shaped from side-to-side. The latter cartilage shape creates a bigger interarytenoid distance to be 'filled' to create complete closure during phonation. Thus both the cricoid and arytenoid cartilages in the female larynx might be combining to produce the glottal chink.

In the frequency domain, for male speakers the abrupt cessation of glottal airflow during phonation causes a relatively strong excitation of the higher harmonics, and the skewed glottal waveform produces a relatively weak fundamental frequency component (Klatt & Klatt 1990:822). The slope, or steepness, of the glottal frequency spectrum is usually given as a uniform 12dB/octave. Monsen & Engebretson (1977) maintain that the slope typically differs from octave to octave (p991), usually increasing in steepness with each successive octave (p986), although their study of glottal source characteristics considered only five female and five male speakers. The absolute slopes of the female and male spectra are roughly equal; but when the slope of the female spectrum is normalised against the male for frequency and intensity – in order to view the harmonic relationships – it is typically steeper octave by octave (p986).

For female speakers, two main processes can be seen to be in operation during phonation. Firstly, closure of the vocal folds is propagated along their length, yielding a quasi-sinusoidal volume velocity waveform from the glottis. The more sinusoidal/symmetrical the waveform, the stronger the fundamental component in the frequency domain, and the weaker (more attenuated) the higher frequency components. Secondly, because the vocal folds are never completely closed at the rear of the larynx, considerable turbulent aspiration noise is created in the voice source. Although weak in intensity, it may be of sufficient strength to replace harmonic excitement of the higher harmonics (Klatt 1986; Ladefoged & Antoñanzas-Barroso 1985 – both cited Klatt & Klatt 1990:827). This is assumed to produce the characteristic breathy voice commonly associated with female speakers (see below for a discussion of the perceptual cues to breathy voice). A number of voice quality perception experiments have found breathiness to be a quality of women's voices (e.g. Söderston *et al.* (1989 – cited Söderston & Hammarberg 1992:23)).

Perceptual cues to breathy voice

Klatt & Klatt (1990:852) identified a number of potential cues to the perception of breathiness in the voice. As already indicated above, two of these cues are the high relative strength of the first harmonic, and therefore a greater spectral tilt[10], and the presence of aspiration noise, especially at higher frequencies. In addition, the bandwidth of F_1 is likely to increase, due to low frequency losses as a result of the partially open glottis[11]. Also, extra poles (formants) and zeros (energy gaps) may appear in the vocal tract transfer function because the glottal chink allows interaction with the subglottal cavities (i.e. the trachea).

Breathy phonation is used to form linguistic contrasts in a number of languages, includ-

[10]Spectral tilt is defined as the angle of decline of the formant amplitudes with increasing frequency (M Barry 1986:2; Huber 1989:477). The increased spectral tilt is a consequence of the high amplitude of the first harmonic and the attenuation of the higher frequencies.

[11]For the GLOVE text-to-speech synthesiser designed by KTH, Sweden, the bandwidth of F_1 is adjusted according to the size of the glottal opening (Carlson *et al.* 1991:483), which, subjectively, "seems to add to the naturalness of the [synthesised] speech". However they cite Nord et al. (1986) who failed to find a significant perceptual effect, and so they wonder whether an average bandwidth would produce just as good a perceptual result.

ing many Indo-Aryan languages such as Gujarati, Nepali, Marathi and Hindi; Niger-Kordofanian languages such as Igbo and Tsonga; Sino-Tibetan languages such as Newari; Mazatec of Mexico; and !Xóõ of Southern Africa (Henton & Bladon 1985:221). Some evidence for cues to breathiness comes from research into the acoustic-phonetic contrasts between breathy and non-breathy vowels in these languages. Klatt & Klatt (1990:823-5) surveyed some of this research, for Gujarati, Hmong and !Xóõ (Pandit 1957; Fischer-Jorgensen 1967; Ladefoged 1983; Ladefoged & Antoñanzas-Barroso 1985; Bickley 1982; Chasaide 1987; Chasaide & Gobl 1987; Huffman 1987). The most salient cue identified was the relative strength of the fundamental component in the frequency spectrum. Other possible cues were the presence of noise, especially at high frequencies, and an increased first formant bandwidth. From a laryngoscopic examination of one subject, Fischer-Jorgensen (1967 – cited Klatt & Klatt 1990:823-4) observed a wider opening in the posterior glottis for breathy vowels than non-breathy vowels. From an examination of the vowels of seven speakers of Gujarati, she also found that the glottal waveforms for breathy vowels showed an increased airflow, more sinusoidal waveshape, and greater open quotient.

Klatt & Klatt (1990:835-7) correlated the results of a perceptual listening test on the breathiness of female and male speakers with the values of a number of potential acoustic cues. The only correlations of any particular strength were for the relative amplitude of the first harmonic, and the presence of noise in the F_3 region of the spectrum. Interestingly, although their ten female speakers were on average rated more breathy than the six male speakers, the difference was very small. There was also a great deal of variation between and within speakers: some male speakers rated more breathy than some of the female speakers; and for each of the stimuli presented (two sentence types, one excised vowel per speaker) individual speakers exhibited considerable variation.

Other evidence for the potential cues to breathiness in the voice come from attempts to improve the voice source model in speech synthesis systems. Klatt & Klatt (1990) used the results of their analysis into natural speech to attempt a synthesis of a breathy female voice using the Klatt synthesiser. They presented voices synthesised from different combinations of the breathiness cues to a panel of listeners for rating. The strongest cue was the amplitude of aspiration noise. However, the sample perceived as most breathy was the one in which all the cues were included: "It is as if the perceiver is aware of *all* of the systematic changes that go into breathy phonation and uses these expectations during perception in such a way that no single cue is as effective as all in combination" (p853). As further evidence of this, Klatt & Klatt found that if only the first harmonic amplitude was raised then some subjects heard an increase in nasality in the voice, rather than the intended breathiness. And yet, when also accompanied by an increase in aspiration noise, the synthesised voice is never heard as nasalised. It is "likely that single-cue manipulations can create somewhat unnatural stimuli that result in perceptual ambiguity" (p853). For his simulation of vocal fold behaviour, Titze (1989a) introduced a glottal chink for synthesised female speech (p1704). During phonation, the folds close relatively abruptly at first, "creating a sudden partial closure [of the anterior glottis]"; the posterior then closes more gradually. His model produces a glottal volume velocity waveform with a greater open quotient for the female simulation, and a small amount of continuous DC airflow. This produces a volume velocity waveform that is more symmetrical than for the male simulation. The simulated male folds close abruptly rather than quasi-sinusoidally. Hermes (1991:497-8) notes that the synthesis of breathiness does not merely entail the addition of a stationary noise signal: in listening tests this tends to result in the noise source separating (streaming) out from the speech sound, forming a distinct percept of its own. Hermes synthesised breathiness using a simple source-filter model, with the source consisting of a combination of lowpass-filtered pulses and highpass-filtered noise bursts.

Perceptual tests showed that the best synthesis (the most complete integration of the pulses and noise into a single percept) occurred when the pulses and noise bursts were of equal energy and coincided in time with each other. This ties in with the description of the production of breathiness above in that the noise comes from the same source as the pulses (i.e. the airstream from the lungs) and its intensity is also controlled to a large extent by the opening and closing of the vocal folds. He further found that the perception of breathiness increased as the cut-off frequency of the highpassed noise was lowered.

Analyses of acoustic correlates of breathy voice

Direct evidence from acoustic analyses of a speaker sex difference in the use of breathy voice is fairly scanty. Most studies have assumed that the relative amplitude of the fundamental component is a sufficient correlate of breathiness, and have investigated only this parameter. Only Klatt & Klatt (1990) have attempted to contrast the aspiration noise cue in women's and men's speech, and they conclude that there is a tendency for female speakers to have a greater degree of aspiration noise in the region of F_3 (p852; see p830-2 for experimental details). However, Nittrouer *et al.* (1990:763) dispute this, questioning Klatt & Klatts' use of a subjective rating system to determine the amount of aspiration noise. Klatt & Klatt located the position of F_3 visually from a spectrogram, excised the third formant with a bandpass filter, and redigitised the waveform; the filtered waveform was then rated for nonperiodicity. Nittrouer *et al.* cite studies by Milenkovic (1987) and Yumoto *et al.* (1982) which found no significant differences in total signal-to-noise and harmonic-to-noise ratios respectively between women's and men's speech. However, as they point out, these noise measures are weighted towards the lower frequencies, while aspiration noise is found mainly in the higher frequencies (Nittrouer *et al.* 1990:763). Henton & Bladon (1985:222) posit an acoustic measure based on the ratio of harmonic energy to noise energy, but did not investigate further. Therefore, in the absence of a better, more objective measure of aspiration noise, Klatt & Klatts' result is still interesting, especially as it backs up other studies indicating aspiration noise as a potential cue to breathiness.

Studies examining the relative amplitude of the fundamental component have used the difference in amplitude between the first and second harmonics[12] as a measure of this. Henton & Bladon (1985:222) adopt an argument of convenience to justify its use, namely that it requires no special processing. Klatt & Klatt (1990:828-9) examined this and other measures of relative amplitude (H_1 relative to A_1, and H_1 relative to RMS amplitude) and found that, for group measures, there is little to choose between them. Because of the difficulties of measuring the others (for example, accurately measuring the amplitude of F_1) they too chose to use H_1–H_2. Both Nittrouer *et al.* (1990) and Günzburger (1991) used the H_1–H_2 measure because Henton & Bladon and Klatt & Klatt used it. The issue of the choice of measure with which to assess the relative amplitude of the first harmonic will be returned to in Section 4.1.2.

The results from the above-mentioned studies are summarised in Table 3.8. All four studies show considerable average differences between the female and male speaker groups for the measure H_1–H_2 (see the final column, labelled 'Differences'). Thus, we could conclude that female speakers of American English (Klatt & Klatt 1990; Nittrouer *et al.* 1990), of the Dutch equivalent of RP (Günzburger 1991), and of British RP and modified Northern British English (Henton & Bladon 1985) typically have a greater first-to-second harmonic amplitude difference than men. The result for US English speakers is

[12] This measure will be referred to as H_1–H_2, where H_1 is the amplitude of the first harmonic or fundamental component, and H_2 is the amplitude of the second harmonic. The amplitude of the first formant will be referred to as A_1.

Source	Lang. spoken	Speaker sex	No. of speakers	Mean (SD) (dB)	Difference (dB)
Henton & Bladon (1985)	MD English	f	12	7.6 (–)	
		m	13	1.6 (–)	6.0
Henton & Bladon (1985)	RP English	f	20	6.4 (–)	
		m	16	0.8 (–)	5.6
Klatt & Klatt (1990)	US English	f	10	1.9 (2.3)	
		m	6	-3.8 (1.9)	5.7
Nittrouer *et al.* (1990)	US English	f	4	-0.2 (3.5)	
		m	4	-4.8 (2.1)	4.6
Günzburger (1991)	'RP' Dutch	f	13	3.9 (5.2)	
		m	13	-0.6 (3.9)	4.5

Table 3.8: Survey of studies comparing measures of H_1–H_2 (the difference in amplitude between the first and second harmonics) for female and male speakers. The 'Difference' column gives the difference between the female and male H_1–H_2 scores. Note: MD = Modified Northern; RP = Received Pronunciation; 'RP' Dutch = Dutch equivalent of RP; f = female; m = male.

supported by the research of Monsen & Engebretson (1977:986-7). However, while there are similarities between the female-male difference from the different studies, there is a great deal of inconsistency in the mean values of H_1–H_2, ranging from -0.2dB to 7.6dB for the female speakers, and -4.8dB to 1.6dB for the males. There are a number of possible explanations for this, including the choice of stimuli, the measurement techniques, the type and number of subjects, between- and within-speaker variability, and cultural differences. These sources of variability will now be discussed.

The main criterion for the choice of which vowels to investigate in these studies was that the lower harmonics should be relatively free from the influence of F_1. For vowels such as /iy/ and /uw/, the first formant is often near, or even below, the second harmonic (Ladefoged *et al.* 1988 – cited Nittrouer *et al.* 1990:766), causing its amplitude to be increased. Thus the open vowels were used in these studies to measure H_1–H_2, although the different researchers were inconsistent in their choice: the two US studies used /aa/, Henton & Bladon used the vowels /ae/, /ah/, /ax/, /er/, while Günzburger used /ae/. Examination of the results suggests that the H_1–H_2 measure is subject to variation due to both the vowel uttered, and its consonantal context. Henton & Bladons' data, reproduced in Table 3.9, suggest that H_1–H_2 is dependent upon the vowel uttered, although mainly for the female speakers. For instance, the mean H_1–H_2 for the female RP speakers ranged from 3.3-8.4dB across the four vowels in their study, but only 0.2-1.0dB for the male RP speakers. Nittrouer *et al.* investigated the influence of the preceding consonant on the relative first harmonic amplitude of /aa/. They found a tendency for H_1–H_2 to increase in value from the voiced stop (they considered /d/ in their study) to the voiceless fricative (/s/ and /sh/) to the voiceless stop (/t/ and /k/) contexts (p774). Of the issues surrounding the recording of the stimuli, the recording level is obviously of great importance in the measurement of harmonic amplitudes, although the effects of this are offset somewhat because we are interested only in *relative* amplitudes.

Probably the most compelling reason for treating these results with caution is that the between-speaker differences, where reported, are considerable. In the Klatt & Klatt study, the mean H_1–H_2 scores for each female speaker ranged from -1.6dB to 7.1dB, and the males from -5.4dB to -0.3dB (see Table III, p829). Similarly, the female speakers in Günzburger's study ranged from -2.5dB to 14.5dB, and the males from -5.0dB to 5.0dB (estimated

Accent		/ae/	/ah/	/ax/	/er/	Mean
RP	f	8.4	6.4	6.2	3.3	–
	m	1.0	0.8	0.2	0.4	–
diff. (f-m)		7.4	5.6	6.0	2.9	5.5
MN	f	10.6	7.6	–	6.3	–
	m	2.4	1.6	–	2.5	–
diff. (f-m)		7.9	6.0	–	3.8	5.9

Table 3.9: Female and male H_1–H_2 data (to nearest 0.1dB) for four vowel sounds. The two groups of subjects were speakers of the Received Pronunciation (RP) and Modified Northern (MN) accents of British English. Also included are the differences between the female and male averages. After Henton & Bladon (1985:224).

from Figure 2, p64). Breathiness, as measured using H_1–H_2, does not appear to be a consistent feature of speaker sex for either women or men, and some men are clearly more breathy than some women. In addition to the between-speaker differences, Klatt & Klatts' suggested from their data that while most speakers appear to have reasonably consistent levels for H_1–H_2 in a given situation, some speakers exhibit considerable variation in different contexts, even in the declarative-type sentences used in their study (p830). This led Klatt & Klatt to conclude that, "within each gender there is much ... variation in acoustic manifestations of breathiness, with some males being more breathy than many females. In addition, it is likely that any individual is capable of adopting a fairly wide range of speaking styles that differ in degree of breathiness" (p852). With regard to the comment on speaking styles, Günzburger found that when her subjects were asked to adopt a 'sexy' voice (i.e. breathier than their normal voice), nearly all of them increased the relative amplitude of the first harmonic, and the group means for the H_1–H_2 measure rose by 1.5dB for the female speakers and 1.7dB for the male speakers.

It is also worth noting the effect of basing a study on data from a relatively few number of speakers (see the fourth column of Table 3.8), particularly the degree of between-subject variability in the H_1–H_2 measure. The mean H_1–H_2 for the thirteen female and thirteen male speakers in Günzburger's study ranged over 17dB and 10dB respectively; while Klatt & Klatts' ten female and six male speakers ranged over 8.7dB and 5.1dB respectively. Given such variation, the use of such small numbers of subjects can only lessen the possibility of a representative population sample. Finally, if a breathy voice quality is a learned characteristic, or is affected by acculturation, then it is highly likely that we would find noticeable differences between subjects from different countries, and even within countries, as evidenced by the differences in H_1–H_2 for the speakers of modified Northern and RP English in Henton & Bladons' study (see Table 3.9). Klatt & Klatt (1990:826) state that, impressionistically, Swedish women sound less breathy than American women. They quote a study which would appear to back this up, in that only two out of the six female subjects exhibited any DC airflow in the nominally closed phase of the glottal cycle (Karlsson 1985). This indicates that breathiness is not a universal marker of women's speech. In other words it is not a feature that has been adopted by women of all cultures. However, a study by Söderston *et al.* (1989 – cited Söderston & Hammarberg 1992:23) found perceived breathiness was common amongst (young) Swedish women, while a posterior glottal opening was found by Söderston & Lindestad (1990 – cited Söderston & Hammarberg 1992:23) to be consistent in Swedish women.

The experimental differences highlighted above make cross-study comparisons hazardous, but it is clear that there is a tendency for women's voices to display a higher relative

first harmonic amplitude than men. And insofar as this measure can be said to be an adequate acoustic correlate of breathiness in the voice, then there is also a tendency for women to be more breathy-voiced than men. However, the amount of between-subject variability reported in this research, and the potential for within-speaker variability in different speaking styles, shows that this should be treated as no more than a tendency, and that "it is unwise to make sweeping generalisations with regard to sex typing, as well as the behaviour of particular individuals" (Klatt & Klatt:1990:852).

Summary

From the evidence available in the literature, a degree of breathiness appears to be a characteristic of the female voice. Voice quality perception studies have indicated that women do tend to have a breathy voice; while articulatory studies have fairly convincingly pinpointed the mechanisms for its production, namely the presence of an opening in the posterior glottis during phonation and its subsequent effects on the voice source. There are, however, suggestions that this voice quality is a learned behaviour (Klatt & Klatt 1990:825). The main acoustic correlates of this breathiness are the high relative amplitude of the first harmonic, and the presence of aspiration noise in the higher frequencies. The relative strength of the first harmonic is generally considered the most important cue to the perception of a breathy voice quality.

There are claims that breathiness impairs intelligibility and that it is indicative of vocal pathology, in that an increase in noise in the voice is considered indicative of laryngeal disorder (Nittrouer *et al.* 1990:762)[13]. If this were true, it raises questions of whether women are adopting a speech behaviour that not only makes them less likely to be understood, but inflicts physical damage on the voice source. However, if a breathy voice quality is a vocal marker of the female population then it is unlikely to degrade intelligibility nor produce laryngeal disorders. These claims are examined in more detail below.

Henton & Bladon (1985:225) claim that breathiness reduces the intelligibility of speech and question why women should choose to produce breathy speech if this is the case. This is countered by Javkin *et al.* (1991), who tested the intelligibility of synthesised breathy speech. They simulated breathiness by controlling the three parameters associated with breathy voice production: spectral tilt, open quotient of the glottal pulse, and the amplitude of aspiration noise. From listening tests using different combinations of the source parameters they concluded that intelligibility was not affected[14].

[13]Nittrouer *et al.* (1990:762) note that breathiness is usually measured using two different signal characteristics, depending on whether the area of application is clinical or linguistic: "Consequently, it is not clear that the label 'breathy' applies to speech signals with the same acoustic (and therefore, articulatory) characteristics in clinical and linguistic descriptions." They discuss whether the amount of aspiration noise and relative amplitude of the fundamental are dependent upon each other (i.e. greater open quotient, more symmetrical waveform), or whether they are independent (noise is a major consequence of many laryngeal disorders and greater relative F_0 amplitude is cited by a number of studies as being the most important perceptual cue to breathiness). According to Söderston & Hammarberg (1992:23): "Incomplete vocal fold closure is one parameter the speech pathologist tries to reduce with vocal exercises in voice therapy of patients with voice disorders as well as in voice training of normal speakers who need to improve their voices to prevent voice problems." However, when they studied the degree of posterior glottal closure in eight young women who were each given forty hours of formal voice training, they found that the degree of closure did not decrease with any significance for any of the women, whilst the degree of perceived breathiness actually increased. Moreover, Klatt & Klatts' findings argue powerfully for the dependence stance (p822-3).

[14]On a cautionary note, Javkin *et al.* synthesised a male voice (p543). Also, it is unclear how natural the noise component in their synthesised speech was. Although they indicate (p541) that this parameter is changed during a word, they give no further details of its dynamics. Hermes (1991) noted that the addition of a stationary noise source may result in the noise simply 'streaming out', in other words forming

Henton & Bladon (1985:225) make a number of claims as to why breathiness adversely affects intelligibility. They say high-pitched breathy voices are unlikely (due to the lower longitudinal tension of the vocal cords) and that breathy-voiced speakers are liable to sound monotonous. However, if breathiness is indeed a feature of women's voices then clearly this contention holds little weight. They also cite studies showing that the perception of speech produced against a background of noise is degraded, stating that this "is essentially what is happening when voiced speech is overlaid on a breathy background" (p225). This is patently untrue as it ignores the fact that the aspiration noise produced during breathy speech results in an excitation of the vocal tract, and is therefore modulated by the resonances there (Javkin *et al.* 1991:540). Javkin *et al.* also note that breathiness could actually *increase* intelligibility, particularly for female speakers, in that the noise may to some extent 'fill in' the relatively wide spectral spacing between the harmonics (p540; see also Klatt & Klatt 1990:822-3).

Considering the role of breathiness in forming linguistic contrasts in various languages, Ladefoged (1983:351 – cited Henton & Bladon 1985:221) claims that breathy voice "would be considered strongly stylistically marked or pathological if used by speakers of English [who do not make such a linguistic contrast]". However, while extreme breathiness would probably be indicative of pathological speech (Perkins 1971; Hanson & Emanuel 1979 – both cited Henton & Bladon 1985:221; Günzburger 1991:62), there is no reason why a degree of breathiness should not be used by women paralinguistically.

---

a separate percept. He found that a periodic noise source synchronised with the glottal pulses caused the noise component to be fully integrated into a single percept.

3.1.3 The vocal tract and the formant frequencies

The formant frequencies of women and men are generally cited as being between 15% (Vaissière 1985:202) and 18% (Fant 1973:93) greater than those of men. This difference is a consequence of the longer male vocal tract, which results in lower frequency resonances being propagated in the supralaryngeal vocal cavity (Daniloff *et al.* 1980:273). According to Chiba & Kajiyama (1941 – cited Fant 1973:46), the average female vocal tract is 15% shorter than the male.

However, the actual difference in formant frequencies is dependent upon both the phoneme being articulated and which formant is being considered. The reason for this disparity (i.e. of the variability in the female-male difference for different phonemes) is that the male vocal tract is not simply a scaled-up version of the female vocal tract (Laver 1988:99). A difference has been found between the female and male length-ratios of the pharyngeal and oral tracts (Klatt & Klatt 1990:825), although there is scant published evidence for this. Fant (1973:88) states that adult males have "a relatively greater pharynx length and more pronounced laryngeal cavities" than females, while the oral cavities are about the same length in both sexes[15]. However, this oft-quoted statement is based solely on data supplied by Chiba & Kajiyama (1941:188-93) from a study of very few speakers. Fant cites Chiba & Kajiyama as saying the length of the mouth cavity (incisor to pharynx wall) of a girl of eight was 30% shorter than an adult female and 56% shorter than an adult male. They are also quoted as saying that the relative overall vocal tract lengths are 1.0 for males and 0.87 for females (and 0.80 for a boy of nine and 0.70 for a girl of eight). This would appear to be flimsy evidence to be basing such a claim on. Hunter & Garn (1972) found that the average difference in size of the ramus (a vertical dimension at the rear of the mandible) between women and men is proportionately greater than for other facial dimensions. This would appear to indicate a proportionately greater depth to the supra-laryngeal tract in men, more specifically in the oral cavity.

The data from measurements of formant frequencies of different vowels tend to substantiate the notion of a disproportionate sexual dimorphism in the supralaryngeal vocal tract. If such a disproportion does exist, since a person's formant frequencies are the product of the shape of their vocal tract, and given that the first two formants in particular can be variously dependent upon the front and back parts of the vocal tract (i.e. of the oral and pharyngeal cavities respectively), we would expect the relative formant frequency differences between the sexes to be inconsistent for different vowels. This effect can be seen by referring to Tables 3.10 and 3.11, which show the scaling factors representing the percentage differences between female and male formant frequencies for a number of vowels. The scaling factors were calculated as follows:

$$k_n = \left[\frac{F_n(female)}{F_n(male)} - 1\right] \times 100$$

where k_n indicates the scaling factor for the n-th formant. Table 3.10 shows the formant data derived from Peterson & Barneys' (1952) study on U.S. women and men. Table 3.11 shows the formant data for the equivalent vowels derived from Fant's (1959) study of Swedish women and men. The mean scaling factors for all formants and all phonemes from the two studies are 17% and 15% respectively, i.e. according to these results, on

[15] Interestingly, Fant (1973:93) argued that: "The scaling of children's data from female data comes closer to a simple factor independent of vowel class." In other words, the configuration of women's vocal tracts is more like those of children than men's. Fant based this statement on a comparison of the child/male and female/male formant scale factors for different vowels from Peterson & Barneys' (1952) data, the graphs of which look similar (although obviously the child/male scale factor is much greater than the female/male) (see p89).

Formant		iy	ih	eh	ae	aa	aw	uh	uw	ah	axr	Mean
F_1	f	310	430	610	860	850	590	470	370	760	500	580
	m	270	390	530	660	730	570	440	300	640	490	500
% diff.		15	10	15	30	16	4	7	23	19	2	15
F_2	f	2790	2480	2330	2050	1220	920	1160	950	1400	1640	1690
	m	2290	1990	1840	1720	1090	840	1020	870	1190	1350	1420
% diff.		22	25	27	19	12	10	14	9	18	21	19
F_3	f	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960	2780
	m	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690	2390
% diff.		10	20	21	18	15	12	20	19	16	16	17

Table 3.10: Formant frequency data (to nearest 10Hz) for ten vowels articulated by 28 female (f) and 33 male (m) U.S. English speakers in a /hVd/ context. Also included are the percentage differences between the male and female formants and their averages. Note that the overall means given in the final column are computed from the vowel means. After Peterson & Barney (1952).

average female formant frequencies are 15-17% greater than male formant frequencies. If we consider the scaling factors formant-by-formant and phoneme-by-phoneme, we can discern the differences referred to above, and also that the disparities can be large. With Peterson & Barneys' data, the scaling factors range from 2% (F_1 of /er/) to 30% (F_1 of /ae/), and with Fant's data they range from -3% (F_2 of /uw/[16]) to 30% (F_1 of /ae/ again, and F_3 of /uw/). This indicates that only certain formant frequencies from certain vowels would be of use when seeking to differentiate between female and male speakers. Fant (1973:93) summarises the deviations from the average scaling factor across all vowels as follows:

a) the first and second formants of all rounded back vowels have relatively low scale factors;
b) the first formants of any close or highly rounded vowel, i.e. high front vowels, also have relatively low scale factors;
c) the first formant of very open front or back vowels has a substantially higher than average scale factor.

If we now compare the consistency of the phoneme-by-phoneme figures between the two studies, we can see a number of discrepancies in that there is agreement on the mean frequencies of only some formants. The mean between-study differences are 30Hz for female speakers and 20Hz for male speakers for F_1, 200Hz and 140Hz for F_2, and 140Hz and 40Hz for F_3. There are a number of possible explanations for this, the chief one being that Fant's phoneme names were derived from matching the F-patterns of his subjects' Swedish phonemes with the nearest U.S. English equivalents. Thus even when the 'same' vowel phoneme was uttered, the formant frequencies would not necessarily be the same due to the effects of the Swedish pronunciation. Moreover, Peterson & Barney measured their formant frequencies from phonemes uttered in a /hVd/ context, while Fant's speakers were asked to produce phonemes sustained over four seconds. This would have a significant effect on vowel quality, and subsequently on the frequencies of the formants.

Fant (1960 – cited in Günzburger & de Vries 1989:143) suggested that the female-male difference in formant frequencies was more consistently pronounced for F_3 than for F_1 and

[16]Thus, for the speakers investigated by Fant, the second formant of /uw/ was actually higher on average for male speakers than it was for female speakers. This was also true for the first formant of /uh/.

Formant		iy	ih	eh	ae	aa	aw	uh	uw	ah	Mean
F_1	f	280	370	550	790	860	520	410	340	570	520
	m	260	330	440	610	680	490	420	310	530	450
% diff.		9	10	25	30	27	6	-1	11	8	16
F_2	f	2520	2540	2140	1830	1200	840	1180	690	1290	1580
	m	2070	2050	1800	1550	1070	830	1070	710	1100	1360
% diff.		22	24	19	18	12	2	10	-3	17	16
F_3	f	3460	2960	2860	2920	2920	2840	2710	2900	2750	2920
	m	2960	2510	2390	2450	2520	2560	2320	2230	2430	2490
% diff.		17	18	20	19	16	11	17	30	13	17

Table 3.11: Formant frequency data (to nearest 10Hz) for nine vowels articulated by 7 female (f) and 7 male (m) Swedish speakers from 4 sec sustained vowel productions. Also included are the percentage differences between the male and female formants and their averages. Note: (1) The overall means given in the final column are computed from the vowel means; (2) Although the vowels uttered were Swedish, the vowel names given are comparable U.S. English vowels selected by F-pattern match (Fant 1973:86); (3) Fant did not give the figures for the female speakers. These were calculated from the male formant frequencies and the female-male percentage difference. After Fant (1973), originally published in Fant (1959).

F_2, proposing a possible systematic sex-related difference. Certainly the female-male F_3 scaling factors reproduced from Fant's 1959 study in Table 3.11 do not fall below 11%, while the F_1 scaling factors for four phonemes and F_2 scaling factors for two phonemes are less than 10%. However, Peterson & Barney's data indicate consistently high scaling factors for both F_2 and F_3. Günzburger & de Vries suggest that further evidence for a systematic difference in F_3 comes from studies of male-to-female transsexuals, who while adopting a female speaking mode caused an upward shift in the third formant while not appreciably affecting the first two formants (Günzburger 1989 – cited in Günzburger & de Vries 1989:143).

A sex difference has also been reported in the width of the vowel formant bandwidths. However, quantitative statements range from male bandwidths being approximately half as wide as those for females (Vaissière 1985:202), to male bandwidths "appear" to be narrower (Bladon 1985:30), although neither source provides any empirical evidence or reference to such.

Measuring the formant frequencies of female speakers

One factor contributing to the lack of research into women's speech is that it has traditionally been regarded as difficult to measure. The main reason for this is that most acoustic studies tend to focus on the formant frequencies as cues to phonetic contrasts; but the higher fundamental frequencies of women (and children) make it more difficult to estimate formant frequency locations (Karlsson 1991:113-4, Klatt & Klatt 1990:820). Automatic formant tracking can be "particularly difficult for female voices (Klatt & Klatt 1990:825, Vaissière 1985:196).

Due to the high F_0 produced by women (80-90% higher than men), the harmonics of their fundamental in the frequency spectrum are more widely spaced. There is subsequently less chance that a harmonic will fall within the bandwidth of a formant, and less chance that

the harmonic will excite the formant – therefore the formants will be less well specified (Karlsson 1991:114). Or, to put it another way, "male vowels have ... a lower fundamental frequency which leads to a denser sampling of the spectrum by more partials" (Bladon 1985:30). Thus spectrograms of female speakers tend to be more difficult to decipher than male speakers. This is particularly troublesome for women's articulations of high vowels and voiced consonants: F_1 is often very close to the fundamental making the first formant difficult to measure (Karlsson 1991:114) (see Section 4.1.2 for a more detailed look at these points). Furthermore, if, as has been suggested in Section 3.1.2, one aspect of the female voice is a degree of breathiness, then the consequences of breathy voice for the frequency spectrum render the formant frequencies even more difficult to measure, e.g. a stronger first harmonic may be confused for F_1, an increase in the bandwidth of F_1 could make it harder to locate a local maximum, and coupling with the trachea can produce extra peaks (Klatt & Klatt 1990:853).

And yet, women's speech is no less intelligible than men's. For example, Tielen measured the intelligibility of female and male voices under different noise conditions and found no differences between the sex groups. Strange *et al.* (1976) found that, for vowels spoken in isolation, women's vowels were identified accurately 74% of the time, against 67% for men. Indeed, Margulies (1979 – cited Klatt & Klatt 1990:825) found female speakers to be significantly more intelligible than male speakers. Comparing the intelligibility of five female and five male speakers in various noise conditions, the women achieved an average score of 73%, against 56% for the men. Clearly human beings do not experience the same problems in perceiving women's speech.

This begs the question, are we missing something? Or perhaps we should be more forthright, and ask if we are fundamentally mistaken in trying to apply to female speech the models of speech production derived from research into the male voice. Certainly, some researchers have suggested that formants are not as important as has been assumed in the human perception of speech, and that we should look to improving our understanding of how the auditory system deals with the acoustic speech signal (e.g. Bladon 1982). There is without doubt room for improvement in the male-orientated model of speech production, in that it could take into account the acoustic consequences of the female vocal apparatus. For example, from some informal observations Klatt & Klatt (1990:820) noted there is a "possibility that vowel spectra obtained from women's voices do not conform as well to an all-pole model, due perhaps to tracheal coupling and source/tract interactions (Fant 1985, Klatt 1986)". This can result in spurious poles and zeros entering into the vocal tract transfer function. Furthermore, the nonuniform differences in laryngeal anatomy found by Titze (1989a) (see Section 3.1.1) suggest there are significant differences in vocal source characteristics. The sexual dimorphism in the length-ratios of the vocal tracts' cavities cause a particular problem. "The prognosis for solving [the speaker sex] problem looked bleak in 1975 when Fant demonstrated that ... formant frequency differences are not consistent from vowel to vowel, nor from formant to formant" (Bladon 1985:29). Previous attempts to account for the female-male differences have relied on simple scaling techniques, such as multiplying the formants by a factor of 1.175 (and the fundamental by 1.7) (Huber 1989:477). When these parameters were incorporated into speech synthesis algorithms, the results were not very convincing (see Klatt 1987:746-7).

Summary

There is unfortunately very little quantitative data on the formant frequencies in the literature, particularly for female speakers. There have been a number of studies of speakers of different languages, but only of male speakers – for example, Fant's (1973) study of 24

Swedish men, Fischer-Jørgenson's (1972) study of Danish men, Jørgenson's (1969) study of 6 German men, and Pols *et al.*s' (1973) study of 50 Dutch men. This makes it very difficult to establish patterns in the differences between the sexes for different vowel types, for example to substantiate the between-vowel deviations from the average scaling factor summarised by Fant. Furthermore, it is almost impossible to gauge the variability in the formant frequencies between speakers. If the average female-male difference for the first three formants taken together is approximately 15%, then there is likely to be a great deal of overlap in the values attained by individual women and men. However, given the differences reported in the scaling factors in both the Peterson & Barney and Fant studies, it is probable that the formant frequency behaviour of women and men is very different for some vowels, and very similar for others.

3.1.4 The vocal apparatus and the effects of aging, height and weight

The vocal apparatus and the effects of aging in childhood

A. The larynx and fundamental frequency

Studies of the anatomy of children indicate there are no significant anatomical differences between girls and boys until they reach puberty (e.g. Crelin 1973 – cited Lieberman & Blumstein 1988:131)[17]. More specifically, anatomical measurements of the larynges of pre-pubescent boys and girls show the larynges are likely to have the same dimensions, given the same overall height and weight (Kirchner 1970; Negus 1962 – both cited Ingrisano *et al.* 1980:62; Kahane 1975 – cited Bennett 1983:139). In addition, using the circumference of the throat as an indicator/correlate of larynx size, Günzberger *et al.* (1987:50) found no significant sex differentiation in this measure for 17 seven and eight year-olds.

Therefore, given the lack of anatomical distinction, we would expect there to be no systematic differences in the fundamental frequency of pre-pubescent children matched for height and weight (Sachs 1975 – cited Smith 1985:59; Robb & Saxman 1985 – cited Lieberman & Blumstein 1988:131). However, from the studies of children's fundamental frequency characteristics, it is extremely difficult to come to any firm conclusion about this, as the reported differences between girls and boys create an ambiguous picture. For example, the average SFF of girls is variously given as being greater than that of boys (Hasek *et al.* 1980; Ingrisano *et al.* 1980), less than that of boys (Fairbanks *et al.* 1949a,b; Sachs *et al.* 1973), or roughly the same (Bennett 1983; Günzburger *et al.* 1987; Weinberg & Bennett 1971a; Vuorenkoski *et al.* 1978). Furthermore, from an examination of Tables 3.12 and 3.13, there is clearly great disparity in the mean SFFs reported for each age group.

It is actually extremely difficult to assess the degree of sex differentiation, if any, that exists in the SFFs of children. One of the biggest problems is deciding what groups of children should be compared, as children, even of the same age, will be at varying stages of physical and mental development. One solution to this problem would be to pair off girls and boys whose heights and weights match, a strategy adopted by Sachs *et al.* (1973). From a group of children of mixed ages they matched nine pairs of girls and boys and measured the fundamental frequency of sustained productions of the vowels /aa/, /iy/ and /uw/, and found the boys F_0 to be significantly higher than the girls. Another strategy would be to compare the fundamental frequencies of children whose sex was perceived accurately (or indeed inaccurately) from their speech, a method used by Günzburger *et al.* (1987), Ingrisano *et al.* (1980) and Weinberg & Bennett (1971a). Weinberg & Bennett (1971a:1210,1211), measuring F_0 from 30 seconds of spontaneous speech from 66 five and six year-olds, found the group means and ranges of speaking F_0 to be very similar (if anything, the five year-old boys were slightly higher). Günzberger *et al.* (1987:51) measured the fundamental of 17 children, both for sustained productions of the vowels /aa/, /iy/ and /uw/ and for sentences. As with Weinberg & Bennett they found great similarity in the group averages, and the boys' average F_0 in sentences was higher than the girls, though not at a level of significance. Further, both Weinberg & Bennett (1971a) and Günzberger *et al.* (1987) found an extensive overlap in the F_0 distributions of the children whose sex was accurately and consistently identified (see Section 3.2.1 on child sex perception for more details). However, in contrast to this, Ingrisano *et al.* (1980:66-7) selected a group of four to five year-old children whose sex was correctly identified more than 90% of the time (see Section 3.2.1) and found the girls' group average F_0 was higher

[17] Tanner (1978 – cited Smith 1985:59) states that boys on average are slightly larger than girls from birth. However, while these slight differences in average size may result in correspondingly small differences in average F_0, the data reported upon here indicate that this is not significant.

Source	Age (yrs)	n	Lang. spoken	Median F_0 (Hz)	Mean F_0 (Hz)	SD (Hz)	SD (st)	Range (Hz)
Cornut *et al.* (1971)	5	–	–	–	279	–	–	–
Cornut *et al.* (1971)	5	–	–	–	283	–	–	–
Weinberg & Bennett (1971a)	5	18	US Engl.	–	248	–	2.9	212-295
Hasek *et al.* (1980)	5	15	US Engl.	–	258	28.7	–	202-306
Ingrisano *et al.* (1980)	5	3	US Engl.	–	286	–	–	233-323
Cornut *et al.* (1971)	6	–	–	–	276	–	–	–
Cornut *et al.* (1971)	6	–	–	–	272	–	–	–
Weinberg & Bennett (1971a)	6	19	US Engl.	–	247	–	3.3	218-274
McGlone & McGlone (1972)	6	10	US Engl.	–	249	–	–	–
Bennett & Weinberg (1979)	6	30	US Engl.	–	245	27.5	–	–
Hasek *et al.* (1980)	6	15	US Engl.	–	254	22.7	–	210-313
Herbert (1942)	7	–	US Engl.	–	273	–	–	–
Fairbanks *et al.* (1949a)	7	15	US Engl.	–	281	–	–	–
Cornut *et al.* (1971)	7	–	–	–	293	–	–	–
Cornut *et al.* (1971)	7	–	–	–	278	–	–	–
Hasek *et al.* (1980)	7	15	US Engl.	–	262	35.8	–	195-303
Hammond (1947)	8	–	US Engl.	–	287	–	–	–
Fairbanks *et al.* (1949a)	8	15	US Engl.	–	288	–	–	–
Cornut *et al.* (1971)	8	–	–	–	294	–	–	–
Cornut *et al.* (1971)	8	–	–	–	281	–	–	–
McGlone & McGlone (1972)	8	10	US Engl.	–	276	–	0.6	–
Vuorenkoski *et al.* (1978)	8	22	–	–	253	–	–	–
Hasek *et al.* (1980)	8	15	US Engl.	–	264	24.6	–	215-306
Bennett (1983)	8	10	US Engl.	–	235	12.3	–	221-258
Günzburger *et al.* (1987)	8	6	Dutch	–	243	–	–	–
Günzburger *et al.* (1987)	8	6	Dutch	–	245	–	–	–
Hasek *et al.* (1980)	9	15	US Engl.	–	247	22.4	–	210-281
Bennett (1983)	9	10	US Engl.	–	222	8.3	–	209-236
Vuorenkoski *et al.* (1978)	10	28	–	–	252	–	–	–
Hasek *et al.* (1980)	10	15	US Engl.	–	254	19.8	–	234-303
Bennett (1983)	10	10	US Engl.	–	228	9.4	–	215-239
Duffy (1970)	11	6	US Engl.	266	266	–	–	–
Bennett (1983)	11	10	US Engl.	–	221	13.4	–	200-244
Duffy (1970)	13	6	US Engl.	264	260	–	–	–
Duffy (1970)	13	6	US Engl.	250	245	–	–	–
Michel *et al.* (1966)	15	44	US Engl.	–	208	–	–	159-260
Hollien & Paul (1969)	15	89	US Engl.	–	216	–	1.5	159-260
Duffy (1970)	15	6	US Engl.	238	237	–	–	–
Michel *et al.* (1966)	16	115	US Engl.	–	207	–	–	144-255
Hollien & Paul (1969)	16	185	US Engl.	–	214	–	1.5	154-256
Michel *et al.* (1966)	17	148	US Engl.	–	208	–	–	127-263
Hollien & Paul (1969)	17	193	US Engl.	–	212	–	1.7	127-263

Table 3.12: Survey by age of studies reporting girls' mean SFF values.

Source	Age (yrs)	n	Lang. spoken	Median F_0 (Hz)	Mean F_0 (Hz)	SD (Hz)	SD (st)	Range (Hz)
Weinberg & Bennett (1971a)	5	15	US Engl.	–	252	–	2.9	217-293
Hasek *et al.* (1980)	5	15	US Engl.	–	248	27.9	–	186-313
Ingrisano *et al.* (1980)	5	4	US Engl.	–	273	–	–	246-273
Weinberg & Bennett (1971a)	6	14	US Engl.	–	247	–	2.9	204-274
Bennett & Weinberg (1979)	6	23	US Engl.	–	238	19.9	–	–
Hasek *et al.* (1980)	6	15	US Engl.	–	263	31.1	–	228-332
Fairbanks *et al.* (1949b)	7	15	US Engl.	–	294	–	–	–
Hasek *et al.* (1980)	7	15	US Engl.	–	234	17.9	–	199-264
Fairbanks *et al.* (1949b)	8	15	US Engl.	–	297	–	–	–
Vuorenkoski *et al.* (1978)	8	28	–	–	259	–	–	–
Hasek *et al.* (1980)	8	15	US Engl.	–	236	28.5	–	195-313
Bennett (1983)	8	15	US Engl.	–	234	19.8	–	204-270
Günzburger *et al.* (1987)	8	11	Dutch	–	246	–	–	–
Günzburger *et al.* (1987)	8	11	Dutch	–	260	–	–	–
Hasek *et al.* (1980)	9	15	US Engl.	–	230	13.1	–	211-254
Bennett (1983)	9	15	US Engl.	–	226	16.4	–	198-263
Curry (1940)	10	–	–	269	263	–	–	–
Hollien & Malick (1962)	10	6	US Engl.	223	210	–	–	–
Hollien *et al.* (1965)	10	6	US Engl.	247	235	–	–	–
Hollien & Malick (1967)	10	6	US Engl.	237	226	–	–	–
Vuorenkoski *et al.* (1978)	10	32	–	–	247	–	–	–
Hasek *et al.* (1980)	10	15	US Engl.	–	229	33.8	–	173-293
Bennett (1983)	10	15	US Engl.	–	224	14.7	–	208-259
Bennett (1983)	11	15	US Engl.	–	216	15.0	–	195-259
Curry (1940)	14	–	–	241	232	–	–	–
Hollien & Malick (1962)	14	–	US Engl.	162	158	–	–	–
Hollien *et al.* (1965)	14	–	US Engl.	199	185	–	–	–
Hollien & Malick (1967)	14	–	US Engl.	184	189	–	–	–
Curry (1940)	18	–	–	137	133	–	–	–
Hollien *et al.* (1965)	18	–	US Engl.	–	115	–	–	–
Hollien & Malick (1962)	18	–	US Engl.	–	121	–	–	–

Table 3.13: Survey by age of studies reporting boys' mean SFF values.

than the boys', while the boys' range of F_0 was more restricted. This trend was also observed in intonation curves. However, their sample was very small, consisting of only three girls and four boys[18].

Further causes of the variable SFFs reported for children lie with the different research methodologies pursued in these studies, the numbers of children comprising the sample groups, and the unpredictable responses of the children, who may feel variously intimidated, bored or playful depending upon the experimental situation[19]. In order to compare and contrast the research reported upon in Tables 3.12 and 3.13, a summary giving details of the experimental conditions under which the fundamental frequency data was obtained is produced below:

- Bennett (1983) – longitudinal study; children wore headset-mounted microphone; generally one utterance of 12 syllable declarative sentence, giving a sample of several vowels and diphthongs; also measured were voiced consonants, vowel/consonant transitions, sentence-final F_0 fall. Average age, when study began: girls 8:1 (7:2 to 8:11); boys 8:3 (7:5 to 8:9).

- Bennett & Weinberg (1979) – 9 syllable declarative sentence. Age range: 6:1 to 7:10.

- Cornut *et al.* (1971) – Children asked to repeat sentences after researcher.

- Duffy (1970) – Portion of the Rainbow passage; subjects in anechoic room. Average ages (±1 month): 11:1, 13 (at least 6 months prior to menarche), 13:1 (mean of 8 months after menarche), 14:11.

- Fairbanks *et al.* (1949a,b) – Recording made in researcher's laboratory; subjects read simple passage, middle portion of which was recorded (53 syllables); children practised passage, with researchers correcting words, then passage read twice. Average age: girls 7:0 (6:10 to 7:2), 7:11 (7:9 to 8:1); boys 7:0 (6:10 to 7:2), 8:0 (7:10 to 8:1).

- Günzburger *et al.* (1987) – Experiment 1: sustained vowels /aa/, /iy/, /uw/. Experiment 2: sentence. Average ages: 7:6 to 8:9.

- Hasek *et al.* (1980) – 5 second production of single vowel; practice given.

- Hollien & Paul (1969) – Rainbow passage; subjects consisted of the 307 girls from Michel *et al.*s' (1966) study, plus an extra 160 girls; F_0 sampled every 33 msec. Average ages (overall range: 15:1 to 17:11): 15:6, 16:6, 17:6.

- Hollien & Malick (1962) – Rainbow passage; African American boys.

- Hollien & Malick (1967) – Rainbow passage; Northern U.S. white boys.

- Hollien *et al.* (1965) – Rainbow passage; Southern U.S. white boys.

- Ingrisano *et al.* (1980) – Children chosen were those best identified (over 90%) in a speaker sex identification task; children imitated a recording of a trained female speaker reading six sentences (of the declarative grammatical type), four of which were used in the analysis; F_0 measurements from syllabic nuclei. Average ages: 4:8 (4 to 5).

[18] They also cite Kaida and Eguchi & Hirsh (1969) as providing some evidence that boys' F_0 is lower than girls'.

[19] Researchers with experience of eliciting the speech of adults may argue that this does not apply solely to children.

- McGlone & McGlone (1972) – Experiment 1: one-word identification of each of set of eleven pictures. Average age: 6:0 (5:1 to 6:10). Experiment 2: same passage as Fairbanks *et al.* (1949a,b), recordings made individually in researcher's laboratory with one researcher present. Average age: 8:0 (7:6 to 8:6).

- Michel *et al.* (1966) – subjects taken from cheerleading conference; Rainbow passage read in a room ("prior to any cheering", p47); presumably all voiced parts analysed for F_0. Average age of all subjects: 16:4.

- Vuorenkoski *et al.* (1978) – Single, isolated, sustained vowel. Note, values given for F_0 in table (from Bennett (1983)) are different from those given in Hasek *et al.* viz. for girls 256, 256; boys 261, 247.

- Weinberg & Bennett (1971a) – 30 seconds of spontaneous speech.

This represents a very mixed bag of experimental methodologies used by different researchers and vocal responses elicited from the children. It therefore comes as no surprise when comparing the results within each age group that there is such disparity in average F_0. For example, the average F_0 of eight-year olds is given as being between 234Hz and 297Hz for girls, and between 235Hz and 294Hz for boys. Given this lack of agreement between studies, and the very wide range in reported average F_0, it is likely that there exist no systematic differences between pre-pubescent girls and boys.

Kent (1976:423) tentatively illustrated the developmental course of the mean fundamental frequency of both sexes from birth to adulthood using a number of sources of data (see Figure 3.4). He assumed there was no differentiation between pre-pubescent girls and boys and concluded there was a gradual decrease in SFF from approximately 300Hz to 260Hz between the age of three and puberty. The data show that a sex difference is apparent in average fundamental frequency only from the age of 13 onwards (Kent 1976:423)[20], with the average fundamental frequency for girls continuing its gradual decrease into puberty, and a much more rapid decrease for boys. During puberty, the male vocal apparatus matures much more rapidly than the female, the larynx expanding and the vocal folds lengthening (Luchsinger & Arnold 1965; Alexander 1971:249), causing boys to have much longer vocal tracts overall than girls (Smith 1985:63). In boys, the vocal cords grow by approximately 10mm, while in girls, this is only about 3-4mm. This sudden maturation accounts for the drop in F_0 experienced by boys (Kent 1976:423)[21].

B. The supra-laryngeal tract and the formant frequencies

Anatomical measurements of the lower face appear to indicate that there are no systematic differences in the supra-laryngeal vocal tracts of pre-pubescent girls and boys. For example, Krogman (1962 – cited Hunter & Garn 1972) says disproportionate sexual dimorphism in the face is postpubertal. In a study of the dimensions of the mandible in over 800 children, Walker & Kowalski (1972) found that upto about the age of twelve "mandibular morphology and growth in the two sexes exhibit remarkable parallelism" (p116)[22]. This indicates there is no disproportionate sexual dimorphism in the oral cavity. However, Bennett & Weinberg (1979:188) contend there is evidence, albeit indirect, for

[20]Minor differences may exist before then, especially given that girls mature earlier than boys, and given that children in general will begin puberty at different ages.

[21]The cause of the voice 'breaking' or 'cracking' is that the vocal folds often expand unevenly (Alexander 1971:249-50).

[22]While their study involved white children of European descent from a single US city, it is probably not unreasonable to suppose that this holds true for children in general.

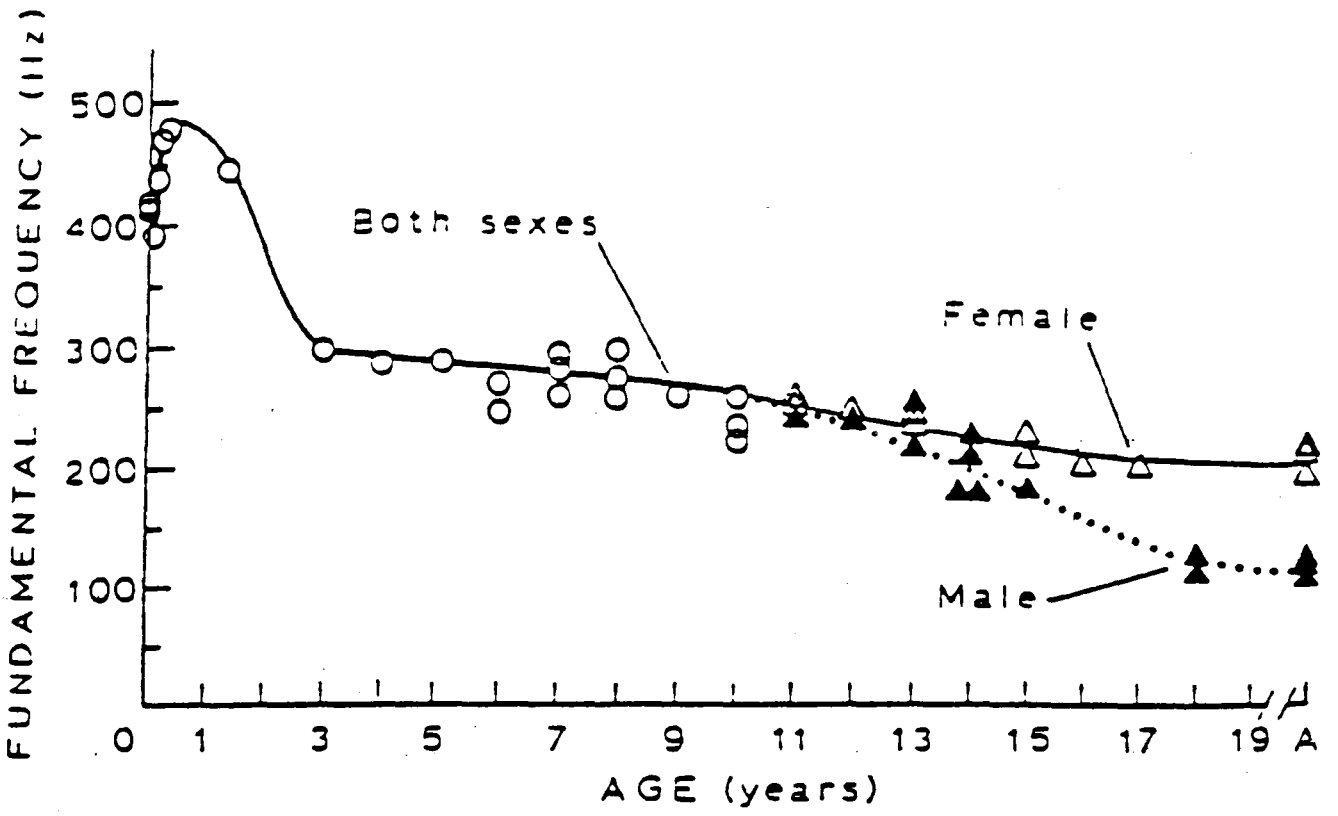


Figure 3.4: Developmental course of the mean fundamental frequency of both sexes from birth to adulthood. From Kent (1976:423).

sexual dimorphism in the lower part of children's vocal tracts. They re-examined cephalometric (i.e. head dimension) data from a study by King (1952) and found the pharynx in boys was on average 2-8% longer than in girls, between the ages of one year and ten years.

As with the results reported for children's fundamental frequency, the few studies detailing girls' and boys' formant frequency characteristics do not reveal a clear picture. From a group of seventeen children, Günzberger *et al.* (1987:51) found no systematic differences between the formants of the girls and boys for the three sustained vowels /aa/, /iy/ and /uw/. Concern was expressed, however, over the ability of the analysis technique (LPC) to accurately measure vowels with a high F_0 and low F_1. In contrast, Sachs *et al.* (1973:77-8) found that boys' formants were significantly lower than girls for sustained /iy/ and /uw/ vowels. More importantly, this result was obtained from their group of 9 pairs of children matched for height and weight. Assuming a direct correlation between general body size and the size of the vocal tract, and assuming there is no sexual dimorphism in vocal tract size, it is possible to conclude that either the girls or the boys, or both, are deliberately altering their formant frequency characteristics. Ingrisano *et al.* (1980:67), in their study, noted a tendency for the boys' first and second formants to be lower than the girls. Furthermore, the girls exhibited greater variability in their formants. However, they did not report the frequency values and carried out no statistical testing of their findings. Bennett & Weinberg (1979:182) found the boys' averages for F_1 and F_2 (950Hz and 2250Hz) were much lower than the girls' averages (1040Hz and 2350Hz) for five /ae/ vowels taken from a spoken sentence[23]. However, the standard deviations for either sex were well over 100Hz.

As stated above, there is a rapid maturation of the vocal apparatus amongst boys when puberty is reached. Anatomical measurements have indicated that the disproportionate sexual dimorphism found in the growth of the larynx is a feature of the supralaryngeal tract as well. Despite the pubertal growth spurt occuring earlier for girls (at about the age of

[23]They measured the formants of 30 girls and 23 boys aged between 6:1-7:10 years.

twelve, compared to fourteen for boys), the changes in the male vocal apparatus are much more marked: the average male vocal apparatus increases 10cm in length, while that of girls increases by only 3-4cm (Smith 1985:63-4)[24]. Furthermore, from their re-examination of King's (1952) data, Bennett & Weinberg (1979:182) found that at sixteen years, i.e. well into puberty for most children, the difference in pharynx lengths between girls and boys was approximately 13%. However, a study by Boerman (1967 – cited Hunter & Garn 1972) of fifteen to seventeen year-olds showed no disproportionate sexual dimorphism in the ramus (a vertical dimension at the rear of the mandible). Given that Hunter & Garn (1972) found the average difference in size of the ramus between women and men to be proportionately greater than for other facial dimensions, this result appears unusual in that the children in Boerman's study were old enough for the majority of them to have entered puberty and should be showing some evidence of this disproportion in their vocal apparatus. It is therefore possible that the increase in size of the ramus is greater for male speakers, but that this growth occurs in late puberty.

In summary, while there is no systematic difference in the dimensions of oral cavities of pre-pubertal girls and boys (as indicated by the size of the mandible), it is possible there is a sexual dimorphism in the pharyngeal cavity. It is reasonable to assume that these anatomical differences are translated into formant frequency differences. Indeed, on balance the formant frequencies of boys have been shown to be lower in general than girls. One other explanation for this, which will be explored in more detail in an examination of the perception of the sex of children in Section 3.2.1, is that children position their articulators to deliberately accentuate the sexual differences between them, i.e. girls want to sound 'more like girls', and boys want to sound 'more like boys'.

The vocal apparatus and the effects of aging in adulthood

Stoicheff (1981) reported that the average female SFF was relatively steady for speakers in their twenties, thirties and forties, and from the age of fifty onwards, with a drop in SFF inbetween. This finding is essentially in agreement with an unpublished study by Kelley (1977 – cited Stoicheff 1981:439), and with the studies by Benjamin (1986) and Krook (1988)[25] (see Table 3.14). Stoicheff examined this relationship further, looking at the influence of the menopause on average SFF. Her subjects reported whether they were premenopausal, were experiencing the menopause or had completed the menopause, and were grouped accordingly (see Table 3.15). Looking at women aged 40 to 59 years, she found that, while the average SFFs of premenopausal and menopausal women were approximately the same, for postmenopausal women it was substantially lower. Given that the average SFFs of pre- and postmenopausal women of all ages were roughly equivalent to the those in the smaller 40-59 year age group, implying that SFF is relatively steady before and after menopause completion, Stoicheff concluded that "it is not chronologic age but the age of menopause completion that influences the SFF" (p439). If we consider the SFF data reported by Benjamin and Krook, as well as that of Stoicheff, it would appear that the menopause is at least partly responsible for a drop of 15-20Hz in the SFF of women.

The physical effect of aging on the vocal apparatus produces changes such as ossification and calcification of laryngeal cartilages, thinning of mucosa, loss of tissue elasticity, muscle atrophy, reduced collagenous connective tissue, and reduced secretory function (see Kahane (1981) and Kent & Burkart (1981) for a review of the physiological changes

[24]Smith says this results in average lengths of 23cm for adult males and 17cm for adult women.

[25]Note that Krook's data show an increase in average SFF with advanced age, i.e. from the age of 70 onwards.

Source	20-29	30-39	40-49	50-59	60-69	70-79	80+
Benjamin (1986)	197	—	—	—	—	180	—
Krook (1988)	196	195	191	182	181	188	188
McGlone & Hollien (1963)	—	—	—	—	196	—	199
Saxman & Burke (1967)	—	196	189	—	—	—	—
Stoicheff (1981)	224	213	221	199	200	202	—

Table 3.14: Survey of longitudinal studies reporting mean SFF (Hz) by age for female speakers. Note some adjustment of the reported age ranges has been made in some cases to conform to the age ranges in this table layout – see Table 3.2 for the exact ranges.

	Pre-		Menopausal		Post-	
Age	*n*	SFF	*n*	SFF	*n*	SFF
40-59 years	14	216	14	220	10	193
All ages	53	218	14	220	44	199

Table 3.15: Mean SFF (Hz) for premenopausal, menopausal and postmenopausal women aged between 40 and 59 years (from Stoicheff 1981:439). The figures given for speakers of all ages are for comparison. All women in their twenties and thirties were premenopausal, while all women 60 and above were postmenopausal.

associated with aging). However, as well as these typical results of aging, the vocal apparatus of female speakers is also affected by the hormonal changes associated with the menopause, bringing about an "abrupt physical change" (Benjamin 1986:41). Gilbert & Weismer (1974 – cited Stoicheff 1981:440) suggested that the fall in SFF associated with the menopause is due to an increase in the testosterone-estrogen ratio which may alter the tissue of the vocal folds. Honjo & Isshiki (1980 – cited Stoicheff 1981:440) observed edema[26] of the vocal folds in 74% of women aged 69-85 years, which they suggested was caused by general endocrine change during the menopause. They posited that it is the resulting increase in vocal fold mass which causes the lowering of the fundamental frequency in women.

Stoicheff (1981:440) also suggests that the menopause is the cause of an age-related increase in SFF variability. She found the average standard deviation was significantly greater for the postmenopausal group ($p < 0.05$). She suggests this could be due to "decreased control over laryngeal adjustments for frequency of vibration of the vocal folds" (p440), perhaps arising out of the increase in vocal fold mass observed by Honjo & Isshiki. However, the general relationship between F_0 variability and age, for women and men, is less clear. The previous studies (Charlip 1968; Endres *et al.* 1971; McGlone & Hollien 1963; Mysak 1959; Wilcox & Horii 1980 – all cited Benjamin 1986:35-6) have produced conflicting results. Benjamin suggests this may be due to the different tasks and different measures of variability used.

The course of SFF change with age for male speakers is less clear than it is for female speakers (see Table 3.16). On the one hand, both Mysak (1959) and Hollien & Shipp (1972) found a substantial increase in mean SFF with advanced age (i.e. over 70 years). On the other hand, Wilcox & Horii (1980) found little difference between the mean SFFs of speakers in their early twenties and speakers in their sixties and seventies. Meanwhile, Benjamin (1986) found a slight decrease in mean SFF for speakers in their seventies.

[26] Edema is an excess accumulation of fluid in tissue or cavity.

Source	20-29	30-39	40-49	50-59	60-69	70-79	80+
Benjamin (1986)	110	—	—	—	—	103	—
Mysak (1959)	—	113			124		141
Wilcox & Horii (1980)	124	—	—	—	122		—
Hollien & Shipp (1972)	120	112	107	118	112	132	146

Table 3.16: Survey of longitudinal studies reporting mean SFF (Hz) by age for male speakers. Note some adjustment of the reported age ranges has been made in some cases to conform to the age ranges in this table layout – see Tables 3.3 and 3.4 for the exact ranges.

A number of studies have demonstrated that listeners are capable of relatively precise estimations of speaker age (Mysak 1959; Charlip 1968; Hollien & Shipp 1972; Shipp & Hollien 1969; Ptacek & Sander 1966; Ptacek *et al.* 1966; Hartman & Danhauer 1976; Helfrich 1979 - cited Künzel 1989:118), suggesting that age-relevant information is carried by the acoustic speech signal. Given the studies reported above, it would seem that F_0 is not used as a cue in determining a speaker's age, unless it is to differentiate between pre- and postmenopausal women, and perhaps younger and older men.

The vocal apparatus and the effects of height and weight

One might assume that physical attributes of a person such as their height and weight are well represented in the acoustic speech signal; in other words that general anatomical characteristics are in some way correlated with the physical size of the vocal apparatus, and are represented in the speech signal as discernible extralinguistic features. Laver & Trudgill (1979) concluded as much: "A tall, well-built man will tend to have a long vocal tract and large vocal folds. His voice quality will reflect the length of his vocal tract by having correspondingly low ranges of formant frequencies, and his voice dynamic features will indicate the dimensions and mass of his vocal folds by a correspondingly low range of fundamental frequencies" (p9 – cited Künzel 1989:118). However, the results of studies investigating the identification of speaker height and weight do not necessarily support this conclusion.

A number of speaker height and weight identification studies carried out by Norman Lass and his colleagues appeared to show that listeners were capable of relatively accurate estimates of height and weight, to within 1.4 inches and $4lb$ respectively (Lass *et al.* 1978a, 1979a, 1979b, 1980b, 1980c; Lass & Davis 1976 – all cited Graddol & Swann 1983:353). However, Cohen *et al.* (1980) demonstrated that Lass' promising results were a result of a twofold averaging process. The results were achieved by comparing the mean of the measured heights and weights of the speakers with the mean of the listener estimates. The identification performance of individual listeners was not considered. In fact, only four out of fourteen weight judgements deviated by less than $10lb$ from the measured physical data (cited in Künzel 1989:118). Graddol & Swann (1983:353) say that when the actual heights and weights were rearranged in ascending order, it was clear the sizes of tall or heavy speakers were underestimated and those of short or light speakers were overestimated (cited in Graddol & Swann 1983:353)[27]. Gunter & Manning (1982) used isolated vowels as stimuli (cited in Graddol & Swann 1983:353), thus excluding potential cues to identification, and found their listeners could not perceive speaker height and

[27]See also a reply by Norman Lass (Lass 1981).

weight accurately (cited in Künzel 1989:118)[28]. However, it remains possible that we are able to gauge *general* body size from the acoustic speech signal. Great accuracy has been observed in the ability of listeners to assign one of three characteristic body types to speakers (Bonaventura 1935; Fay & Middleton 1940; Moses 1940, 1941 – all cited Laver & Trudgill 1979:9).

Is there simply insufficient correlation between body size and the structure of the vocal apparatus, thereby imparting too few cues to the acoustic speech signal? Some height and weight identification studies have used stimuli with parts of the frequency spectrum filtered out to suppress potential cues, with significant effects on identification scores (Gunter & Manning 1982; Lass *et al.* 1980c, 1980d – cited Künzel 1989:118). A number of investigators have found no relationship between height and weight and SFF for either female or male speakers (Lass & Brown 1978; Majewski *et al.* 1972 – both cited Graddol & Swann 1983:353; Künzel 1989). Graddol & Swann (1983) found a correlation between male average SFF and height, although this was not particularly strong. Their results suggest the strength of this relationship is increased by more dynamic intonational characteristics. One of the stimuli they used was an isolated vowel phonated at the "lowest possible pitch" (p355), which they hoped would be a good indicator of the physical constraints of the larynx and therefore be related to body size. The male speakers showed a slight but significant correlation between this F_0 measure and height, whereas the female speakers did not, suggesting that the "women were not using a SFF that reflected the size of their vocal apparatus" (p364).

In summary, firstly there is little evidence that listeners can identify the height or weight of speakers of either sex; and secondly, there is not much evidence of any sex-related differences in how height and weight affect the acoustic speech signal, in particular for SFF. This latter conclusion is somewhat surprising, in that one might assume that the sexual dimorphism in height presumably would have great bearing on the size of the female and male vocal apparatus. This suggests that factors other than physical size play a significant role in shaping a person's SFF.

[28] According to Künzel, they found deviations from the measurements of upto 19*lbs* and 3 inches

3.2 The perception of speaker sex

Intuition tells us that the perception of the sex of a particular speaker is due to the perceived average pitch of that person's voice. Günzburger & de Vries (1989:143) cite fundamental frequency, and its perceptual counterpart pitch, as the main cue for distinguishing between speaker sex, the next most important cue being the formant frequencies (which have been investigated mainly for F_1 and F_2). However, little is known about the acoustic cues used by listeners to decide whether they are hearing a girl or a boy (Bennett & Weinberg 1979:179); while, as was shown in the earlier sections of this chapter, F_0 is not the only cue that has the potential to be sex-differentiating.

What follows is a look at two distinct groups of speakers – children before the effects of puberty, and mature adults – in an attempt to seek out the cues used in the perception of speaker sex. In Section 3.2.1 the apparent biological similarity of the vocal apparatus of children of different sexes is contrasted with the relatively high rates of accurate sex identification reported in the literature. Adult sex perception is looked at in Section 3.2.2. In both sections explanations of biology and acculturation are examined to explain the consistently high rates of speaker sex identification.

3.2.1 Pre-pubescent speaker sex recognition

Important clues to the mechanisms used for the successful perception of speaker sex come from studies into the identification of the sex of pre-pubescent child[29] speakers (Lieberman & Blumstein 1988). The justification for this statement arises out of the relatively high rates of accurate sex identification, despite the anatomy and physiology of girls and boys being very similar before the advent of puberty. While the identification rates are much lower than for adults, it would seem prudent to look closely at the reasons for why it is possible to tell girls from boys on the basis of their speech alone.

As was shown in Section 3.1.4, until the onset of puberty there are no significant differences in anatomy or physiology between children of similar ages[30]. Thus when they are matched for height and weight, we would expect girls and boys to exhibit no systematic differences in either average fundamental frequency or average formant frequencies (Sachs 1975; Robb & Saxman 1985 – cited Lieberman & Blumstein 1988:131). Consequently, one may presume that it would not be possible to accurately identify the sex of a child from her or his speech alone. However, as has already been stated, this proves not to be the case.

Overall, the studies suggest a child's sex can be determined from her/his voice with an accuracy of around 70-80%, far more than chance. This indicates that, for some children at least, their speech contains characteristics apparently in contradiction with their anatomical and physiological development, which "could point in the direction of culturally determined differences in men's and women's speech habits, which probably have taken root at an early age" (Günzberger *et al.* 1987:48).

Below, we will first review the child sex identification studies that have appeared in the literature, and follow this with a consideration of some of the possible factors accounting for the high rates of accurate identification.

[29] From now on, we will refer to prepubescent children simply as children (or girls or boys).

[30] Although there was a suggestion that the pharynx is longer in length in boys than in girls.

Source	Age range (years)	Lang. spoken	Sex	*n*	Mean ID score (%)	Pooled score (%)
Bennett & Weinberg (1979) (1) Phonated vowel	6:1–7:10	US Engl.	f	31	63	65
			m	26	68	
Bennett & Weinberg (1979) (2) Whispered vowel	6:1–7:10	US Engl.	f	31	65	66
			m	26	67	
Günzberger *et al.* (1987) (1) Normally-sighted listeners	7:6–8:9	Dutch	f	6	52	55
			m	11	57	
Günzberger *et al.* (1987) (2) Visually-impaired listeners	7:6–8:9	Dutch	f	6	66	60
			m	11	54	
Sachs (1975)	4:0–10:4 4:0–11:10	US Engl.	f	9	72	66
			m	9	59	

Table 3.17: Survey of studies on the identification of the sex of pre-pubescent child speakers using isolated vowels as stimuli. The pooled score in the final column represents the combined female and male identification rates. More information about each study's experimental conditions is summarised in Table 3.18.

Review of identification studies

We will now examine the studies into the perception of the sex of child speakers. The review is split into two parts, the first dealing with those studies which used isolated vowels as stimuli, the second part with studies which used longer sections of connected speech.

The use of isolated vowels as stimuli is of interest because typically the only cues available to the speaker's sex are relatively steady-state fundamental and formant frequencies. It is possible that some sex-dependent phonetic variation will be present in the vowel, but most studies try to control for this (e.g. by having practice sessions, or having phoneticians screen the stimuli – see for example Bennett & Weinberg 1979:180). Moreover, the length of a vowel stimulus is very short: even for the sustained vowels experimenters generally elicit from their subjects, the stimulus is only 0.5 to 1 second long. Thus the listener has only a short period of time in which to pick up on the (few) available cues.

The majority of studies, however, have used longer sections of speech, generally periods of spontaneous speech elicited in some way from the children, or sentences the children are prompted into reading. As well as containing more cues to the speaker's SFF and formant frequencies, connected speech contains many suprasegmental features which could act as cues to the speaker's sex, e.g. intonation patterns, formant transitions, syllable rate. Again, there are methods for controlling certain aspects of the stimuli the experimenter may wish to enhance or remove, depending upon the focus of the investigation. For instance, Ingrisano *et al.* (1980:63) had their child subjects imitate prerecorded sentences that were controlled for the prosodic features of time, stress and F_0, thereby hoping to minimise any sex-identifying prosodic cues in the stimuli they presented to their listeners.

A. Studies using isolated vowels as stimuli

The results of the child sex identification studies which used vowels as stimuli are presented in Table 3.17. The mean rates of child sex identification range from the 52% reported by Günzberger *et al.* (1987) for 6 girls, to the 72% reported by Sachs (1975) for 9 girls.

Günzberger *et al.* (1987) used two groups of listeners, the first group formed of normally-sighted adults, and the second of visually-impaired teenagers who were presumed to have above average auditory abilities (p52). The pooled results (i.e. the combined identification

Source	No. of listeners	Age (years)	Forced choice?	Type of stimuli elicited from each subject
Bennett & Weinberg (1979) (1) and (2)	29 f	17–24 (Av. 20:6)	Yes	Phonated and whispered /ae/ normal and monotone sentence
Günzberger *et al.* (1987) (1) Normally-sighted listeners	38	adult	Yes	5 sentences, 3 sustained vowels
Günzberger *et al.* (1987) (2) Visually-impaired listeners	14	14:6–19:5	Yes	5 sentences, 3 sustained vowels
Ingrisano *et al.* (1980)	76 f 48 m	adult	Yes	4 sentences
Meditch (1975)	209 f 183 m	18–23 18–23	–	Spontaneous speech (approx. 2min)
Murry *et al.* (1975)	8 f	adult	Yes	6 × 15sec of crying (pain, startled, hunger)
Sachs *et al.* (1973)	83	young adult	–	Short sentence (approx. 5sec)
Sachs (1975) (1) and (2)	56 f 19 m	young adult	Yes	(1) 3 sustained vowels (2) backwards sentences from Sachs *et al.* (1973) (approx. 5sec)
Weinberg & Bennett (1971b)	61	adult	–	Spontaneous speech (approx. 30sec)

Table 3.18: Experimental information related to Tables 3.17 and 3.19 – specifically, information about the listeners (all listeners were speakers of U.S. English, apart from the Dutch listeners in Günzberger *et al.*s' study), the type of test, and the type of stimuli used in the test. A dash (–) indicates the information was not given.

rate for girls *and* boys) from both groups of listeners reached statistical significance ($p < 0.05$), although the rates for girls listened to by the first group and boys by the second group did not (p51,53). Sachs (1975) took eighteen of the children from the Sachs *et al.* (1973) study and formed them into nine pairs matched for height and weight, thus hoping to control for general size differences as their subjects ranged so much in age. The pooled identification rate of 66% was significantly greater than chance ($p < 0.01$) (p157). While Bennett & Weinberg (1979) did not report the results of any statistical testing, the pooled identification scores from experiments using both phonated and whispered vowels were comparable to Sachs (1975), and involved a much larger sample of children. The identification scores from these studies suggest that we are able to perceive of the sex of children from vowel stimuli at a slightly better than chance level. However the group scores for girls and boys showed great between-study inconsistency. For example, in the Sachs study, the boys were far more likely to be identified accurately (72%) than the girls (59%); in the Günzberger *et al.* study, only the girls mean identification score reached the level of significance for the visually-handicapped listeners, and only the boys score did for the normally-sighted listeners.

The major cue used in the identification appears to be the formant frequencies. Bennett & Weinberg presented two types of vowel stimuli to their listeners: whispered vowels, which have no fundamental frequency present, and normally phonated vowels. They found no difference between the identification scores for the two vowel types, which suggests that the vowel formant frequencies provided the major cues to the sex identification, while the fundamental frequency had little bearing (Bennett & Weinberg 1979:183). Moreover, no systematic relationship was found between individual children's F_0 and listeners' correct identifications in Günzberger *et al.*s' (1987:51,53) study. They also found that the mean

F_0 of the children's vowels was almost the same for girls and boys (243Hz and 246Hz).

While it is apparent that the sex of children can be perceived from isolated vowel stimuli at a level greater than chance, the group identification scores discussed above are quite low and "mask potentially large variations in the prominence of sexual vocal attributes among individual children" (Bennett & Weinberg 1979:187). A closer examination of the individual identification rates reveals that some children have a very well-developed vocal sexual identity from a very young age, while other children remain sexually ambiguous, and still other children are consistently mistaken for the opposite sex. For instance, Sachs (1975) reported that six out of nine boys and four out of nine girls were correctly and consistently identified, while two boys and four girls were consistently thought to be of the opposite sex. The fact that the taller children (regardless of sex) were identified as boys and the shorter children as girls, suggests that the listeners "may have been influenced primarily by vocal tract size ... or by some other correlate of age of child in their decisions about the sex of the child" (p158). Bennett & Weinberg found nearly three-quarters of the girls and more than half of the boys were identically classed in each of the two vowel experiments. Given that the similarity between the girls' and boys' mean F_0s, it would seem that F_0 did not play a significant part in identifying the sex of those children.

B. Studies using connected speech as stimuli

Whereas the mean rates of child sex identification using vowels were in the range 55% to 70%, the rates if identification for connected speech ranged from 70% to 80% (the results of the studies are presented in Table 3.19). Clearly it is easier to identify the sex of children from connected speech than from isolated vowels, an unsurprising result given the variety of potential cues available and the greater number of speech sounds from which to make a judgement.

In a study using recorded samples of 30 seconds of spontaneous speech from 66 five- and six-year-old girls and boys (whose mean fundamentals were very similar), Weinberg & Bennett (1971b) reported a correct sex identification rate from 61 adult listeners of 74%. From sentences read by six girls and eleven boys, Günzberger *et al.* (1987) found a correct identification rate of 74% for their group of normally-sighted listeners, although the rate for their visually-impaired listeners was somewhat less at 65%. Both Weinberg & Bennett and Günzberger *et al.* reported little difference in the F_0 mean of either sex. Sachs *et al.* (1973) found an even higher rate using 12 girls and 14 boys aged between four and fourteen. From the same sentence spoken by each child, listeners accurately identified the child's sex 81% of the time. In their sample, the boys' average F_0 was significantly higher than that of the girls', indicating that F_0 was not used as a perceptual cue.

Sachs (1975) attempted to show that the cues to the identification of speaker sex were not limited simply to vowel formant patterns, and that perception is perhaps linked to an amalgamation of a range of cues spread over a child's utterance. It is possible that the sex-identifying cues in a child's voice lie in the suprasegmental information found in a whole utterance (e.g. intonation contour, formant frequency range, or in the temporal information associated with intonation, speaking rate), which could have been a factor in the Weinberg & Bennett and Sachs *et al.* studies, each of whom used stimuli that were of at least sentence length. Thus Sachs used the sentence recordings used by Sachs *et al.*, but presented them backwards to a new group of listeners to try to confuse the temporal cues. This resulted in the accuracy of identification dropping to 59%. It is however possible that the confusing nature of backwards speech renders any conclusions drawn from this experiment about the dominance of intonational or speaking rate cues over segmental ones somewhat contentious.

Source	Age range (years)	Lang. spoken	Sex	n	Mean ID score (%)	Pooled score (%)
Murry *et al.* (1975)	0:3–0:5	US Engl.	f	4	–	51
			m	4	–	
Meditch (1975)	3:0–5:0	US Engl.	f	6	74	79
			m	5	85	
Ingrisano *et al.* (1980)	4:0–5:0 (Av. 4:8)	US Engl.	f	19	–	71
			m	17	–	
Weinberg & Bennett (1971b)	5:0–6:0	US Engl.	f	29	71	74
			m	37	78	
Bennett & Weinberg (1979) (1) Normal sentence	6:1–7:10	US Engl.	f	30	69	70
			m	23	71	
Bennett & Weinberg (1979) (2) Monotone sentence	6:1–7:10	US Engl.	f	30	63	72
			m	23	81	
Günzberger *et al.* (1987) (1) Normally-sighted listeners	7:6–8:9	Dutch	f	6	76	74
			m	11	71	
Günzberger *et al.* (1987) (2) Visually-impaired listeners	7:6–8:9	Dutch	f	6	70	65
			m	11	60	
Sachs *et al.* (1973)	4:0–10:4 4:0–11:10	US Engl.	f	12	–	81
			m	14	–	
Sachs (1975) (backwards sentences)	4:0–10:4 4:0–11:10	US Engl.	f	9	69	59
			m	9	49	

Table 3.19: Survey of studies on the identification of the sex of pre-pubescent child speakers using connected speech (sentences or spontaneous speech) as stimuli. The pooled score in the final column represents the combined female and male identification rates. More information about each study's experimental conditions is summarised in Table 3.18.

Bennett & Weinberg (1979) presented listeners with normally-phonated and monotone sentences elicited from their sample of children and found that 73% of the girls and 87% of the boys were identically classed for the two sentence-types. Such high rates of consistency provide compelling evidence that the listeners were not relying on intonation patterns to convey the primary information about the children's sex. They also found that the girls were significantly better identified from the normally-phonated sentences as opposed to the monotone sentences, while for the boys it was the other way round. Taking into account the fact that on average the boys were much more accurately identified in the monotone condition than the girls, and the fact that "the presence of monotonicity brought about a small but significant increase in the number of male judgements in response to both boys and girls", it appears that a decrease in F_0 variation adds to the perception of a male voice quality. In other words, a less animated intonation could have been perceived as being more consistent with 'maleness'.

B. An explanation of the high rates of accurate identification

The size of the vocal tract cavities is highly correlated with the overall size of children (Smith 1985:60), and so Weinberg & Bennett, noting that the boys in their study tended to be larger on average (although the differences in size were small), inferred that the boys had larger vocal tracts. Thus they speculated that differences between the formant frequencies of girls and boys arise out of the boys having larger vocal tract dimensions (p1212), and may explain why children's sex can be perceived in their voices. However, there is evidence that listeners may be basing their judgements on perceived notions of femininity and masculinity, and identifying deeper-voiced children as boys. Sachs *et al.* (1973) paired off nine sets of girls and boys who were matched for height and weight and measured F_1 and F_2 for three vowels produced by each of them. They found that, in general, the most 'boy'-like voices (i.e. the best identified boys and worst identified girls) had lower formant frequencies, and the most 'girl'-like voices (i.e. the best identified girls and worst identified boys) had higher formant frequencies[31] (p79). Furthermore, Sachs (1975), using the same nine pairs of children, found that the taller (or older) children tended to be heard as boys, and the shorter (or younger) children tended to be heard as girls (p157-60). Sachs suggests that the wide range of children in the sample (from four to twelve years) may have guided the listeners' judgements, leading them to mislabel young children as girls, and old children as boys. Subjective evaluations of the vocal aspects of sex in adults have classified the female voice as 'immature' and 'child-like', and the male voice as 'mature' (see Meditch (1975:433) for references). Thus listeners may compare "the individual children's speech to an 'ideal type', their conception of the masculine and feminine speech patterns. Because adult males and females evaluate the male-associated pattern as being more 'mature', [the listeners] may have been listening for speech which was the least 'childlike' to label 'male'. They were listening for an absence of child speech features" (p433).

Sachs *et al.* argue that anatomical factors do not account for their finding that the boys from their nine matched pairs on average had lower formant frequencies than the girls (p77-8). Moreover, they argue that if sex differentiation is based solely on formants as related to head size, then all small children would be labelled girls, while all large children would be labelled boys (p155-6). They note that two of the girls in their sample who were consistently identified as boys, and who had formant frequencies close to the boys average, were neither the tallest nor the heaviest of the children. One of the girls was described by a neighbour as "athletic, strong, and competitive", and the other as "a tomboy, very sports-minded, a real tough kid but well liked" (p79), the inference being that the girls'

[31] Counter-intuitively, the 'boy'-like voices had the highest F_0s, while the 'girl'-like voices had the lowest.

personalities were more 'boy'-like, leading them to adopt vocal characteristics associated with masculinity. These two girls were also misidentified as boys in both test conditions in the Sachs (1975:168) study. Sachs (1975) also noted a middle-sized range of children of either sex who were consistently and correctly identified (p158). Thus Sachs *et al.* concluded that the formant frequency differences between girls and boys are accentuated by modifications to the vocal tract, either by pronouncing vowels with phonetic variations, or by altering the configuration of the lips. It is well known that the resonating chambers of the vocal tract can be altered in size and shape by moving the head, throat, jaws, lips and tongue.

Lieberman & Blumstein (1988:131) repeat Sachs *et al.*s' assertion, adding that boys "automise a set of speech motor patterns in which they round their lips to drive their formant frequencies lower". Certainly research by Mattingly (1966) suggests that adults deliberately alter their formant values. He reanalysed Peterson & Barneys' (1952) formant frequency data (see Table 3.10) and found that the degree of separation between the formants of adult women and men was greater than would be expected from vocal tract size alone. Children may be "modelling their behaviour on the average difference that exists between adult males and females" (Lieberman & Blumstein 1988:131). The proposition is that children undergo a process of gender conditioning which shapes their vocal characteristics as they mature, even before anatomical development takes this out of their hands. Smith (1985:61-2) considers that: "On the basis of what is currently known, it is at least ... plausible to hypothesise that pre-pubescent children are encouraged to conform to sex-associated behavioural norms, which encompass speech behaviour among many other things." Sexual stereotyping and conditioning imposes behavioural norms on children from a very young age, almost from birth (see Smith 1985:62-3 for a review of this area), defining their status as girl or boy. This affects a variety of behaviours, including speech characteristics. Children are aware of both the sex-related differences in speech, from their contact with adults, and the apparent importance of the gender roles they are expected to conform to. It is therefore entirely possible that they will adjust their voice quality accordingly.

Conclusions

The major acoustic cue to the identification of the sex of children appears to be the formant frequencies. The formants are the over-riding factor in sex identification from isolated vowels, although it is likely that other (suprasegmental) cues are also used when identifying sex from connected speech. The level of fundamental frequency does not appear to be a sex-identifying cue in children's voices. In most of the studies reviewed here, the mean F_0s of the girls and boys were very similar, and in one case the girls' mean F_0 was significantly *higher* than the boys'. This is consistent with the conclusions of the review of the studies of children's fundamental frequencies in Section 3.1.4. "Regardless of which position is taken on the relationship of *average* F_0 in male and female children, the extensive overlap in F_0 values across suggests that alternative explanations for identification accuracy should be explored" (Ingrisano *et al.* 1980:65). Direct backing for this statement comes from Weinberg & Bennett (1971b:1211), who found that the distributions of the mean F_0s of the twenty-five girls and twenty boys consistently identified correctly overlapped extensively. The intonation contour does not appear to be a primary cue, as the comparison of identification rates for monotone and normally-phonated sentences by Bennett & Weinberg (1979) shows.

However, the drawing of conclusions was not clear-cut. Significant factors clouding the outcome of the identification studies were the influence of sexual stereotypes on the adults

who were required to identify the sex of the children, and the extent of anatomical and social development of the children used as subjects. Looking at the identification scores for individual children, the sexual identity of some children is as prominent as it is in adults, while on the other hand, "the voices of many [children] do not yet evidence the pronounced sexual dimorphism that is so evident among adult men and women" (Bennett & Weinberg 1979:187). It is therefore more fruitful to examine the identification scores of individual children (particularly the consistently well-identified and misidentified children) to find the acoustic and biological correlates of sexual identity.

Finally, it was suggested that children alter their speech patterns in order to conform to gender stereotypes. As children are aware from a very early age of the given importance of males acting like 'men' and females acting like 'women', it is likely they will attempt to adopt gender-specific vocal characteristics in order to 'fit in'. The results reported here indicate that this happens as early as the age of three or four (see the identification scores from Meditch (1975) and Ingrisano *et al.* (1980) in Table 3.19). It is also likely that a number of acoustic cues are used in combination when deciding the sex of a child, although while children may learn from adults some of the prosodic differences applied by women and men, the results of the sex identification studies reported here suggest that the most significant alterations are in the formant frequencies. The general lack of anatomical sexual dimorphism in children means that it is only by making adjustments to the shape of their vocal tracts that they can attempt to mimic the formant frequency differences between women and men. Consequently, while speaker sex ID by children exceeds chance, the levels of accurate identification are not as impressive as those from the identification of the sex of adults (Ingrisano *et al.* 1980:62).

3.2.2 Post-pubescent speaker sex recognition

In adults then, there are very obvious sex-related differences in the size and shape of the vocal folds and vocal tract, resulting in approximate average differences of 90% between fundamental frequencies and 17% between formants (see Section 3.1). With the onset of the maturation of the vocal apparatus, so the accuracy of speaker sex identification increases such that near perfect scores are witnessed in most studies involving adult speakers. Sachs (1975) showed that the perception of speaker sex from isolated vowels spoken by children (see above) is little better than chance; contrast this with the findings of Schwartz (1968), Ingeman (1968), Schwartz & Rine (1968), RO Coleman (1971) and Lass *et al.* (1976, 1978b) below, which show very accurate identification of the sex of adult speakers from such isolated stimuli, and the importance of the anatomical differences between adult females and males becomes apparent. In fact, even when the speech signal is altered significantly, the accuracy of identification is retained, which would tend to show the durability of speaker sex cues (see Coleman (1971) and Lass *et al.* (1978b, 1980a) below).

It is likely then that once humans pass into adulthood, then the anatomical divergence caused by puberty allows some of the perceptual cues used to determine that speaker's sex to become more pronounced and others, especially pitch, to have effect. A number of researchers have attempted to ascertain the nature of these sex-differentiating cues to determine which ones are necessary and sufficient for correct identification. What follows is a review of the literature on such studies. To leave the text relatively uncluttered, the experimental details pertaining to each study - i.e. number, sex, age and geographical location of speakers and listeners, and the type of auditory stimulus used - have been placed in Table 3.20.

While perceived pitch is *probably* the most important cue, it is by no means the only one, and indeed is not even a necessary one. Both Schwartz (1968) and Ingeman (1968) recorded women and men articulating voiceless fricatives for use as auditory stimuli in their identification experiments (see Table 3.20 for experimental details). Despite the absence of the fundamental (due to the voiceless nature of these speech sounds), speaker sex identification was close to 100% accurate for certain of the fricatives in both studies, showing that F_0 is not a *necessary* cue for identification. Their results are summarised in Table 3.21. Schwartz carried out a statistical analysis on his data, which revealed that only the identifications for /s/ and /sh/ were greater than chance. Both Schwartz and Ingeman considered that a diminished formant structure caused the accuracy of identification to drop for certain of the fricatives, although the smallness of their samples mitigate against any binding conclusions. Schwartz posited that the relatively flat and broadband spectra for /f/ and /th/ were responsible for the failure to perceive clear-cut sex differences; Ingeman goes further, noting that accurate identification is reduced as the proportion of the vocal tract in front of the constriction diminishes - thus /h/ is most easily identified.

Schwartz & Rine (1968) and Coleman (1971) also presented findings supporting this conclusion. Schwartz & Rine had listeners identify sex from two whispered vowels, /iy/ and /aa/ (see Table 3.20 for experimental details). In whispered speech the fundamental is again absent, the excitation of the vocal tract being caused by the flow of air through the open glottis. Pronunciations of /aa/ and /iy/ produced correct identification of sex 100% and 95% of the time respectively. Coleman had female and male speakers produce /iy/, /uw/ and a prose passage using a laryngeal vibrator, through which the fundamental was maintained at a constant 85Hz for all speakers. Sex identification was in excess of 80% accurate, even though the artificial (and presumably highly intrusive) F_0 was nearly a third of that normally produced by women. Interestingly, the judgements of speaker

Source	No. of speakers	Age of speakers	No. of listeners	Age of listeners	Stimuli
Schwartz (1968)	9 female 9 male	young adults	10	–	4 isolated voiceless fricatives
Schwartz & Rine (1968)	5 female 5 male	–	8	young adults	2 isolated vowels: whispered
Ingeman (1968)	6 female 8 male	–	5 female 5 male	–	9 isolated voiceless fricatives
Weinberg & Bennett (1971a)	15 female 18 male	–	18	–	–
Lass *et al.* (1976)	10 female 10 male	19 - 24	15	–	6 isolated vowels: normal, filtered, whispered
Lass *et al.* (1978b)	5 black female 5 black male 5 white female 5 white male	–	30 white female	–	4 sentences: played forwards, backwards, time-compressed
Lass *et al.* (1979c)	5 black female 5 black male 5 white female 5 white male	–	10 female 10 male	–	4 isolated vowels, 4 monosyllabic words, 4 bisyllabic words, 4 sentences
Lass *et al.* (1980a)	5 black female 5 black male 5 white female 5 white male	–	14 female 14 male	18 - 28	4 sentences: unfiltered, low-pass filtered, high-pass filtered

Table 3.20: Summary of experimental conditions in various speaker sex identification studies.

Source	/h/	/sh/	/s/	/z/	/zh/	/th/	/ch/	/dh/	/f/
Schwartz (1968)		90	93			69			74
Ingeman (1968)	91	77	75	73	67	61	60	55	54

Table 3.21: Speaker sex identification rates (percentages) from presentations of isolated voiceless fricatives.

sex correlated highly (r = 0.70) with the speaker's average vocal tract resonances. In fact. the higher the formant frequencies, the more female-like the voice was judged to be, and vice versa. These findings, which bypass the perceptual information imparted by the fundamental frequency, show that a speaker's vocal tract characteristics are potent determinants of her/his sex.

Further evidence is supplied by a study using female and male adult esophageal speakers by Weinberg & Bennett (1971a). Their group of listeners correctly identified the speakers' sex 90% of the time. Using a correlation matrix they discovered that neither F_0, phonation time nor phonation rate accounted for this accuracy, rather that: "Spectral characteristics ... were highly related to the recognition performance of the listeners, suggesting that the judgements may have been related to some measure of vocal tract size" (p147).

The studies of RO Coleman (1976) and Lass *et al.* (1976) purport to show the greater importance of perceived pitch in judgements of speaker sex. In the first of two experiments, Coleman presented backwards the prose recordings of 20 female and 20 male speakers to his subjects. In this way he hoped to reduce the influence of any sex-related differences in rate, juncture and inflection, so that judgements would be made solely on vocal tract resonance and F_0 information. By means of rank-order correlation coefficients computed between measures of formant values, F_0 and judgements of the degree of femaleness and maleness in the voices (with degree of correlation indicative of the contribution of each of the vocal characteristics to listener judgements), he concluded that "listeners were basing their judgements of the degree of maleness or femaleness in the voice on the frequency of the laryngeal fundamental" (p173-4). However, his ranking scheme appears to have been based on incrementally ordering each person's fundamental, formant values and sex judgements from 1 to 40, with similar values being accorded the same rank. Thus no indication of the measure's absolute value is conveyed, merely a position in the that measure's 'league table'.

For his second experiment, Coleman chose 5 each of the female and male speakers to represent extremes in vocal tract resonances. Each then spoke a five second passage using a laryngeal vibrator, once with the frequency set at 120Hz (a typical male value) and another at 240Hz (a typical female value). Forced-choice sex identifications were then obtained from 18 female and 7 male listeners, the premise being that "consciously or unconsciously, [the listeners would] give preference to the more perceptually prominent of the two vocal cues" (p169). For the situation in which the low formant values (i.e. from the male speakers) were combined with the low F_0, and the high formant values with the high F_0, not unexpectedly the speakers were consistently identified as male and female respectively. However, the combinations of low formant values with high F_0, and high formant values with low F_0 showed that the male characteristics dominated in both cases. It may well be that the laryngeal vibrator used, designed for the male larynx, failed to produce a sufficiently natural sounding female fundamental which reduced its perceptual prominence. Coleman also noted the possibility of a sex-dependent listening behaviour on the female-dominated panel. With the combination of male F_0 and female vocal tract, the women labelled the voice male 80% of the time, compared to 45% for the men. Certainly, Coleman's conclusion that "in natural speech, the degree of male and female voice quality in the voice is a function of the frequency of the laryngeal fundamental", and that "individual vocal tract characteristics ... contribute little or nothing to the perception of this particular voice quality" (p179) is somewhat contentious, especially in the light of the earlier experiments by Schwartz (1968), Ingeman (1968), Schwartz & Rine (1968) and Coleman (1971) and later ones by Lass *et al.* (1980a).

Lass[32] *et al.* (1976) had 10 women and 10 men articulate the vowels /iy/, /er/, /ae/, /aa/, /aw/ and /uw/ in isolation in both normal (voiced) and whispered speaking conditions. A third set of stimuli was created by low pass filtering the speech at 255Hz to remove as much of the formant information as possible[33] (see Table 3.20 for experimental details). In this way they hoped to present stimuli in which either the vocal tract resonance characteristics or the fundamental frequency were absent so that their relative importance could be assessed. The fifteen listeners reported accurate identifications of 96% for the voiced condition, 91% for the filtered condition, and 75% for the whispered condition. The listeners were also required to indicate their confidence in making their judgements by means of a seven-point rating scale (Coleman 1971), ranging from *1* for a guess to *7* for complete confidence. Their mean confidence ratings were 5.6, 5.4 and 3.0 for the voiced, filtered and whispered tapes respectively. Lass *et al.* concluded that "the laryngeal fundamental appears to be a more important acoustic cue for speaker sex identification than the speaker's resonance characteristics" (p678). However, this contrasts sharply with the results reported by Schwartz & Rine (1968). Tartter (1989:1678) suggests that the differences in identification rates between whispered and phonated vowels may be in part due to the lack of experience in listening to whispered speech. Simple familiarity with the type of production might be an important factor in vowel identification (Broad 1976 – cited Kallail & Emanuel 1985). Nevertheless, Lass *et al.* used a much larger data set than Schwartz & Rine (1800 separate stimuli against 160). However, another argument that might be used against Lass *et al.*s' conclusion is that the quality of whispered speech is particularly degraded, and in particular the fact that the fundamental frequency is not the only intrinsic information removed from the vowel. Nolan (1983:125) says about whispered speech (in relation to the importance of F_0 in speaker recognition tasks), "clearly ... characteristic source-spectrum features, and possibly vocal tract transmission information, are ... being lost." Whispered speech spectra are flatter and have less overall power than normal ones (Tartter 1989:1678), and so many of the cues from the vocal tract may be lost.[34]

Of more interest then is the study by Lass *et al.* (1980a) which investigated the effects on speaker sex (and race) judgements of filtering portions of the broadband speech spectrum. The 20 speakers recorded a total of four sentences each, from which three master tapes were constructed: an unfiltered tape, a 255 Hz low-pass filtered tape and a 255 Hz high-pass filtered tape (see Table 3.20 for experimental details). Lass *et al.* found that sex identification was both high and consistent for all three conditions: the mean was 96.25%, with a range of 95.36% (for the low-pass filtered tape) to 97.01% (for the unfiltered tape). A similar, statistically insignificant range of differences was found for the female and male identification figures taken separately. And listener confidence in their decisions (based on the same seven-point scale outlined above), very nearly 7.0 in all cases. The test-retest

[32]Norman Lass carried out a number of experiments during the seventies to try to find the specific acoustic cues for the identification of speaker sex, as well as such factors as the race, height and weight of speakers. He has provided us with data on the effects on sex identification of whispered speech (Lass *et al.* 1976), filtered speech (Lass *et al.* 1976, 1980a), temporal speech alterations (Lass *et al.* 1978b), phonetic complexity (Lass *et al.* 1979c), and the race of the speaker (Lass *et al.* 1978b, 1979c, 1980a).

[33]This decision was based on the information provided by Peterson & Barney (1952), which showed that, for their population set, the lowest *average* value for F_1 is for a male /iy/ and occurs at 270Hz. One obvious criticism of such a decision is that an average value covers a range of different speakers, some of whom will be below and some above this level, and that this will apply to vowels other than /iy/ as well. In other words, formant information would be included in the stimulus. However, comparing the scores for /iy/ against vowels with much higher first formants, this would not appear to have affected the identification rates.

[34]Although not mentioned in the paper, Smith (1985:65) says the listeners were all female. Figure 2 in Lass *et al.*'s paper shows that male vowels were identified more accurately than female. Could this be an indication of sex-related listening behaviour?

agreements were also consistently high. Lass *et al.* conclude that these results indicate "that sex identification can be made accurately from acoustic information available in different portions of the broadband speech sp ctrum" (p110).

Lass *et al.* (1978b – cited Smith 1985:65-6) played three tapes of ten women and ten men each speaking four short sentences. Even though one tape was time-compressed to 40% of the original recording time and another was played backwards (the final tape was heard unaltered), speaker sex identification was nearly perfect under all three conditions (cf. the backwards condition in Sachs' (1975) experiment with childrens' voices). Lass *et al.* concluded that these results tend "to support the hypothesis that a perceptually realistic signal [i.e. one that at least approximates natural speech] is not essential for ... sex identification" (p289).

In an investigation of the effects of phonetic complexity on speaker race and sex identifications, Lass *et al.* (1979c) presented four kinds of auditory stimuli spoken by 20 speakers (see Table 3.20): isolated vowels, monosyllabic words, bisyllabic words and sentences. The 10 female and ten male listeners judged each stimulus for the sex and race of the speaker and rated the confidence in their decision using the same seven-point scale mentioned above. Lass *et al.* found no relation between identification rates and phonetic complexity, the scores being consistently high in every case. Overall, the mean identification rate was 97.6%, with little variation across the different types of stimuli. One might tentatively conclude that all the necessary cues for the identification of speaker sex are contained within a single segment. In addition, listener confidence in making the identifications was 6.90 across all four levels of phonetic complexity, ranging from 6.70 to 7.00. A proportion of the stimuli were retested to check for listener reliability, resulting in a 98% test-retest agreement.

Included in Lass's studies on the cues important to the perception of speaker sex were experiments on the cues for speaker race perception (Lass *et al.* 1978b, 1979c, 1980a). All three studies show that the fact that a speaker is black or white has no effect on the perception of speaker sex.

Smith (1985:66-7) raises a pertinent point about the manner in which the results of these studies have been reported, namely that in the majority of cases we are told only of the numbers of correct judgements. This can have the effect of obscuring "consensual listener judgements" (p66), those identifications the listeners consistently make, whether in error or not. While accuracy scores are useful from the point of view of enabling us to see the overlap between actual sex differences and vocal types, we are unable to discover the precise nature of vocal sex stereotypes causing incorrect identification. Obviously, the higher the accuracy score, the greater the degree of (accurate) consensus of judgement and reflection of true sex-identifying vocal features; and in a number of the studies reported here this is the case. However, the diminished scores reported by Sachs (1975) and Lass *et al.* (1976) "cannot automatically be interpreted as being due to the absence of sex stereotypes for these variables. Reliable patterns of sex classification that were inaccurate would have gone undetected" (Smith 1985:67). For example, the 75% reported by Lass *et al.* for the whispered condition indicates at least this much consensus among the listeners; but they could have been responding with 100% agreement as to what they *perceived* was the sex of the speaker if they each followed the same stereotypical vocal cue.

Summing up, the data show that speaker sex identification is an extremely accurate process, especially for adult speakers. But while there is significant anatomical stratification in adults, the data of Mattingly (1966) and the ability of humans to differentiate between the sex of children suggest that anatomy is not the only factor enabling us to perceive the sex of the person talking to us. Lieberman & Blumstein (1988:131) contend that while

English, RP	1.20
French	0.97
English, Middle Northern British	0.90
English, General American	0.88
Swedish (long vowels)	0.73
Dutch, standard	0.63
Dutch, Utrecht dialect	0.56

Table 3.22: Average normalising displacement (Bark) for female/male speakers, based on reported F_1 and F_2 values. The displacement factor represents the shift applied to the Bark-scaled female spectrum to bring it in line with the male spectrum.

adult males have, on average, longer vocal tracts than females, "this group difference does not necessarily typify an individual who wants to sound male. Boys and many adult men who have short supralaryngeal vocal tracts ... learn to round their lips and execute other articulatory manoeuvres that yield lower formant frequencies (Mattingly, personal communication [to the authors]; Goldstein 1980)." Thus, they believe there is a significant amount of social conditioning that defines the sexuality in our voices, such that gender is conveyed in the speech signal through these learned manoeuvres that shift the formant frequencies[35].

Further evidence for the influence of social conditioning affecting our speech patterns comes, indirectly, from Bladon (1985). For his speaker sex normalisation experiment he gathered data from a number of speech communities. He found that the normalising displacement factor necessary to map female onto male speech differed across both language communities and socioeconomic groups - see Table 3.22. In other words, putting aside the possibility that vocal apparatus differs across adjacent countries and socioeconomic class, there is clearly a factor other than anatomical differences shaping vocal characteristics. On the basis of this, Bladon considers it "inescapable that we must implicate a sex-linked, socially motivated, learned characteristic in some communities males speak more like (or, more unlike) females than considerations of sexual dimorphism alone would predict" (p36).

Edwards (1979 – cited Smith 1985:67-8) provides a more disciplined examination. He recorded 20 working class and 20 middle class ten-year-old girls and boys reading the same short passage. Overall accuracy of 83.6% correct sex identification was reported for the 14 student teacher listeners, but the errors in identification prove more interesting. The working class girls and the middle class boys were inaccurately identified more consistently than the other speech groups: we may infer that something about the speech of the working class group caused it to be perceived as more masculine, while the middle class group appeared more feminine. In a separate experiment, five of the listeners rated the voices according to some subjective speech-related criteria: they classified the working class children as having lower, rougher and more masculine voices overall. If we can assume that a regional accent is an indication of a more working class background, then this is borne out by the data in Table 3.22 which shows the distance between female and male speech is less for the Middle Northern British accent than it is for RP. Several researchers have commented on the tough masculine feel of so-called working class speech (e.g. Trudgill 1975). With regard to Edwards' study, it is perhaps possible that a degree of stereotyping was being used by the (probably) middle class student teachers, and whether

[35]They actually state that: "There is considerable overlap in the average fundamental frequencies of phonation for adult male and female speakers (Lieberman 1967)" (p131). Having searched through this reference there appears to be no trace of evidence backing up this claim. However, the data of Hudson & Holbrook (1981) (see Section 3.1) show that, during the normal course of speaking, this is so.

this obscures the sex perception issue. When using utterances as large as short passages the researcher has less control over the medium when trying to test a specific speech cue; the speaker has more obvious voluntary control over how the passage is read, bringing in more class-related speech inflections. As perceived by a middle class person, the speech of a working class child may well appear 'rougher' in comparison with her/his own.

On a different tack, but still related to the issue of learned speech patterns, is the study of Spencer (1988) into the ability of male-to-female transsexuals (MTS) to adopt female speech characteristics. The voices of 8 MTS and two control groups of 7 women and 7 men reading the first two sentences of the Rainbow Passage were played to 46 listeners, who were asked to identify the sex of each speaker, and to rate the 'femaleness' and 'maleness' of the voice on a seven-point scale. While the sex of the speakers in both control groups was identified correctly 100% of the time, the categorisation of MTS as female or male depended upon the fundamental being greater or lower than 160Hz, respectively. Moreover, there was a broadly linear relationship between the subjective ratings of voice sex quality and F_0, i.e. the greater the fundamental, the greater the perception of 'femaleness' in the voice[36]. Two points of interest arise from this. Firstly, it is apparent that consistent gender assignments can be made even when the scale ratings indicate a lack of the correct sex vocal quality. As Spencer says: "Evidently, a voice need not be entirely representative of males or females in general to elicit a particular sex judgement" (p38).

Secondly, as Spencer says: "If vocal fundamental frequency were the only cue used by this panel of listeners, all speakers having vocal fundamental frequencies of 160Hz or higher would have been perceived to have highly representative female voices" (p38). It would appear then that merely raising the pitch of the voice is not sufficient to convey 'femaleness', probably because of the information about sex transmitted by the larger male vocal tract cavities. Spencer notes that the four MTS speakers identified as female were all under 5 ft. 8 in. in height, indicating they have shorter vocal tracts: it is possible that this may have been combined with manipulation of the vocal tract shape (cf Mattingly 1966). They had also attempted to adopt what they felt was a female intonation pattern. Finally, despite professing to be satisfied with their female voices, it is clear that "not every male who wants to sound female is able to effect the change to a satisfactory degree" (p40).

3.3 Conclusions

The data reported upon in this chapter suggest there is more to the acoustic-phonetic characterisation of the female voice than an increase in fundamental frequency alone. It is clear from anatomical and acoustic measurements (see Section 3.1) that a number of voice source and vocal tract characteristics differ significantly between women and men. For example, Titze (1989a) suggests differential scaling factors for both the fundamental frequency and the mean glottal airflow, sound power, glottal efficiency and amplitude of vibration, while Fant (1975) suggests that the differences between female and male formant frequencies are dependent upon the vowel being articulated. It is also clear (see Section 3.2) that there is a socially-conditioned effect upon the speech signal that is intended to accentuate the differences between female and male speech. We learn to alter our vocal characteristics to conform to vocal sex stereotypes – women to sound more 'feminine', men to sound more 'masculine'. The degree of sexual dimorphism, or the distinguishability of women's and men's voices, and how it pervades all aspects of the acoustic speech signal, is attested to by the consistently high rates of correct identification,

[36] From a visual inspection of Figure 1 (Spencer 1988:37), this relationship would not, in general, appear to apply to the control groups, i.e. perception of 'femaleness' and 'maleness' is only loosely related to F_0.

whatever the experimental stimuli used, in sex perception tests (see Section 3.2).

It was shown in Section 3.1.4 that a number of physiological factors effect the acoustic-phonetic parameters of the voice. Smokers have been found to have significantly lower SFFs than non-smokers. In a comparison of 15 smokers and 15 non-smokers aged 30-54 years, Gilbert & Weismer (1974 – cited Stoicheff 1981:437) found the average SFF of the smokers was 164Hz, compared to 183Hz for the non-smokers. Furthermore, laryngeal examinations found vocal fold thickening in thirteen of the smokers, but only one of the non-smokers.

And yet the speech research community has failed to come to grips with this issue. It is clear from attempts at synthesis that a model of femaleness in the voice - i.e. the production of a naturally sounding female voice - has not been fully described. Discussing his speech synthesiser, Klattalk, Klatt (1987:756) says "in spite of an ability to modify [the synthesis parameters of] average F_0, F_0 range, spectral tilt, glottal open quotient, and breathiness [as well as scaling formant frequencies for different vocal tract sizes], a truly feminine voice quality remains elusive."

Similarly, automatic speech recognisers often perform poorly in the recognition of women's voices when compared to men's. Several studies have reported inferior recognition rates for women's voices. In the first of two experiments, Noyes & Frankish (1989) found a statistically significant difference between the recognition scores for females and males for a simple digit string recognition task. Their second experiment involved the more complex task of inputting destinations for parcels: again, poorer recognition rates for women were reported. Waterworth (1984) reported two studies in which female speakers' recognition scores were significantly below those of males. Doddington & Schalk (1981) reported that 6 out of the 7 recognisers they tested exhibited this same problem of reduced recognition rates for women. Pallett (1985:373) makes a telling remark: "Some speech recognition systems employ features that may have been optimised for ... adult males." In other words, these recognisers may have incorporated models of male laryngeal and supralaryngeal characteristics into their algorithms. Noyes & Frankish (1989:112) note pessimistically that: "Since the existence of a gender difference is now well-established, perhaps future speaker-dependent systems will have to be designed so that they recognise only male or only female speech, in order to achieve greater accuracy."

Thus while it is clear that we experience few difficulties in understanding the speech of either sex (see Section 3.1.3), and perceive with ease the sex of the speaker, the field of speech synthesis has failed to capture the essence of that difference and the field of speech recognition has failed to iron out the difference, one of the reasons being the lack of study into the effects of between- and within-speaker variability. This chapter has indicated the need for a more complex approach to the variability in the voice due to speaker sex.

To sum up:

1. The combined effects of biology and acculturation have ensured that there is a unique femaleness in women's voices, and a unique maleness in men's voices. In other words, the acoustic-phonetic characteristics of the female voice are, in general, different from those of the male. This ensures that the perception of speaker sex is a very easy task; and could even be described as a trivial task.

2. However, and somewhat in contrast to the first point, there is no simple relationship between the characteristics of female and male voices. The female vocal apparatus is **not** a scaled down version of the male vocal apparatus. There appear to be fundamental differences in the anatomy of both the voice source and vocal tract.

In addition, there is evidence that the vocal apparatus is used in different ways by women and men, principally to accentuate the acoustic-phonetic differences between them. Thus a speaker's sex is signalled in the voice by a number of parameters.

3. The notion of an 'average' or 'ideal' speaker typifying each sex is at best inadequate. Individual speakers do **not** conform to group averages. The group averages quoted in the literature are highly misleading, disguising as they do:

   - the wide range of values individuals utilise during everyday speech;
   - the different ranges of values adopted by individuals in response to different situations (speaking at a committee meeting, becoming animated on meeting a well-loved friend, responding to a stimulus from a computer);
   - the fact that while an individual may be close to their sex's average for one sex-differentiating parameter, they may be close to the opposite sex's group average for another.

Chapter 4

An Analysis of the Acoustic-Phonetic Markers of Speaker Sex

This chapter presents:

1. A method for the automated acoustic-phonetic and statistical analysis of a large speech database; and
2. The results of a large-scale study of the acoustic-phonetic markers of speaker sex.

The study used vowel phone data from the TIMIT CD-ROM, and investigated the acoustic-phonetic measures of fundamental frequency, the relative amplitude of the first harmonic and the formant frequencies for their correlations with speaker sex and their variability both between- and within-speakers. Also investigated are the practicalities of a large-scale analysis.

The method of investigation is discussed in Section 4.1. Here the choice of speech data is examined, and the data described; the signal processing techniques used to analyse the data and measure the values of the acoustic-phonetic measures are described; and the structure and contents of the TIMIT database are described in detail. The automated analysis procedure is described in Appendix B. Briefly, four main stages were involved in the analysis procedure: the preparation of a database of input speech (the requirements for which are fulfilled by the TIMIT database); the establishment of structures to control the analysis of the data; the formation of a database of analysed speech, including the extraction of data from the input database, and the signal processing of the data to measure its frequency characteristics; and the statistical analysis of the output database. Also included in the description of the third stage is an evaluation of the algorithms designed to measure the fundamental frequency, relative amplitude of the first harmonic and formant frequencies.

The results of the study are presented in Section 4.2 for the fundamental frequency, the relative amplitude of the first harmonic and the formant frequencies. Regrettably, only the fundamental frequency and relative first harmonic amplitude were subjected to a thorough analysis. A lack of time, combined with a lack of confidence in the results, led to a less thorough examination of the formant frequencies. Where possible the results are compared with the relevant data from the literature. The results are reported in the following form: an analysis of the overall data, an analysis of the data by phone, and an

analysis of the data by speaker variable (i.e. by age, dialect, etc.). Particular attention is paid to the distribution of the mean and the range of values produced by each speaker to facilitate the analysis of between- and within-speaker variability in Section 4.3. Also discussed in this section are the sex-differentiating potentials of the three acoustic-phonetic measures. The conclusions are presented in Section 4.4, which includes remarks on speaker characterisation in general, on the characterisation of speaker sex, and on the automatic analysis of speech databases.

Note that throughout this chapter reference is made to SPEAKER MEANS and SLICE MEANS, where a SPEAKER MEAN is the mean value produced by a particular speaker for a particular acoustic-phonetic measure, and a SLICE MEAN is the mean value of an individual speech slice for a particular acoustic-phonetic measure.

4.1 Method

This section describes the method used in the analysis of the data on the TIMIT database. The database consists of recordings of ten read sentences, from each of 420 speakers. The digitised acoustic waveform of each sentence is stored, together with its phonetic and orthographic transcriptions, in its own directory under a hierarchical directory structure. The analysis required the extraction and acoustic-phonetic analysis of particular vowel phones from the speech waveform files. The procedure developed to achieve this is detailed in Appendix B, while this section describes the background to the analysis of the data.

The segments of speech used as input data to the analysis are described in Section 4.1.1. The data comprised all instances of the vowel phones /aa/, /ae/, /ao/, /iy/, /uw/ and /ux/ contained in the TIMIT database, a total of nearly 16,000 speech segments. This section discusses the reasons why these particular vowels were chosen, the establishment of the core data set, and gives some statistics on the lengths of the speech segments and the number uttered per speaker. Section 4.1.2 describes the signal processing techniques used to analyse the input speech data and measure the values of the acoustic-phonetic parameters. Briefly, the fundamental frequency was measured using cepstral analysis; the relative amplitude of the first harmonic by locating the first two harmonics and comparing their amplitudes; and the formant frequencies using an algorithm developed by the Centre for Speech Technology Research (CSTR) at Edinburgh University. Finally, the source of the speech data, the TIMIT CD-ROM, is described in full in Section 4.1.3. The first part to this section consists of an outline of the structure of the database held on the CD-ROM, in particular the directory structure it uses to organise the speech waveform files; a description of the TIMIT notation used to label the phones, and adopted in this thesis; and the information it provides about the extralinguistic attributes of the speakers (e.g. age, height). There is also a discussion of why the type of speech data available on the database unavoidably limits the scope of the conclusions this thesis is able to draw.

4.1.1 Discussion of the speech data used in the analysis

A. An examination of the choice of data

The speech data chosen for this study were the vowels /aa/, /ae/, /ao/, /iy/, /uw/ and /ux/. There were two principal reasons for this choice of data, centered around the use of vowels, and the use of long segments, or slices, of speech.

The use of vowel phones as input data

The acoustic-phonetic parameters investigated in this study are all features of the frequency domain, and vowels tend to have more readily measurable frequency characteristics. The particular relevance of vowel phonation to the ease of measurement is that, on the one hand, the quasi-periodic vibration of the vocal folds provides a strong fundamental frequency component and prominent harmonic structure in the frequency domain, and secondly, the subsequent excitation of the vocal tract produces prominent resonances, realised as the formant frequencies. The frequency characteristics of vowels therefore are more easily accessible, and would thus present fewer problems of accurate measurement to an automated parameter estimation procedure.

Furthermore, for reasons which are explained below, the particular subset of vowels chosen for analysis were all relatively longer in duration. The criterion used in the selection of these vowels was that the large majority of segments should be over 1000 samples in length. The other vowels represented on the database tended to be much shorter, and were therefore deemed unsuitable for the large-scale analysis.

It should be noted that /ao/ is actually a diphthong, and as such its inclusion in the data set must be called into question. Unfortunately, it was only at a very late stage that the author realised its true nature. Diphthong formant frequency characteristics are much more dynamic than those of monophthongal vowels because diphthongs involve "a change in quality within the one vowel, ... [although, as] a matter of convenience they can be described as movements from one vowel to another" (Ladefoged 1975:69). They may also be described as movements from a vowel to a glide (Lieberman & Blumstein 1988:223). One would expect, therefore, that it would not be possible to find a central, steady-state portion of the diphthong from which to measure a representative, 'average' formant frequency. Furthermore, one would expect this intra-vowel frequency dynamism to be reflected in the s.d.s of the formant frequencies for each measured segment. However, relatively large s.d.s were not found in the formant frequency measurements of the /ao/ segments. Part of the explanation for this may lie in the fact that the second element of a diphthong "is often so brief and transitory that it is difficult to determine its exact quality" (Ladefoged 1975:69). It would appear possible then that the measurement window applied to each segment to capture its steady-state characteristics (see below) excluded this second, transitory element. In the light of this it was decided to retain the results for /ao/, rather than recalculate the data without them.

The use of long speech slices as input data

It was decided to sample a range of points throughout each vowel phone's length to provide a mean value representative of the phone's overall frequency characteristics, rather than sample a single value from the centre of the phone. There are a number of potential errors associated with the sampling of only the midpoint of a phone:

- Real-life phonation is full of perturbations, or fluctuations, in the nominally periodic

vibration of the vocal folds. Using single sample values means the measurement of the parameters is vulnerable to the frequency characteristics of the perturbation. Parameter values could therefore be unrepresentative of the phone as a whole. Jitter (fluctuations in the period of phonation) and diplophonia (a double glottal pulse where during phonation two pulses have appeared together) would cause problems of this type.

- Even in a segment of speech as small as a phone, the fundamental frequency is never, or rarely, single-valued. Intonation characteristics change rapidly during the course of a sentence, causing sudden or gradual rises and falls in F_0 within phones.

- A preliminary investigation of the TIMIT data revealed that a number of the phones began and ended with different but relatively steady values of F_0, with a rapid transition between the two values taking place in the middle of the phone. A sample from the midpoint of the phone could reflect any point in the transition phase, whereas sampling a number of points would be more likely to include the peripheral steady-states. Furthermore, this would yield a large s.d. for the phone's F_0, enabling such phones to be more easily detected.

The type of signal processing analysis used in this study required segments of speech of at least 1000 samples to satisfy the windowing constraints. Particularly for the analysis of fundamental frequency, which was based on performing cepstra, longer sections of speech include more waveform periods. This greater amount of information about the waveform's frequency characteristics causes more prominent, and therefore more easily measurable, features from the signal processing. For example, for the fundamental frequency analysis, this results in prominent cepstral F_0 peaks in the quefrency domain. The requirement for speech slices of this length is discussed in more detail in Appendix B.4.

It was further decided to include only the 'steady-state' middle portions of the vowel sounds in the analysis, i.e. to omit the beginnings and ends of the phones, in an attempt to reduce the coarticulation effects from adjacent phones in the sentences[1]. While it is unlikely that any vowel phone in connected speech is completely free of the influences of adjacent speech sounds, due to the positioning of the articulators in the formation of the preceding phones and the anticipatory movements of the articulators to produce the following phones, a preliminary analysis showed that the edges of phones were often so corrupted that the signal processing algorithms had great difficulty producing accurate results. For instance, a voiceless fricative can interfere with the voicing of the initial portion of a following vowel sound. The use of long speech segments enabled the relatively steady-state values from the middle of the phones to be measured.

B. Establishment of the core data set

Here we describe the removal of some of the speech slices from the analysis, either because their lengths were considered to be too short, or because the signal processing algorithms could not cope with them due to insufficient periodicity in their wavelengths[2].

[1] It should be noted that coarticulation effects are not necessarily limited to the initial or closing portions of speech sounds, and that the effects of articulating a phone are not necessarily limited to the phones it is adjacent to.

[2] Note that a further amendment to the core data set resulted from the shortening of slices with partially-corrupted wavelengths. A description of the procedure used in selecting which slices were to be reduced in length, and the numbers of slices involved, is given in Appendix B.6.

Phone	FEMALE					MALE				
	No. slices removed		No. slices analysed		Total	No. slices removed		No. slices analysed		Total
/aa/	15	(1.7)	844	(98.3)	859	35	(1.9)	1818	(98.1)	1853
/ae/	3	(0.3)	1126	(99.7)	1129	19	(0.8)	2468	(99.2)	2487
/ao/	43	(5.2)	782	(94.8)	825	90	(4.5)	1918	(95.5)	2008
/iy/	417	(19.5)	1720	(80.5)	2137	982	(21.2)	3643	(78.8)	4625
/uw/	23	(13.1)	153	(86.9)	176	72	(16.2)	372	(83.8)	444
/ux/	84	(18.5)	371	(81.5)	455	223	(21.4)	818	(78.6)	1041
TOTALS	585	(10.5)	4996	(89.5)	5581	1421	(11.4)	11037	(88.6)	12458

Table 4.1: Summary of the number (and percentage) of female and male speech slices removed and the number (and percentage) analysed for each vowel. The columns headed 'No. slices removed' list the slices removed from the analysis because they were considered to be too small (i.e. contained less than 1000 samples). The columns headed 'No. slices analysed' list the total number of slices considered for study. The columns headed 'Total' give the total number of slices on the data base for each vowel phone.

Removal of short slices

For the six vowel phones, the total number of slices available on the database was 18039, of which 30.9% were from female speakers and 69.1% from male speakers. Thus there were 2.2 male slices for every female slice. However, in an attempt to keep the above-mentioned coarticulation effects to a minimum, the shorter productions of the vowel phones were removed from the analysis. The procedure for the automatic removal of short speech slices is described in Appendix B.2. A fairly arbitrary cut-off length of 1000 samples was chosen for the rejection of the shorter vowels, although it was one that allowed a steady-state of sufficient length to be relatively free from coarticulation effects[3]. A digital sampling rate of 16000 samples/second was used for the TIMIT CD-ROM, with the result that vowels had to be at least 62.5msec in length. For a (female) F_0 of 200Hz, this meant at least 12.5 periods were available for signal processing. A (male) F_0 of 120Hz yields at least 7.5 periods. The numbers and percentages of slices removed and slices thus available for analysis are given in Table 4.1. A total of 2006 slices (or 11.1% of the total) were removed from the analysis.

Rejection of analysed slices

In addition to the slices removed because of their small length, a number of slices were rejected because they were unsuitable for signal processing. After the extraction of acoustic-phonetic parameters had been performed on the speech data, the output was thoroughly checked to ensure the signal processing software had performed its job properly (the checking procedure is described in Appendix B.6). This highlighted a number of phones which were judged to be insufficiently periodic to produce meaningful frequency characteristics, and which the analysis software failed to cope with. A total of 295 slices (or 1.6% of the total) were removed from the analysis. Table 4.2 lists the number of slices rejected, the total number of slices removed (i.e. the slices removed because of their length plus the rejected slices), and the total number of slices left to make up the core data set. Thus the core set of speech data, or the data used in the acoustic-phonetic analysis, consisted of a total of 15738 slices. of which 31.1% were from female speakers and 68.9% from male

[3]Note that a number of slices from the original pilot study consisting of slightly less than 1000 samples were nevertheless included. From a visual inspection of their waveforms, they appeared to suffer negligible coarticulation. Thus 14 additional female and 43 additional male /aa/ phones were included in the analysis.

Phone	FEMALE						MALE					
	No. slices removed		Total no. removed		Total no. analysed		No. slices removed		Total no. removed		Total no. analysed	
/aa/	14	(1.6)	29	(3.4)	830	(96.6)	24	(1.3)	59	(3.2)	1794	(96.8)
/ae/	30	(2.7)	33	(2.9)	1096	(97.1)	56	(2.3)	75	(3.0)	2412	(97.0)
/ao/	15	(1.8)	58	(7.0)	767	(93.0)	46	(2.3)	136	(6.8)	1872	(93.2)
/iy/	32	(1.5)	449	(21.0)	1688	(79.0)	58	(1.3)	1040	(22.5)	3585	(77.5)
/uw/	2	(1.1)	25	(14.2)	151	(85.8)	1	(0.2)	73	(16.4)	371	(83.6)
/ux/	5	(1.1)	89	(19.6)	366	(80.4)	12	(1.2)	235	(22.6)	806	(77.4)
TOTALS	98	(1.8)	683	(12.2)	4898	(87.9)	197	(1.6)	1618	(13.0)	10840	(87.0)

Table 4.2: Summary of the number (and percentage) of female and male speech slices removed and the number analysed for each vowel. Note the percentages given are of the total number of slices on the database. The columns headed 'No. slices removed' list the slices rejected because their waveforms were considered unsuitable for the acoustic analysis. The columns headed 'Total no. removed' list the total number of slices not used in the acoustic analysis (i.e. the number considered too small plus the number rejected). The columns headed 'Total no. analysed' give the total number of slices forming the core data set.

speakers. Hence 2.2 male slices were analysed for every female slice.

C. A description of the slices in the core data set

Here we give details of the mean lengths of the phones, and the numbers and types of phone uttered per speaker.

Length of the slices

Table 4.3 summarises the lengths of female and male speech slices in the core data set, and Figure 4.1 gives a visual representation. There was little difference between the lengths of female and male productions of the various phones, although in general the female slices were slightly longer. The shortest phone under consideration was /iy/, with an overall mean length of 97msec. This was reflected in the number of short /iy/ phones removed. The longest phone was /ae/, with an overall mean of 149msec. The mean lengths of the other phones were similar, with a mean for these four phones of 125msec.

Number of phones uttered per speaker

Table 4.4 gives the number of phones of each type per speaker in the core data set. These observations are important for the statistics carried out on the acoustic-phonetic parameter means for each speaker, as it gives some idea of the numerical basis for the means. The table shows there was little difference between the average numbers of each phone uttered by the female and male speakers, and that on average the /uw/ and /ux/ phones were not well-represented per speaker. Moreover, only 64% of the speakers used the /uw/ phone in their read speech. Interestingly, while this did not appear to be linked to dialect region (insofar as the regions can be considered heterogeneous), nine of the ten black female and all of the twelve black male speakers used /uw/, while only 60% of the white speakers did.

Phone	Mean slice length (s.d.)			
	FEMALE		MALE	
/aa/	129	(43)	123	(38)
/ae/	152	(46)	147	(42)
/ao/	131	(43)	125	(40)
/iy/	99	(34)	96	(31)
/uw/	116	(46)	120	(48)
/ux/	126	(47)	119	(44)
TOTALS	123	(46)	120	(43)

Table 4.3: Summary of the lengths of speech slices (in msec) used in the analysis by vowel phone. Note this is for the core data set only, i.e. it includes only samples of 62.5msec or more.

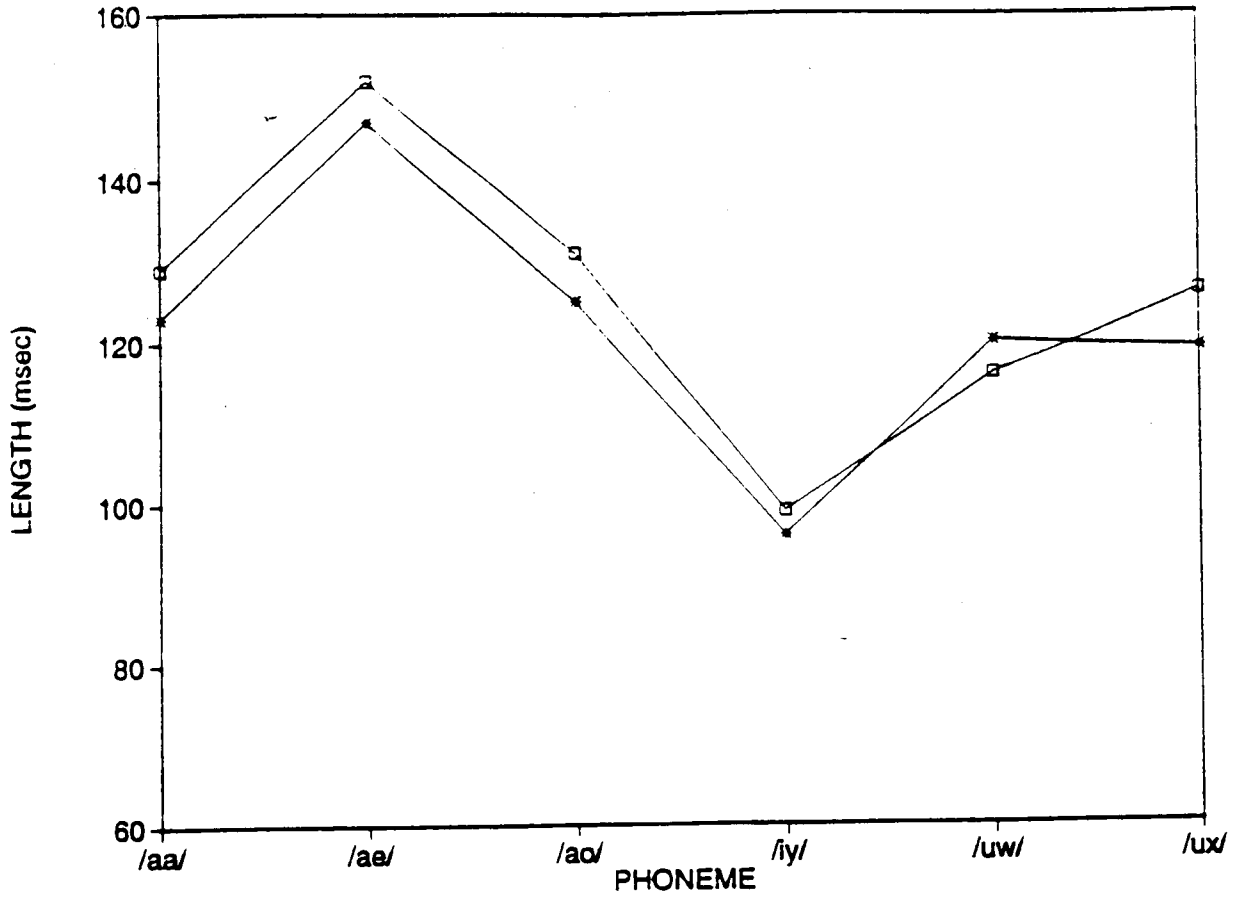


Figure 4.1: Lengths of analysed slices for female (asterisks) and male (squares) speakers.

Phone	FEMALE				MALE			
	n	Mean	(s.d.)	Range	n	Mean	(s.d.)	Range
/aa/	130	6.4	(2.2)	2–12	290	6.2	(2.3)	1–14
/ae/	130	8.4	(2.5)	3–16	290	8.3	(2.4)	3–16
/ao/	130	5.9	(2.4)	1–11	290	6.5	(2.2)	2–14
/iy/	130	13.0	(3.6)	6–28	290	12.4	(3.9)	3–28
/uw/	76	2.0	(1.3)	0–7	186	2.0	(1.2)	0–7
/ux/	128	2.9	(1.4)	0–7	280	2.9	(1.4)	0–7
ALL	130	37.7	(5.8)	24–55	290	37.4	(5.7)	22–54

Table 4.4: Summary of the number and type of phones uttered per speaker on the TIMIT database. The column headed 'n' lists the number of speakers who spoke at least one instance of the phone. The column headed 'Mean' lists the mean number of each phone uttered per speaker. The column headed 'Range' lists the smallest and largest number of each phone uttered per speaker.

4.1.2 Techniques for measuring the acoustic-phonetic parameters

This section outlines the signal processing techniques used in the measurement of the acoustic-phonetic parameters. The fundamental frequency was measured using cepstral analysis. A cepstrogram was produced for each slice, and a simple peak-picking algorithm located the cepstral F_0 peaks in the quefrency domain. The F_0 peak represents the slice's fundamental frequency. The relative amplitude of the first harmonic was measured by locating the peaks of the first two harmonics in the frequency domain, again via a simple peak-picking algorithm, and subtracting the amplitude of the first harmonic from the second. The formant frequencies were measured using a formant frequency estimator developed at the Centre for Speech Technology Research (CSTR), at the University of Edinburgh. The implementation of these techniques into automated parameter measurement software is described in Appendix B.4, and an evaluation of the software's accuracy is presented in Appendix B.6.

A. Cepstral analysis

A definition of the cepstrum

In Autumn 1959, while researching into seismology, Bogert observed quasi-periodic ripples in the spectrograms of seismic signals, and noted that this was characteristic of any signal's spectrum and its echo. Tukey then suggested that by taking the inverse spectrum of the log of the spectrum, the 'frequency' of the ripples could be determined. To provide a name to describe this procedure, he flipped round the first four letters of 'spectrum', and came up with the term 'cepstrum'. In the paper they subsequently published (Bogert *et al.* 1963), they also coined the terms 'quefrency', 'saphe', and 'rahmonic' to describe the cepstrum's parameters. While they found this technique unsuitable for seismic signals, Schroeder read the paper and realised that the spectra of voiced speech also contained ripples. Then Noll (1964) published a paper detailing his proposal for an algorithm utilising cepstral analysis for the detection of voiced and unvoiced segments of speech (see Noll 1967).

The basic model of voiced speech production consists of an acoustic chamber (the vocal tract) excited by a quasi-periodic train of pulses (the waveform emanating from the voice source)[4]. This interaction of two systems is reflected in the frequency domain, whereby

[4]Note that for unvoiced speech production, the vocal tract is excited by random noise.

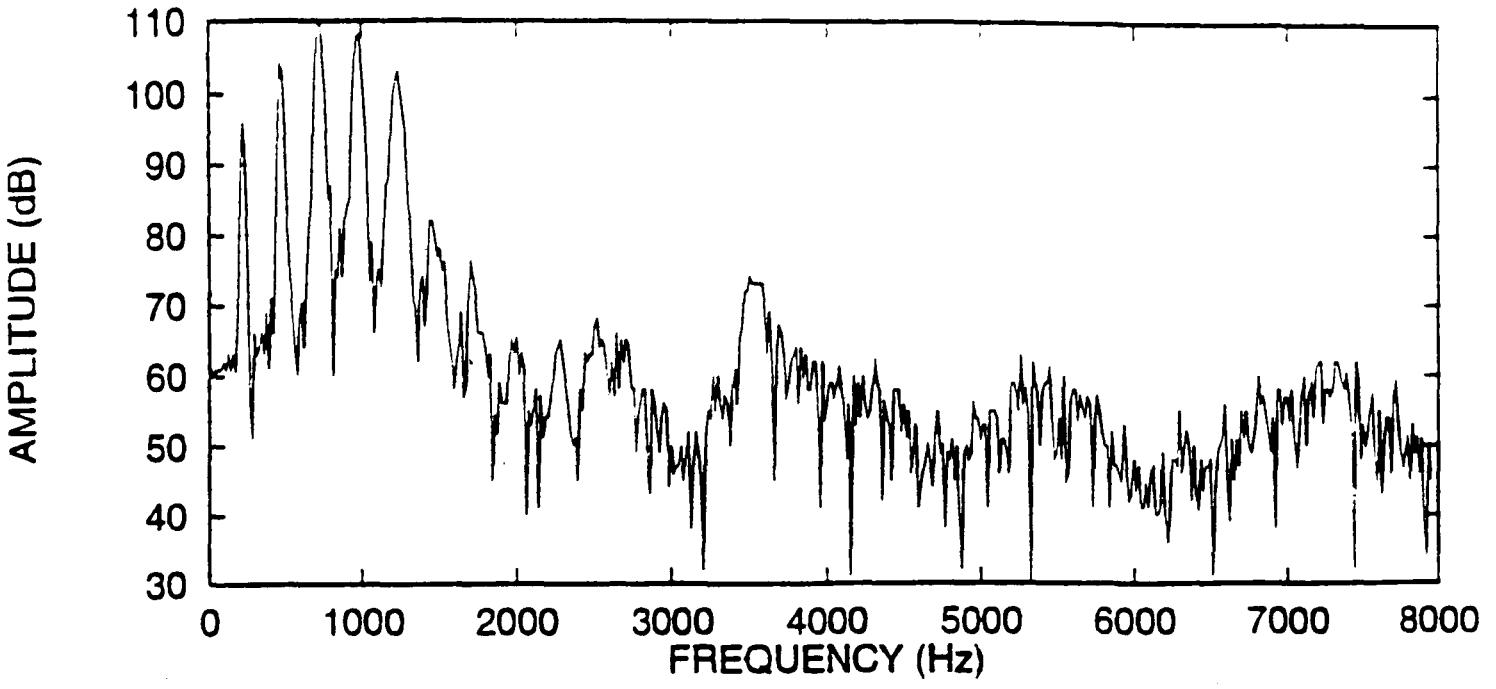


Figure 4.2: Spectrum of a segment of voiced speech (from the centre of one of speaker **fclt0**'s productions of an /ao/ vowel). The fundamental frequency for this segment is approximately 250Hz. Size of analysis window is 1024 samples (64msec).

the harmonics of the fundamental frequency (arising from the vibrating vocal folds) are superimposed onto the spectral envelope (or frequency characteristics) of the vocal tract (which is formed from the vocal tract's resonant frequencies). If we consider the spectrum of a segment of voiced speech shown in Figure 4.2, we can see that the harmonics create ripples along the frequency axis with a 'period' equal to the fundamental frequency, which in this case is approximately 250Hz. The rippling effect is particularly evident in the first five harmonics, upto 1250Hz. The 'frequency' of these ripples is known as the quefrency (whose units are those of time), and is equal to the period of F_0. Noll's hypothesis was that the 'periodicity' of the harmonic structure would allow the two elements of the speech production system to be separated out using cepstral analysis.

Mathematically, voiced speech sounds may be considered as being the result of a convolution of the voice source with the vocal tract response. Thus their separation is a deconvolution of these two signals. If $h(t)$ is the impulse response of the vocal tract and $g(t)$ the quasi-periodic glottal source signal, then the convolution of these terms to produce the output speech signal $s(t)$ is expressed as:

$$s(t) = \int_{-\infty}^{\infty} h(\tau)g(t-\tau)d\tau$$

and the frequency domain representation is:

$$S(f) = H(f)G(f) \tag{4.1}$$

Computing the cepstrum of a voiced speech segment carries out the separation, a technique more generally described as homomorphic deconvolution[5]. The cepstrum is derived from

[5]Theoretically, homomorphic systems obey a generalised principle of superposition, whereby the principle of superposition obeyed by linear systems is applied to nonlinear systems (see Oppenheim (1967) for a detailed discussion of generalised superposition). A homomorphic system for deconvolution takes inputs defined by convolution and transforms them into an additive combination of corresponding outputs, thereby achieving the required separation of components. Oppenheim & Schafer (1968) and Oppenheim *et al.* (1968) cover the mathematics required for the application of this technique to speech signals, while Rabiner & Schafer (1978:335-85) provide a thorough summary.

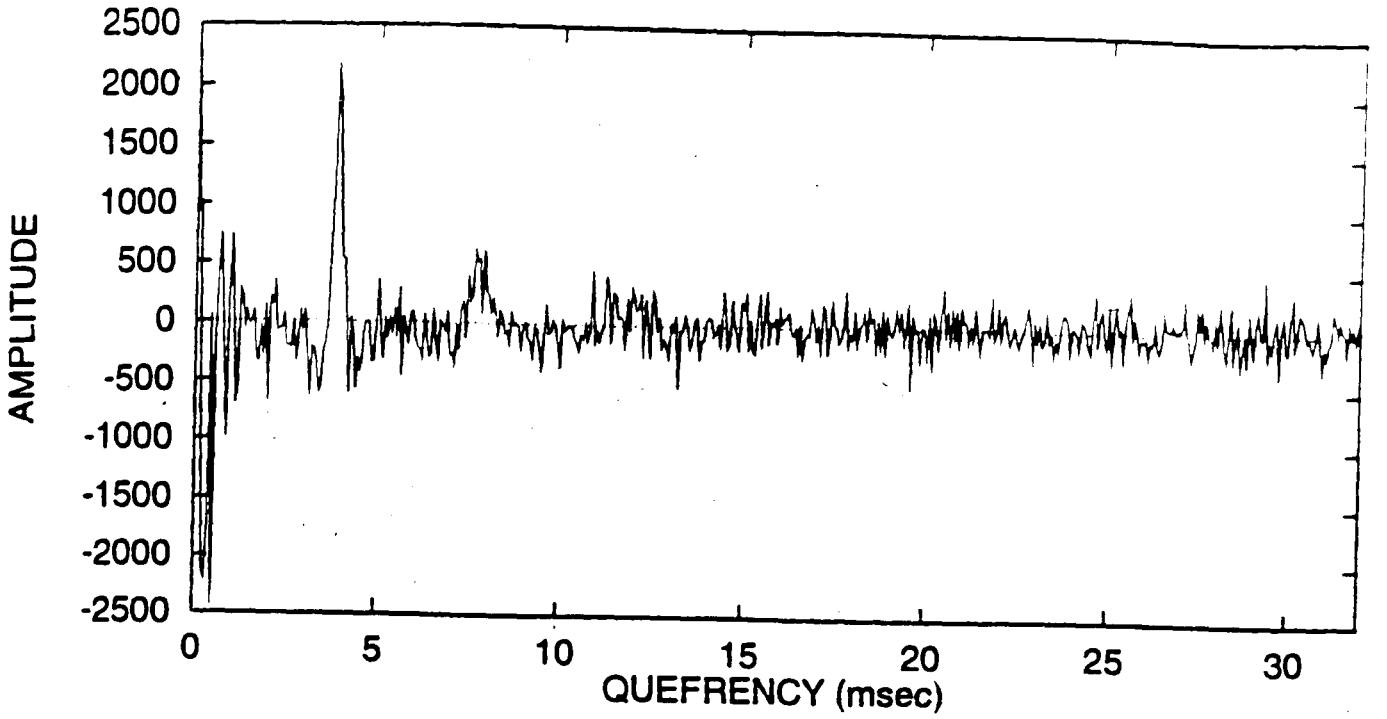


Figure 4.3: Cepstrum of a segment of voiced speech.

Equation 4.1 by first taking the magnitude and logarithm:

$$\log |S(f)| = \log |H(f)| + \log |G(f)|$$

and finally by performing an inverse Fourier transform to yield:

$$C_S(\tau) = C_H(\tau) + C_G(\tau) \tag{4.2}$$

where $C_S(\tau)$ is the cepstrum of the speech signal $s(t)$, and $C_H(\tau)$ and $C_G(\tau)$ are the cepstra of the vocal tract impulse response and glottal source respectively.

The cepstrum of the speech depicted in Figure 4.2 is shown in Figure 4.3. The slowly varying envelope of the spectrum (roughly, its formant structure) has been transformed into the low quefrency peaks at the extreme left hand end of the cepstrum ($C_H(\tau)$ in Equation 4.2). In contrast, the rapidly varying harmonic structure of the spectrum has been transformed into a single, easily isolated peak ($C_G(\tau)$ in Equation 4.2) at a quefrency of approximately 4msec. The quefrency of this peak is equivalent to the period of the fundamental frequency, and therefore F_0 is simply its inverse.

Advantages of using the cepstrum

For a long time, cepstral analysis has been one of the best methods for determining F_0 (Hermes 1992:4; Parsons 1986). This is primarily due to the soundness of the speech production model upon which it is based: in general, the ripples caused by the harmonics of the fundamental frequency behave in a sufficiently periodic way to produce a strong F_0 peak in the cepstral domain.

For relatively noise-free speech, an F_0 tracking algorithm based on cepstral analysis is fairly easy to implement on a computer, using the Fast Fourier Transform (FFT) algorithm developed by Cooley & Tukey (1965). Voiced speech recorded in noise-free environments, such as that contained on the TIMIT database, yields a strong, easily detected cepstral F_0 peak. This can be located by a simple peak picking algorithm[6]. Obviously, computing cepstra is expensive in terms of processing time, as two FFTs have to be performed, but the robustness of the method is more of an advantage than its slowness is a disadvantage.

[6]Noll (1967) claimed that cepstral analysis is resistant to noise, arguing that additive narrowband noise

B. Measurement of the relative amplitude of the first harmonic

Choice of an acoustic correlate of breathiness

Studies examining the degree of breathiness in speech have used the relative amplitude of the fundamental component in the frequency spectrum as a measure, or acoustic correlate, of this. More specifically, they have used the difference in amplitude between the first and second harmonics[7] . Klatt & Klatt (1990:828-9) examined this and other measures of relative amplitude (H_1 relative to A_1, and H_1 relative to overall rms amplitude of the spectrum) and found that, for group measures, there is little to choose between them. There are of course problems in using any of these measures: H_2 is partly dependent upon the location of zeroes in the source spectrum; A_1 is hard to estimate because there is no guarantee that the formant will be defined by a harmonic centred on the formant frequency; the rms amplitude is primarily dependent upon the amplitude of F_1 and therefore suffers from the same unpredictable variability (Klatt & Klatt 1990:854). However, Klatt & Klatt (1990:828-9) consider that the difficulties of obtaining accurate measurements of H_2 outweigh the difficulties of measuring A_1 or the overall rms amplitude. Further backing for the use of H_1–H_2 comes from Bickley (1982 – cited Henton & Bladon 1985:222), who found that the first harmonic amplitude was consistently enhanced relative to the other harmonics when comparing breathy and non-breathy vowels in Gujarati. Perceptual tests showed the salience of this acoustic correlate.

Since all four studies from the literature examining breathiness (see Section 3.1.2) used the H_1–H_2 measure to assess the relative amplitude of the first harmonic (and therefore to compare the incidence of breathy voice in women and men), it seemed pertinent to follow suit. Thus software was developed to locate and measure the amplitudes of the first two harmonics, involving a simple peak-picking algorithm.

Choice of vowels for analysis

Only the open vowels, /aa/, /ae/, /ao/, were used for the harmonic amplitude difference analysis, the reason being that only in the open vowels is the frequency of the first formant high enough to avoid interfering with the lower harmonics (Henton & Bladon 1985:223). Indeed, in some vowels F_1 can even lie below the second harmonic (Ladefoged *et al.* 1988 – cited Nittrouer *et al.* 1990:766). The three other vowels, /iy/, /uw/, /ux/, were omitted from the analysis as it was possible that interaction with F_1 would increase the amplitudes of the first two harmonics. This is in line with the previous studies of relative first harmonic amplitude: Henton & Bladon (1985:223) used /ae/, /ah/, /ax/, /er/; Klatt & Klatt (1990:826) used /aa/; Günzberger (1991:63) used /ae/; and Nittrouer *et al.* (1990: 766) used /aa/.

---

obscures only a few spectral peaks, while lower levels of additive noise have the effect of filling in the gaps between the peaks, not destroying them. However, Parsons (1986) argues that additive noise (i.e. noise from a source other than the speaker's vocal apparatus) causes the performance of the cepstrum to deteriorate rapidly, as the signal now being analysed is $H(f)G(f) + N(f)$ (c.f. Equation 4.1), where $N(f)$ is the noise spectrum. In other words, the multiplicative spectral property upon which the cepstrum depends has been lost. The low-lying portions of the spectra of noisy speech are filled with noise, limiting periodicity, and the cepstral F_0 peak will be broadened and corrupted.

[7] As before, this measure will be referred to as the relative amplitude of the first harmonic or H_1–H_2, where H_1 is the amplitude of the first harmonic or fundamental component, and H_2 is the amplitude of the second harmonic. The amplitude of the first formant will be referred to as A_1.

C. Formant frequency estimator

For the measurement of the formant frequencies, an automatic estimator (version 1.2, dated 27-4-90) designed by Alan Crowe and colleagues at the Centre for Speech Technology Research (CSTR), University of Edinburgh, U.K., was used. The principles behind the estimator are briefly described below (see Crowe 1988 for more details).

A description of the principles behind the CSTR formant frequency estimator
Contrary to the impression often given in the textbooks, the construction of a formant frequency estimator is not easy (Crowe 1988:683). Although the spectral analysis of a segment of speech will often reveal three distinct peaks representing the first three formant frequencies, generally the situation is much more complex. Spurious peaks are likely to appear, due to, for example, the harmonics of the fundamental frequency or from nasal resonances, while a common occurrence is for two formants to be so close together that they merge into one peak. Linear prediction spectra or cepstrally smoothed spectra can be used to reduce the effects of spurious peaks, but these naturally reduce the resolution of the spectrum and make the problem of merged formants worse.

The CSTR formant frequency estimator performs its task in three stages: spectral estimation, transformation from continuous to discrete data, and tracking[8]. Crowe considered that the transformation stage used in other estimators was inadequate. Generally the transformation involves simply locating the (smoothed) spectrum's local maxima, leaving the spectral estimation and tracking stages to compensate for the inadequacies of this (for example, the spectrum can be adaptively smoothed to reduce the number of spurious peaks before the local maxima are found). Crowe tried a different approach to the transformation using the principle of nonlocality, whereby human beings are able to decide whether a particular local maximum is a spurious peak, a true formant, or indeed two merged formants, by considering the layout of the spectrum as a whole. This is tackled in the estimator by using a technique called generalised centroids, where the spectra under analysis are treated as multi-modal distributions, and where individual local spectral maxima can be considered as being composed of more than one peak.

Evaluating the performance of the formant frequency estimator
No formal evaluation of the formant estimator's performance has been carried out, although it appeared to be reasonably insensitive to the FFT parameter settings[9] (Alan Crowe, personal communication). While its performance was checked for two male speakers, producing 'plausible looking' formant tracks, it performed badly on the one woman's speech examined. Clearly there was a need to assess the estimator's performance under different conditions, particularly its reaction to different phones and to female speech. This is discussed below, while the evaluation exercise is reported in Appendix B.6.

The estimator allows the user to define a search space, consisting of lower and upper frequency bounds, within which it will locate its three best choices for F_1, F_2 and F_3. However, different phones produce a range of formant frequencies, and so care must be taken to set the bounds of the search space so that only the first three formants are included. For example, the lowest F_1 value from Peterson & Barneys' (1952) data is

[8]Note that here, 'tracking' does not imply that the computations are based on information about formant locations from previous frames. The CSTR estimator calculates its formant frequencies on a frame-by-frame basis, so that the estimator can respond to rapid transitions. Here, tracking refers to the search for the true formants in a single analysis frame.

[9]The settings used were a pre-emphasis of the speech waveform (6dB/octave), followed by a 512-sample Hamming-window.

270Hz for males, and the highest F_3 value is 3310Hz for females (see Table 3.10). If the lower limit is too low, the search space may include the first harmonic (i.e. the fundamental frequency component of the spectrum), especially if the input speech has a high F_0; if the upper limit is set too high, the search space may include F_4. If either situation occurs, the estimator may become confused, for example computing as F_1 an amalgamation of F_0 and F_1, or picking out F_4 as F_3. Furthermore, if the bounds of the search space are set such that one of the target formants is excluded – i.e. F_1 at the lower end, or F_3 at the upper end – then the estimator will be forced to select a spectral feature other than a formant, perhaps a prominent harmonic. Therefore some compromise must be reached which takes into account the varied formant frequencies of the different phones.

There are two particular problems to be dealt with in the measurement of the formant frequencies of women's speech, both of which involve the relatively high female SFF. Firstly, for phones with a low F_1, such as /iy/ and /uw/, a high F_0 can interfere with the amplitude and apparent location of the first formant. Secondly, the widely-spaced harmonics arising from the relatively high F_0 ensure that the formants of female speakers are sampled less frequently than they would be by the harmonics produced by the lower male F_0. Consequently the female formants may have insufficient definition to be recognisable by the estimator, or the estimator might skew the formant to one side to fit the peak of the harmonic rather than that of the formant.

These two problems cause great difficulties for any automatic analysis of formant frequencies, and the situation is made worse by the varied formant frequencies attained by individual speakers. For example, if the mean frequency of the third formant of /iy/ is given as 3300Hz, then assuming a roughly Normal distribution of formant frequency values we can expect approximately half the F_3 values to be in excess of this value. We must therefore set the upper bound of the search space high enough to encompass the full range of frequencies used by speakers, yet at the same time avoiding the inclusion of F_4 in the search space. However, any third formants with frequencies in excess of the upper search limit will be underestimated by the estimator, while any speakers with a particularly low F_3 may find their F_4 being measured instead. While the results of a small-scale study can be checked individually, this is less feasible for a study involving many speakers and/or a large quantity of speech data. Thus an analysis of between- and/or within-speaker formant frequency variability will almost inevitably compromised, unless steps are taken to assess the results thoroughly.

4.1.3 The TIMIT database

The speech data used for this study came from the National Institute of Standards and Technology (NIST) production of the DARPA TIMIT Acoustic-Phonetic Speech Database Training Set[10], held on CD-ROM. The TIMIT database was designed for the development and evaluation of automatic spech recognition systems. The Training Set, as opposed to the Testing Set released on a subsequent CD-ROM, is for use as system training material. The database holds read speech data from a total of 130 female and 290 male speakers from eight major dialect regions of the USA. The speech was recorded using a Sennheiser head-mounted microphone in a quiet environment, and was digitised at a 20kHz sampling rate, before being downsampled to 16kHz for distribution.

The design of the text corpus was handled by the Massachusetts Institute of Technology (MIT), S.R.I. International (SRI), and Texas Instruments, Inc. (TI). The speech data was recorded at TI, and was transcribed at MIT and verified by NIST. The CD-ROM was prepared for production by NIST.

The rest of this section describes the structure of the TIMIT database, and discusses the limitations placed on any conclusions this study is able to reach due to the type of speech material and speakers represented on the database.

The Structure Of The Database

This section is divided into five parts, and describes the make-up of the speech data and how the data is organised on the database. The first part describes the directory and file structure of the database, including the way the directories are divided by dialect region, sex, speaker and sentence, and the sentence files which store the speech data and its transcriptions. The second part looks at the file storage format of the speech data. In the third part, the notation used to name the phones is described. The fourth part lists the types of sentence spoken by the subjects. Finally, the fifth part covers the information provided on the database about the speakers, including speaker sex, age, height, ethnic group, educational background and dialect region.

A. The directory hierarchy

The speech data from the 420 speakers is stored in a hierarchical directory structure on a CD-ROM (see Figure 4.4 for a visual representation). The directory structure will now be described from the top down, with reference to the figure.

The speakers are ordered into eight dialect regions covering the whole of the USA. The dialect regions (with their TIMIT identification code) are: New England (dr1), Northern (dr2), North Midland (dr3), South Midland (dr4), Southern (dr5), New York City (dr6), Western (dr7) and Army Brat (dr8). The final 'region', dr8, covers those people who moved around during their childhood, and who are therefore more likely to have a mixture of dialects represented in their voices. The numbers of speakers comprising the sample for each dialect are listed in Table 4.5. The dialect regions also serve to divide the speakers into directories in the database (row 1 in Figure 4.4). Thus the directory **dr3** contains all the speech data for the North Midland region.

Each speaker is given a unique identification code which acts as the name of their directory (row 2 in Figure 4.4). The first letter in the code gives the sex of the speaker (i.e. **f** or **m**); the next three letters are the initials of the speaker's name; and the number on the end

[10]From now on the database will be referred to as either the TIMIT database or the TIMIT CD-ROM.

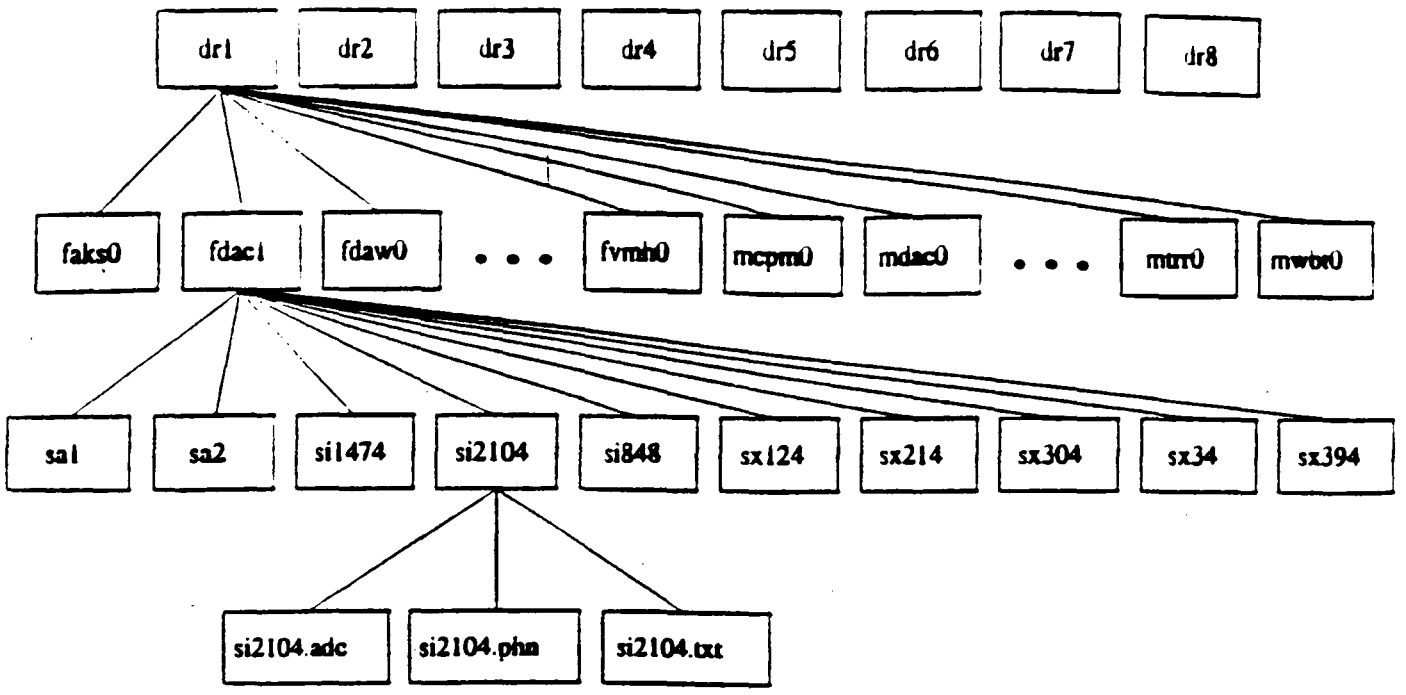


Figure 4.4: The TIMIT CD-ROM directory structure. The top level shows the dialect directories; the second level the speaker directories; the third level the sentence directories; and the bottom level the sentence data files.

allows for any speakers whose initials are identical (a '0' indicates the first speaker with a given set of initials, a '1' for the second, etc.). For example, **fdac1** is a female speaker who is the second person on the database with the initials D.A.C. The directory **fdac1** contains the ten sentences uttered by that speaker.

Each sentence is also given a unique identification code, although some sentences are spoken by more than one person. Again the identification number is also the name of the directory containing the relevant speech data (row 3 in Figure 4.4). The first two letters of the identification code indicate the sentence type, followed by the sentence number.

Finally, within each sentence directory are three files (row 4 in Figure 4.4). If we take the sentence **si2104** as an example, the files contain the following data:

- **si2104.adc** – The digitised speech waveform.
- **si2104.phn** – A phonetic transcription of the sentence, as spoken by that speaker. Each phone in the transcription is accompanied by the sample numbers of where it starts and ends in **si2104.adc**. See Figure 4.5.
- **si2104.txt** – An orthographic version of the sentence (from the original prompt to the speakers), together with the first and last sample numbers of the speech waveform in **si2104.adc**. See Figure 4.6.

From now on, these files will be referred to as **adc**, **phn** and **txt** files. Note that although this sentence may be spoken by other speakers, for obvious reasons the files **si2104.adc** and **si2104.phn** are unique to the speakers, in this case to speaker **fdac1**. The file **si2104.txt** will be the same for all speakers who say this sentence.

B. The speech data file format

The digitised speech data in the **adc** files is stored in VAX binary representation preceded by header giving information on sampling rate, number of samples in the file, etc. A utility written by Dr. Martin Cooke at the department of Computer Science, University

```
0 2680 h#
2680 5013 aa
5013 6480 hh
6480 12520 aa
12520 14560 hh
14560 16120 iy
16120 18900 th
18900 23144 ao
23144 25107 tcl
25107 25433 q
25433 27697 ey
27697 28324 dcl
28324 29953 l
29953 32880 ah
32880 36060 sh
36060 37651 dcl
37651 37890 d
37890 39094 ix
39094 40609 v
40609 42787 ao
42787 43970 r
43970 47480 s
47480 51387 iy
51387 56664 pau
56664 57273 q
57273 58739 ae
58739 59951 q
59951 61538 l
61538 66230 ae
66230 69560 s
69560 70920 tcl
70920 71609 t
71609 73520 h#
```

Figure 4.5: The contents of the file **dr1/fdac1/si2104/si2104.phn** on the TIMIT CD-ROM, a phonetic transcription of the sentence **si2104** as spoken by **fdac1**. Each line contains the start and finish sample numbers from the sentence **si2104.adc**, and the identifier of the phone being uttered (see Table 4.7 for the full list of phone identifiers).

```
0 73626 Ahah, he thought, a lush divorcee at last.
```

Figure 4.6: The contents of the file **dr1/fdac1/si2104/si2104.txt** on the TIMIT CD-ROM, an orthographic transcription of the sentence **si2104** as spoken by **fdac1**. The two numbers represent the first and final sample numbers of the sentence **si2104.adc**, making this sentence 4.6 seconds long.

of Sheffield, called **get-rep**, converts the speech data into ASCII format. For a given sentence, **get-rep** extracts the sentence from the TIMIT database, removes the header, and applies the format conversion. The conversion was necessary because the processing of the speech data was carried out in ASCII.

C. The phone notation

The transcriptions held in the **phn** files are fairly narrow and are acoustically-orientated. Therefore the speech sounds will be referred to from now on as phones rather than phonemes. The notation used to represent the standard IPA phonetic symbols is the ASCII CMU/ARPAbet (see Figure 4.7).

D. A description of the sentences

The speakers on the TIMIT database each read a total of ten sentences, which were divided into three sentence-types. The identification codes and descriptions of the sentence-types are as follows:

- **sa** – SRI dialect calibration (shibboleth) sentence (2 per speaker). Designed to allow comparison between the dialect regions represented on the data base.

- **si** – TI random contextual variant sentence (3 per speaker). Designed to provide examples of phones in all possible left and right contexts.

- **sx** – MIT phonetically compact sentence (5 per speaker). Chosen to provide alternative contexts and multiple occurrences of the same phonetic sequence in different word sequences.

The full identification code for each sentence consists of the two letter sentence-type code followed by the number of the sentence. Thus the example used above, **si2104**, is number 2104 of the random contextual variant type. Every speaker read the same two calibration sentences, the orthographic transcription forms of which are:

sa1 "She had your dark suit in greasy wash water all year"
sa2 "Don't ask me to carry an oily rag like that"

Obviously, each person's realisation of the sentences could differ greatly. The other sentences were chosen at random for each speaker. For the **si** sentences there were 1890 possible choices (**si453** to **si2342**); and for the **sx** sentences, there were 450 choices (**sx3** to **sx452**).

E. The speaker information

The TIMIT database holds a total of 130 female and 290 male speakers from all over the USA. The numbers of speakers of each sex representing the eight dialects are listed in Table 4.5. As can be seen from column 7, the male speakers outnumbered the female speakers by more than 2 to 1.

Also held on the TIMIT CD-ROM is a file of information about the attributes of the speakers, **trnspkr.log**, a portion of which is reproduced in Figure 4.8. The information consists of, for each speaker:

Phone	Example	Phone	Example
iy	b**ea**t	en	butt**on**
ih	b**i**t	ng (eng)	si**ng**
eh	b**e**t	m (em)	**m**om
ae	b**a**t	n (nx)	**n**on
ix	rose**s**	hh (hv)	**h**ay
ax	**th**e	b	**b**ob
ah	b**u**tt	d	**d**ad
uw (ux)	b**oo**t	dx	bu**tt**er
uh	b**oo**k	g	**g**ag
ao	**a**bout	p	**p**op
aa	c**o**t	t	**t**ot
er (axr)	b**ir**d	k	**k**ick
ay	b**i**te	z	**z**oo
oy	b**oy**	zh	mea**s**ure
aw	b**ough**	v	**v**ery
ow	b**oa**t	f	**f**ief
ey	b**ai**t	th	**th**ief
w	**w**et	s	**s**is
r	**r**ed	sh	**sh**oe
y	**y**et	dh	**th**ey
l	**l**ed	ch	**ch**urch
el	bott**le**	jh	**j**udge

Phone	Allophones	Description
cl	pcl, tcl, kcl, qcl	unvoiced closure
vcl	bcl, dcl, gcl	voiced closure
epi	–	epinthetic closure
q	–	glottal closure
sil	–	silence
pau	–	between silence
h#	–	begin/end silence

Figure 4.7: List of phones used in the TIMIT database, from Lee and Hon. The phones in brackets in the upper table are allophones. Note this is the 'collapsed set': Lee and Hon decided the original set of phones was too detailed, and that HMMs would not be able to distinguish the phones quite so well.

Dialect region		No. female speakers (%)		No. male speakers (%)		Ratio f : m	Total no. speakers (%)	
dr1	New England	15	(40.5)	22	(59.5)	1 : 1.5	37	(8.8)
dr2	Northern	18	(27.7)	47	(72.3)	1 : 2.6	65	(15.5)
dr3	North Midland	15	(22.7)	51	(77.3)	1 : 3.4	66	(15.7)
dr4	South Midland	19	(29.2)	46	(70.8)	1 : 2.4	65	(15.5)
dr5	Southern	26	(40.0)	39	(60.0)	1 : 1.5	65	(15.5)
dr6	New York City	11	(34.4)	21	(65.6)	1 : 1.9	32	(7.6)
dr7	Western	18	(27.3)	48	(72.7)	1 : 2.7	66	(15.7)
dr8	Moved around	8	(33.3)	16	(66.7)	1 : 2.0	24	(5.7)
	TOTALS	130	(31.0)	290	(69.0)	1 : 2.2	420	(100.0)

Table 4.5: The number (and percentage) of speakers by dialect region and sex represented on the TIMIT CD-ROM. The percentages given in columns 4 and 6 are the female and male representation for that dialect (i.e. not of the total number of speakers). The ratio of female to male speakers for each dialect region is given in column 7. The figures in the final column are the percentages of the total number of speakers for each dialect.

- An identification number.
- The speaker's initials (three letters for each speaker).
- The date of the recording session for the speaker's sentences (in the U.S. style, i.e. month/day/year).
- The speaker's sex.
- The speaker's dialect region.
- The original recording session number.
- The speaker's birthdate (month/day/year).
- The speaker's height (feet and inches).
- The speaker's ethnic group[11].
- The speaker's highest level of formal education.

Thus it was possible to define certain variables with which to analysis speaker characteristics. These variables were age, height, ethnic group, dialect region and education. The variables could then be used to separate the speakers into distinct groups. For example, in the data analysis the age variable consists of six groups: 20-29 years, 30-39 years, 40-49 years, 50-59 years, and 60 years and over. While it is possible to tabulate the information into speaker groups by hand, for 420 speakers this would be extremely tedious. Thus UNIX shell scripts were written to analyse the information for the different variables automatically. This produced the following data on the speakers:

Age of speakers The 130 female speakers had a mean age of 29.9 years (s.d. 9.6), ranging from 21 to 85 years. The 290 male speakers had a mean age of 29.6 years (s.d. 7.3), ranging from 20 to 85 years. The numbers of speakers in each age group are summarised

[11]The description used on the TIMIT CD-ROM is speaker 'race' (see the header for the TIMIT file **trnspkr.log** reproduced in Figure 4.8). 'Ethnic group' is the preferred term here. Similarly, the categorisation of speakers in groups labelled white, black, Hispanic, Spanish, oriental and native American is that used by TIMIT.

*ID	Init	RDate	Sex	DR	RS	BDate	Ht	Race	EDU
1	CMM	1/09/86	F	2	1X	01/30/60	5'2"	WHT	HS
2	EDW	1/15/86	F	4	1X	01/17/60	5'7"	BLK	HS
3	JDM1	3/7/86	M	4	37	05/17/61	6'0"	WHT	BS
4	ALK	1/28/86	F	3	11B	07/23/38	5'6"	WHT	HS
5	GXP	2/3/86	M	4	15	05/15/50	5'8"	BLK	BS
6	CRH	1/17/86	F	4	1X	02/24/58	5'4"	BLK	HS
7	MWH	1/16/86	M	3	1X	06/28/59	6'0"	WHT	BS
8	HPG	1/17/86	M	3	1X	9/19/47	5'9"	WHT	BS
9	KCL	2/13/86	M	4	21	01/12/47	5'6"	WHT	BS
10	CRC	1/31/86	M	5	14B	08/06/47	5'7"	WHT	MS
11	RTK	1/14/86	M	3	1X	8/08/58	6'0"	WHT	MS
12	CMJ	1/16/86	M	6	1X	2/06/58	5'7"	WHT	MS
13	EFG	1/16/86	M	2	1X	2/26/56	5'8"	WHT	BS
14	JLS	1/17/86	M	4	1X	6/01/60	5'9"	WHT	BS
15	PRK	1/17/86	M	4	1X	3/10/58	5'11"	WHT	MS
16	SKP	1/20/86	F	5	5	12/15/61	5'8"	WHT	BS
17	PGL	1/20/86	M	2	5	6/29/61	6'0"	WHT	BS
18	KJL	1/20/86	M	7	5	11/18/55	6'1"	WHT	MS
19	EXM	1/20/86	F	5	5	12/03/54	5'4"	BLK	MS
20	RWS	1/20/86	M	1	5	3/10/59	6'2"	WHT	BS

Figure 4.8: Part of the contents of the file **trnspkr.log**, showing the speaker information for the first twenty speakers on the CD-ROM. The column headed 'ID' lists the speaker's identification number; 'Init' the speaker's initials; 'RDate' the recording date; 'Sex' the speaker's sex; 'DR' the speaker's dialect region; 'RS' the recording session number; 'BDate' the speaker's birth date; 'Ht' the speaker's height (in feet and inches); 'Race' the speaker's ethnic group; and 'EDU' the speaker's highest formal educational level. Note the dates given in columns 3 and 7 use the U.S. convention, i.e. month/day/year.

Age (years)	No. female speakers (%)		No. male speakers (%)		Ratio f : m	Total no. speakers (%)	
20 – 29	82	(30.9)	183	(69.1)	1 : 2.2	265	(63.1)
30 – 39	32	(27.1)	86	(72.9)	1 : 2.7	118	(28.1)
40 – 49	10	(43.5)	13	(56.5)	1 : 1.3	23	(5.5)
50 – 59	4	(36.4)	7	(63.6)	1 : 1.8	11	(2.6)
$\geq$ 60	2	(66.7)	1	(33.3)	1 : 0.5	3	(0.7)
TOTALS	130	(31.0)	290	(69.0)	1 : 2.2	420	(100.0)

Table 4.6: The distribution of the ages of the speakers on the TIMIT CD-ROM. The percentages given in columns 4 and 6 are the female and male representation for that age range (i.e. not of the total number of speakers). The ratio of female to male speakers for each age range is given in column 7. The figures in the final column are the percentages of the total number of speakers for each age range. Note: Of the speakers aged over 60 years, the two females are 67 and 85, while the male is 85.

in Table 4.6. Note that in the sixty years and over age groups, the two female speakers are 67 and 85 years old, and the single male speaker is 85 years old.

Height of speakers The 130 female speakers had a mean height of 5ft 5.2in (s.d. 2.5in), ranging from 5ft to 6ft. The 290 male speakers had a mean height of 5ft 10.8in (s.d. 2.8in), ranging from 5ft 2in to 6ft 8in. The numbers of speakers in each height group are summarised in Table 4.7.

Ethnic group of speakers 22 speakers were described as black, 10 of them female and 12 male. 376 speakers were described as white, 117 of them female and 259 male. Of the other speakers, 2 males were described as Hispanic, 1 male as Spanish, 2 males as native Americans, 2 males as oriental, 1 male as 'OTH'; the rest, including the three remaining females, had no designation. Only the black and white speakers were used in the variable analysis of ethnic group, realising a total of 398 speakers. The numbers of speakers in each category are summarised in Table 4.8.

Educational level of speakers The categories of highest educational level attained by the speakers were: high school, associate degree, bachelor's degree, master's degree and doctorate degree. No educational information was given for eight speakers. The numbers of speakers at each level are summarised in Table 4.9.

The limitations of the data

This section looks briefly at the limitations placed upon any conclusions this study is able to reach as a result of the type of speech material and speakers represented on the database. The TIMIT database was chosen as the source of speech data for this study because of its size, structure and availability, not for how accurately it samples the world population of female and male speakers. It therefore suffers from all the limitations of a restricted sample.

From the data supplied with the TIMIT database, and analysed above, it is possible to say that the majority of the speakers represented on the TIMIT database are aged between 20 and 39 years (88% of the women and 93% of the men are in their twenties and thirties), are white (90% of the women and 89% of the men are white), and have been educated to degree level (73% of the women and 88% of the men hold an associate, bachelor's, master's or doctorate degree). The latter statistic, referring to the speakers'

Height (ft, in)	No. female speakers (%)		No. male speakers (%)		Ratio f : m	Total no. speakers (%)	
$\leq$ 5'1"	7	(100.0)	–	(0.0)	–	7	(1.7)
5'2" – 5'3"	25	(96.2)	1	(3.8)	1 : 0.0	26	(6.2)
5'4" – 5'5"	37	(92.5)	3	(7.5)	1 : 0.1	40	(9.5)
5'6" – 5'7"	32	(55.2)	26	(44.8)	1 : 0.8	58	(13.8)
5'8" – 5'9"	24	(27.6)	63	(72.4)	1 : 2.6	87	(20.7)
5'10" – 5'11"	4	(4.9)	77	(95.1)	1 : 19.3	81	(19.3)
6'0" – 6'1"	1	(1.5)	67	(98.5)	1 : 67.0	68	(16.2)
6'2" – 6'3"	–	(0.0)	42	(100.0)	–	42	(10.0)
6'4" – 6'5"	–	(0.0)	7	(100.0)	–	7	(1.7)
$\geq$ 6'6"	–	(0.0)	4	(100.0)	–	4	(1.0)
TOTALS	130	(31.0)	290	(69.0)	1 : 2.2	420	(100.0)

Table 4.7: The distribution of the heights of the speakers on the TIMIT CD-ROM. The percentages given in columns 4 and 6 are the female and male representation for that height range (i.e. not of the total number of speakers). The ratio of female to male speakers for each height range is given in column 7. The figures in the final column are the percentages of the total number of speakers for each height range.

Ethnic group	No. female speakers (%)		No. male speakers (%)		Ratio f : m	Total no. speakers (%)	
Black	10	(45.5)	12	(54.5)	1 : 1.2	22	(5.5)
White	117	(31.1)	259	(68.9)	1 : 2.2	376	(94.5)
TOTALS	127	(31.9)	281	(68.1)	1 : 2.2	398	(100.0)

Table 4.8: The distribution of the black and white speakers on the TIMIT CD-ROM. The percentages given in columns 4 and 6 are the female and male representation for that ethnic group (i.e. not of the total number of speakers). The ratio of female to male speakers for each ethnic group is given in column 7. The figures in the final column are the percentages of the total number of speakers for each ethnic group.

Educational level	No. female speakers (%)		No. male speakers (%)		Ratio f : m	Total no. speakers (%)	
No designation	1	(12.5)	7	(87.5)	1 : 7.0	8	(1.3)
High School	34	(54.8)	28	(45.2)	1 : 0.8	62	(14.8)
Associate Degree	6	(42.9)	8	(57.1)	1 : 1.3	14	(2.3)
Bachelor's Degree	62	(27.8)	161	(72.2)	1 : 2.6	223	(36.0)
Master's Degree	24	(24.0)	76	(76.0)	1 : 3.2	100	(16.1)
Doctorate Degree	3	(23.1)	10	(76.9)	1 : 3.3	13	(2.1)
TOTALS	130	(31.0)	290	(69.0)	1 : 2.2	420	(100.0)

Table 4.9: The distribution of the educational levels of the speakers on the TIMIT CD-ROM. The percentages given in columns 4 and 6 are the female and male representation for that level (i.e. not of the total number of speakers). The ratio of female to male speakers for each level is given in column 7. The figures in the final column are the percentages of the total number of speakers for each level.

educational background, also implies that the socioeconomic background of the speakers is predominately middle class. Furthermore, 91% of the women ranged in height from 5'2" to 5'9", and 95% of the men ranged between 5'6" and 6'3". Clearly, the TIMIT database is only truly representative of one section of population; namely, U.S. citizens who are, in the main, white, middle class, university educated, and relatively young. Furthermore, male representation on the database is substantially greater than that for women, such that the male speakers outnumber the female speakers by two to one. As a result, this thesis cannot claim to be more than a study of that particular subset of the world's population.

Furthermore, the database consists solely of read sentences, which tend to possess a less dynamic intonation contour than spontaneous speech. The frequency characteristics measured from read speech will therefore not be truly representative of that person's normal speaking patterns.

4.2 Results

This section presents the results of the analysis of the acoustic-phonetic markers of speaker sex. The first three parts to this section report the results for each marker in depth: the fundamental frequency in Section 4.2.1, the relative amplitude of the first harmonic in Section 4.2.2, and the formant frequencies in Section 4.2.3. The final part, 4.2.4, summarises the results.

The in-depth reporting of the results takes the following form: an analysis of the overall data, an analysis of data by phone, and an analysis of data by speaker variable. Where possible the results are compared with the relevant data from the literature. Particular attention is paid to the distribution of the mean and the range of values produced by each speaker to facilitate an analysis of between- and within-speaker variability.

Note about significance testing of the results

Partly to do with a lack of time, the results which follow in this section have not been tested for statistical significance. However, the main reason for this is due to doubts over the applicability of significance testing itself. For a comprehensive critique of the use of significance testing, see Atkins & Jarrett (1979)[1]. Briefly, while significance tests are intended to be an objective way of "drawing conclusions from [small samples of] quantitative data" (Atkins & Jarrett 1979:87), all too often the tests are applied without considering whether the sample satisfies the rather strict conditions required by the test, particularly the requirement for a Normal distribution of the sample and that the sample is truly representative of the overall population. Moreover, results found to reach the rather arbitrary levels of significance set by researchers are often reported as if they are in themselves scientifically valid, whereas the most a test can realistically achieve is to *indicate the possibility* of a significant result.

[1] Atkins L, Jarrett D (1979) "The significance of significance tests." In Irvine J, Miles I, Evans J (1979:87-109) *Demystifying Social Statistics*. Pluto Press: London.

Sex	n	Mean (s.d.)	Min	Q1	Q2	Q3	Max
f	130	208 (23)	146	193	206	220	270
m	290	120 (17)	82	108	118	129	183

Table 4.10: Female and male mean SFF data (to nearest 1Hz). Note that this data is computed from the SPEAKER MEANS.

4.2.1 Fundamental frequency

Analysis of overall data.

The mean SFF[12] for the female speakers was 208Hz (with a s.d. of 23Hz); for the male speakers it was 120Hz (17Hz) (see Table 4.10). Thus, the mean female SFF was 73% greater than the male. The median SFFs were 206Hz and 118Hz respectively. These results compare favourably with those reported in the literature.

A. Distribution of SFF

The s.d.s reported above indicate that the mean SFFs of women and men are, in general, very different. This will now be examined in more detail. The SPEAKER MEANS (i.e. each individual's mean SFF) ranged from 146Hz to 270Hz for the female speakers, and from 82Hz to 183Hz for the males. The distribution of the SPEAKER MEANS is illustrated in histogram form in Figure 4.9. It is clear from the histogram that there is very little overlap in mean SFF between the sexes. The overlap region, stretching from 146-183Hz as defined by the lowest female mean SFF and highest male mean SFF, consists of 11.5% (15) of female and 6.9% (20) of male speakers. The number and percentage of speakers whose mean SFF fell within the overlap region is as follows:

		Mean SFF (Hz)									
		140-45	145-50	150-55	155-60	160-65	165-70	170-75	175-80	180-85	185-90
f	n	0	1	1	1	2	1	3	4	5	5
	%	0.0	0.8	0.8	0.8	1.5	0.8	2.3	3.1	3.8	3.8
m	n	12	10	4	1	3	3	0	1	2	0
	%	4.1	3.4	1.4	0.3	1.0	1.0	0.0	0.3	0.7	0.0

If we define an arbitrary cut-off of 170Hz in order to differentiate between the sexes, we find there is an even greater distinction: only 4.6% (6) of the female speakers are below the cut-off, and only 1.0% (3) of the male speakers are above it. By placing the cut-off at 165Hz, these figures become 3.8% (5) and 2.1% (6) respectively.

Another way of looking at the distribution of SFF is to consider the SLICE MEANS (i.e. the mean F_0 of single speech segments). In this way we can take into account the actual frequencies attained by the speakers rather than their overall averages. The mean SFFs computed from the SLICE MEANS were, not surprisingly, the same as from the SPEAKER MEANS, but with larger s.d.s of 36Hz and 23Hz respectively, indicating that female and male use of SFF is, in general, very different. The histogram in Figure 4.10 illustrates the distribution of SFF, with the female speakers ranging from 63Hz to 384Hz, and the male speakers from 40Hz to 277Hz. Again there was very little overlap between the sexes, even though the histogram is more representative of the range of SFF used by women and men. The following table examines the region of greatest overlap in more detail, with the number and percentage of SLICE MEANS at 5Hz intervals:

[12]This mean was computed from the SPEAKER MEANS, i.e. it was the mean of each individual's mean SFF.

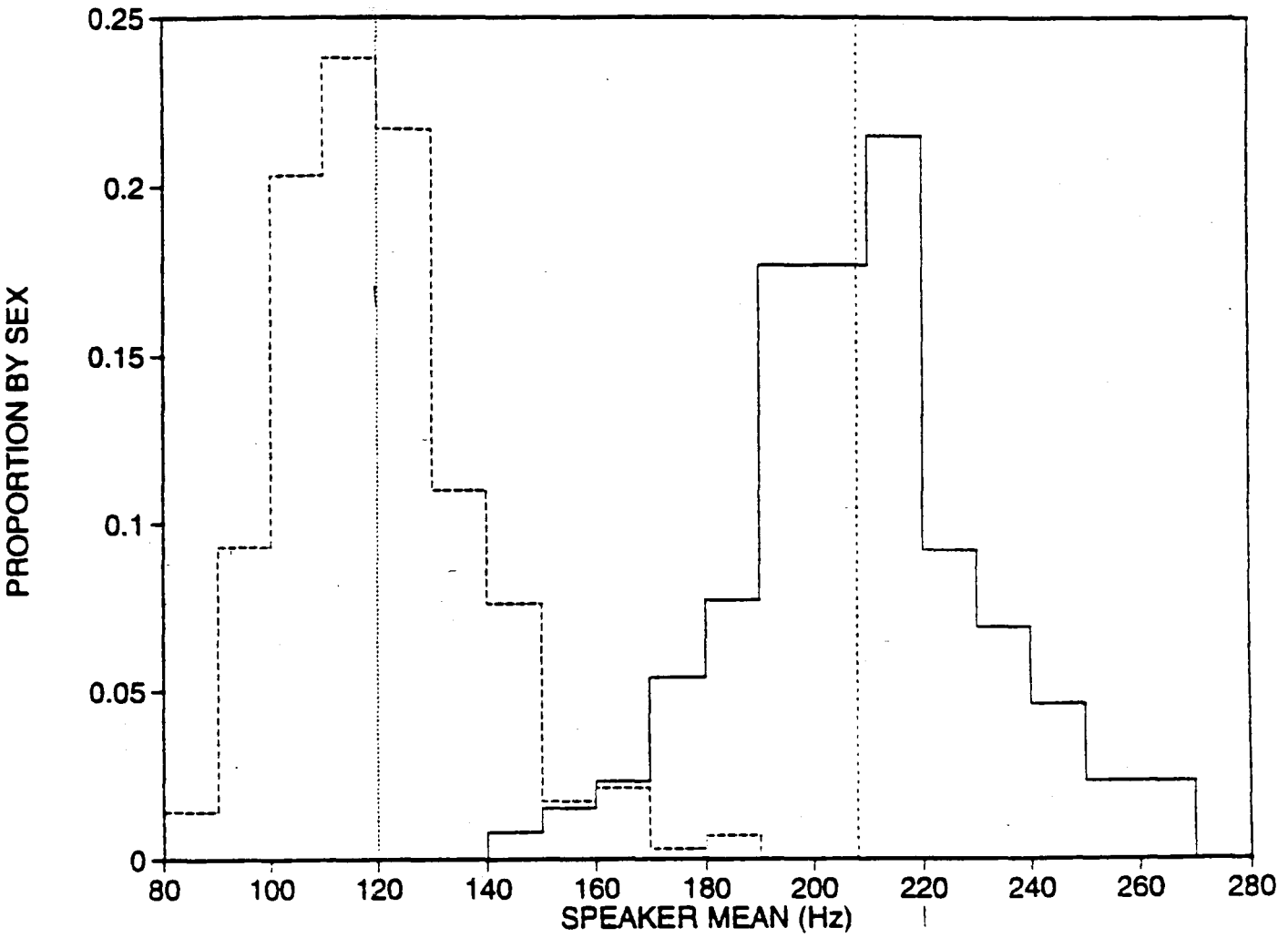


Figure 4.9: Distribution of F_0 SPEAKER MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The bars of the histogram represent a 10Hz interval. The y-axis represents the proportion of speakers of one sex having a particular mean SFF. The data used to produce the histogram is given in Table 4.11. The overall mean SFFs for each sex are shown as single, vertical lines.

SFF (Hz)	Female		Male	
	n	%	n	%
80–90	–	–	4	1.4
90–100	–	–	27	9.3
100–110	–	–	59	20.3
110–120	–	–	69	23.8
120–130	–	–	63	21.7
130–140	–	–	32	11.0
140–150	1	0.8	22	7.6
150–160	2	1.5	5	1.7
160–170	3	2.3	6	2.1
170–180	7	5.4	1	0.3
180–190	10	7.7	2	0.7
190–200	23	17.7	–	–
200–210	23	17.7	–	–
210–220	28	21.5	–	–
220–230	12	9.2	–	–
230–240	9	6.9	–	–
240–250	6	4.6	–	–
250–260	3	2.3	–	–
260–270	3	2.3	–	–

Table 4.11: Number and percentage of female and male SFF SPEAKER MEANS at 10Hz intervals. A dash (–) indicates there were no speakers with a mean SFF in that interval. This data was used to plot the histogram in Figure 4.9.

		Mean SFF (Hz)									
		140-45	145-50	150-55	155-60	160-65	165-70	170-75	175-80	180-85	185-90
f	n	33	42	63	84	95	134	193	224	248	256
	%	0.7	0.9	1.3	1.7	1.9	2.7	3.9	4.6	5.1	5.2
m	n	441	338	264	213	127	116	91	73	48	35
	%	4.1	3.1	2.4	2.0	1.2	1.1	0.8	0.7	0.4	0.3

If we consider the 170Hz cut-off again, 11.4% (559) of the segments uttered by female speakers are below it, and 3.2% (345) of the male segments are above it. The 165Hz cut-off yields figures of 8.7% (427) and 4.4% (463) respectively. However, from Figure 4.10 it is clear that a better cut-off point would be 160Hz, which yields 6.5% (318) of female segments and 5.6% (612) of male segments. Applying this new cut-off to the SPEAKER MEANS, we find 2.3% (3) of female speakers have a mean SFF below the cut-off, and 3.1% (9) of male speakers are above it.

However, the 318 female segments below the 160Hz cut-off were uttered by 50.8% (66) of the women, and the 612 male segments by 36.2% (105) of the men. It is clear then that despite the lack of overlap between the SFFs of women and men, a large proportion of the speakers here at least occasionally use a SFF on the other side of the cut-off. Taking this further, the 2.2% (109) of female slices below 140Hz were uttered by 26.2% (34) of the speakers, and the 1.7% (182) of male slices above 180Hz uttered by 13.8% (40) of the speakers.

B. Range of SFF

The mean range of SFF, computed from the mean of each speaker's range, was 10.6st (4.5st) for the female speakers, and 9.7st (3.8st) for the male speakers. The Hertz equiv-

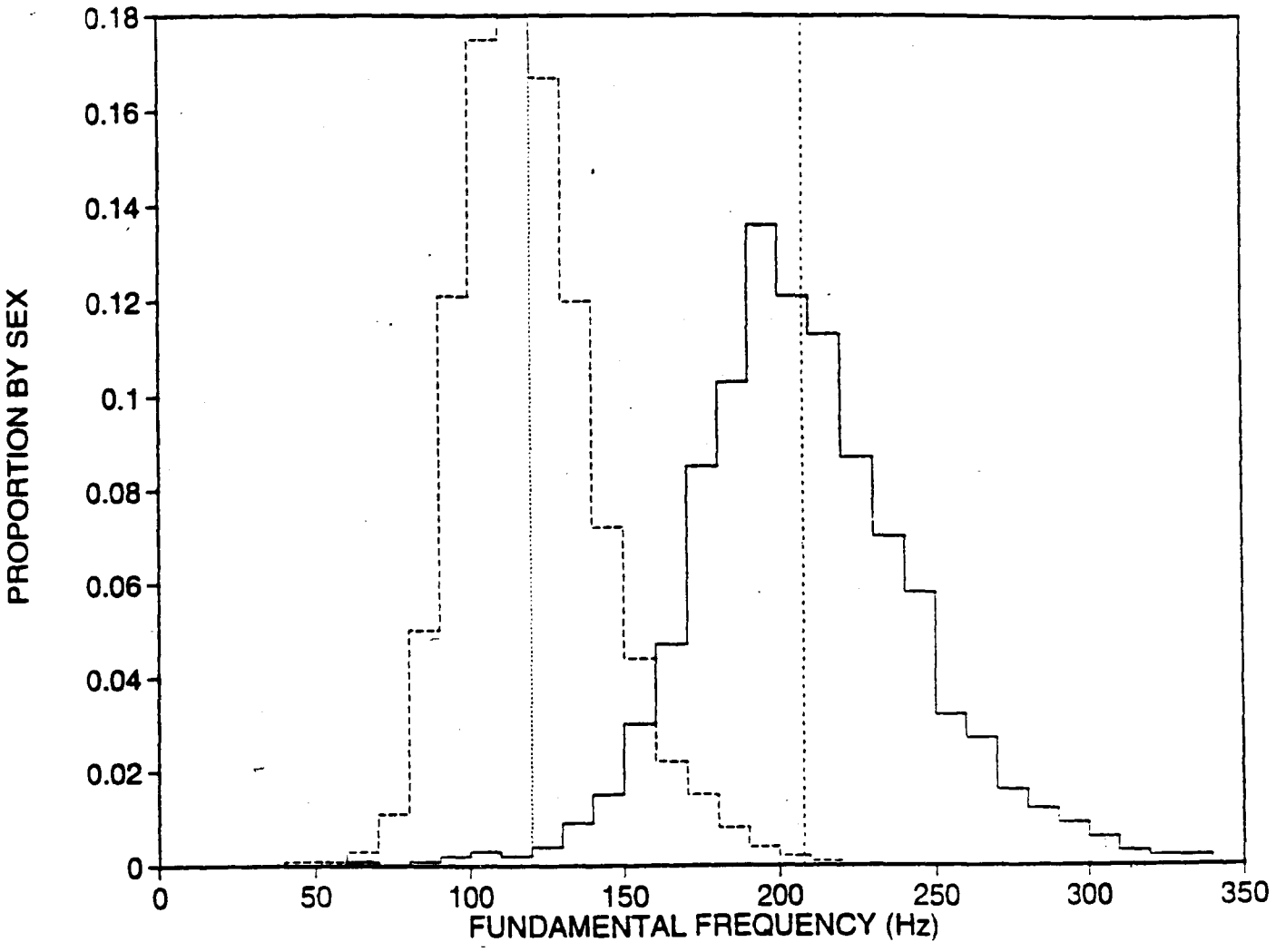


Figure 4.10: Distribution of SLICE MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The bars of the histogram represent a 10Hz interval. The y-axis represents the proportion of speakers of one sex having a particular mean SFF. The data used to produce the histogram is given in Table 4.12. The overall mean SFFs for each sex are shown as single, vertical lines. Note: Some of the SLICE MEANS are not represented by a histogram bar as there were insufficient means to register on the histogram: for the female speakers, there two slices at 70-80Hz, two at 360-70Hz, and one at 380-90Hz; for the male speakers, there were one slice at 30-40Hz, four at 220-30Hz, four at 230-40Hz, two at 240-50Hz, and one at 270-280Hz.

SFF (Hz)	Female		Male	
	n	%	n	%
40–50	–	–	13	0.1
50–60	–	–	15	0.1
60–70	4	0.1	31	0.3
70–80	2	0.0	125	1.2
80–90	6	0.1	537	5.0
90–100	11	0.2	1313	12.1
100–110	14	0.3	1898	17.5
110–120	10	0.2	1956	18.0
120–130	18	0.4	1811	16.7
130–140	45	0.9	1297	12.0
140–150	75	1.5	779	7.2
150–160	147	3.0	477	4.4
160–170	229	4.7	243	2.2
170–180	417	8.5	164	1.5
180–190	504	10.3	83	0.8
190–200	666	13.6	48	0.4
200–210	595	12.1	23	0.2
210–220	553	11.3	16	0.1
220–230	427	8.7	4	0.0
230–240	345	7.0	4	0.0
240–250	284	5.8	2	0.0
250–260	157	3.2	–	–
260–270	133	2.7	–	–
270–280	80	1.6	1	0.0
280–290	60	1.2	–	–
290–300	45	0.9	–	–
300–310	30	0.6	–	–
310–320	16	0.3	–	–
320–330	11	0.2	–	–
330–340	8	0.2	–	–
340–350	3	0.1	–	–
350–360	–	–	–	–
360–370	2	0.0	–	–
370–380	–	–	–	–
380–390	1	0.0	–	–

Table 4.12: Number and percentage of female and male F_0 SLICE MEANS at 10Hz intervals. A dash (–) indicates there were no slices with a F_0 in that interval. This data was used to plot the histogram in Figure 4.10.

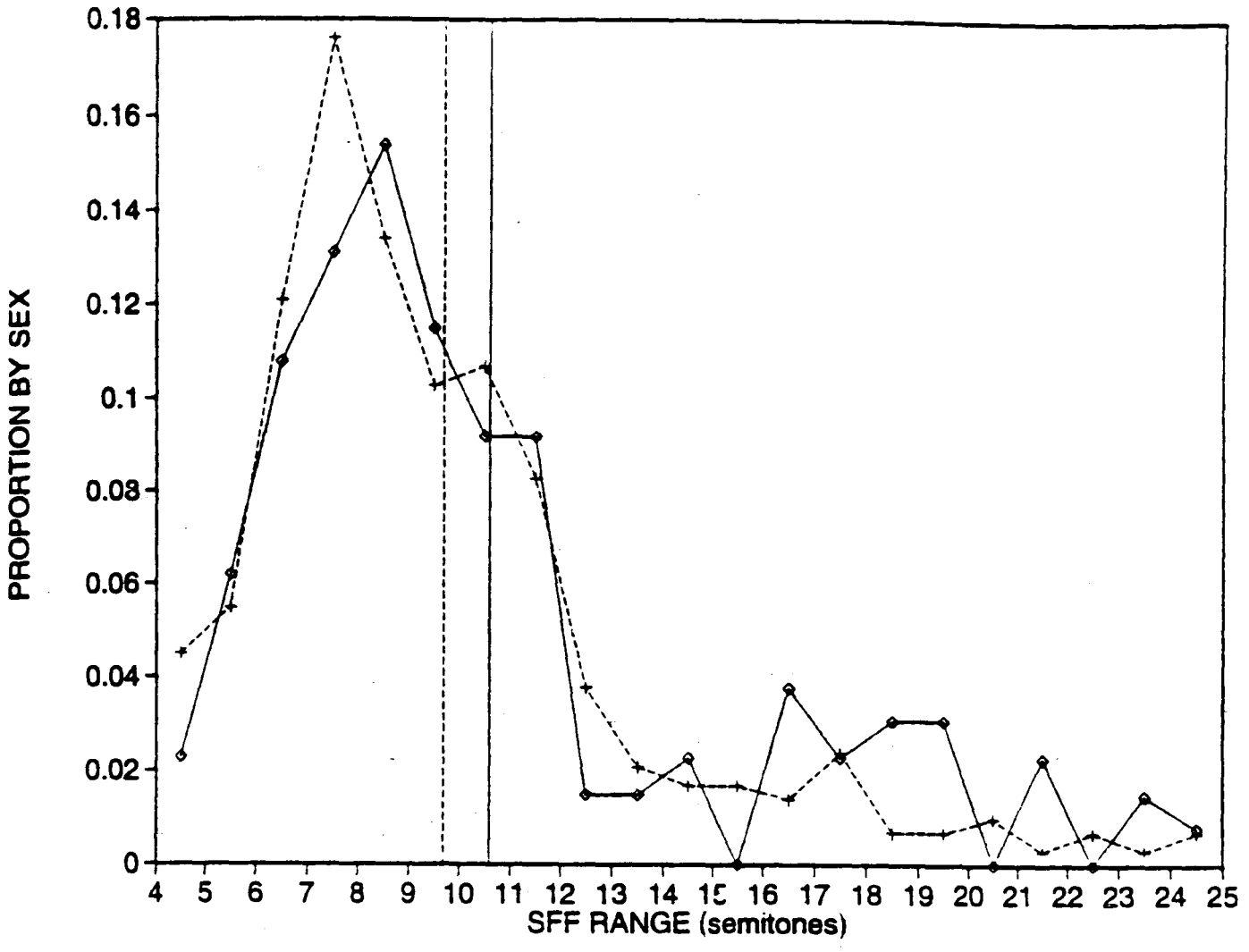


Figure 4.11: Distribution of SFF range for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of ranges in a 1st interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean SFF. The data used to produce the histogram is given in Table 4.13. The mean SFF ranges are shown as single, vertical lines.

alents were 122Hz (40Hz) and 66Hz (24Hz) respectively[13]. Although the female mean range in Hertz is almost twice that of the male mean range, comparing the values on the semitone (st) scale reveals that on average the women in this study had similar SFF ranges to the men.

The range of SFF used by individuals will now be investigated further, with a look at the distribution of SFF ranges, illustrated in histogram form in Figure 4.11. The figure reveals that the distribution of ranges is highly-skewed, with most of the speakers of both sexes having a range less than 12st. Numerically, this is 76% (100) of the women and 81.4% (236) of the men. 2.3% (3) of the female and 4.5% (13) of the male speakers had SFF ranges lower than 5st, with the smallest female range at 4.8st (58Hz), and the smallest male range at 4.0st (26Hz), indicating rather more men than women had very restricted SFF ranges. Of the speakers with ranges greater than 12st, they were spread fairly evenly upto the maximum range for both sexes of 24.7st (equivalent to 227Hz for the female speaker, and 141Hz for the male).

These data show that when the range is measured in semitones, there was a slight tendency for male speakers to have a smaller range of SFF than female speakers, as evidenced by

[13]Computing the mean SFF range by subtracting the mean range minimum from the mean range maximum gave ranges of 10.2st or 122Hz, and 9.4st or 65Hz, respectively. The mean range minima and maxima were 151-273Hz for the female speakers, and 89-154Hz for the male speakers.

Range (st)	Female		Male	
	n	%	n	%
4–5	3	2.3	13	4.5
5–6	8	6.2	16	5.5
6–7	14	10.8	35	12.1
7–8	17	13.1	51	17.6
8–9	20	15.4	39	13.4
9–10	15	11.5	30	10.3
10–11	12	9.2	31	10.7
11–12	12	9.2	24	8.3
12–13	2	1.5	11	3.8
13–14	2	1.5	6	2.1
14–15	3	2.3	5	1.7
15–16	–	–	5	1.7
16–17	5	3.8	4	1.4
17–18	3	2.3	7	2.4
18–19	4	3.1	2	0.7
19–20	4	3.1	2	0.7
20–21	–	–	3	1.0
21–22	3	2.3	1	0.3
22–23	–	–	2	0.7
23–24	2	1.5	1	0.3
24–25	1	0.8	2	0.7

Table 4.13: Number and percentage of female and male SFF ranges at 1st intervals. A dash (–) indicates there were no speakers with an SFF range in that interval. This data was used to plot the histogram in Figure 4.11.

the lower mean and the mode (see Figure 4.11)[14]. Having said that, there is also much similarity between the two sexes.

As reported above, in the discussion on the distribution of SFF, the average SFF ranges in terms of Hertz of female and male speakers were quite distinct. Consider the following table showing the mean limits to the range of SFF used by women and men:

	Range limits (Hz)			
	Minimum		Maximum	
f	151	(35)	273	(35)
m	89	(19)	154	(27)

This provides further proof that on average there was little overlap between the sexes in their use of SFF. In addition, bearing in mind that these are the *mean* range limits, the similarity of the female maximum and male minimum serves to illustrate again the existence of a band of frequencies used occasionally by both female and male speakers.

Analysis of data by phone.

Examining the mean SFFs[15] for each phone, we find they follow very similar patterns for each sex: the lowest mean SFF was measured for /ao/, followed by /aa/ and /ae/, then

[14] If we consider intervals of 0.5st, as opposed to the 1st intervals in the histogram, the female mode becomes 8.5-9st, while the male mode is 7.5-8st.

[15] The statistics given in this section were computed from the SPEAKER MEANS, as opposed to the SLICE MEANS.

Phone	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
/aa/	130	203	(23)	131	190	201	215	256
/ae/	130	204	(25)	135	190	204	219	273
/ao/	130	195	(27)	93	182	196	211	261
/iy/	130	215	(24)	151	199	215	229	286
/uw/	76	219	(35)	153	191	217	242	310
/ux/	128	215	(27)	145	197	217	230	281
TOTALS	130	208	(23)	146	193	206	220	270

Table 4.14: Female mean SFF data (to nearest 1Hz) by vowel phone. Note that all the means are computed from the SPEAKER MEANS.

Phone	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
/aa/	290	119	(17)	78	108	116	128	179
/ae/	290	118	(17)	83	106	116	128	182
/ao/	290	115	(17)	72	103	113	124	193
/iy/	290	122	(18)	84	109	120	132	191
/uw/	186	126	(24)	76	109	124	138	206
/ux/	280	126	(21)	88	112	124	135	223
TOTALS	290	120	(17)	82	108	118	129	183

Table 4.15: Male mean SFF data (to nearest 1Hz) by vowel phone. Note that all the means are computed from the SPEAKER MEANS.

/iy/, and finally /uw and /ux/ (see Tables 4.14 and 4.15 for the figures. and Figure 4.12 for a visual representation). The difference between /aa/, /ae/, /ao/ and /iy/, /uw/, /ux/ was more pronounced for the female speakers. This data suggests that SFF is to some extent dependent upon the vowel phone being spoken, with a difference between the lowest and highest female mean SFFs of 24Hz, and 11Hz for the male mean SFFs.

For both sexes, the s.d. for each phone was also similar, indicating that the distribution of SPEAKER MEANS around a phone was similar. For the female speakers, most of their mean SFFs for a particular phone were clustered within 25Hz of the overall mean for that phone, and within 18Hz for the males. Figure 4.12 shows a clear separation between the sexes of the s.d. intervals, providing further evidence for a difference in female and male use of SFF. The exception in s.d. for both sexes was /uw/, which had a noticeably larger s.d. than the other phones. This cause of this, a wider distribution of SFF SPEAKER MEANS, can be seen in Figure 4.13, which compares the distribution of female SPEAKER MEANS for /aa/, /uw/ and for all phones. The histogram shows how the SPEAKER MEANS for all phones were clustered around the female group mean SFF of 208Hz, as evidenced by the relatively tall and thin profile. This pattern was broadly repeated for /aa/, with the profile shifted to the left to reflect the lower mean SFF. However, the profile for /uw/ is wider and flatter, showing how its SPEAKER MEANS are more widely distributed. The explanation may lie in the relatively few number of /uw/ phones spoken per speaker. On average, only two phones were spoken per speaker (see Table 4.4), with the consequence that the computation of most of the SPEAKER MEANS for /uw/ were based on very few values. Thus for some speakers, the value of their mean SFF would have been based upon a single phone at, say, the beginning or end of a sentence; given the effects of the intonation contour, this may have been the root of the wide distribution of SPEAKER MEANS for /uw/.

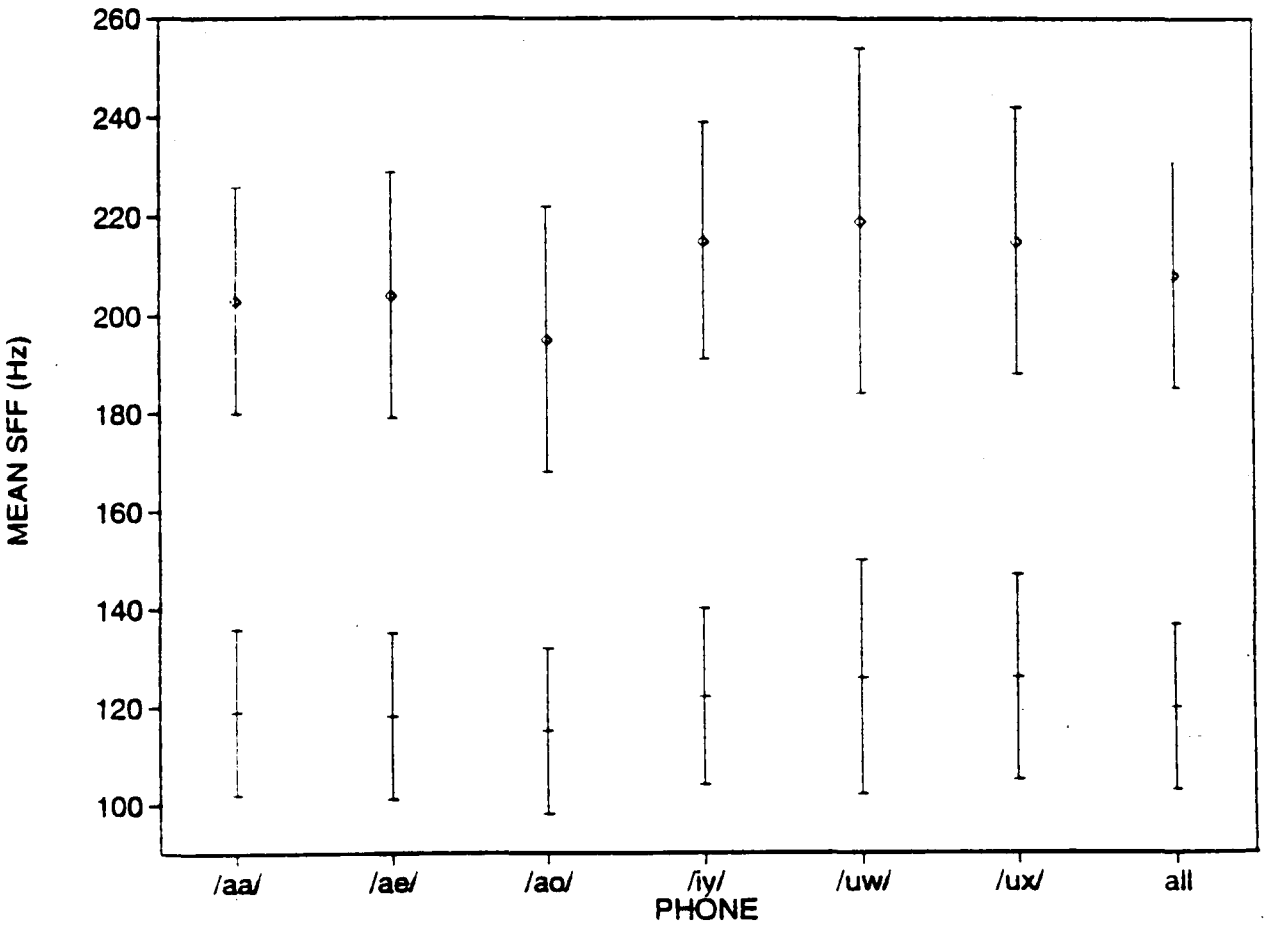


Figure 4.12: Mean SFF (Hz) by phone for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean. The mean for all speakers is on the right.

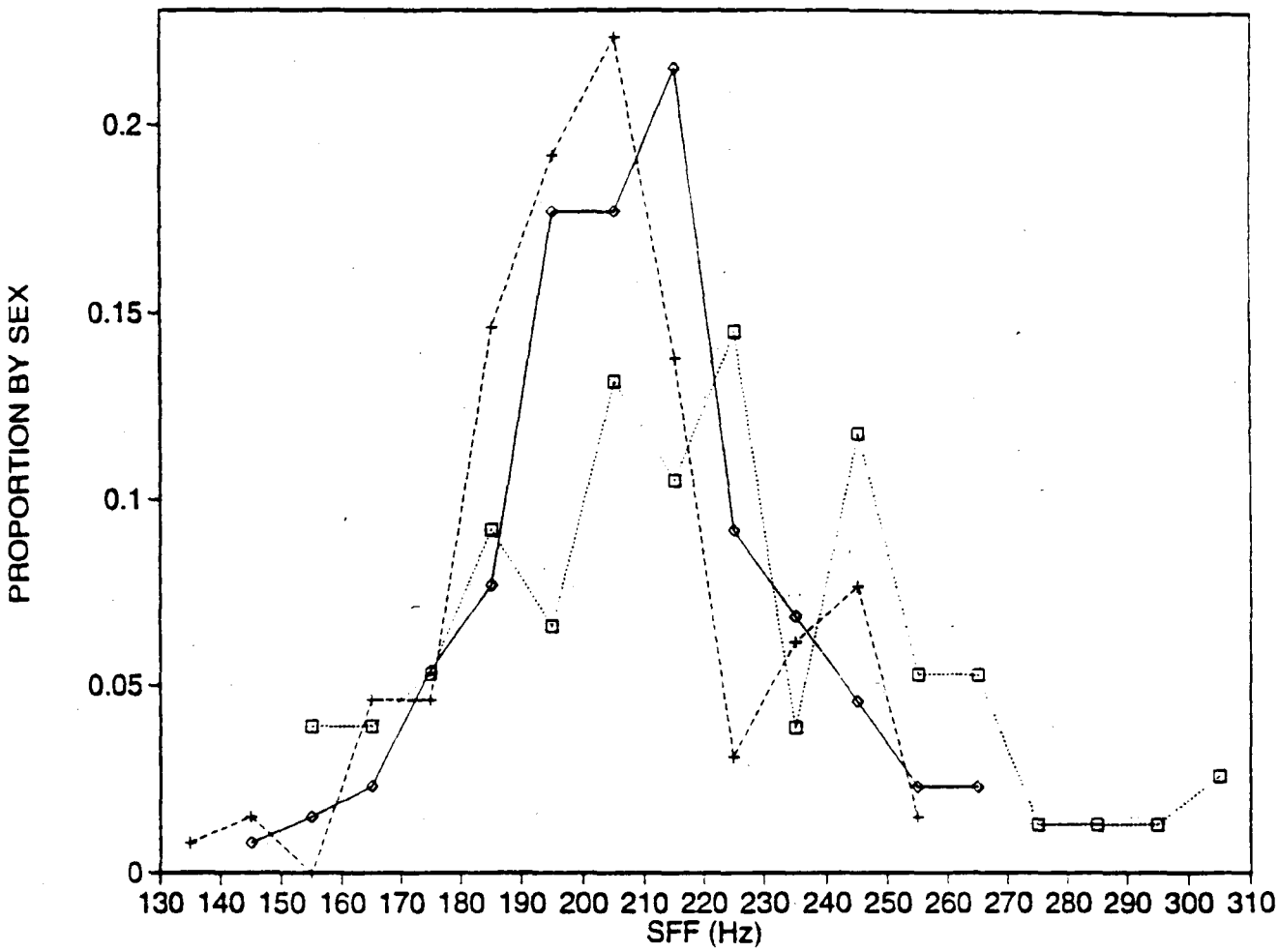


Figure 4.13: Distribution of SFF SPEAKER MEANS (Hz) for female speakers for all phones (solid line, diamonds), /aa/ (dashed line, plusses) and /uw/ (dotted line, squares). The data points represent the proportion of SPEAKER MEANS in a 10Hz interval, and are plotted at the midpoint of the interval.

Age	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
20 – 29	82	211	(21)	172	196	209	221	270
30 – 39	32	210	(21)	155	198	211	220	256
40 – 49	10	189	(22)	151	173	191	201	228
50 – 59	4	184	(46)	146	154	169	213	249
≥ 60	2	187	(37)	161	–	187	–	213

Table 4.16: Female mean SFF data (to nearest 1Hz) by age.

Age	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
20 – 29	183	120	(17)	82	108	118	129	183
30 – 39	86	119	(15)	91	108	117	128	168
40 – 49	13	114	(22)	86	99	114	123	166
50 – 59	7	129	(31)	100	103	115	149	180
≥ 60	1	130	(–)	–	–	130	–	–

Table 4.17: Male mean SFF data (to nearest 1Hz) by age.

Analysis of data by speaker variable.

Note that the means in each analysis variable group were computed from the SPEAKER MEANS (as opposed to the SLICE MEANS).

A. Analysis by age :
For both sexes, the age groups 20-29 and 30-39 each produced almost identical mean SFFs (and s.d.s) (see Tables 4.16 and 4.17 and Figure 4.14). As would be expected, these means are almost identical to the overall mean SFF, since it was the speakers in these age groups who made up the bulk of the sample. There is then a substantial drop in mean SFF for female speakers in their forties, and a small drop for men of the same age. It is difficult to draw any conclusions about SFF for speakers aged 50 and above as their representation in the sample was so small.

The means for the female speakers appear to follow the pattern of SFF change with age reported in the literature, if not the values[16]. The drop in the SFF of middle-aged women found by Stoicheff (1981) and others is in evidence here, and is of the same order (i.e. 20Hz). However, the figures here suggest this decrease occurs to women in their forties, rather than their fifties (see Section 3.1.4). Stoicheff linked the drop in SFF to the completion of the menopause; if this is the case, then the female TIMIT speakers were experiencing the menopause at an earlier age than Stoicheff's group of subjects. Certainly, the number of women with a relatively low SFF is much greater in the 40-49 age group than in the younger women. Thus, for the women aged 20-29 and 30-39, only 12% and 13% of the speakers have mean SFFs less than 190Hz; while the figure is 50Figure 4.15, which shows the mean SFFs of the sixteen women aged over 40 years. Only four of these women are above the overall mean SFF. However, the drawing of conclusions is severely hampered by the few speakers in this study over the age of 40 years, and furthermore there is no information on menopause completion for the TIMIT speakers. Consequently, while the large drop in mean SFF between the 30-39 and 40-49 age groups may indeed be

[16] Given the wide range of values for the mean female SFF reported in the literature, this is perhaps to be expected. The TIMIT data may instead provide support for the age-related trend in SFF.

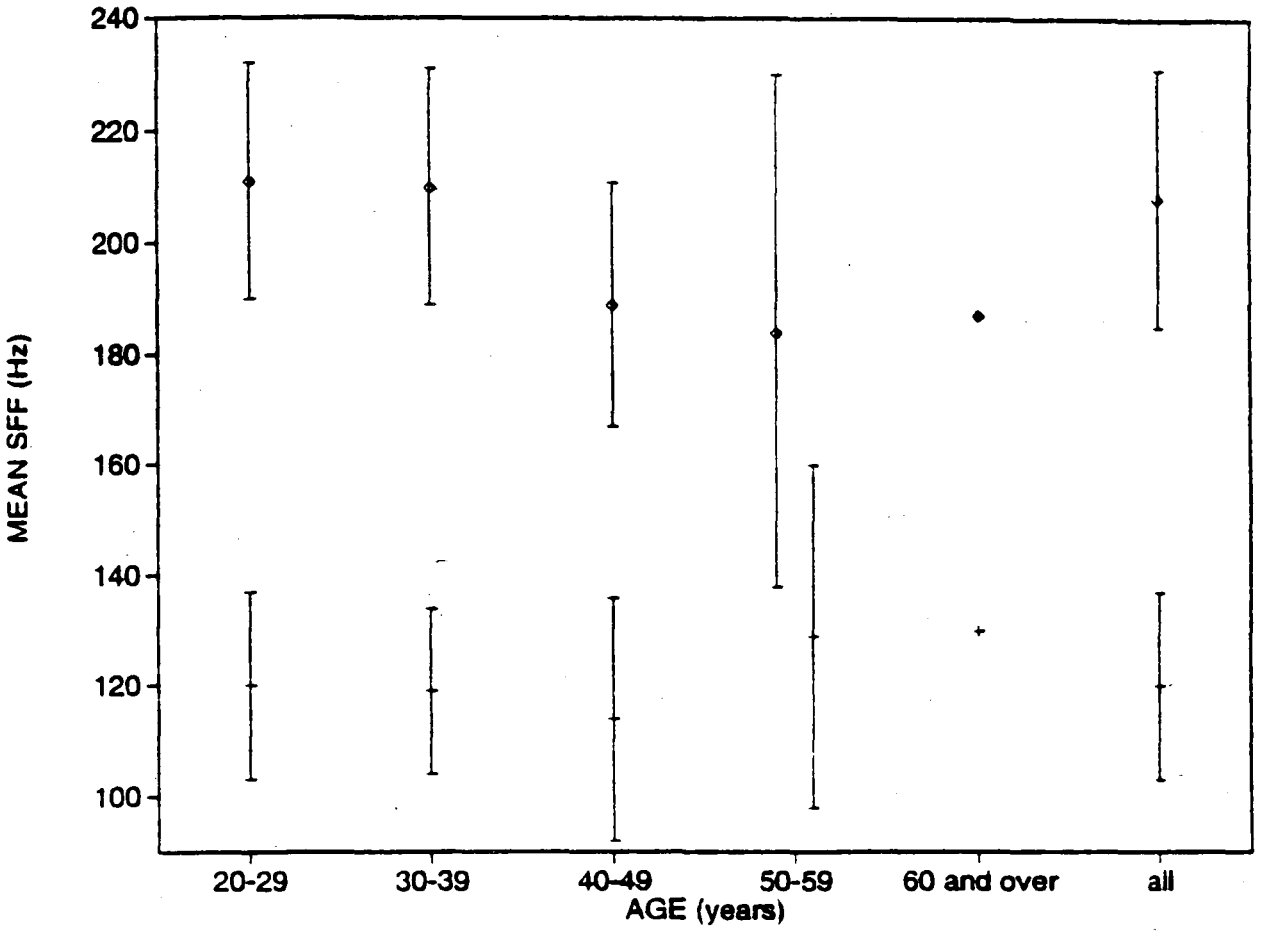


Figure 4.14: Mean SFF (Hz) by age for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean (note, no s.d. intervals are given for the age group '60 and over' as there were so few speakers). The mean for all speakers is on the right.

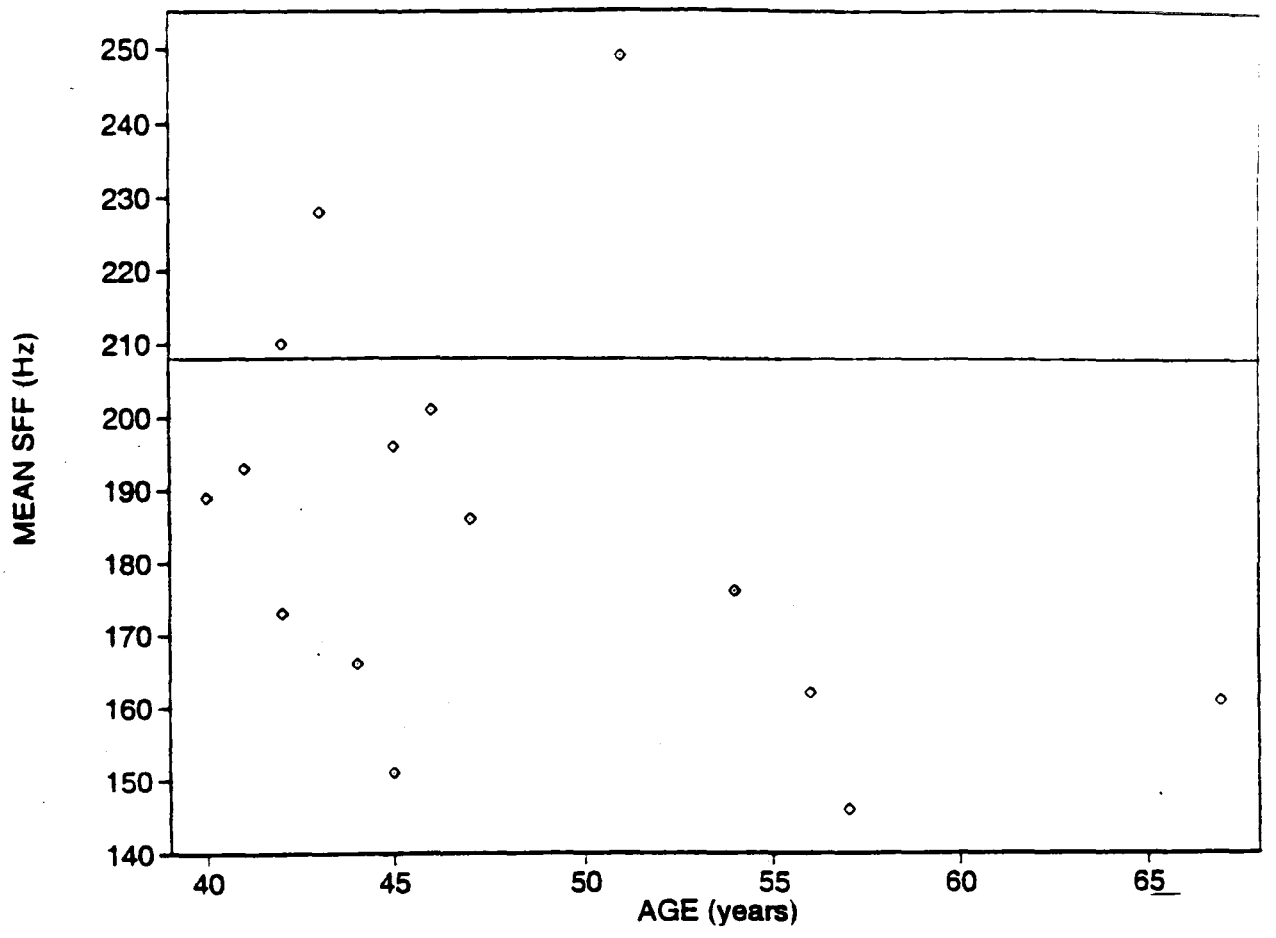


Figure 4.15: Mean SFFs (Hz) of the female speakers over 40 years old. The solid line at 208Hz represents the mean SFF of all the female speakers. Note: The 85 year-old speaker fkfb0 has been left off the graph – her mean SFF was 213Hz.

menopause-related, there may be other factors involved.

The pattern of SFF change with age for the male speakers indicates SFF is steady through the 20s and 30s, drops slightly in the 40s, and rises in the fifties. However, again there were few speakers over 40 years to lend these statistics much weight. Figure 4.16 shows the mean SFFs of the twenty-two men aged over 40 years. The men in their 40s show a tendency for lower SFFs, with two-thirds of them below the overall male mean SFF. However, if we consider the seven male speakers in the age group 50-59, we can see how the lack of data can leave a confused picture. Although the mean SFF for this age group is 129Hz, we find this is composed of SPEAKER MEANS of 100Hz, 101Hz, 108Hz, 115Hz, 148Hz, 149Hz and 180Hz, i.e. over half of these speakers have a mean SFF that is actually below the overall male mean SFF of 120Hz. Thus, although the group mean indicates an upwards trend in SFF at this age, closer inspection of the data reveals that there is no trend.

B. Analysis by height :

From an examination of Figure 4.17, showing the mean SFFs (and s.d. intervals) for the different height groups, there appears to be some correlation between height and SFF for both female and male speakers. However, the small numbers of speakers representing some of the height groups renders trends based on these groups unreliable. If we consider only those height groups composed of more than ten speakers – leaving us with the four female groups in the range 5'2"–5'9" (see Table 4.18) and the five male groups in the range 5'6"–6'3" (see Table 4.19) – then the mean SFF for the male speakers remained fairly constant with increasing speaker height, while there was a noticeable drop for the

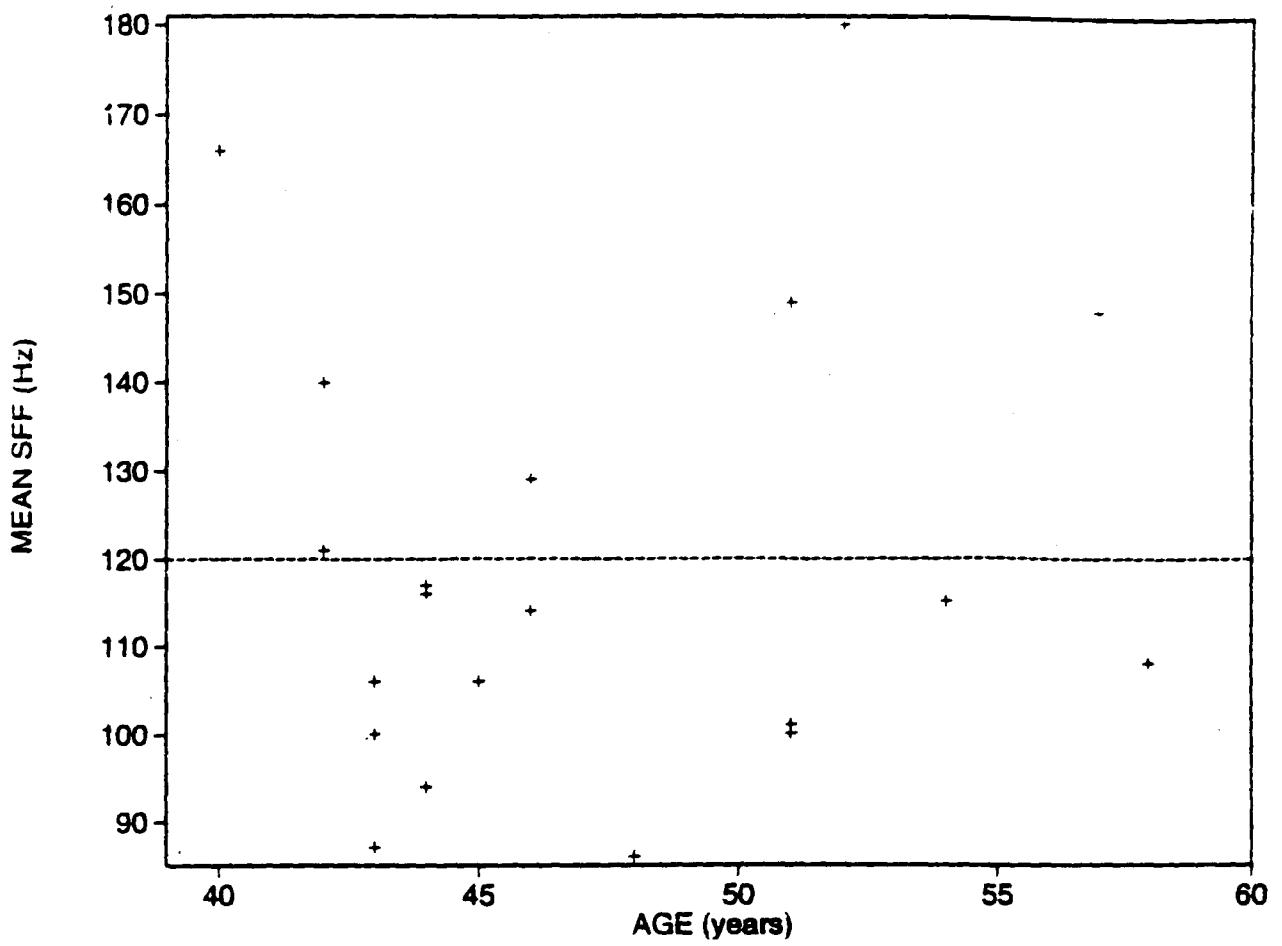


Figure 4.16: Mean SFFs (Hz) of the male speakers over 40 years old. The dashed line at 120Hz represents the mean SFF of all the male speakers. Note: The 85 year-old speaker mrjm1 has been left off the graph – his mean SFF was 130Hz.

Height	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
$\leq$ 5'1"	7	222	(37)	161	201	224	250	270
5'2" - 5'3"	25	210	(22)	151	201	215	221	246
5'4" - 5'5"	37	212	(20)	166	200	212	228	264
5'6" - 5'7"	32	206	(25)	146	192	202	215	266
5'8" - 5'9"	24	197	(17)	172	183	197	209	236
5'10" - 5'11"	4	212	(28)	183	193	207	231	251
6'0" - 6'1"	1	172	(-)	-	-	172	-	-

Table 4.18: Female mean SFF data (to nearest 1Hz) by height.

Height	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
5'2" - 5'3"	1	137	(-)	-	-	137	-	-
5'4" - 5'5"	3	134	(26)	111	-	123	-	161
5'6" - 5'7"	26	123	(17)	96	114	121	130	183
5'8" - 5'9"	63	119	(16)	82	107	118	129	166
5'10" - 5'11"	77	121	(20)	86	108	118	131	180
6'0" - 6'1"	67	120	(16)	92	108	118	132	166
6'2" - 6'3"	42	117	(15)	96	106	116	125	168
6'4" - 6'5"	7	112	(16)	91	99	112	124	135
$\geq$ 6'6"	4	112	(5)	107	109	111	114	118

Table 4.19: Male mean SFF data (to nearest 1Hz) by height.

female speakers. These height groups will now be examined in more detail.

Figure 4.18 shows the distribution of SPEAKER MEANS for the four height groups covering the majority of female speakers. The Figure shows that speakers in the height groups 5'2"-5'3" and 5'4"-5'5" had a similar distribution of SFF, with the greater proportion of them having a mean SFF of 200-220Hz. For speakers with heights between 5'6" and 5'9", there was clearly a trend for lower mean SFFs (i.e. 180-200Hz), although nearly as large a proportion had mean SFFs of 200-220Hz.

Figure 4.19 shows the distribution of mean SFF for male speakers in the height groups between 5'6" and 6'3". While the distributions for the five height groups are generally similar, a comparison of the profiles for the groups 5'5"-5'6" and 6'2"-6'3" indicates the existence of a downward trend in mean SFF with increasing speaker height. This downward trend is further suggested by the mean SFFs of the male speakers in the tallest height groups, as only two out of the seven speakers in the 6'4"–6'5" group and none of the speakers in the $\geq$6'6" group had a mean SFF greater than the overall male mean[17].

C. Analysis by ethnic group :

There is little difference between the mean SFFs of black and white women and men, although the means for the black women and men were both slightly lower than for the white (see Tables 4.20 and 4.21). This is in agreement with the findings of Hitch & Holbrook (1981). The only major difference between the speaker groups is that the black women had a much smaller distribution of SFF SPEAKER MEANS than the white women (see the s.d.s in Table 4.20). However, as the black women's means were distributed evenly between 180-225Hz, the same region in which 79% of the white women's means were concentrated,

[17]Of the smaller heights, there was only one speaker representing the group 5'2"–5'3", and in the 5'4"–5'5" group the SPEAKER MEANS were 111Hz, 123Hz and 161Hz.

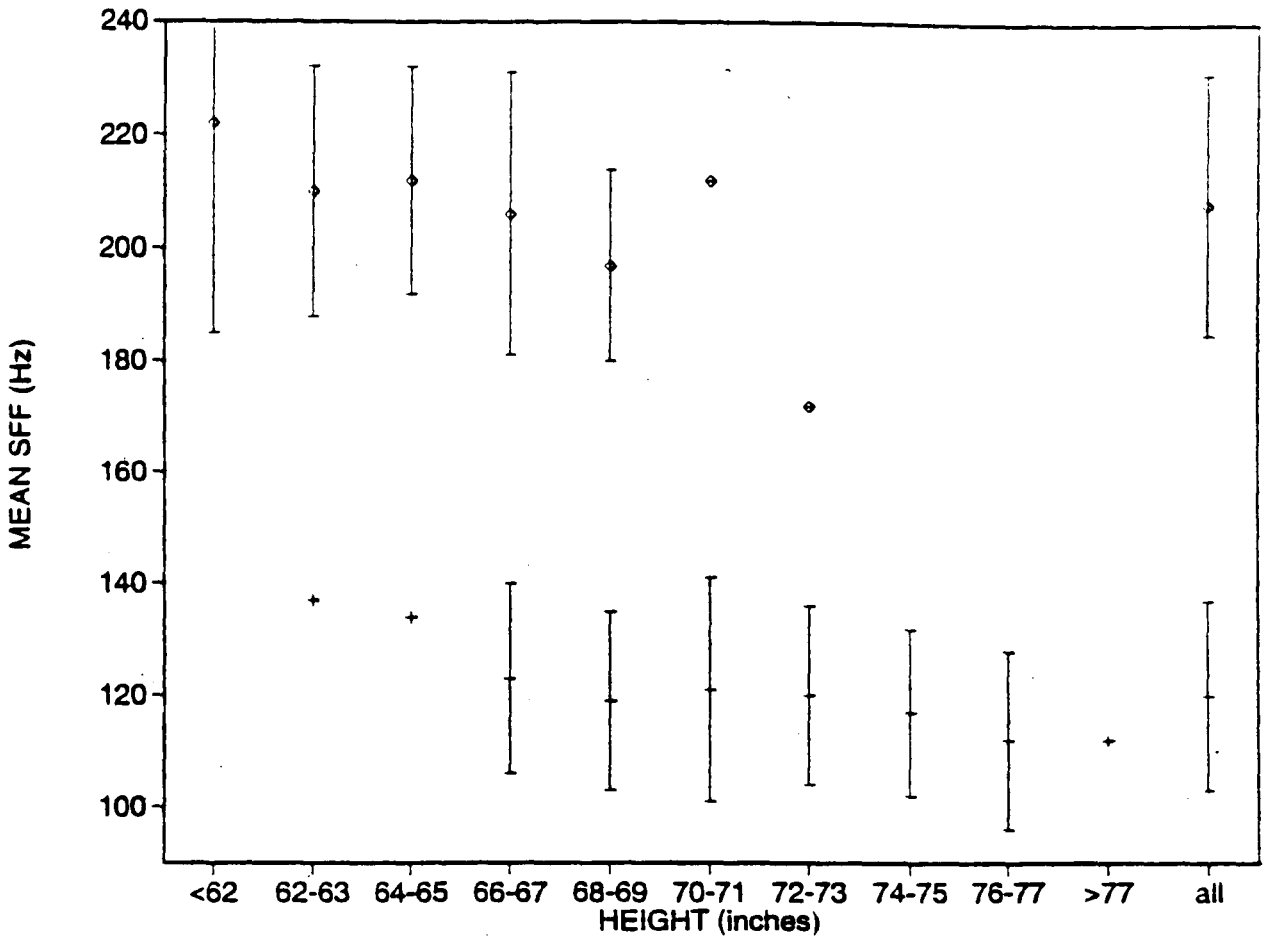


Figure 4.17: Mean SFF (Hz) by height for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean (note no s.d. intervals are given for some of the height groups as there were so few speakers). The mean for all speakers is on the right.

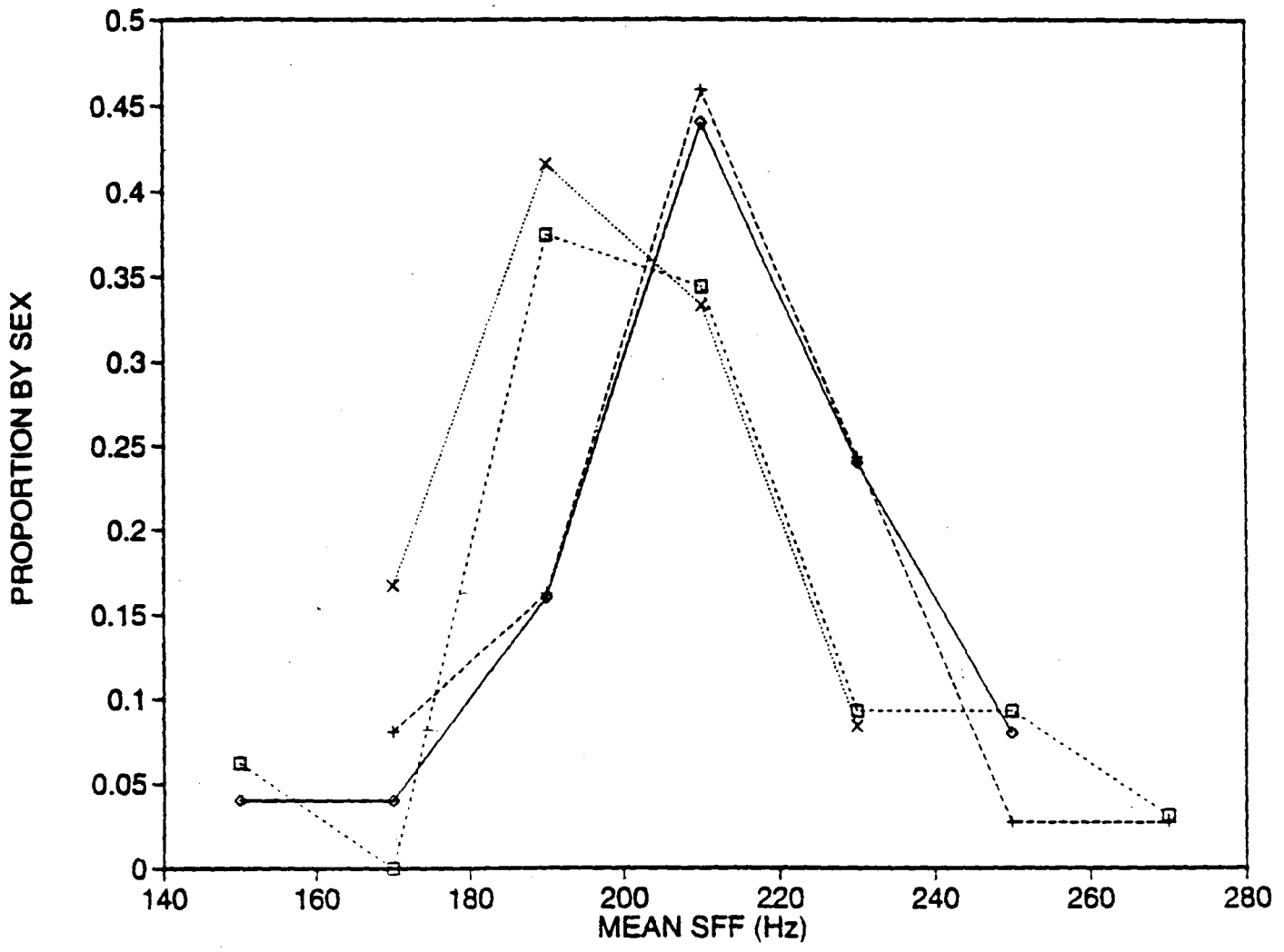


Figure 4.18: Distribution of SFF SPEAKER MEANS (Hz) for female speakers for the height groups 5'2"-5'3" (solid line, diamonds), 5'4"-5'5" (dashed line, plusses). 5'6"-5'7" (dashed line, squares) and 5'8"-5'9" (dotted line, crosses). The data points represent the proportion of SPEAKER MEANS in a 20Hz interval for that height group, and are plotted at the midpoint of the interval.

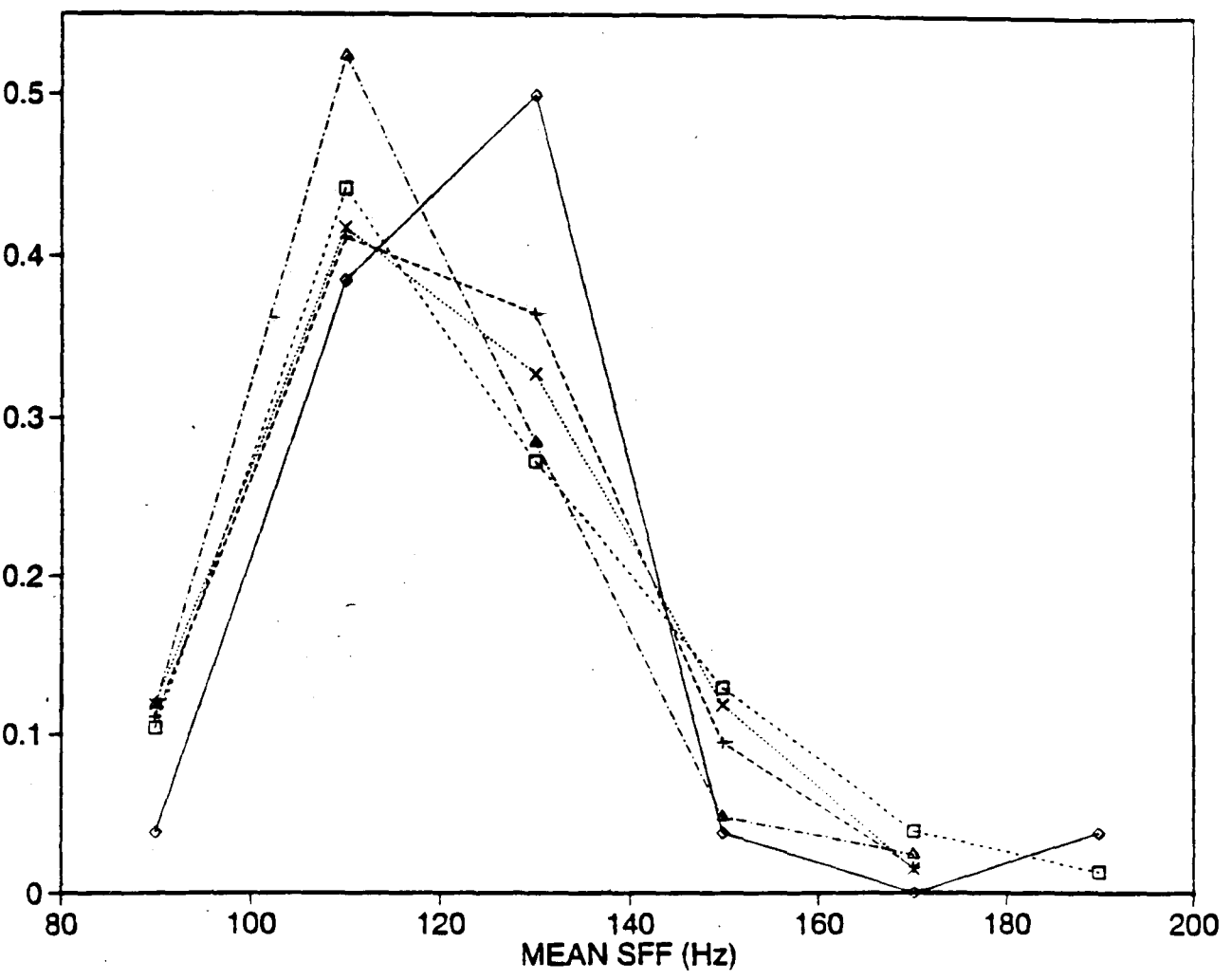


Figure 4.19: Distribution of SFF SPEAKER MEANS (Hz) for male speakers for the height groups 5'6"-5'7" (solid line, diamonds), 5'8"-5'9" (dashed line, plusses), 5'10"-5'11" (dashed line, squares), 6'0"-6'1" (dotted line, crosses) and 6'2"-6'3" (dotted-and-dashed line, triangles). The data points represent the proportion of SPEAKER MEANS in a 20Hz interval for that height group, and are plotted at the midpoint of the interval.

Ethnic group	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
black	10	201	(14)	183	192	199	216	222
white	117	208	(24)	146	193	208	221	270

Table 4.20: Female mean SFF data (to nearest 1Hz) by ethnic group.

Ethnic group	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
black	12	117	(17)	95	101	116	129	146
white	259	120	(17)	82	108	118	129	183

Table 4.21: Male mean SFF data (to nearest 1Hz) by ethnic group.

the explanation for the smaller s.d. probably lies in there only being ten speakers in the sample, i.e. the sample of black female speakers was insufficient to produce a wide range of SPEAKER MEANS.

D. Analysis by dialect :
For the female speakers, the means for the dialect regions ranged from 198Hz to 218Hz (see Table 4.22, and Figure 4.20). By way of contrasting the different distributions of SFF SPEAKER MEANS within the dialect regions, the following table lists the numbers of SPEAKER MEANS at 10Hz intervals for the dialect regions which produced the lowest and highest mean SFFs:

SFF	dr3	dr4
150–160	–	1
160–170	–	–
170–180	4	–
180–190	2	–
190–200	1	2
200–210	4	3
210–220	2	5
220–230	1	2
230–240	–	3
240–250	1	1
250–260	–	1
260–270	–	1

While the representation of speakers in each dialect group was relatively small (as the 130 female speakers were spread across eight dialect regions), they are of a sufficient number to indicate that female speakers from different dialect regions can be expected to differ in mean SFF. Thus, if we consider the regions dr3 and dr4, three-quarters of the speakers from dr3 had a mean SFF less than 210Hz, compared to only a third of the speakers from dr4.

There was little difference between the dialects for the male speakers (see Table 4.23). The mean SFFs ranged from 117Hz to 125Hz, although for dr8, the group comprised of speakers who moved around the country during childhood, the mean was 112Hz.

For both sexes, dr4 produced the highest mean SFF, while dr3 produced the lowest female SFF and joint lowest male SFF (discounting dr8). Otherwise, the dialects did not produce a pattern of SFF repeated across the sexes

Dialect	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
dr1	15	204	(29)	146	187	210	219	264
dr2	18	202	(21)	155	192	201	210	266
dr3	15	198	(20)	172	181	201	210	242
dr4	19	218	(26)	151	206	216	234	270
dr5	26	210	(22)	166	194	208	220	256
dr6	11	207	(22)	176	192	198	225	245
dr7	18	208	(22)	162	196	207	218	251
dr8	8	216	(20)	192	197	216	229	248

Table 4.22: Female mean SFF data (to nearest 1Hz) by dialect.

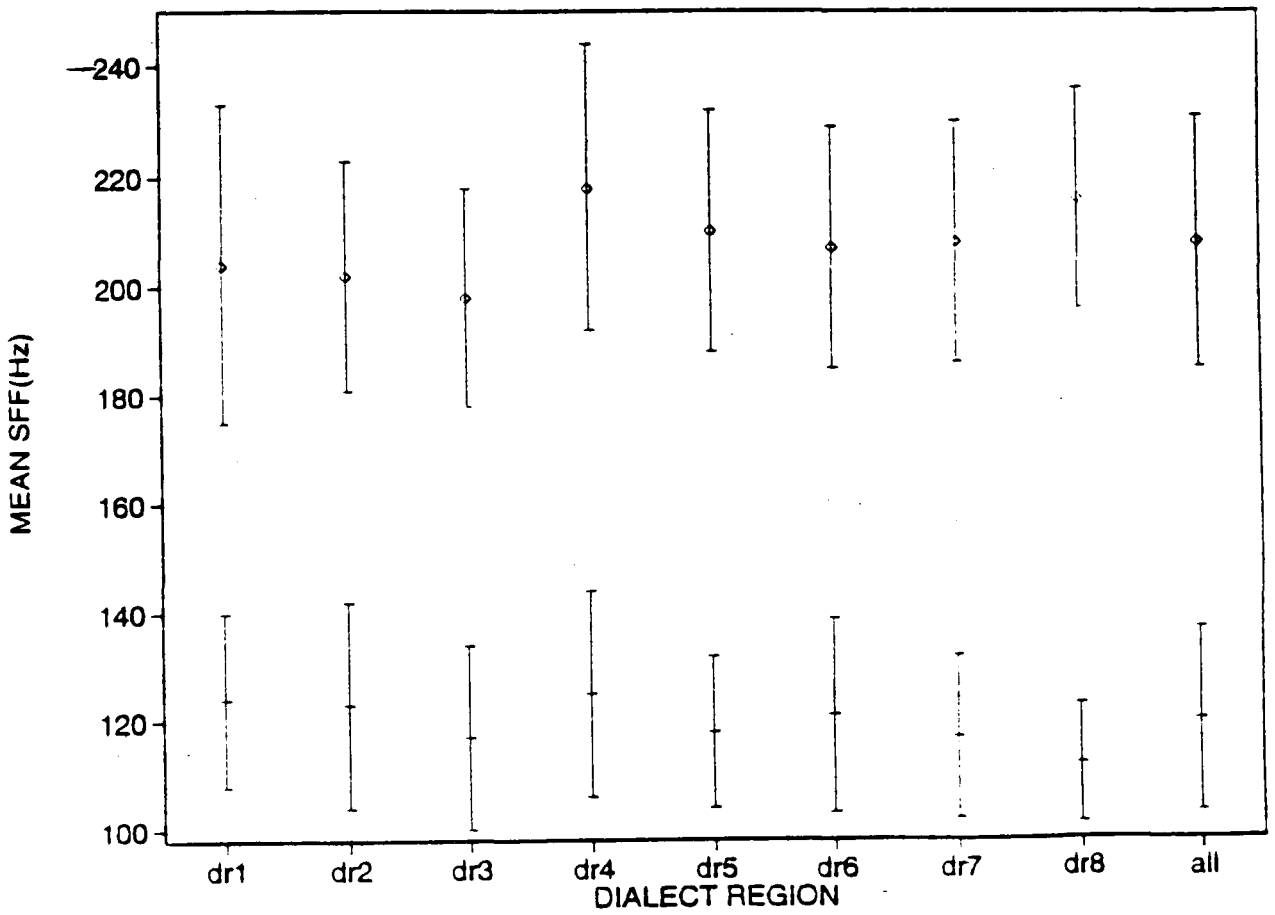


Figure 4.20: Mean SFF (Hz) by dialect for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean. The mean for all speakers is on the right.

Dialect	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
dr1	22	124	(16)	101	111	122	133	166
dr2	47	123	(19)	95	108	120	133	180
dr3	51	117	(17)	86	106	116	125	177
dr4	46	125	(19)	92	110	119	137	183
dr5	39	118	(14)	92	110	117	125	146
dr6	21	121	(18)	92	108	120	130	168
dr7	48	117	(15)	82	106	118	126	155
dr8	16	112	(11)	98	104	110	121	135

Table 4.23: Male mean SFF data (to nearest 1Hz) by dialect.

Sex	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
f	130	-4.8	(2.3)	-9.1	-6.3	-5.1	-3.7	1.8
m	290	-6.1	(2.0)	-11.7	-7.5	-6.4	-5.0	0.8

Table 4.24: Female and male mean H_1–H_2 data (to nearest 0.1dB). Note that this data is computed from the SPEAKER MEANS.

4.2.2 Relative amplitude of the first harmonic

Analysis of overall data

The mean H_1–H_2 for the female speakers was -4.8dB (2.3dB); for the male speakers it was -6.1dB (2.0dB) (see Table 4.24. The figures are substantially different to those reported in the literature, being both much lower and much closer in value. The closest figures are, interestingly, from those studies using U.S. English speakers as subjects (Klatt & Klatt 1990, Nittrouer *et al.* 1990 – see Table 3.8), and then only for the male speakers. In contrast to the previous studies, most of which reported a positive value for the female mean H_1–H_2, the analysis of the TIMIT data produced values which were very much negative values. In other words, for the vast majority of speakers measured for this study, the amplitude of their first harmonic was on average less than their second. What is most interesting is that the female-male difference in H_1–H_2 of 1.4dB is much less, indicating a much smaller tendency for women's speech to be breathier than men's than has previously been reported in the literature.

A. Distribution of H_1–H_2

The distribution of individual mean H_1–H_2 scores is illustrated in histogram form in Figure 4.21, with the female speakers ranged from -9.1dB to 1.8dB, and the males from -11.7dB to 0.8dB. The histogram reveals a substantial overlap between the female and male SPEAKER MEANS, much more so than reported in the literature. 78.5% of the female speakers and 76.9% of the male speakers had a mean H_1–H_2 value within the range -8dB to -3dB. However, male speakers were more likely to have a more negative mean H_1–H_2: 57.9% (168) of the men produced a mean H_1–H_2 less than -6dB, while only 31.5% (41) of the women did. Furthermore, if we define an arbitrary division between breathy and non-breathy speakers, given by the mean H_1–H_2 for all speakers of -5.5dB[18], then we find 56.9% (74) of the female speakers and 31.7% (92) of the male speakers were 'breathy'.

The observations made about the distribution of the SPEAKER MEANS also hold for the SLICE MEANS. The mean H_1–H_2 values for female and male speakers computed from the SLICE MEANS were slightly closer at -4.9dB (3.6dB) and -6.1dB (2.7dB) respectively. The histogram produced in Figure 4.22 illustrates the tendency the female speakers showed for a more positive H_1–H_2. Applying the non-breathy/breathy 'division' of -5.5dB to the SLICE MEANS, we find 54.3% (1463) of the female segments and 36.6% (2227) of the male segments were in excess of it. While the bulk of the segments had a H_1–H_2 in the central overlap region (60.0% (1616) of the female and 65.8% (4002) of the male scores were in the region -8dB $< H_1$–$H_2 <$ -3dB), the female speakers were more likely to produce extreme values. While this was to be expected at the more positive end of the distribution (see the right-hand side of Figure 4.22 and the lower portion of Table 4.26), this was also

[18]To account for the greater male representation in this sample of speakers, this overall mean is formed from the mean of the female and male means. If the mean was computed from all 420 H_1–H_2 SPEAKER MEANS, it would be weighted rather more towards the overall male mean.

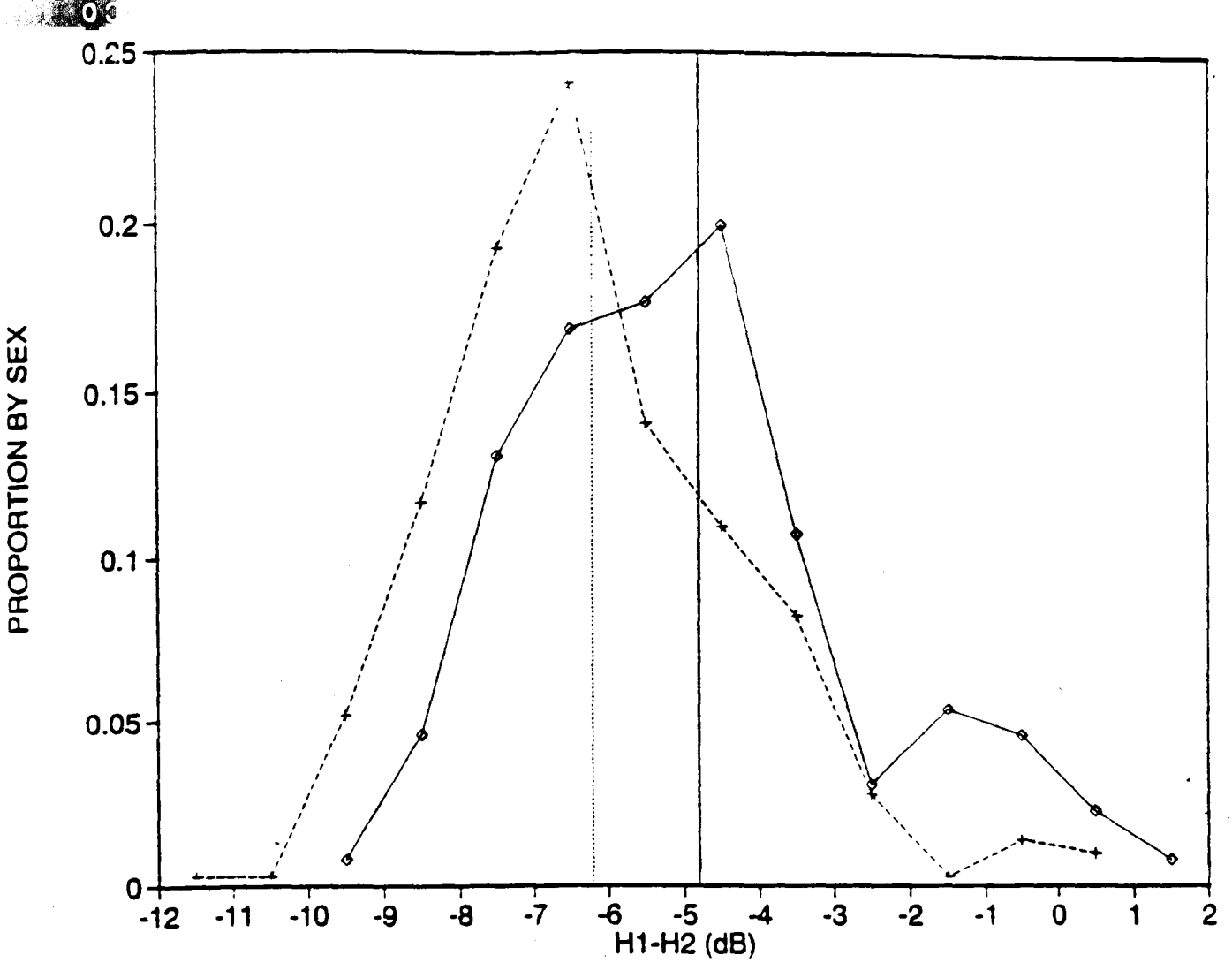


Figure 4.21: Distribution of H_1–H_2 SPEAKER MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of SPEAKER MEANS in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean H_1–H_2. The data used to produce the histogram is given in Table 4.25. The overall mean H_1–H_2 scores for each sex are shown as single, vertical lines.

H_1–H_2 (dB)	Female		Male	
	n	%	n	%
-12 – -11	–	–	1	0.3
-11 – -10	–	–	1	0.3
-10 – -9	1	0.8	15	5.2
-9 – -8	6	4.6	34	11.7
-8 – -7	17	13.1	56	19.3
-7 – -6	22	16.9	70	24.1
-6 – -5	23	17.7	41	14.1
-5 – -4	26	20.0	32	11.0
-4 – -3	14	10.8	24	8.3
-3 – -2	4	3.1	8	2.8
-2 – -1	7	5.4	1	0.3
-1 – 0	6	4.6	4	1.4
0 – 1	3	2.3	3	1.0
1 – 2	1	0.8	–	–

Table 4.25: Number and percentage of female and male H_1–H_2 SPEAKER MEANS at 1dB intervals. A dash (–) indicates there were no speakers with a mean SFF in that interval. This data was used to plot the histogram in Figure 4.21.

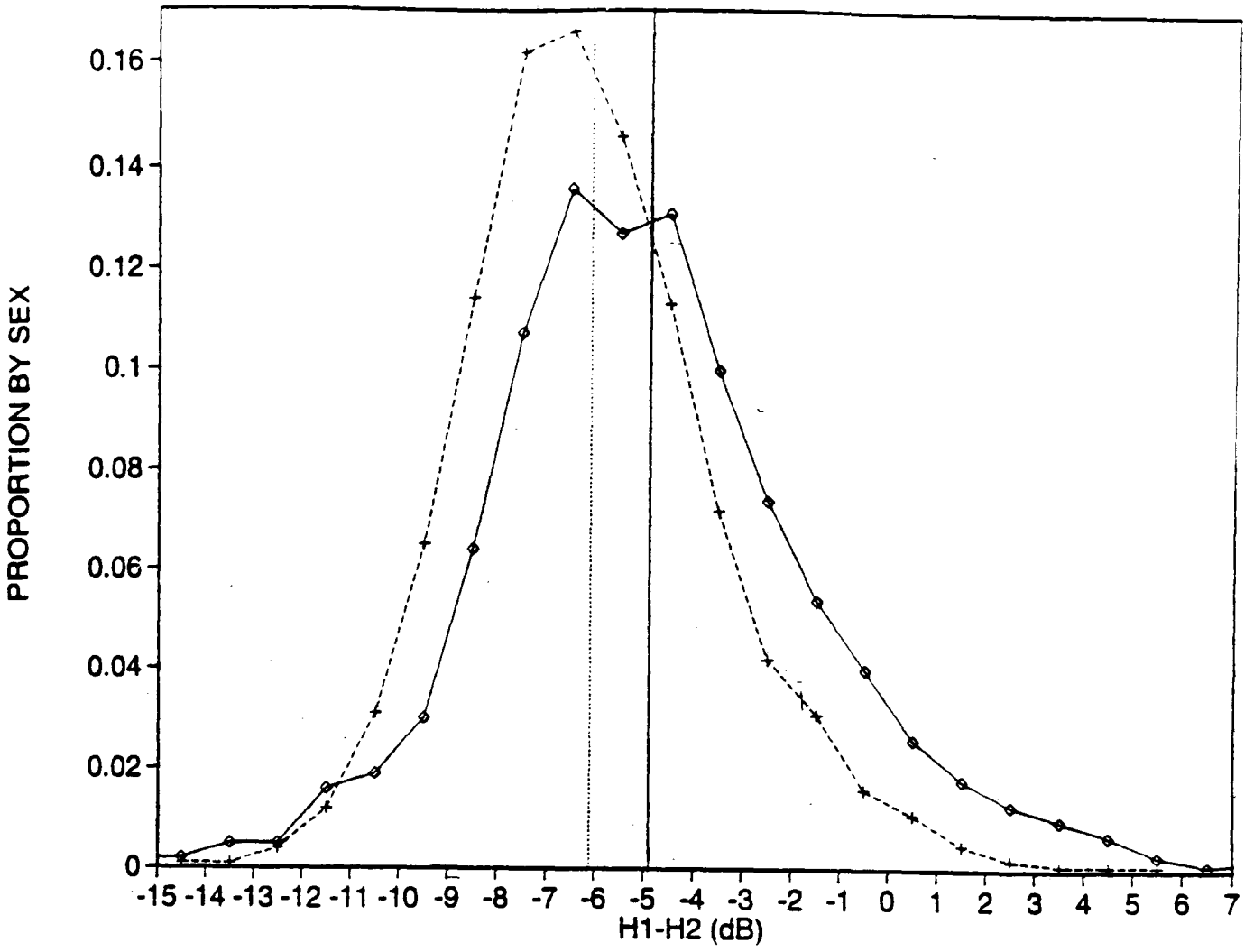


Figure 4.22: Distribution of $\bar{H_1}$–H_2 SLICE MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of SLICE MEANS in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean H_1–H_2. Note: of the slices outside the range of this histogram, 18 female and 7 male slices had a mean H_1–H_2 less than -15dB, and 15 female and 2 male slices were greater than 7dB. The data used to produce the histogram is given in Table 4.26. The overall mean H_1–H_2 scores for each sex are shown as single, vertical lines.

the case for more negative values. Considering the positive extreme first, 8.5% (228) of female slices had a positive value for H_1–H_2, compared to only 2.3% (137) of male slices. Furthermore, the proportions of slices with a H_1–H_2 in excess of 3dB were 2.6% (71) and 0.3% (27) for the female and male speakers respectively. The more interesting result is at the more negative end of the distribution, where for each 1dB division of the histogram in Figure 4.22 less than -11dB, there are a proportionately greater number of female slices than there are male slices; indeed, 3.4% (91) of female slices and 2.0% (119) of male slices are less than -11dB. Positive values of H_1–H_2 were attained by 47.7% (62) of the women, but only 15.9% (46) of the men; while the extreme negative values were spoken by 41.5% (54) of the women, but only 22.8% (66) of the men. This appears to indicate that the women tended to be much more varied in their use of H_1–H_2.

B. Distribution of H_1 and H_2

We will now consider the data concerning the amplitudes of the first and second harmonics (see Tables 4.27 and 4.28). As has been seen from the examination of the H_1–H_2 data, a speaker's first harmonic generally has a lower amplitude than their second harmonic. This is illustrated in Figure 4.23 for H_1 and H_2, which reveals a tendency for women's speech

H_1–H_2 (dB)	Female		Male	
	n	%	n	%
-29 – -28	1	0.0	–	–
-28 – -27	–	–	–	–
-27 – -26	–	–	–	–
-26 – -25	–	–	–	–
-25 – -24	–	–	–	–
-24 – -23	–	–	–	–
-23 – -22	–	–	1	0.0
-22 – -21	–	–	–	–
-21 – -20	1	0.0	–	–
-20 – -19	1	0.0	1	0.0
-19 – -18	3	0.1	–	–
-18 – -17	3	0.1	1	0.0
-17 – -16	3	0.1	1	0.0
-16 – -15	6	0.2	3	0.0
-15 – -14	5	0.2	7	0.1
-14 – -13	13	0.5	8	0.1
-13 – -12	14	0.5	23	0.4
-12 – -11	42	1.6	74	1.2
-11 – -10	50	1.9	191	3.1
-10 – -9	82	3.0	395	6.5
-9 – -8	173	6.4	691	11.4
-8 – -7	289	10.7	986	16.2
-7 – -6	367	13.6	1006	16.6
-6 – -5	341	12.7	888	14.6
-5 – -4	352	13.1	689	11.3
-4 – -3	268	10.0	436	7.2
-3 – -2	199	7.4	256	4.2
-2 – -1	146	5.4	186	3.1
-1 – 0	108	4.0	98	1.6
0 – 1	70	2.6	68	1.1
1 – 2	49	1.8	28	0.5
2 – 3	36	1.3	14	0.2
3 – 4	27	1.0	9	0.1
4 – 5	19	0.7	8	0.1
5 – 6	7	0.3	5	0.1
6 – 7	2	0.1	3	0.0
7 – 8	6	0.2	–	–
8 – 9	3	0.1	1	0.0
9 – 10	3	0.1	–	–
10 – 11	1	0.0	1	0.0
11 – 12	2	0.1	–	–
12 – 13	–	–	–	–
13 – 14	–	–	–	–
14 – 15	–	–	–	–
15 – 16	1	0.0	–	–

Table 4.26: Number and percentage of female and male H_1–H_2 SPEAKER MEANS at 1dB intervals. A dash (–) indicates there were no speakers with a mean SFF in that interval. This data was used to plot the histogram in Figure 4.22.

Sex	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
f	130	90.4	(3.3)	81.7	88.5	90.5	93.1	98.3
m	290	85.2	(3.7)	73.2	83.0	84.9	87.4	96.7

Table 4.27: Female and male mean H_1 data (to nearest 0.1dB). Note that this data is computed from the SPEAKER MEANS.

Sex	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
f	130	95.3	(3.7)	87.4	92.6	95.3	98.1	103.9
m	290	91.2	(4.0)	79.2	88.9	91.3	93.7	104.7

Table 4.28: Female and male mean H_2 data (to nearest 0.1dB). Note that this data is computed from the SPEAKER MEANS.

to contain higher first and second harmonic amplitudes than men's. While the s.d.s of H_1 and H_2 are very similar (for both sexes), the smaller difference between the female and male H_2 means indicate this tendency was greater for the amplitude of the first harmonic. The extent of these tendencies will now be examined in more detail.

The histogram of the H_1 SPEAKER MEANS in Figure 4.24 shows quite a separation between the female and male SPEAKER MEANS. If we set a cut-off point to separate the sexes at 88dB, we find 21.5% (28) of female speakers had a first harmonic amplitude below it, and 21.7% (63) of male speakers were above it. If we do the same for H_2 (see Figure 4.25), we can see much more of an overlap in the histogram than was the case for H_1. This, inevitably, leads to a less decisive cut-off. For a cut-off at 93dB, 29.2% (38) of female speakers were below it, while 31.7% (92) of male speakers were above it. If we now turn our attention to the SLICE MEANS, to see how the harmonic amplitude data for individual speech segments affects the cut-off figures, we find the separation between women's and men's speech to be not so distinctive (see the histograms in Figures 4.26 and 4.27). Using the 88dB cut-off for H_1, 28.6% (771) of the female segments were below it, and 28.6% (1738) of the male segments were above it. While this looks like a promising result, the 771 female segments were spoken by 90.8% (118) of the women, and the 1738 male segments by 82.8% (240) of the men. Similarly, using a cut-off set at 93.45dB for H_2[19], we find 35.7% (961) of female segments were below it and 35.1% (2136) of male segments were above it. These segments were spoken by 96.9% (126) of the women and 84.5% (245) of the men. The result of this is that while there was a tendency for female speakers to have higher amplitude harmonics than male speakers, it is clear that the range of amplitudes attained by the vast majority of these speakers overlapped.

C. Range of H_1–H_2

The purpose here is to investigate whether speakers maintain the relative amplitude of their first harmonic at a consistent level, either as a result of their anatomical makeup or as a paralinguistic signal of personality. Two indicators of range will be considered, the speaker's entire range of H_1–H_2, and the s.d. of their H_1–H_2 values.

The mean range of H_1–H_2 (i.e. the mean of each speaker's range) was 10.8dB (4.5dB) for

[19]This figure is the mean of the female and male group H_2 means, and was used in place of the earlier 93dB cut-off because it gave greater equality between the female scores below and male scores above the cut-off. For the 93dB cut-off, 32.3% (869) of female slices were below it, 38.2% (2324) of male slices were above it.

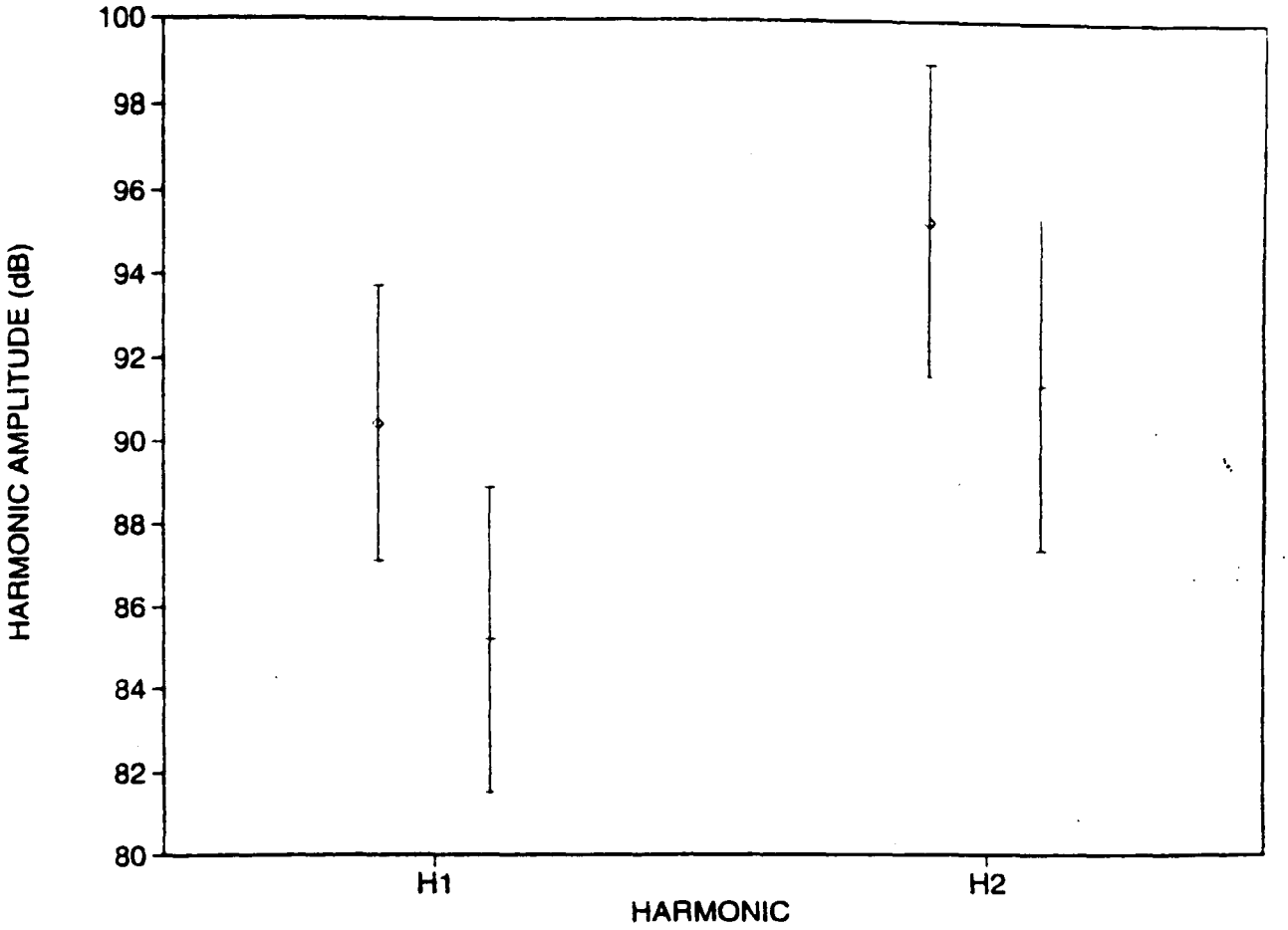


Figure 4.23: Mean H_1 and H_2 (dB) for female (diamonds) and male (plusses) speakers. The s.d.s are represented by vertical lines around the mean.

the female speakers and 7.0dB (2.9dB) for the males. This indicates a substantial trend for female speakers to be more varied in their use of H_1–H_2. However, as can be seen from Figure 4.21, the female speakers showed greater diversity in the range of H_1–H_2 they employed. While there were a few men with relatively large H_1–H_2 ranges, the overwhelming majority (90.3% or 262) had a range less than 10dB, compared to 50.8% (66) of women. The men were also far more likely to have a restricted range (22.1% (64) had a range less than 6dB, compared to only 5.4% (7) of the women). Interestingly, a large number of female speakers had similar H_1–H_2 ranges to the males. In fact, 48.5% of the female speakers had a range within 1 s.d. of the male mean H_1–H_2 range (an interval which contained 78.6% (228) of male speakers). Where the women differed from the men was in the number of speakers who employed a large range of H_1–H_2, with almost half of the female speakers having a range of 10dB or more, compared to less than a tenth of the males. As is illustrated in Figure 4.21, the bulk (42.3% or 55) of these female speakers had a range between 10dB and 17dB (compared to only 7.9% (23) of the male speakers).

Another way of looking at the range of H_1–H_2 is to consider the s.d.s of the H_1–H_2 SPEAKER MEANS. This has the effect of allowing us to examine where the bulk of a speaker's H_1–H_2 SLICE MEANS lay, and it also removes the possibility that the large female ranges were due to single, rogue values (although a 90% range would perhaps be better at this). The histogram of s.d.s produced in Figure 4.29 confirms the observations based on the full H_1–H_2 ranges, with the female speakers much more likely to have a large s.d. of H_1–H_2 than the male speakers. 33.8% (44) of female speakers had a s.d. greater than 3dB, yet only 3.4% (10) of the males did. Looked at another way, the overwhelming majority of the male speakers had a s.d. less than 2.5 dB (90.7% or 263, compared to 45.4% or 59 of the female speakers). The mean s.d.s were 2.7dB (1.0dB) for the female speakers, and

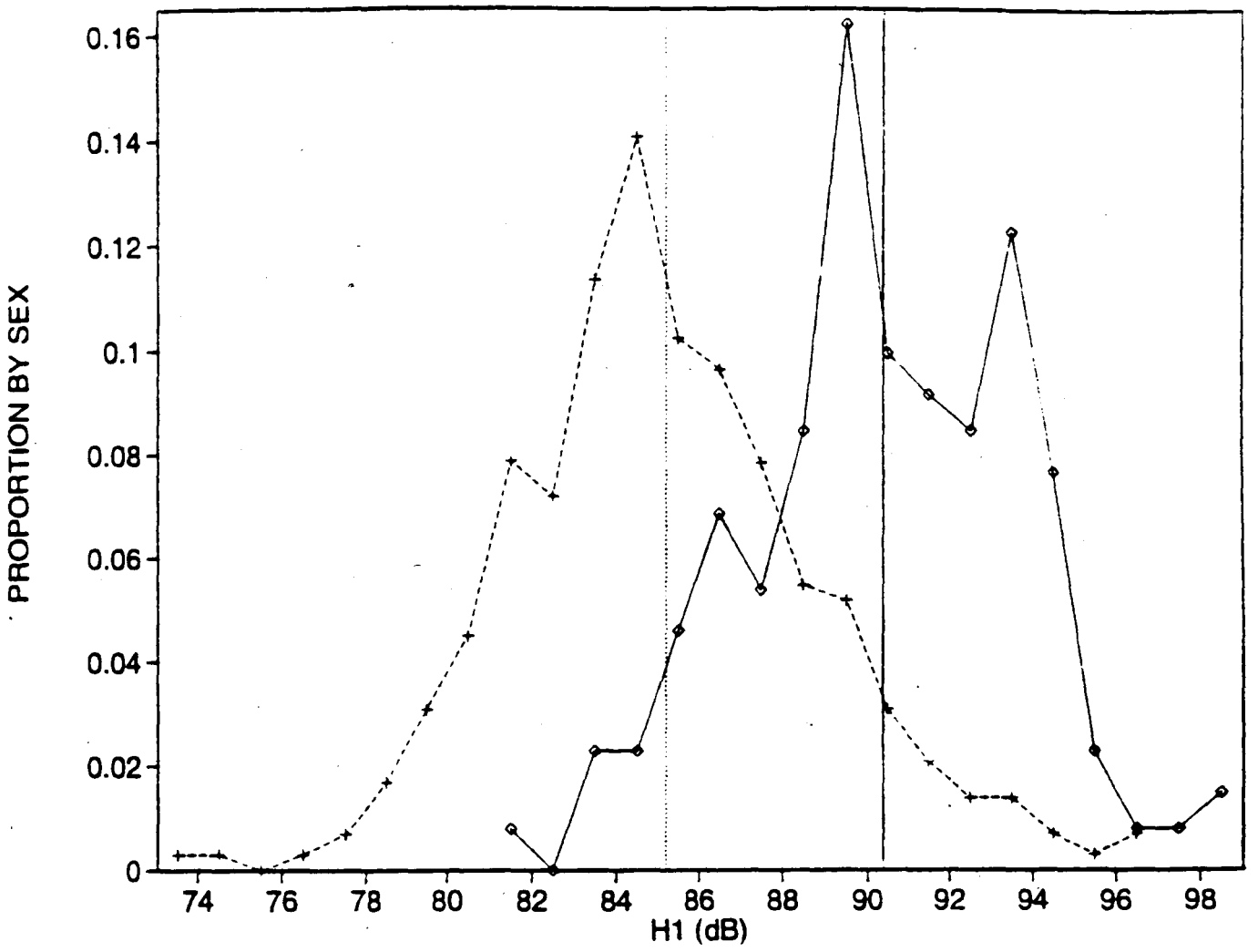


Figure 4.24: Distribution of H_1 SPEAKER MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of SPEAKER MEANS in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean H_1. The overall mean H_1 scores for each sex are shown as single, vertical lines.

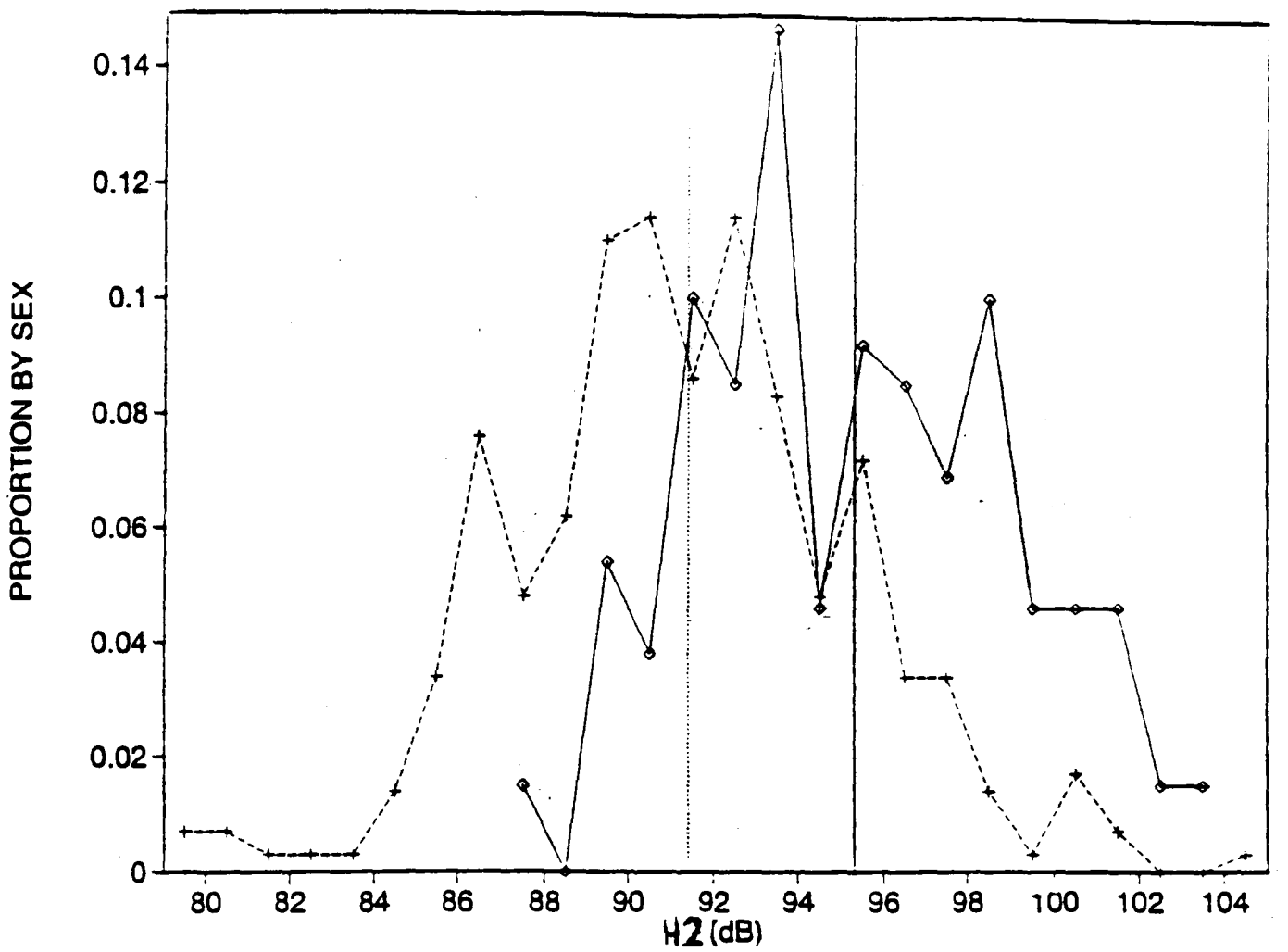


Figure 4.25: Distribution of H_2 SPEAKER MEANS for female (solid line. diamonds) and male (broken line. plusses) speakers. The data points represent the proportion of SPEAKER MEANS in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean H_2. The overall mean H_2 scores for each sex are shown as single, vertical lines.

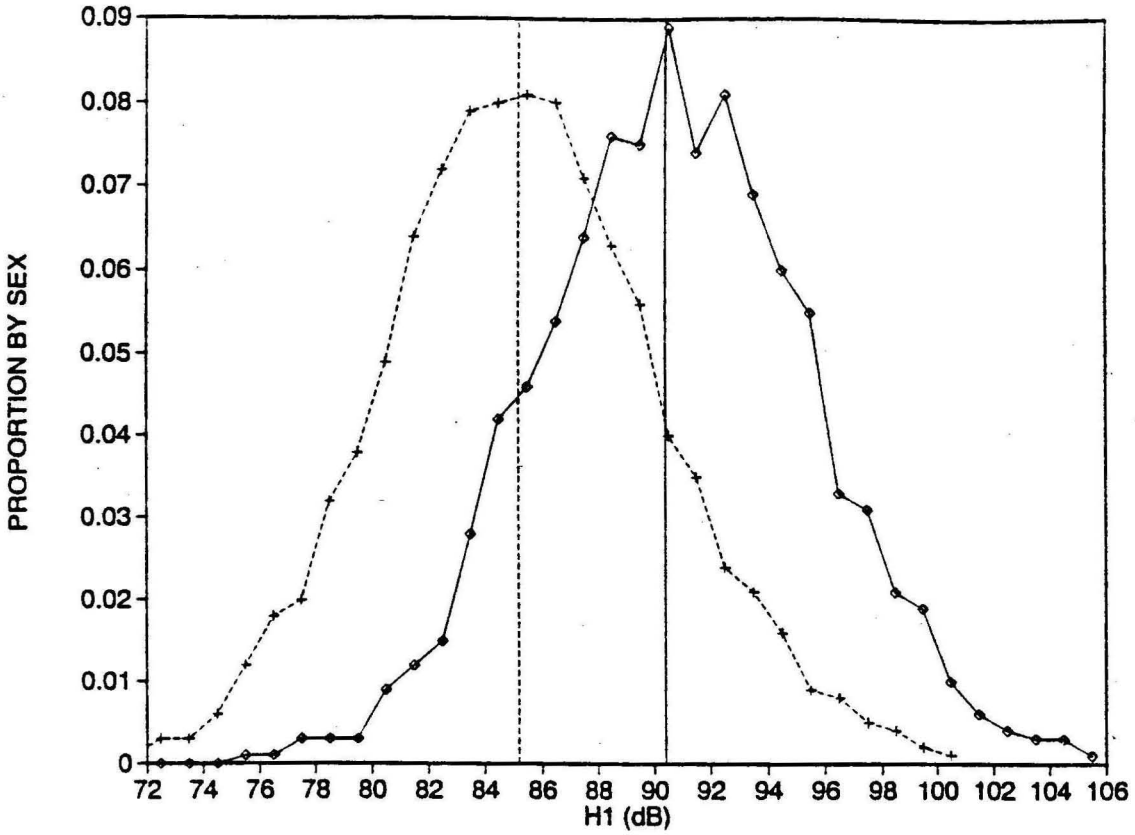


Figure 4.26: Distribution of H_1 SLICE MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of SLICE MEANS in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean H_1. Note: of the slices outside the range of this histogram, 15 female and 37 male slices had a mean H_1 less than 72dB, and 1 female and 2 male slices were greater than 106dB. The overall mean H_1 scores for each sex are shown as single, vertical lines.

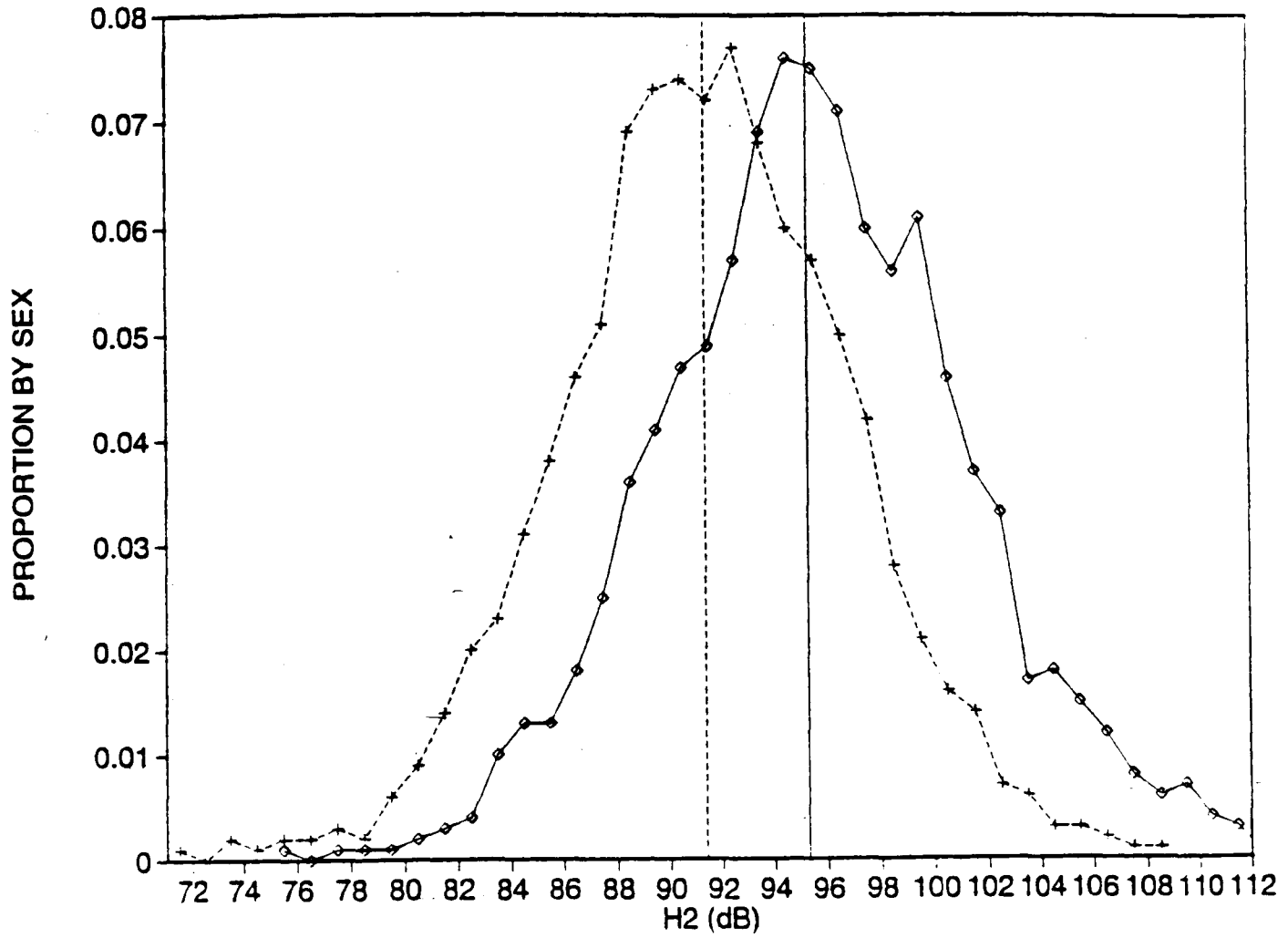


Figure 4.27: Distribution of H_2 SLICE MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of SLICE MEANS in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean H_2. Note: of the slices outside the range of this histogram, 1 female and 9 male slices had a mean H_2 less than 71dB, and 7 female and 2 male slices were greater than 112dB. The overall mean H_2 scores for each sex are shown as single, vertical lines.

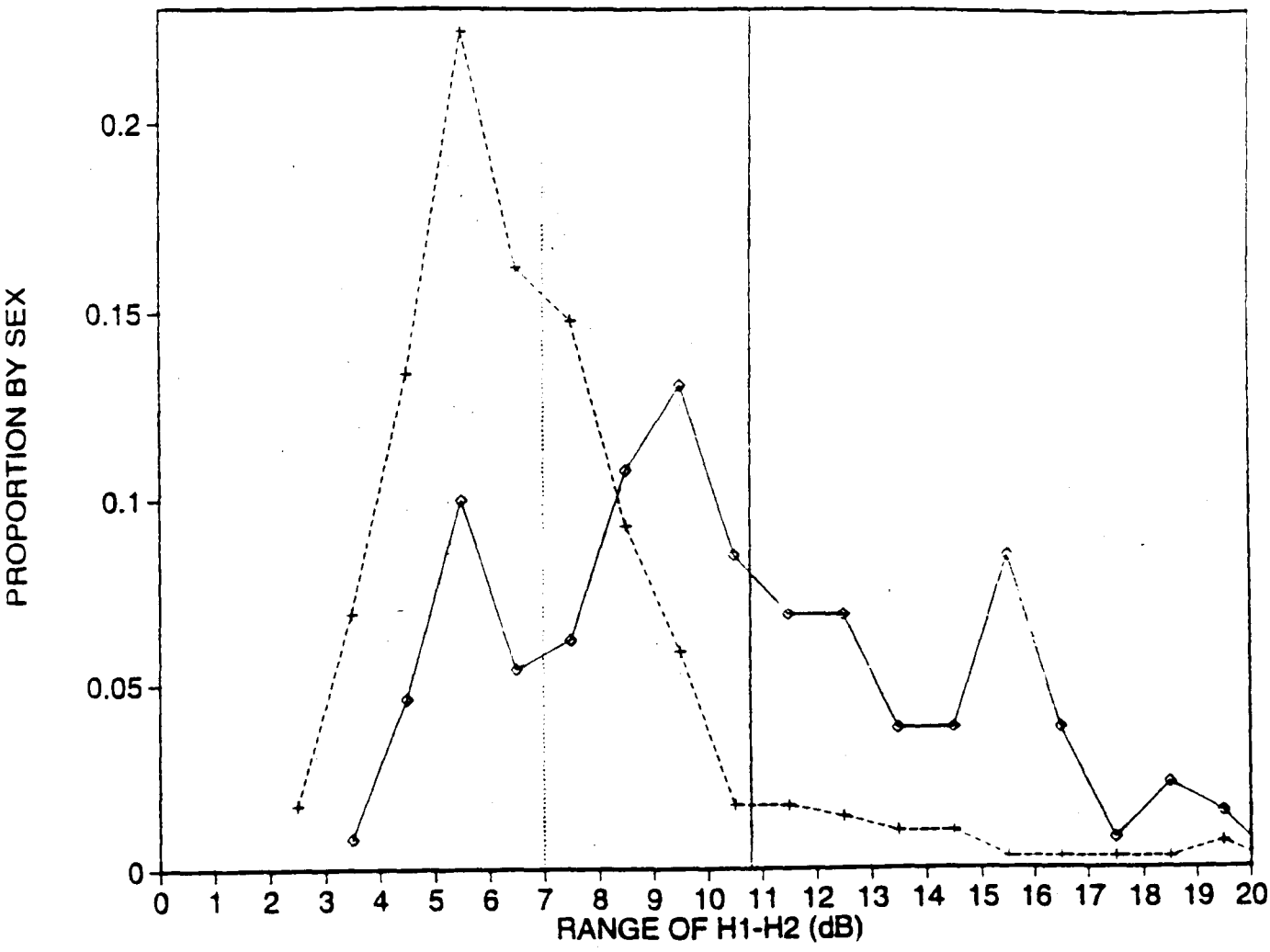


Figure 4.28: Distribution of H_1-H_2 ranges for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of ranges in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular range. Note: Three female and one male speaker, whose ranges were 20dB and over, were left of the histogram. Their ranges were 22.9dB, 26.6dB and 28.1dB, and 24.1dB respectively.

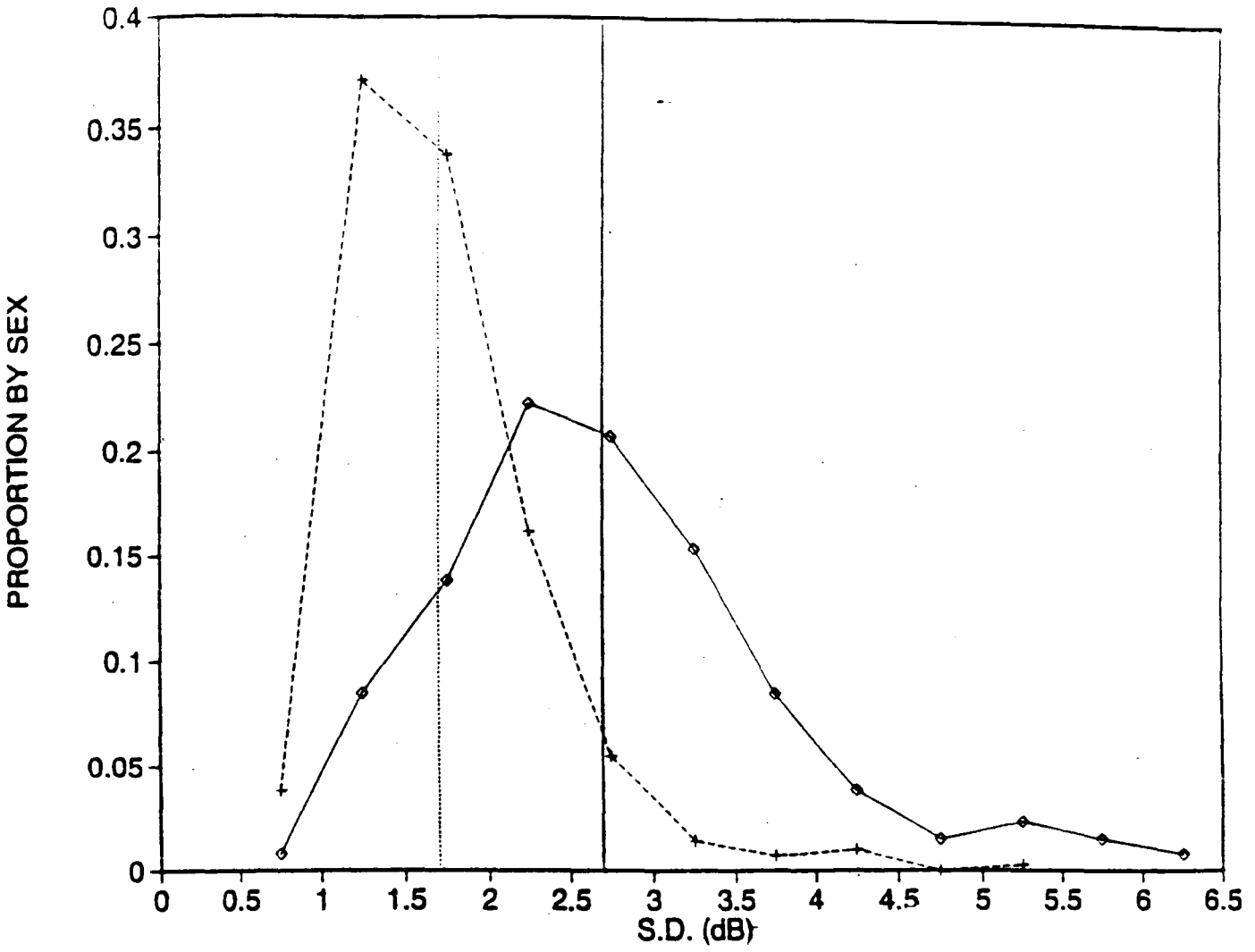


Figure 4.29: Distribution of the s.d.s of the H_1–H_2 SPEAKER MEANS for female (solid line, diamonds) and male (broken line, plusses) speakers. The data points represent the proportion of s.d.s in a 0.5dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular range.

1.7dB (0.6dB) for the males.

Analysis of data by phone

The female speakers exhibited some variation in mean H_1–H_2 between phones. although this was fairly limited (see Table 4.29). The mean and s.d. for /aa/ were exactly the same as those for all three phones, while the biggest contrast was between /ae/ and ao/. This pattern was reflected in the distributions of SPEAKER MEANS for the phones (see the histogram in Figure 4.30), which also shows the similarity between the distributions for /aa/ and /ae/. There was a slight tendency for the female speakers to have a more negative H_1–H_2 for /ao/ than for the other two phones – the proportions of speakers with a mean H_1–H_2 of less than -8dB were 4.6% (6 speakers) for /aa/, 3.9% (5) for /ae/, 20.6% (27) for /ao/, and 14.7% (396) for all phones. Furthermore, /ao/ was the source of the bulk of the extreme negative H_1–H_2 SLICE MEANS – of the 91 slices with a H_1–H_2 less than -11dB, 70.3% (64) came from the /ao/[20]. In contrast, for the male speakers there were no appreciable differences between either the means or the distributions of H_1–H_2 for each phone (see Table 4.30).

[20] As a proportion of the total number of each phone, 1.2% (10) were from /aa/, 1.6% (17) were from /ae/, and 8.3% (64) were from /ao/.

Phone	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
/aa/	130	-4.8	(2.3)	-9.6	-6.6	-5.3	-3.7	3.5
/ae/	130	-4.2	(2.5)	-10.0	-6.1	-4.3	-2.9	3.2
/ao/	130	-5.7	(2.8)	-12.6	-7.4	-5.9	-4.2	1.1
all	130	-4.8	(2.3)	-9.1	-6.3	-5.1	-3.7	1.8

Table 4.29: Female mean H_1–H_2 data (to nearest 0.1dB) by vowel phone. Note that all the means are computed from the SPEAKER MEANS.

Phone	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
/aa/	290	-6.1	(2.0)	-11.4	-7.4	-6.3	-5.0	0.8
/ae/	290	-6.0	(2.1)	-10.4	-7.4	-6.2	-4.6	2.2
/ao/	290	-6.4	(2.2)	-13.9	-7.7	-6.6	-5.1	0.8
all	290	-6.1	(2.0)	-11.7	-7.5	-6.4	-5.0	0.8

Table 4.30: Male mean H_1–H_2 data (to nearest 0.1dB) by vowel phone. Note that all the means are computed from the SPEAKER MEANS.

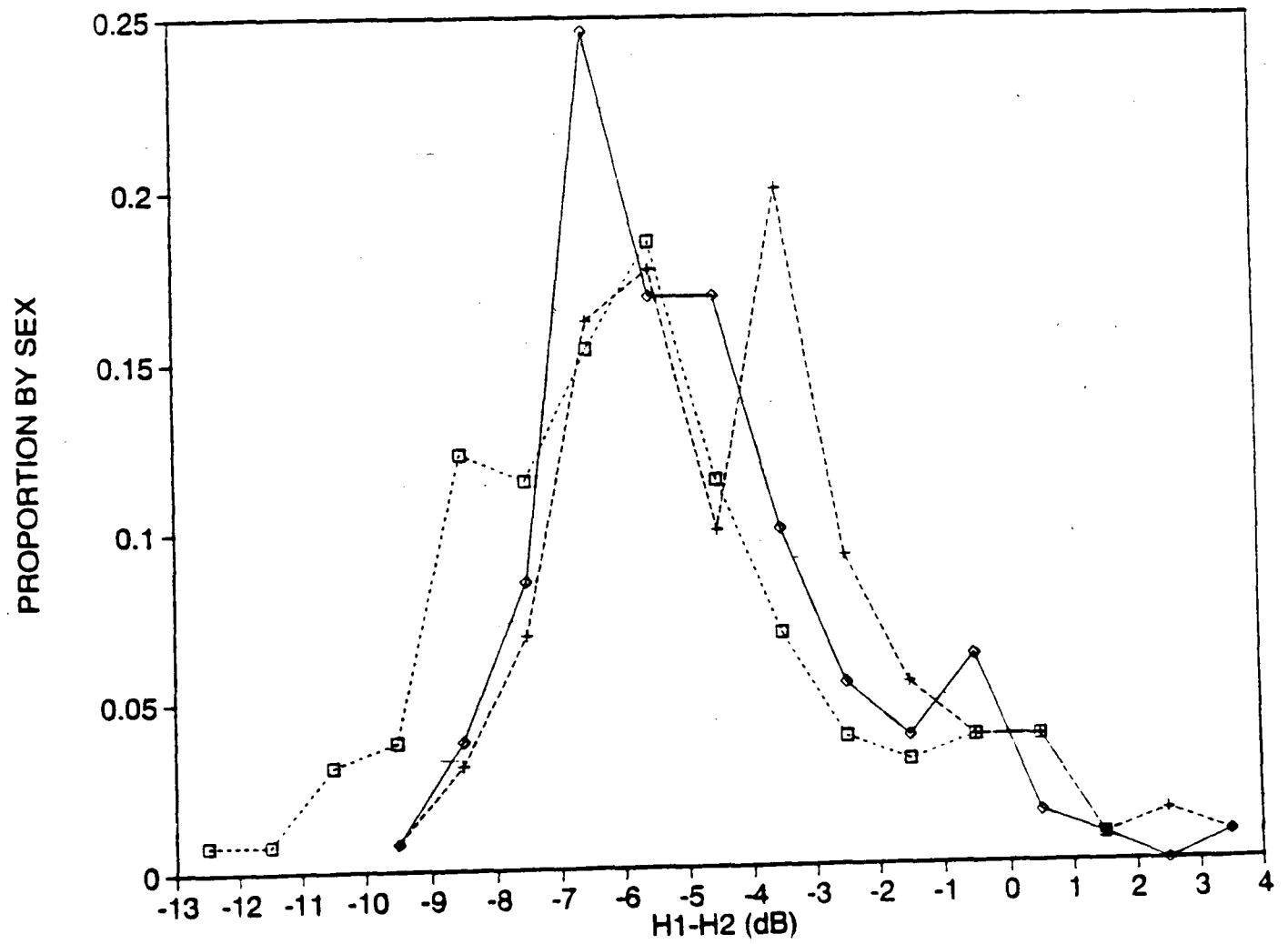


Figure 4.30: Distribution of H_1–H_2 SPEAKER MEANS for /aa/ (solid line, diamonds), /ae/ (broken line, plusses) and /ao/ (broken line, squares) speakers. The data points represent the proportion of SPEAKER MEANS in a 1dB interval, and are plotted at the midpoint of the interval. The y-axis represents the proportion of speakers of one sex having a particular mean H_1–H_2. The overall mean H_1–H_2 scores for each sex are shown as single, vertical lines.

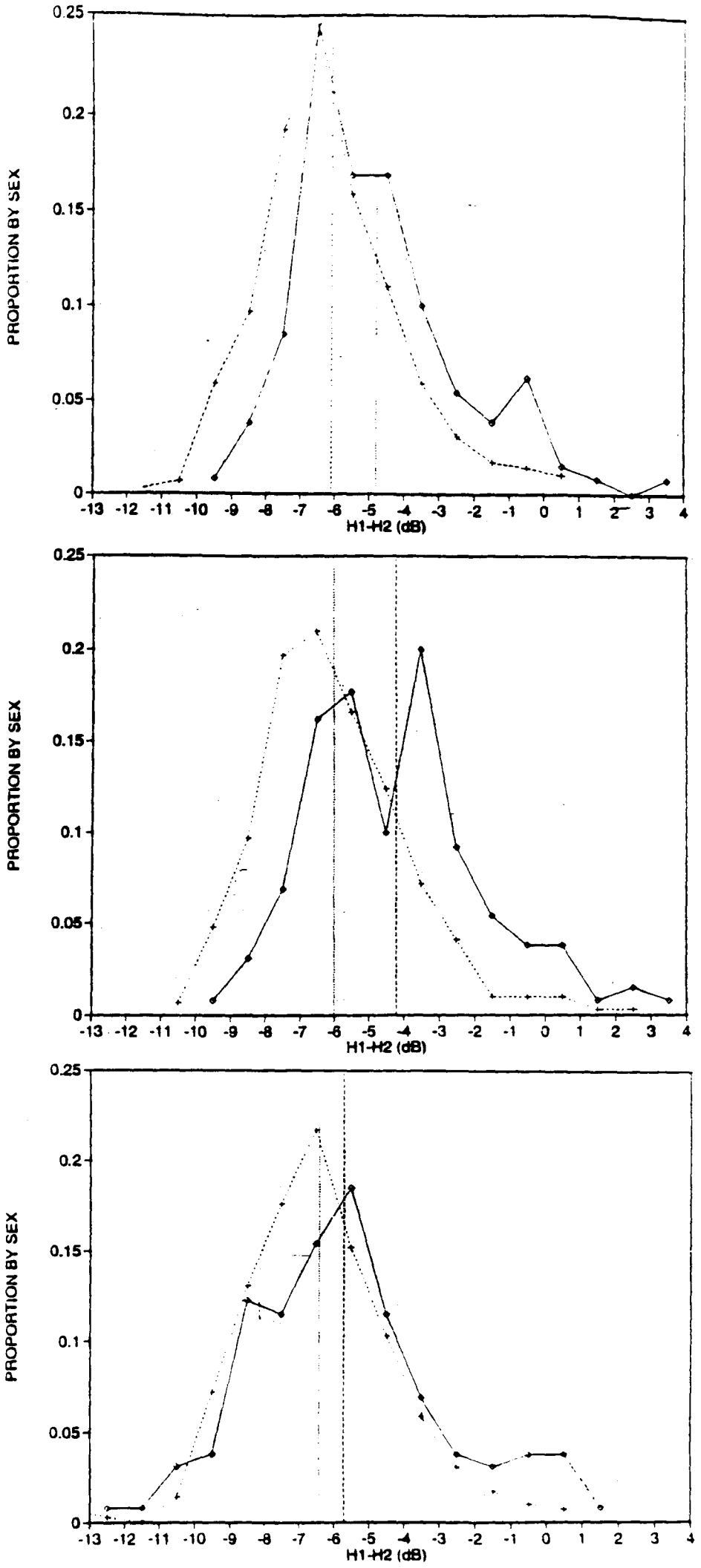


Figure 4.31: Distribution of H_1-H_2 SPEAKER MEANS for the phones /aa/ (top), /ae/ (middle) and /ao/ (bottom) spoken by female (solid line, diamonds) and male (broken line, plusses) speakers. The bars of the histogram represent 1dB. The y-axis represents the proportion of speakers of one sex having a particular mean H_1-H_2. The overall mean H_1-H_2 scores for the female and male productions of the phone are shown as single, vertical lines.

Age	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
20 – 29	82	-4.8	(2.2)	-9.1	-6.1	-5.1	-3.5	0.5
30 – 39	32	-4.8	(2.7)	-8.5	-7.0	-4.8	-3.5	1.8
40 – 49	10	-5.0	(1.8)	-7.3	-7.0	-4.5	-4.1	-1.7
50 – 59	4	-5.7	(0.9)	-6.3	-6.2	-6.1	-5.2	-4.3
≥ 60	2	-5.2	(0.1)	-5.2	–	-5.2	–	-5.1

Table 4.31: Female mean H_1–H_2 data (to nearest 0.1dB) by age.

Age	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
20 – 29	183	-6.3	(1.9)	-10.2	-7.5	-6.4	-5.2	0.8
30 – 39	86	-6.2	(2.1)	-11.7	-7.6	-6.6	-4.9	0.3
40 – 49	13	-5.1	(2.2)	-8.3	-6.7	-5.4	-3.3	-0.9
50 – 59	7	-4.7	(2.0)	-7.0	-6.7	-4.1	-2.9	-2.3
≥ 60	1	-3.4	(–)	–	–	-3.4	–	–

Table 4.32: Male mean H_1–H_2 data (to nearest 0.1dB) by age.

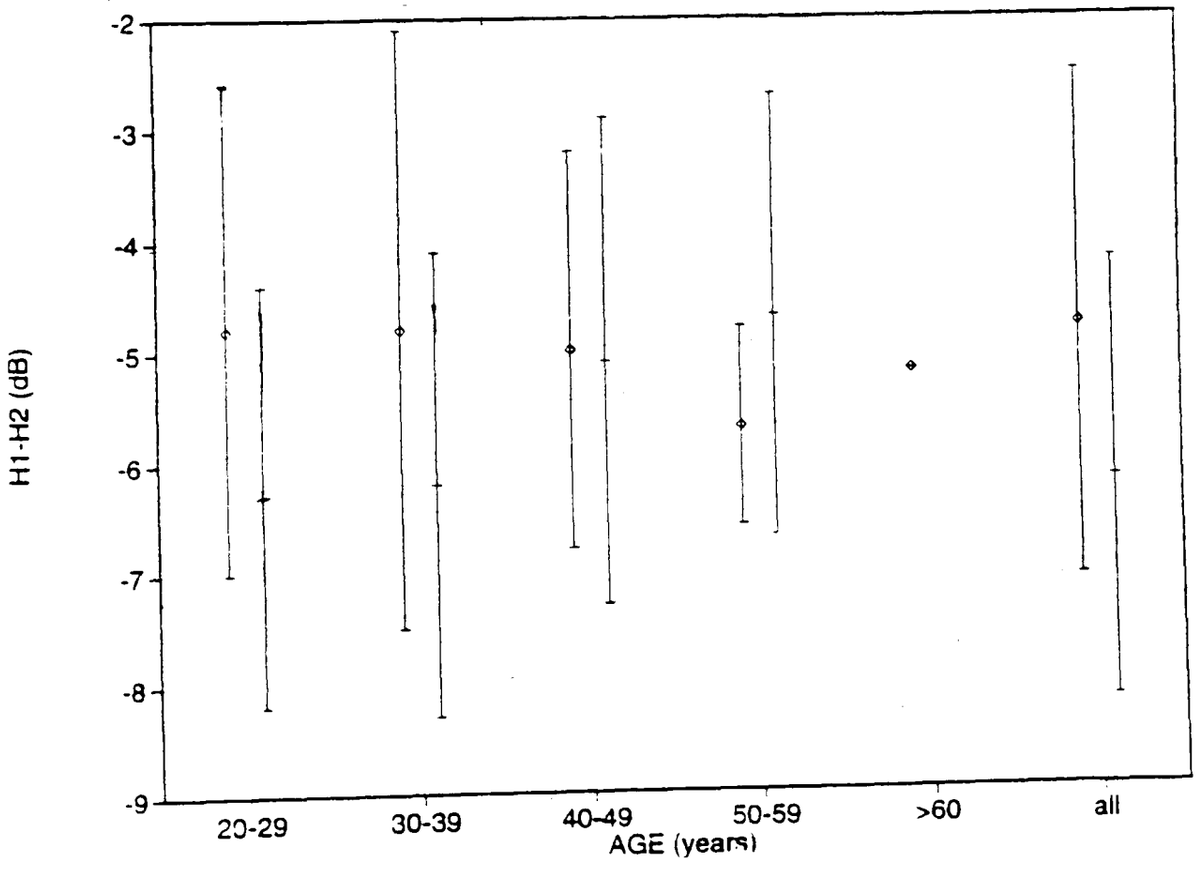


Figure 4.32: Mean H_1–H_2 (dB) by age for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean (note, no s.d. intervals are given for the age group '60 and over' as there were so few speakers). The mean for all speakers is on the right.

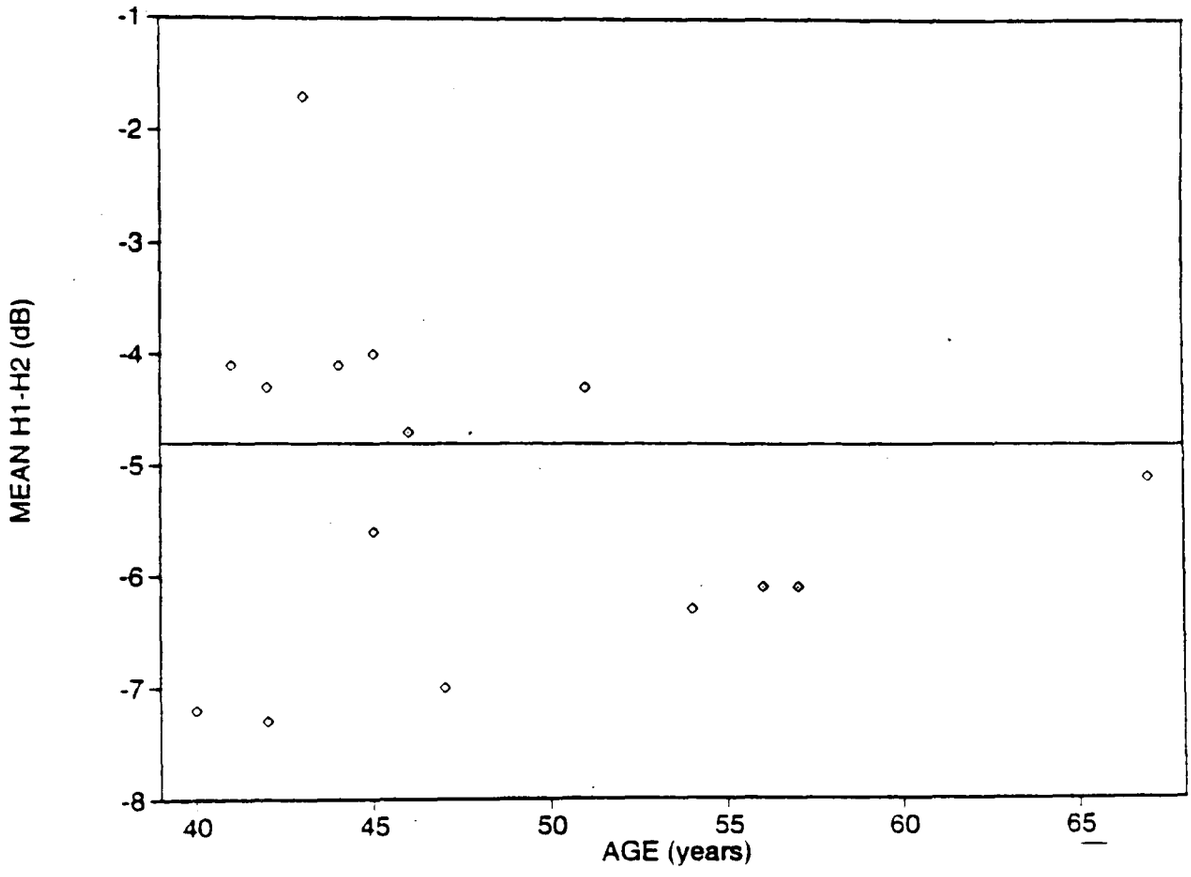


Figure 4.33: Mean H_1–H_2s (dB) of the female speakers over 40 years old. The solid line at -4.8dB represents the mean H_1–H_2 of all the female speakers. Note: The 85 year-old speaker fkfb0 has been left off the graph – her mean H_1–H_2 was -5.2dB.

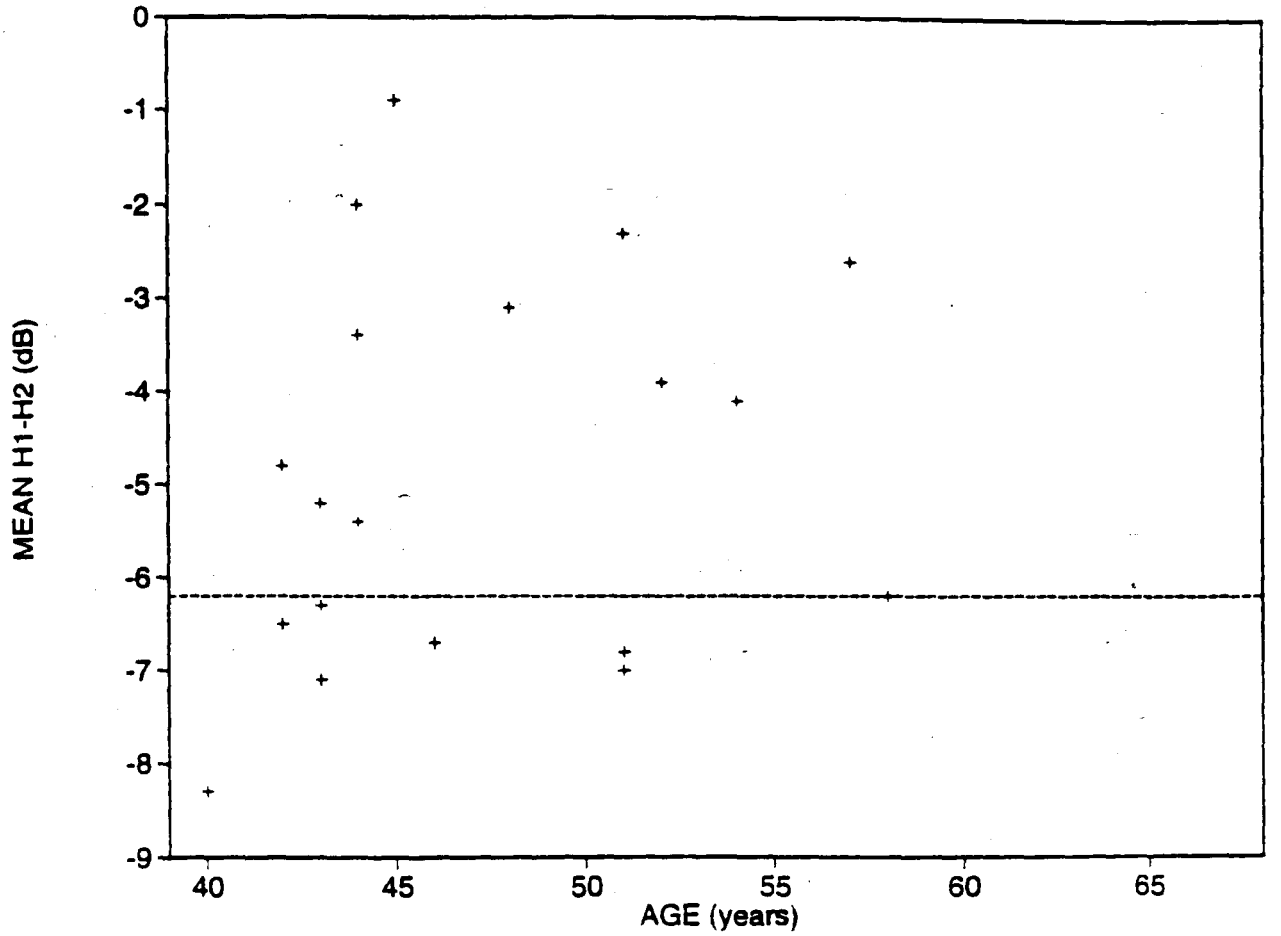


Figure 4.34: Mean H_1–H_2s (dB) of the male speakers over 40 years old. The solid line at -6.2dB represents the mean H_1–H_2 of all the female speakers. Note: The 85 year-old speaker mrjm1 has been left off the graph – his mean H_1–H_2 was -3.4dB.

Height	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
$\leq$5'1"	7	-5.1	(3.0)	-8.6	-7.8	-5.1	-2.8	-0.3
5'2" – 5'3"	25	-4.5	(2.3)	-7.7	-6.1	-4.8	-3.4	0.1
5'4" – 5'5"	37	-4.8	(2.4)	-9.1	-6.4	-4.7	-3.9	1.8
5'6" – 5'7"	32	-4.9	(2.3)	-8.9	-6.1	-5.3	-3.9	0.5
5'8" – 5'9"	24	-5.1	(1.9)	-8.4	-6.9	-5.1	-3.9	-1.1
5'10" – 5'11"	4	-5.2	(1.9)	-7.5	-6.7	-4.8	-3.6	-3.4
6'0" – 6'1"	1	-5.7	–	–	–	-5.7	–	–

Table 4.33: Female mean H_1–H_2 data (to nearest 0.1dB) by height.

Height	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
5'2" – 5'3"	1	-9.3	–	–	–	-9.3	–	–
5'4" – 5'5"	3	-6.5	(1.7)	-8.1	–	-6.7	–	-4.7
5'6" – 5'7"	26	-5.8	(1.9)	-9.7	-7.4	-5.7	-4.2	-1.9
5'8" – 5'9"	63	-6.2	(2.4)	-11.7	-7.8	-6.4	-4.7	0.2
5'10" – 5'11"	77	-6.3	(1.8)	-10.2	-7.5	-6.5	-5.2	-2.3
6'0" – 6'1"	67	-6.2	(2.0)	-9.1	-7.5	-6.6	-5.1	0.8
6'2" – 6'3"	42	-6.0	(1.9)	-9.2	-7.2	-6.0	-5.2	0.3
6'4" – 6'5"	7	-5.8	(1.3)	-7.6	-6.8	-5.8	-5.1	-3.8
$\geq$6'6"	4	-5.6	(1.4)	-7.7	-6.5	-5.0	-4.8	-4.6

Table 4.34: Male mean H_1–H_2 data (to nearest 0.1dB) by height.

of speakers measured.

B. Analysis by height :

While there is a hint of a trend of falling H_1–H_2 with increasing height for the female speakers, and of rising H_1–H_2 for the male speakers (see Tables 4.33 and 4.34, and Figure 4.35), the small numbers of speakers representing some of the height groups renders trends based on these groups somewhat unreliable. If we consider only those height groups composed of more than ten speakers – leaving us with the four female groups in the range 5'2"-5'9" and the five male groups in the range 5'6"–6'3" – we find that for both women and men the difference in mean H_1–H_2 between the height groups is very small. Furthermore, there were no reliable trends between the distributions of H_1–H_2 SPEAKER MEANS in these height groups.

C. Analysis by ethnic group :

The black female and male speakers and the white male speakers all have very similar means (and s.d.s) for H_1–H_2 (see Tables 4.35 and 4.36, and Figure 4.36. However, the white females' mean H_1–H_2 is over 1dB higher. Put another way, the black speakers of either sex had similar mean H_1–H_2 values, while the means for the female and male white speakers mirrored the overall female and male group means (which was to be expected, as the vast majority of the speakers on the TIMIT database were white). The distribution of H_1–H_2 means for the black speakers is given in Table 4.37.

D. Analysis by dialect :

For the female speakers, the group means for the eight dialect regions ranged from a high of -4.0dB for dr4 to a low of -5.4dB for dr5 and dr8, and appeared to fall into two groups.

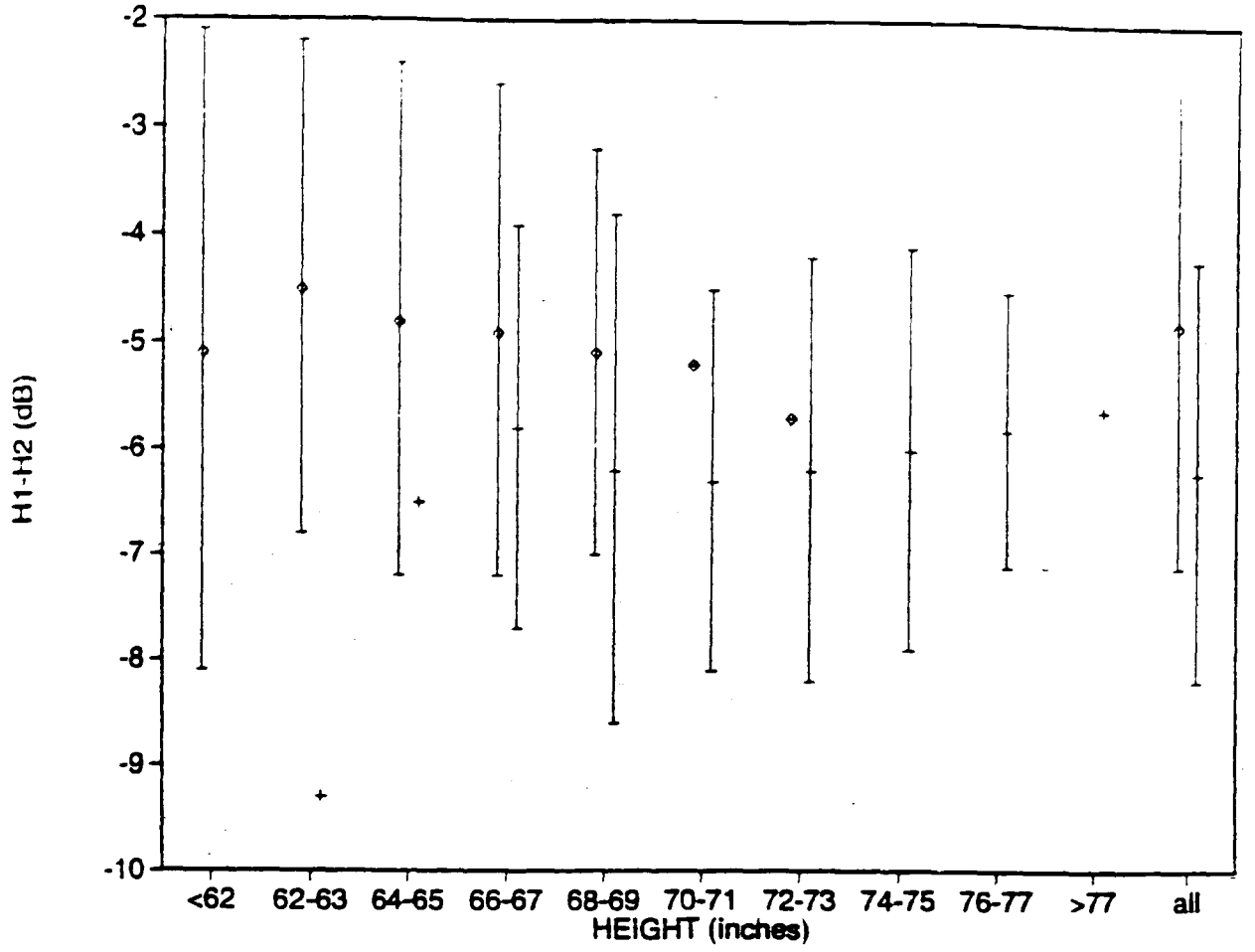


Figure 4.35: Mean H_1–H_2 (dB) by height for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean (note no s.d. intervals are given for some of the height groups as there were so few speakers). The mean for all speakers is on the right.

Ethnic group	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
black	10	-6.0	(1.8)	-8.2	-7.0	-6.4	-5.9	-1.7
white	117	-4.8	(2.2)	-9.1	-6.1	-5.1	-3.6	1.8

Table 4.35: Female mean H_1–H_2 data (to nearest 0.1dB) by ethnic group.

Ethnic group	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
black	12	-6.3	(2.4)	-9.3	-7.5	-6.6	-5.5	-0.4
white	259	-6.2	(2.0)	-11.7	-7.5	-6.4	-4.9	0.8

Table 4.36: Male mean H_1–H_2 data (to nearest 0.1dB) by ethnic group.

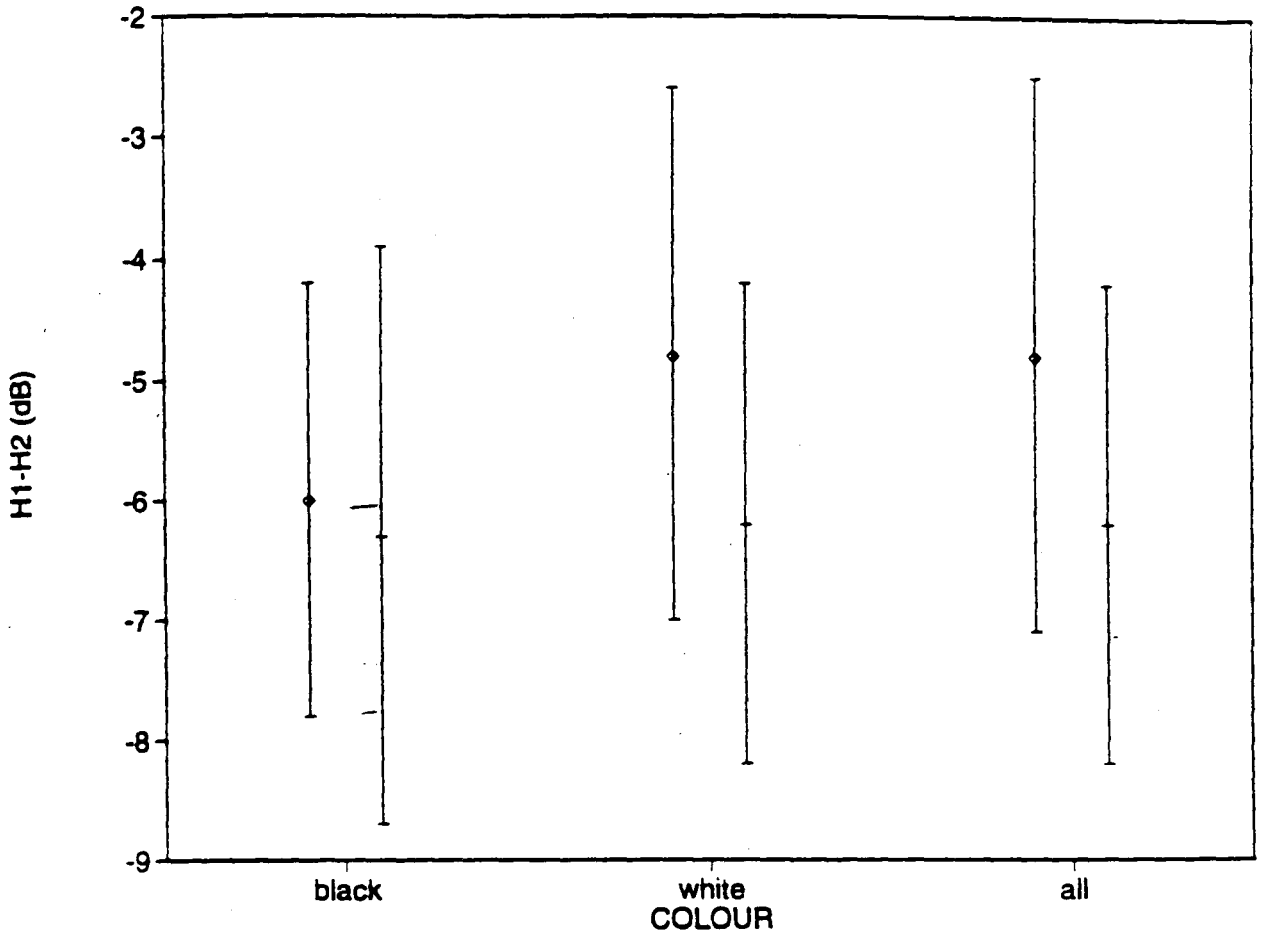


Figure 4.36: Mean H_1–H_2 (dB) by ethnic group for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean. The mean for all speakers is on the right.

H_1–H_2	Female	Male
-10 – -9	–	2
-9 – -8	1	–
-8 – -7	2	3
-7 – -6	4	3
-6 – -5	1	2
-5 – -4	1	1
-4 – -3	–	–
-3 – -2	–	–
-2 – -1	1	–
-1 – 0	–	1

Table 4.37: Number of black female and male H_1–H_2 SPEAKER MEANS at 1dB intervals. A dash (–) indicates there were no speakers with a mean H_1–H_2 in that interval.

Dialect	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
dr1	15	-4.5	(2.3)	-9.1	-5.2	-5.1	-4.0	-0.1
dr2	18	-4.1	(2.4)	-8.9	-6.1	-3.6	-3.0	0.0
dr3	15	-5.2	(2.3)	-8.5	-6.9	-5.4	-4.2	-0.3
dr4	19	-4.0	(2.9)	-8.6	-6.1	-4.4	-2.5	1.8
dr5	26	-5.4	(1.8)	-8.2	-6.8	-6.0	-4.3	-1.1
dr6	11	-5.2	(2.1)	-7.9	-6.7	-5.9	-4.2	-1.5
dr7	18	-5.2	(1.3)	-7.5	-6.0	-5.0	-4.1	-3.1
dr8	8	-5.4	(2.7)	-8.2	-7.7	-5.5	-4.5	0.3

Table 4.38: Female mean H_1–H_2 data (to nearest 0.1dB) by dialect.

Dialect	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
dr1	22	-6.7	(1.9)	-11.7	-7.7	-7.1	-5.4	-3.7
dr2	47	-5.9	(2.0)	-9.2	-7.3	-6.3	-4.5	-0.4
dr3	51	-6.0	(2.3)	-9.7	-7.8	-6.4	-4.3	0.8
dr4	46	-6.1	(1.7)	-9.8	-7.3	-6.3	-5.2	-2.6
dr5	39	-6.3	(1.9)	-9.3	-7.4	-6.6	-5.4	0.3
dr6	21	-5.9	(1.9)	-9.2	-6.8	-6.4	-5.0	-0.9
dr7	48	-6.5	(1.9)	-10.2	-7.7	-6.6	-5.7	0.2
dr8	16	-5.6	(2.1)	-9.1	-7.0	-5.5	-3.8	-1.9

Table 4.39: Male mean H_1–H_2 data (to nearest 0.1dB) by dialect.

The regions dr1, dr2 and dr4 had mean H_1–H_2s of between -4.0dB and -4.5dB: the regions dr3, dr5, dr6, dr7 and dr8 had means between -5.2 and -5.4dB (see Table 4.38). For the male speakers, the group means were spread fairly evenly between the -5.6dB measured for dr8 and the -6.7dB for dr1 (see Table 4.39). The means for the dialect regions, together with intervals representing their s.d.s, are illustrated in Figure 4.37.

While no particular trends were observed between the sexes, the means for dr8 (-5.4dB and -5.6dB for women and men respectively) were almost the same.

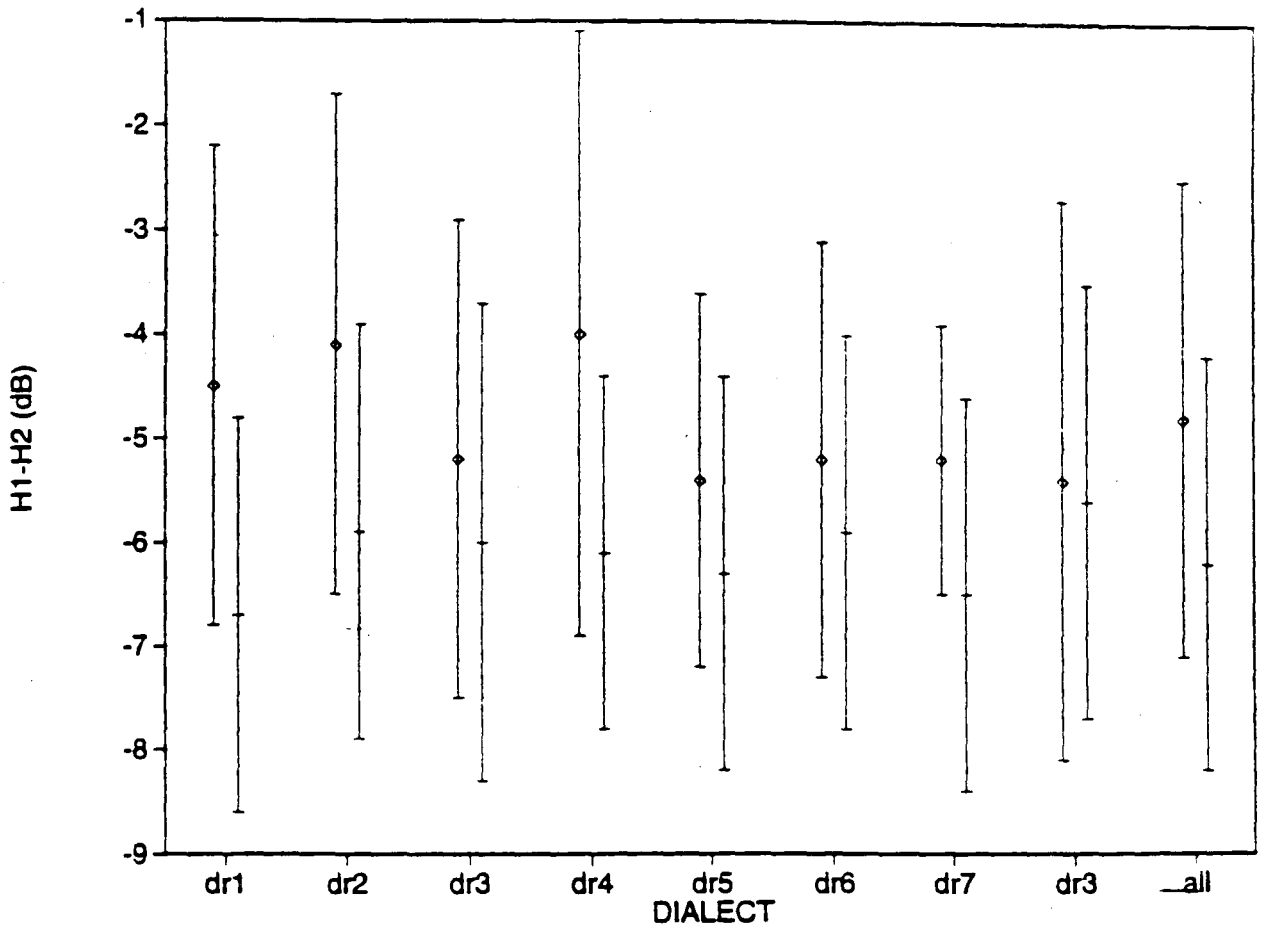


Figure 4.37: Mean H_1–H_2 (dB) by dialect for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean. The mean for all speakers is on the right.

4.2.3 Formant frequency

Reported below are the overall group means for F_1, F_2 and F_3 for both sexes, and an analysis of the formants by phone. These results are compared with the formant frequency data reported in the literature.

The analysis of the formant frequency data is limited in scope due to the lack of confidence in the output of the formant frequency estimator, and due to insufficient time to carry out a complete investigation. The reasons for this are discussed.

Analysis of overall data

The overall group means for first, second and third formant frequencies are given in Table 4.40, and are illustrated in Figure 4.38. The scaling factors ranged from 9-20% for F_1, 12-20% for F_2 and 5-19% for F_3, with overall factors for the formants of 16%, 18% and 11% respectively. This produces a general scaling factor of approximately 15%, which means that on average the female speakers' formant frequencies were 15% greater than the male speakers'. This figure is comparable with the female-male differences reported in the literature. Now due to the different types of phones measured, the overall formant scaling factors given in this study are to some extent incompatible with those given in Tables 3.10 and 3.11 for the Peterson & Barney (1952) and Fant (1959) studies. However, we can directly compare the phones common to each study, namely /aa/, /ae/, /iy/ and /uw/ (and /ux/, an allophone of /uw/, from the TIMIT data), and are given below. By recalculating overall formant frequency means using only the figures for these four phones[21], we can also compare the overall formant scaling factors for the three studies[22]:

all		PB	F	T
F_1	f	600	570	620
	m	490	470	540
% diff.		22	21	15
F_2	f	1750	1560	1820
	m	1490	1350	1570
% diff.		17	16	16
F_3	f	2910	3050	2730
	m	2530	2570	2420
% diff.		15	19	13

The results from the TIMIT study are most similar to Peterson & Barneys' results, especially for F_1 and F_2, although the female-male difference is somewhat less for F_1. Fant's results are noticeably different, especially for F_2 and F_3. However, Fant's results are also rather different to Peterson & Barneys' results, in particular for F_2 and the female F_3. The main reason for this is probably that Fant's subjects were Swedish, uttering Swedish phonemes (see also Section 3.1.3). He derived his vowel phoneme names by finding the nearest equivalent U.S. English phoneme through F-pattern matching. This indicates that these Swedish phonemes are not always a good match for the U.S. phonemes. The implication is that in the direct vowel comparisons, more weight should be attached to the comparison with the Peterson & Barney data. There is however further cause for caution. Peterson & Barney measured their formant frequencies from vowels spoken in a simple

[21] The Peterson & Barney and Fant means were computed by averaging the means for the four phones under consideration. The TIMIT overall means also include the figures for /ux/.

[22] In the following tables, unless stated otherwise the column headed 'PB' refers to Peterson & Barneys' (1952) study, 'F' refers to Fant's (1959) study, and 'T' refers to this study of the TIMIT data. The formant frequencies are rounded to the nearest 10Hz.

		/aa/	/ae/	/ao/	/iy/	/uw/	/ux/	Mean
F_1	f	810	800	730	500	500	480	660
	m	690	670	610	460	440	420	570
% diff.		17	19	20	9	14	14	16
F_2	f	1320	1980	1120	2440	1390	1980	1870
	m	1180	1650	960	2080	1240	1680	1580
% diff.		12	20	17	17	12	18	18
F_3	f	2430	2860	2570	2990	2660	2700	2770
	m	2310	2410	2300	2790	2290	2320	2490
% diff.		5	19	12	7	16	16	11

Table 4.40: Female and male formant frequency data (to nearest 10Hz) for six phones from the TIMIT database. The figures quoted as percentage differences were computed according to Fant's female-male formant scaling factor equation – for example, the Table shows that the second formant of /iy/, the mean for the female speakers was 17% greater than the males.

/hVd/ context, in contrast to the vowels from the TIMIT database, which came from read sentences. Speakers are likely to articulate their vowels more precisely when asked to produce them one after another in isolated stimuli, and are therefore more likely to achieve the vowel formant frequency targets. In connected speech, vowels are prone to reduction, as well as the coarticulatory effects of different phonetic contexts.

Analysis of the data by phone.

Below, in part A, we will compare the formant frequency data for /aa/, /ae/, /iy/, /uw/ and /ux/ from the present study with the studies of Peterson & Barney (1952) and Fant (1959), looking at each of the phones individually. In part B, the distribution of formant frequencies is briefly discussed.

However, we will consider first the variation between the scaling factors for the TIMIT data, with reference to Table 4.40. The first thing to note is the lack of any clear trends: the scaling factors typically differ between formants and between phones. However, comparing with the previous studies, the values of the scaling factors are more consistent here: there were no values over 20%, and none under 5%. This suggests the female-male differences in formant frequency are somewhat less varied than has been found in previous research. For example, the scaling factors in Peterson & Barneys' study ranged from 2% to 30%, and from -3% to 30% in Fant's. Furthermore, both studies reported a number of factors over 20%, while the results from the TIMIT data suggest the female-male difference does not reach such an extreme.

Comparing the scaling factors across the formants, those of F_2 are the most consistent, while those of F_3 are the least. This is at odds with Fant's (1960) suggestion that the female-male difference was more consistent and more pronounced for F_3. The TIMIT results suggest there is more likely to be a difference between the sexes for F_1 and F_2 than for F_3. Finally, comparing the scaling factors across the phones, we find that for some phones the scaling factors show marked consistency for the three formants, e.g. those for /ae/ are all approximately 20%, and the factors for /uw/ and /ux/ are all approximately 14% and 16% respectively (although they are less consistent in value). For other phones there is less consistency, e.g. for /aa/ the factors range from 5% for F_3 to 17% for F_1.

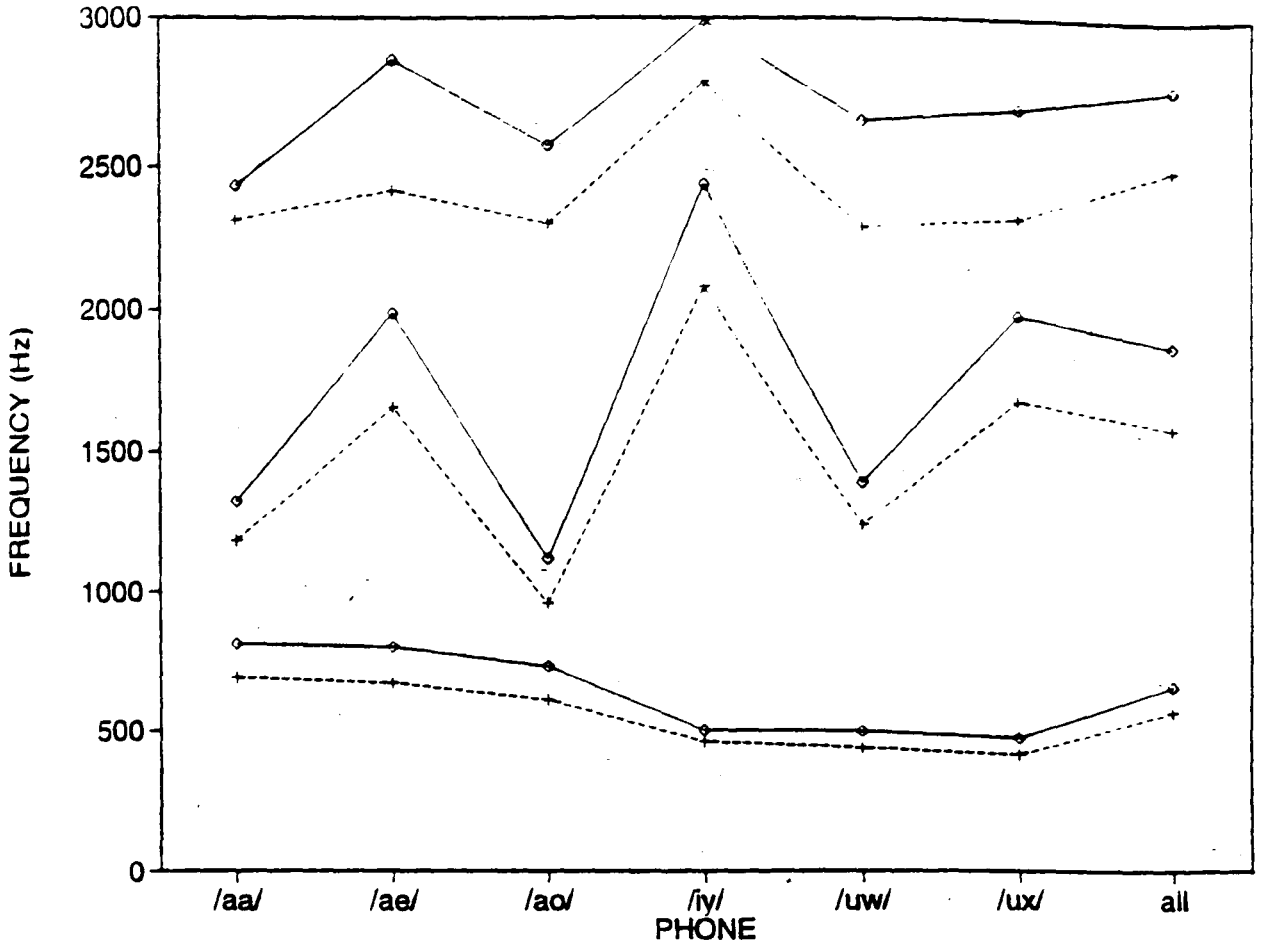


Figure 4.38: Mean F_1, F_2 and F_3 (Hz) by phone for female (diamond) and male (cross) speakers. The means for all the phones are on the right.

As well as showing the most consistent female-male difference, /ae/ also has the largest overall difference. The average scaling factors for each phone place them in the following order of increasing female-male differentiation: /aa/ and /iy/ (11%), /uw/ (14%), /ao/ and /ux/ (16%), and /ae/ (19%).

A. Comparison of the data with other studies

A comparison of the results for /aa/ is given below:

/aa/		PB	F	T
F_1	f	850	860	810
	m	730	680	690
% diff.		16	27	17
F_2	f	1220	1200	1320
	m	1090	1070	1180
% diff.		12	12	12
F_3	f	2810	2920	2430
	m	2440	2520	2310
% diff.		15	16	5

While the TIMIT results for F_1 are similar to both other studies, and are about 100Hz higher for F_2, there is a major discrepancy in the figures for F_3, particularly for the female speakers. The female F_3 mean is nearly 400Hz less than the Peterson & Barney result, and 500Hz less than the Fant. Similarly, the TIMIT F_1 and F_2 scaling factors are much the same as those from the other studies (although Fant's F_1 factor is greater), but the F_3 factor is a third of the others.

In contrast. the results from all three studies for /ae/ are very similar:

/ae/		PB	F	T
F_1	f	860	790	800
	m	660	610	670
% diff.		30	30	19
F_2	f	2050	1830	1980
	m	1720	1550	1650
% diff.		19	18	20
F_3	f	2850	2920	2860
	m	2410	2450	2410
% diff.		18	19	19

While the TIMIT F_2 means are 100-150Hz higher than the Fant means, the only discrepancy in the scaling factors is for F_1.

The TIMIT results for /iy/ are, however, very different, with F_1 much higher and F_3 much lower than the other results:

/iy/		PB	F	T
F_1	f	310	280	500
	m	270	260	460
% diff.		15	9	9
F_2	f	2790	2520	2440
	m	2290	2070	2080
% diff.		22	22	17
F_3	f	3310	3460	2990
	m	3010	2960	2790
% diff.		10	17	7

The only real similarities the other studies show to the TIMIT data are with Fant's F_2 data, which are themselves over 200Hz less than the Peterson & Barney data for both female and male speakers. However, the scaling factors are all of a similar order, showing the difference between the female and male F_2 means to be greater than for the other formants.

The comparison for the phone /uw/ will be presented in a slightly different way due to the inclusion in the TIMIT study of its allophone, /ux/. In the following table, the columns headed 'PB' and 'F' contain the results for /uw/ from the Peterson & Barney and Fant studies, 'T1' and 'T2' refer to the results for /uw/ and /ux/ from the TIMIT data, and 'T3' refers to the combined means for /uw/ and /ux/:

/uw/		PB	F	T1	T2	T3
F_1	f	370	340	500	480	490
	m	300	310	440	420	430
% diff.		23	11	14	14	14
F_2	f	950	690	1390	1980	1760
	m	870	710	1240	1680	1500
% diff.		9	-3	12	18	17
F_3	f	2670	2900	2660	2700	2690
	m	2240	2230	2290	2320	2310
% diff.		19	30	16	16	16

Because of the acoustic similarities between allophones, one would expect their formant patterns to be very similar. This is indeed the case for the two allophones from the TIMIT

		F_1		F_2		F_3	
/aa/	f	810	(60)	1320	(110)	2430	(190)
	m	690	(40)	1180	(90)	2310	(200)
/ae/	f	800	(60)	1980	(150)	2860	(140)
	m	670	(40)	1650	(100)	2410	(110)
/ao/	f	730	(60)	1120	(100)	2570	(260)
	m	610	(40)	960	(70)	2300	(200)
/iy/	f	500	(40)	2440	(150)	2990	(130)
	m	460	(80)	2080	(120)	2790	(150)
/uw/	f	500	(70)	1390	(290)	2660	(280)
	m	440	(50)	1240	(180)	2290	(170)
/ux/	f	480	(40)	1980	(210)	2700	(140)
	m	420	(30)	1680	(170)	2320	(120)
TOTAL	f	660	(40)	1870	(120)	2770	(120)
	m	570	(40)	1580	(100)	2490	(120)

Table 4.41: Female and male formant frequency data (to nearest 10Hz), with s.d.s in brackets, for six phones from the TIMIT database.

data, which had almost identical group means for F_1 and F_3. The acoustic differences between /uw/ and /ux/ are reflected in very different values for F_2. If we now compare the TIMIT means with Peterson & Barneys' results, we find that only with F_3 is there any agreement in formant values. While the TIMIT means for F_1 are approximately 100Hz higher, the data for F_2 for both /uw/ and /ux/ are very different to the Peterson & Barney F_2 mean. At this point we could assume that Peterson & Barney included any allophones of their target phonemes in their measurements, although it is of course possible they coached their subjects to produce only the relevant phones. If we do this for the TIMIT allophones, however, the combined F_2 mean is still way in excess of both the Peterson & Barney and Fant F_2 means.

B. Distribution of the formant frequencies

We will now consider the distribution of the SPEAKER MEANS for the six phones from the TIMIT data. This will tell us how speakers' mean formant frequencies differed between the sexes. To this end the formant frequency group means for the phones are reproduced in Table 4.41 with their s.d.s, and are illustrated in Figures 4.39, 4.40 and 4.41 for F_1, F_2 and F_3 respectively. What is immediately apparent from the Figures is the number of instances where the s.d. intervals around the female and male means do not cross, indicating that in these cases the female speakers' mean formant frequencies were on average different to the means of the male speakers. However, this was not the case for every formant, and certainly not for every phone. The only phone showing this degree of dissimilarity for all three formants was /ae/. The other instances were the first formant of /aa/ and /ao/, the second formant of /iy/ (and almost /ao/), and the third formant of /ux/ (and almost /uw/).

Closer examination of each speaker's formant frequencies highlighted the problems anticipated by the evaluation of the CSTR estimator. The evaluation, reported in Appendix B.6, found that it was impossible to define a search space for the estimator that could take into account within- and between-speaker deviation from the expected range of formant values. A minor consequence of this is an over- or underestimation of the formant frequencies, especially of F_3. More serious is the tendency of the estimator to track spectral features other than the first three formants, particularly when F_0 or F_4 are included

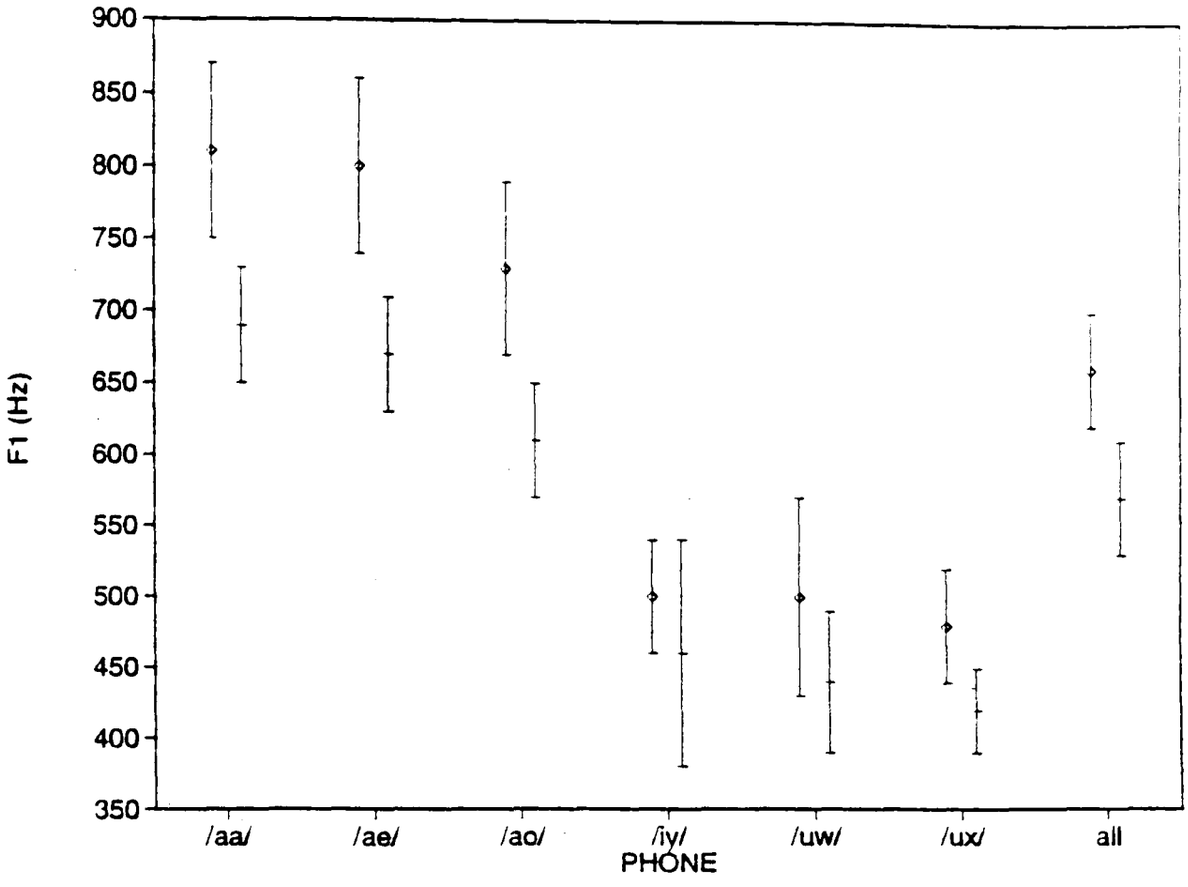


Figure 4.39: Mean and s.d. of F_1 (Hz) by phone for female (diamond) and male (cross) speakers. The means for all the phones are on the right.

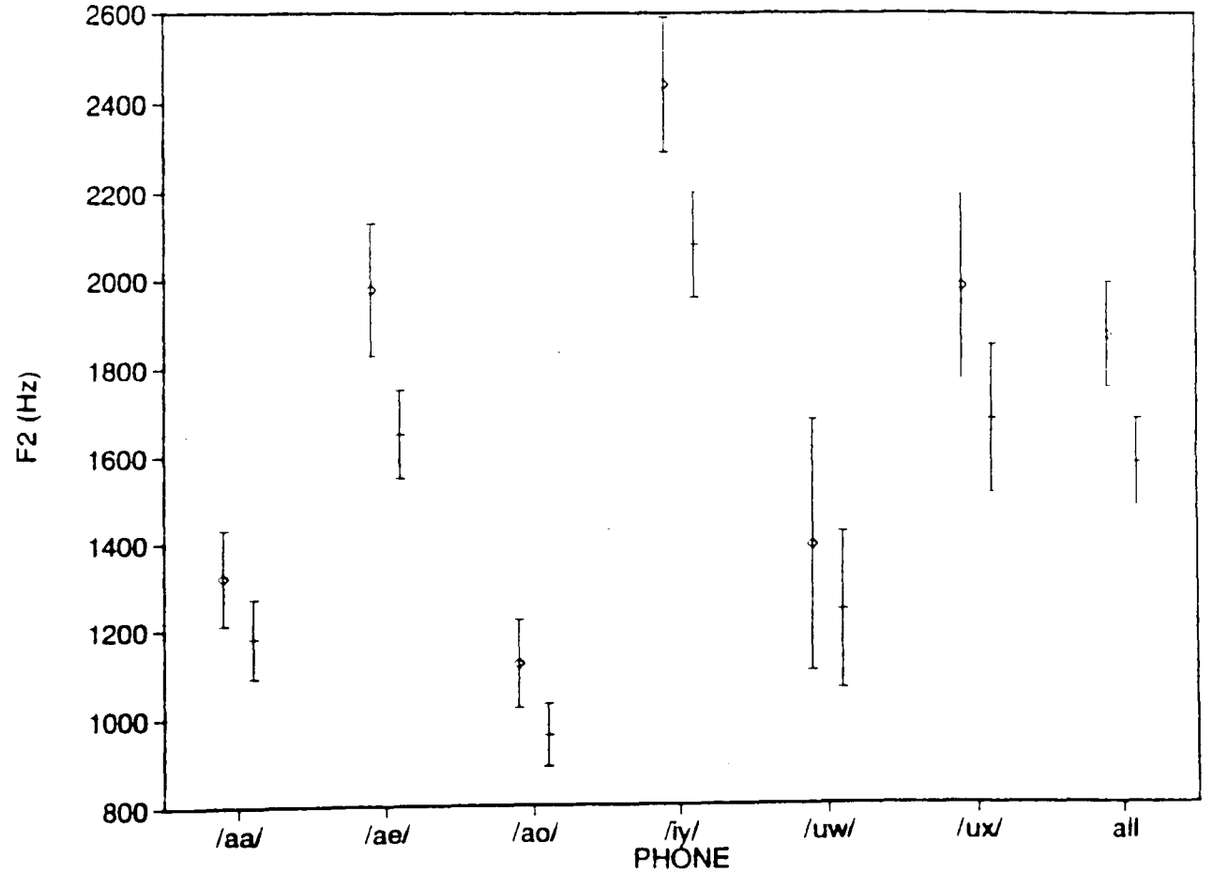


Figure 4.40: Mean and s.d. of F_2 (Hz) by phone for female (diamond) and male (cross) speakers. The means for all the phones are on the right.

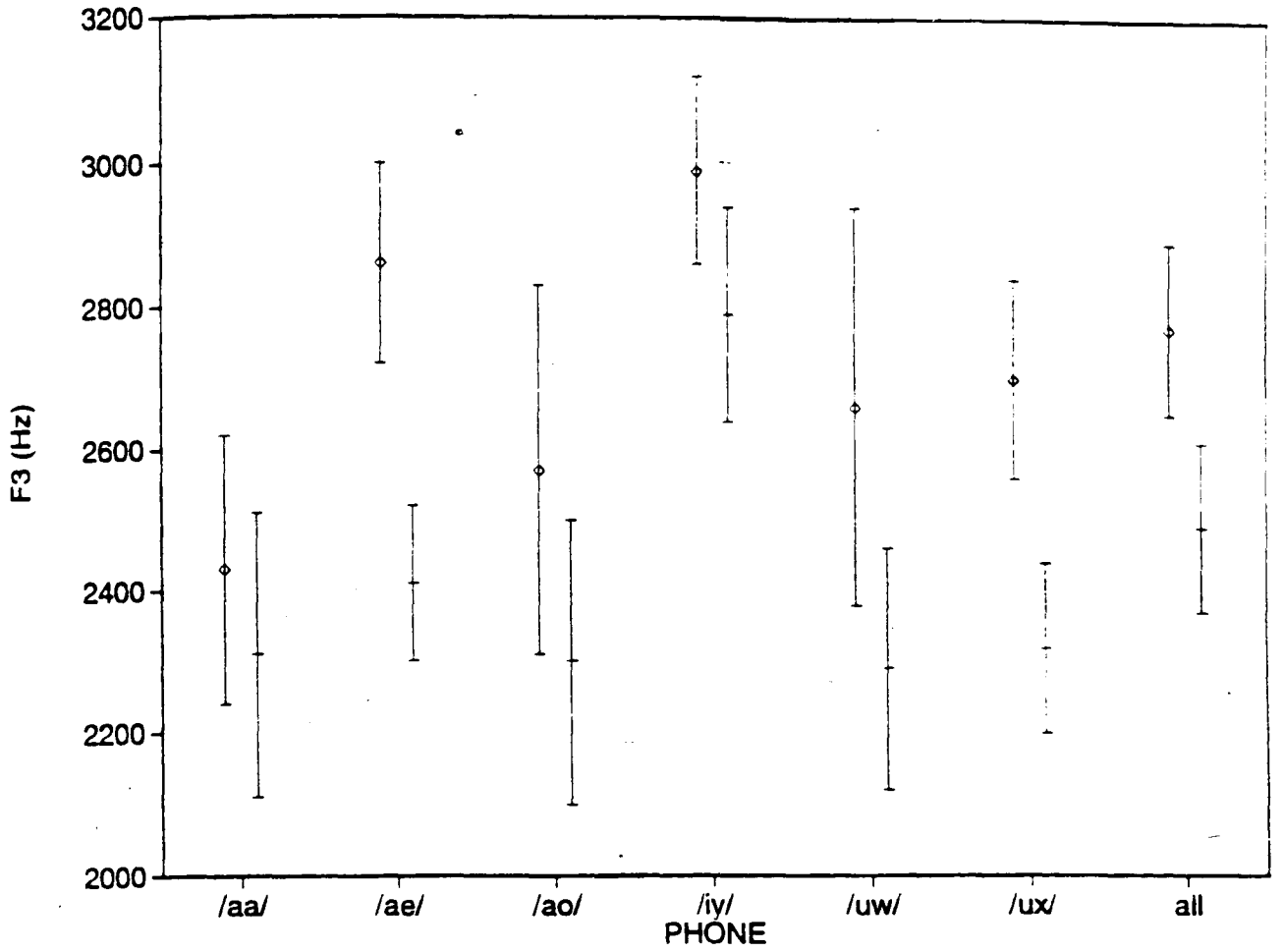


Figure 4.41: Mean and s.d. of F_3 (Hz) by phone for female (diamond) and male (cross) speakers. The means for all the phones are on the right.

in the search space. The evaluation showed that speakers produce a very wide range of formant frequencies (for example, the mean F_3 of /ae/ for speakers **fsem0** and **fcdr1** were hand-measured at 3200Hz and 2300Hz respectively). Without checking at least some of the output for every speaker, it was not possible to be confident about assessing the distribution of the formant frequencies.

4.2.4 A Summary of the results

A. Fundamental frequency

Group means : The mean female and male SFFs were measured at 208Hz and 120Hz respectively. Thus the female group mean was 73% greater than the male.

Distribution of SFF : There was very little overlap between either the mean SFFs of individual women and men, or the fundamental frequencies of individual speech segments. Thus it can be stated with some confidence that in the majority of their speech, the women and men in this study used distinct SFFs.

A small proportion of speakers habitually used a SFF that was atypical for their sex, i.e. women whose mean SFF was relatively low for female speakers, and men whose mean SFF was relatively high for male speakers. However, the SFFs of even these speakers did not approach the group mean of the opposite sex: the lowest mean SFF for a female speaker was measured at 146Hz, compared to the overall male mean of 120Hz; and the highest mean SFF for a male speaker was 183Hz, compared to the overall female mean of 208Hz. In addition, the lowest female mean SFF was still higher than the vast majority of male means (93.1% (270) of the male speakers had a mean SFF below 146Hz), and vice versa for the highest male mean SFF (88.5% (115) of the female speakers had a mean SFF above 183Hz).

The lowest female mean SFF and highest male mean SFF can be used to define a band of frequencies employed by speakers of both sexes during the course of normal speech, i.e. 146-183Hz. Most women (86.9% or 113) uttered at least one vowel phone whose F_0 lay within this region, while rather less men (57.9% or 168) did. We can however define a value which effectively separates female from male SFFs. Thus, only 2.3% of the female speakers had a mean SFF less than 160Hz, while only 3.1% of the male mean SFFs were in excess of this value. Similarly, only 6.5% of the segments of speech uttered by female speakers were under 160Hz, and only 5.6% of the male segments were above it. One word of caution however regarding the extent of the separation of SFF, over half the women and over a third of the men uttered at least one segment whose F_0 lay the other side of this cut-off value, indicating at least some degree of overlap.

Range of SFF : When measured in semitones (st), the mean range of SFF used by women and men were, at 10.6st (122Hz) and 9.7st (66Hz) respectively, very similar. A slight trend was observed for the male speakers to have smaller ranges than the females. The male speakers were twice as likely to have a very restricted range of SFF (less than 5st)[23]. Of the larger SFF ranges (i.e. greater than 12st), the distribution of female and male ranges was very similar, indicating that the female and male speakers were equally likely to have a large range of SFF.

B. Relative amplitude of first harmonic

Group means : The mean female and male H_1–H_2 group means were measured at -4.8dB and -6.2dB respectively, giving a female-male difference of 1.4dB. Thus for the majority of

[23] It is worth noting that the proportions of female and male speakers whose ranges were less than 6st were 8.5% (11) and 10.0% (29) respectively; and the proportions whose ranges were less than 7st were 19.3% (25) and 22.1% (64) respectively. In other words, it was only for speakers with very restricted ranges that differential behaviour between the sexes was found.

speakers, the amplitude of their first harmonic was less than the amplitude of their second.

Distribution of H_1–H_2 : There was a substantial overlap between both the mean H_1–H_2s of individual women and men, and the H_1–H_2 values for individual speech segments. Over three-quarters of all speakers had a mean H_1–H_2 between -8dB and -3dB; moreover, three-fifths of all speech segments had a H_1–H_2 in the same region. It would appear then that the H_1–H_2 measure does not clearly discriminate between the sexes.

That said, there was a noticeable trend for the female speakers to have a more positive H_1–H_2, and they were more varied in their use of H_1–H_2. Examining the first point, and using the midpoint between the female and male group means as a guide, 56.9% (74) of the female speakers had a mean H_1–H_2 greater than -5.5dB, compared to 31.7% (92) of the male speakers. If we now consider the more extreme ends of the distribution of means for individuals (see Figure 4.22 and Table 4.26), the differential behaviour of the sexes becomes more apparent. The proportion of female speakers with a mean H_1–H_2 greater than -2dB (13.1% or 17 speakers) far outnumbered the proportion of males (2.7% or 8 speakers). Similarly, the proportion of male speakers with a mean H_1–H_2 less than -9dB was 5.8% (17), compared to 0.8% (1) of female speakers. In contrast, female use of H_1–H_2 was more varied than male use, in that the women were more likely to produce extreme negative and positive values of H_1–H_2 during the course of their speech.

Distribution of H_1 and H_2 : The female and male group means were 90.4dB and 85.2dB for H_1, and 95.3dB and 91.4dB for H_2. This gives us female-male differences of 5.2dB for H_1, and 3.9dB for H_2. When we consider the mean harmonic amplitudes for individual speakers, the figures suggest a significant trend for women to speak with higher first and second harmonic amplitudes. However, when we take into account the harmonic amplitudes of individual speech segments, the division between female and male speakers is less clear cut, with the vast majority of speakers using a range of harmonic amplitudes that overlap considerably. Overall though, there is still a tendency for the harmonic amplitudes of women to be greater than those of men, and to a greater extent for the first harmonic than the second.

Range of H_1–H_2 : The mean ranges of H_1–H_2 were 10.8dB for women and 7.0dB for men. On the whole, the female and male speakers exhibited differential behaviour in the range of H_1–H_2 they each used, in that the women were far more likely to have a more dynamic range. Overwhelmingly the men had more restricted ranges, with 90% having ranges less than 10dB, and over 20% having ranges less than 6dB (the corresponding figures for the women being 50% and 5%). In contrast, two fifths of the women employed H_1–H_2 ranges of between 10dB and 17dB (compared to less than 8% of the men). This was reflected in the number of women producing more extreme values for H_1–H_2: three times as many female speakers as male produced at least one segment in which the H_1–H_2 was positive (i.e. in which the first harmonic had a higher amplitude than the second); and at the other end of the scale, where one would expect the male speakers to predominate, nearly twice as many female speakers produced at least one segment in which the H_1–H_2 was less than -11dB.

C. Formant frequencies

The purpose in comparing overall female and male formant frequency means, when the means for different phones are so varied, is to gauge the extent to which female speakers

differ on average from male speakers. The means for the different phones are examined below.

Group means : On average, the formant frequencies of the female speakers were approximately 15% greater than the male speakers'. The mean frequencies for the female and male formants from six vowels were measured at 660Hz and 570Hz for F_1, 1870Hz and 1580Hz for F_2, and 2770Hz and 2490Hz for F_3. This gave female-male scaling factors for the three formants of 16%, 18% and 11% respectively. However, the scaling factors differed for each phone, ranging from 9-20% for F_1, 12-20% for F_2 and 5-19% for F_3. The formant with the most consistent, and also largest, scaling factors was F_2, closely followed by F_1.

D. Analysis by phone

There appeared to be some correlation between the phone being spoken and the level of F_0 within that phone. For both sexes, the mean F_0s of the phones placed them in the following order of ascending F_0: /ao/, /aa/ and /ae/, and /iy/, /ux/ and /uw/[24]. The difference between the group mean F_0s of /ao/ and /uw/ was 24Hz for the female and 11Hz for the male speakers. This has implications for the investigation of the SFF of different speaker groups, in that the results could be dependent upon the proportions of the phones being used to measure the SFF. For this study, /uw/ and its allophone /ux/ comprised less than 11% of the total number of speech slices examined. Furthermore, only two-thirds of the speakers actually used /uw/ in their speech. This raises the interesting question of whether by including more instances of these phones the overall female and male group means could have been raised.

While the group means of the phones were noticeably different, the distributions of mean SFFs of individual speakers were very similar in range for the phones[25]. This equality in distribution implies a consistency in F_0 dynamics across speakers when uttering different phones. In other words, while the values attained may be different, speakers may follow a set pattern of lowering and raising their SFF when uttering different phones. Thus speakers may lower their fundamental frequency (relative to their own mean SFF) by a set amount for /ao/, and raise it for /uw/. The reason for this may be physical, in that the manoeuvring of the articulators in the supralaryngeal vocal tract to form the different phones could cause an interaction with the muscles of the laryngeal tract, thereby affecting the fundamental frequency. It is likely that while speakers' SFFs can differ, there is similarity in the control of their vocal musculature, resulting in similar SFF dynamics across different phones.

Only the female speakers showed any sign of a correlation between the phone spoken and their mean H_1-H_2. They were most likely to have a more negative H_1-H_2 for /ao/,

[24]There is always the possibility that this may have been due to the positioning of the phones within the carrier sentences. In declarative sentences, which were used as stimuli for the TIMIT database, F_0 tends to fall through the sentence; thus, due to some statistical anomaly, there may have been a greater proportion of /ao/ phones in the latter halves of sentences and /uw/ phones in sentence-initial positions. This anomaly may have arisen from the inclusion of phones from the TIMIT test sentences in this data analysis, i.e. from the two sentences spoken by all the subjects. Whereas the other sentences were chosen at random, presumably yielding a fairly random distribution of sentence positions, this may have resulted in a certain percentage of each phone type occurring in particular sentence positions. This may have been sufficient to skew the group F_0 means for the phones, especially for those means based on small numbers of phones uttered per speaker, such as /uw/ and /ux/. Unfortunately, time considerations precluded further investigation of this point.

[25]The exception, for both sexes, was /uw/, which had a wider distribution of means than the other phones, although this may have been an artifact of the few instances of this phone used per speaker.

and most likely to have a more positive H_1-H_2 for /ae/, although the differences in the distribution of SPEAKER MEANS between /ae/ and /aa/ were relatively minor. The male speakers showed no appreciable differences in mean H_1-H_2 across the three phones, although as for the women, the relative first harmonic amplitude was slightly less in /ao/ phones.

This tendency for the female speakers to produce different H_1-H_2 values for the three phones meant that the female-male difference was not the same for all phones. Thus while the distributions of individual female and male means for /ao/ were similar (although the female speakers had a slight tendency to have a more positive mean H_1-H_2), the phone evidencing the greatest contrast between the sexes was /ae/. By way of an illustration, 45.4% (59 speakers) of the women had a mean H_1-H_2 more than -4dB for /ae/, compared to 15.2% (44) of the men. In contrast, the figures for women and men for /aa/ were 28.5% (37) and 13.1% (38); for /ao/ they were 22.3% (29) and 12.4% (36); and for all three phones they were 26.9% (35) and 13.8% (40).

Considering all three formants together, the phone having the largest female-male difference was /ae/ at 19%, with /aa/ and /iy/ having the least at 11%. However, the size of the difference depended upon the formant, and could stretch over a large range. For example, the scaling factors for /aa/ ranged from 5% for F_3 to 17% for F_1. For some phones, the scaling factors for some phones were relatively consistent across the three formants: e.g. the factors for /ae/ were 19%, 20% and 19%.

E. Analysis by speaker variable

Due to uncertainties about the accuracy of the formant frequency data, only the fundamental frequency and relative first harmonic amplitude were analysed by speaker variable. The speaker variables considered were those of age, height, ethnic group and dialect region.

(i) Speaker age

Mean SFF was steady for speakers of both sexes in their 20s and 30s. The mean SFFs for these two age groups were approximately 210Hz for the female speakers and 120Hz for the male speakers. That these figures are the same as the overall means for each sex comes as no surprise, as the majority of speakers examined for this study (88% of the women and 93% of the men) were aged between 20 and 39 years. For female speakers over the age of 40, there was a substantial drop in SFF, possibly caused by women completing the menopause (the mean SFFs for the age groups 40-49 and 50-59 were 189Hz and 184Hz respectively). There were no discernible trends for the older men, as the few speakers representing the older age groups had a wide range of mean SFFs.

The mean relative first harmonic amplitude was also steady for speakers of both sexes in their 20s and 30s (and almost the same for female speakers in their 40s). There was some indication that for speakers over the age of 40 H_1-H_2 fell for women and rose for men, to the extent that the group mean for men in their 50s was actually greater than the women's group mean. Thus these results suggest that older men have a more breathy voice quality than older women (insofar as the relative amplitude of the first harmonic is an adequate correlate of the perceptual voice quality of breathiness). It is possible that this is due to aging in the vocal processes, although it is odd that older men should end up having a greater relative first harmonic amplitude. If however social conditioning plays a significant part in dichotomising this parameter along sexual lines, it is also possible that as we get older we attach less importance to sounding as different as possible to the other sex. In other words, men may not feel so inclined to sound less breathy than women, and women may feel less inclined to use breathiness as a marker of their sex.

(ii) Speaker height

There were clear trends for decreasing SFF with increasing speaker height for both sexes, although these were difficult to quantify with any certainty. There were two reasons for this, the first being that the speakers displayed a wide range of mean SFFs within their height groups, resulting in a large overlap between adjacent height groups. Secondly, only the height groups close to the average height for each sex were well-represented on the TIMIT database, specifically women between 5'2" and 5'9", and men between 5'6" and 6'3". Thus while the means for these height groups indicated a downward trend in SFF, it was difficult to extrapolate for shorter or taller people. However, the profiles of the distributions of individual SFF means compared across these groups did indicate a general downward shift in SFF. This conclusion appeals to intuition, in that one would expect taller speakers to have larger vocal folds, and therefore be more likely to produce a lower SFF. However, the wide range of SFFs measured for these speakers attests to the fact that speaker height is only one factor influencing vocal fold size (others being body mass, acculturation, etc.).

No reliable trends were detected between speaker height and H_1–H_2, for either sex.

(iii) Ethnic group of speaker

There were no significant differences in SFF between black and white speakers.

On the other hand, both female and male black speakers exhibited similar use of H_1–H_2, which in turn was similar to white male use. In other words, the black speakers did not appear to use H_1–H_2 as a sex marker. However, black speakers were under-represented on the TIMIT database, the ten female and twelve male black speakers comprising only 5% of the total number of speakers. In contrast, the white speakers accounted for 90% of the sample.

(iv) Speaker dialect region

There was a substantial difference in SFF behaviour between the female speakers comprising the dialect regions of North and South Midlands, the regions showing the lowest (198Hz) and highest (218Hz) mean SFFs. In the absence of explanations due to, for example, disproportionate numbers of short people representing the North Midlands dialect region, this indicates a large acculturation effect defining regional background in these speakers' voices. While the differences between the regions for male speakers were largely negligible, the group means running from 125Hz for dr4 down to 117Hz for dr3 and dr7, the 'region' comprised of speakers who moved around during childhood had a mean of 112Hz.

For H_1–H_2, the women again showed great variability across the dialect regions, and the men less so. The group means for the women fell into two ranges of H_1–H_2: -4.0dB to -4.5dB, and -5.2dB to -5.4dB. The male group means were spread fairly evenly between -5.6dB and -6.7dB.

4.3 Discussion Of Results

In this section, the results reported in Section 4.2, and summarised in Section 4.2.4, are discussed by examining the main conclusions of the literature review of the acoustic-phonetic characterisation of speaker sex, namely:

1. **The acoustic-phonetic characteristics of women's speech are, in general, different to those of men's.** This is examined by comparing the female and male group averages for each of the acoustic-phonetic measures investigated.

2. **Speaker sex is signalled by a number of acoustic-phonetic parameters.** Check if there are any correlations between the APVs, e.g. does a female speaker have to have a high F_0, breathy voice and high formants?

3. **There is no such thing as an average speaker for each sex.** Show the variation in APV values due to between-speaker differences (compare speaker means with the average for the entire sex) and within-speaker differences (compare a speaker's slice means with their speaker mean). Also look at effect of speaker height/age/etc.

Regrettably, only the fundamental frequency and relative first harmonic amplitude were subjected to a thorough analysis. A lack of confidence in the results produced by the estimator, combined with insufficient time with which to carry out a more thorough investigation, led to a less thorough analysis of the formant frequencies (see Part D of Appendix B.6). Consequently, the discussion of the effects of between- and within-speaker differences focusses on fundamental frequency and the relative amplitude of the first harmonic.

1. Are the acoustic-phonetic characteristics of women's speech different to those of men's?

Judged on their group behaviour, the female and male speakers on the TIMIT database displayed a sexual dichotomy for each of the three acoustic-phonetic measures investigated, although the degree to which this dichotomy was in evidence was dependent upon the measure. The greatest difference was found in SFF. Not only was the average female SFF 73% greater than the male, but a comparison of the distributions of the average SFFs of individual speakers showed there was very little overlap between the sexes. Furthermore, even when taking into account the range of F_0 used by individuals, the female and male speakers rarely attained similar values. A value of F_0 set at 160Hz illustrates the strength of the dichotomy: fully 98% of the female speakers' mean SFFs were in excess of this value, and 97% of the male mean SFFs were below it; even more surprisingly, 94% of the F_0s of individual speech segments from female speakers were greater than 160Hz, while 94% of the instances of male speech were less than it.

The marker with the least capability for differentiating between the sexes was the relative amplitude of the first harmonic. The difference between the female and male group means for H_1–H_2 was measured at 1.4dB, considerably less than has been reported in the literature. However, a tendency did exist for female speech to have a higher H_1–H_2 than male speech, although there was a substantial overlap in the value of this measure realised in the voices of individual women and men. Furthermore, female and male speakers exhibited differential behaviour in the range of H_1–H_2 they produced. The men overwhelmingly had restricted ranges, while the distribution of ranges the women produced was much wider.

On average, the formant frequencies of women were 15% greater than those of men, although as has been reported previously in the literature, the female-male difference in

formant frequencies was dependent upon both the phone being uttered and the formant being measured. The difference in group means ranged from 5% for F_3 of /aa/ to 20% for F_1 of /ao/ and F_2 of /ae/. While the overall female-male difference is similar to that reported by both Peterson & Barney (1952) and Fant (1959), the differences for individual phones and formants were less varied, with the majority lying between 9% and 20%. Both the Peterson & Barney and Fant studies reported a number of differences below 9% and above 20%, indicating that the female-male dichotomy in the formant frequencies was more consistent for the TIMIT data. This was particularly true for F_1 and F_2, and the formants of /ae/, /ao/, /uw/ and /ux/. The phone showing the greatest degree of difference between the sexes was /ae/: the female-male differences for F_1, F_2 and F_3 were 19%, 20% and 19% respectively, and the separation between the s.d.s of each sex for the three formants indicated that female and male speakers rarely attained the same frequency values.

In summary, the acoustic-phonetic characteristics of women's speech examined in this study were in general different to those of men's. For each of the three frequency measures, at least a general tendency was found for female speakers to attain different values to male speakers.

2. Is speaker sex is signalled by a number of acoustic-phonetic parameters?

As has already been stated, the most powerful sex-differentiating acoustic-phonetic marker was the SFF. There was very little overlap between the mean SFFs of individual women and those of men, or even between the F_0s of individual speech segments. While a small number of speakers had a SFF atypical of their sex, in that some women had a mean SFF that was relatively low and some men had a relatively high mean SFF, even the SFFs of these speakers did not approach the opposite sex's group mean. These results indicate that SFF is probably the major signaller of sex in the voice. However, the results could give a false impression of the degree of the sexual dichotomy – i.e. of the lack of overlap in SFF between female and male speech – as the input data consisted solely of read speech. The less dynamic intonation associated with read speech would have restricted the range of SFF employed by the subjects when their speech was being recorded. Furthermore, while it has been stated that the recording conditions were relatively stress-free, they may have sufficiently intimidated some people into reducing their normal speaking range further. Moreover, speech from people experiencing a range of emotions would probably produce a greater overlap between the sexes.

The survey of the studies on the perception of speaker sex reported in Section 3.2.2 showed that SFF was not a *necessary* marker of sex, in that an absence of phonation information in the acoustic speech signal did not adversely affect the ability of listeners to identify a person's sex. Furthermore, the survey of child sex perception studies in Section 3.2.1 indicated that children were using acoustic means other than F_0 to successfully signal their sex. More importantly, they showed that the impression of femaleness and maleness in the voice is extremely robust and is formed from a number of dimensions. Thus, while the SFFs of women and men measured from the TIMIT data evidence remarkable differentiation, the results for H_1–H_2 and the formant frequencies clearly show there are other sources of acoustic sexual dichotomy.

Greater differentiation was found in the formant frequencies than in the H_1–H_2 measure. Using the s.d. of the SPEAKER MEANS as a guide to the distribution of formant frequencies (in that 68% of the means are encompassed by the s.d. interval of a normally-distributed sample), for approximately half of the formant frequencies of the phones, the distributions

of female and male speakers are relatively distinct. This can be viewed in Figures 4.39 to 4.41. Thus, speaker sex appears to be signalled powerfully in parts of the formant structure of the frequency spectrum.

Although there was substantial overlap in female and male relative first harmonic amplitudes, female speakers in general could be expected to have a higher value of H_1–H_2. The major difference between the sexes in this measure was the number of women who had a very wide range of H_1–H_2. While the majority of both women and men had a range of less than 10dB (involving 51% of the women, and an overwhelming 90% of the men), over two fifths of the female speakers had a range between 10dB and 17dB compared to less than a tenth of the males. Rather more reliable differentiators of sex were found in the actual amplitudes of the harmonics. The female and male group means were 90.4dB and 85.2dB for H_1, and 95.3dB and 91.4dB for H_2, giving female-male differences of 5.2dB for H_1, and 3.9dB for H_2. Although they were not investigated in great depth, the distributions of the mean amplitudes of individual speakers suggested a substantial trend for women to speak with higher first and second harmonic amplitudes.

To investigate the possibility that the acoustic-phonetic measures are inter-connected – i.e. that speakers with a particular value of SFF also have particular values for H_1–H_2 and the formants – a number of analyses were run to establish whether any correlations existed. Nittrouer *et al.* (1990:770,772) suggested that speakers with a relatively high SFF also have relatively high H_1–H_2s. Thus linear correlation analyses[26] were performed on the SPEAKER MEANS with F_0 and H_1–H_2 as variables. The correlation coefficients for each of the three phones /aa/, /ae/, /ao/, and all three together were:

Sex	F_0 vs. H_1–H_2			
	/aa/	/ae/	/ao/	all
f	-0.001	0.146	-0.208	0.015
m	0.016	0.064	-0.051	0.008

Neither the coefficients nor scatter plots of the data revealed any pattern whatsoever. The strongest correlation here, for the female /ao/, is mostly due to two females with low mean SFFs (both 93Hz) and high H_1–H_2 means (0.4Hz and 0.5Hz): removing these two speakers from the correlation analysis more than halves the coefficient to -0.091. It is possible that the SPEAKER MEANS, which are generally composed of wide ranges of values (for both F_0 and H_1–H_2), disguise a correlation existing for speech segments. Thus correlation analyses were performed on the F_0 and H_1–H_2 SLICE MEANS, the results of which were:

Sex	F_0 vs. H_1–H_2			
	/aa/	/ae/	/ao/	all
f	-0.012	0.096	-0.088	0.030
m	0.035	-0.016	-0.115	-0.032

Again, neither the coefficients nor scatter plots of the data revealed any patterns. The strongest correlation here, for the male /ao/, reveals a very small, counter-intuitive trend for breathiness (as measured by the H_1–H_2 parameter) to decrease with increasing F_0.

However, there does appear to be a correlation between the amplitudes of the harmonics and SFF. Correlation analyses performed on the SPEAKER MEANS with F_0 and H_1 and then F_0 and H_2 as variables, produced the following results:

Sex	F_0 vs. H_1	F_0 vs. H_2
f	0.567	0.492
m	0.645	0.589

[26]The correlation analyses were performed using the |STAT data analysis software package on a Next computer.

As the correlation coefficients and the scatter plots in Figures 4.42 to 4.43 show, for both sexes, increasing SFF indicates increasing harmonic amplitudes, although there was a larger correlation for the male speakers. Furthermore, there appears to be a *general* tendency for speakers with higher fundamentals to have higher harmonic amplitudes. This has already been indicated by the tendency found for female speakers to have higher first and second harmonic amplitudes. Further evidence comes from correlation analyses performed on *all* the SPEAKER MEANS, which produced coefficients of 0.721 using F_0 and H_1 as variables, and 0.595 using F_0 and H_2 as variables.

The presence of a correlation between SFF and any of the formant frequencies was also investigated. Thus, correlation analyses were performed using the SFF and the formant frequency SPEAKER MEANS as variables. As the coefficients reproduced in the table below show, there were no strong correlations between SFF and any of the formants, although the formant frequencies of some of the phones showed a degree of correlation.

Analysis	Sex	/aa/	/ae/	/ao/	/iy/	all
F_0 vs. F_1	f	0.177	0.258	0.155	0.384	0.310
	m	0.224	0.133	0.079	-0.098	-0.018
F_0 vs. F_2	f	0.035	0.084	0.076	0.126	0.208
	m	0.068	0.104	0.025	0.102	0.100
F_0 vs. F_3	f	-0.017	0.219	-0.008	0.258	0.213
	m	0.057	0.102	0.135	0.043	0.118

In summary, the SFF appears to be the most powerful and most consistent acoustic-phonetic measure for the differentiation of speakers of a different sex. However, it is highly unlikely that we base our perceptions of speaker sex on SFF alone, given the high degree of sexual dichotomy to be found in other measures. The formant frequencies appear to be a further powerful source of dichotomy, although more investigation is needed to establish to which formants and to which phones this applies. Finally, the amplitudes of the first two harmonics also appeared to be capable of sex-differentiation.

3. Is there such thing as an average speaker for each sex?

Assessing the extent of speaker variability

Figures for acoustic-phonetic measures are often reported in the literature as if they apply to all speakers of a particular type, for example, the fundamental frequency of men *is* 120Hz. The research reported in this thesis has shown that within any identifiable group of speakers there is generally a substantial amount of both within- and between speaker variation in the value of the measure

The extent of between-speaker variability for a particular acoustic-phonetic measure can be investigated by examining the distribution of SPEAKER MEANS, and by using the standard deviation of the SPEAKER MEANS as a rough guide. For a normally-distributed sample, the interval encompassed by the s.d. of the sample contains 68.3% of the data. While the distributions examined in this thesis are unlikely to ever be *exactly* normal, they are generally close enough to allow the s.d. to serve as a rough guide to the range containing the greatest density of data. Consider, for example, the SFF results. The overall mean female SFF was measured at 208Hz, with a s.d. of 23Hz. From this we can reasonably state that the majority (i.e. approximately 70%) of female SPEAKER MEANS fell within the range 185-231Hz. The remainder of the means fall outside this range (between 146Hz and 270Hz), and may be considered as outliers. However, for the thorough characterisation of female SFF characteristics, it is essential that they are included in any description of female SFF. The approximately 15% of the women who had a mean SFF between

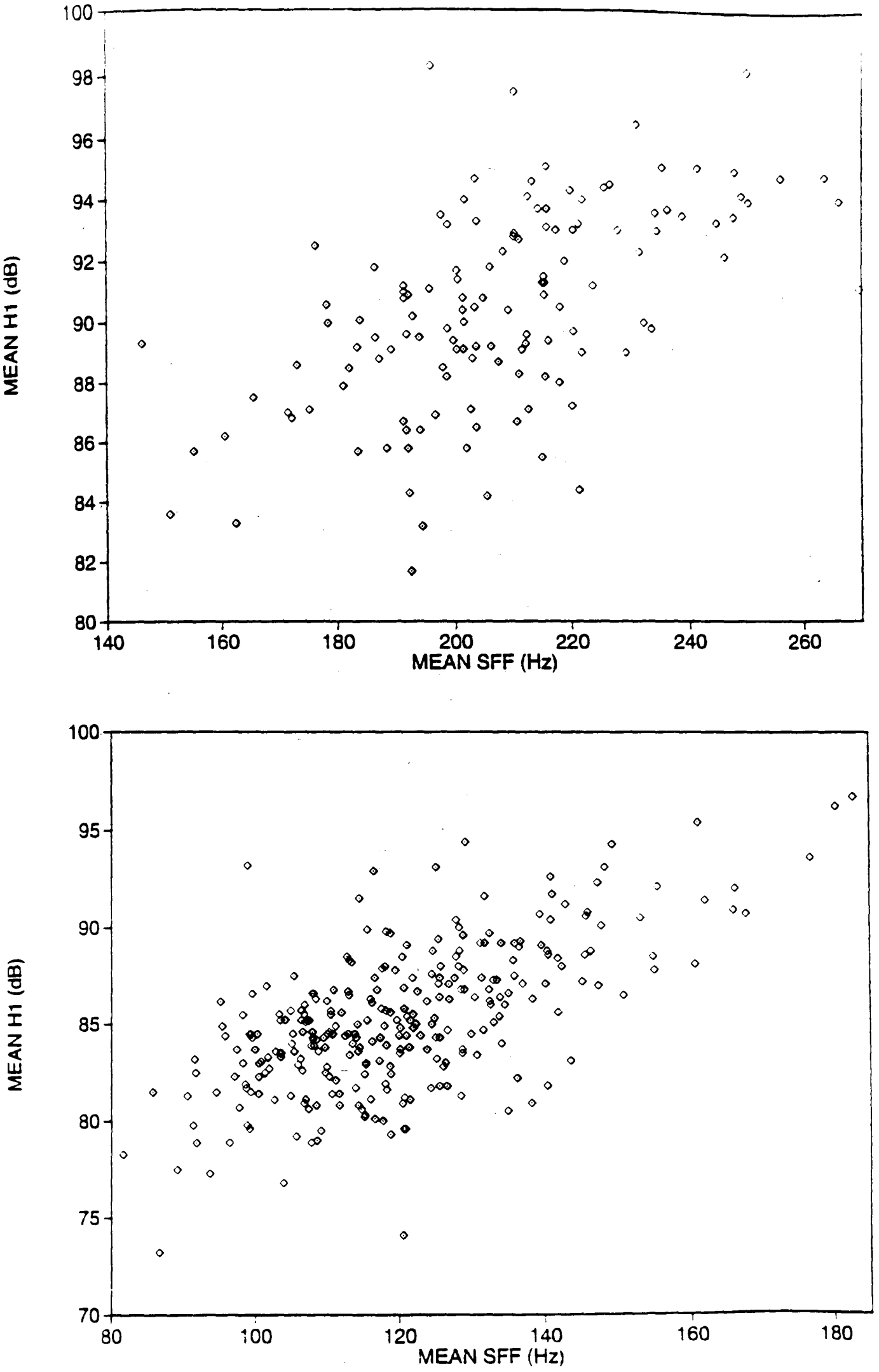
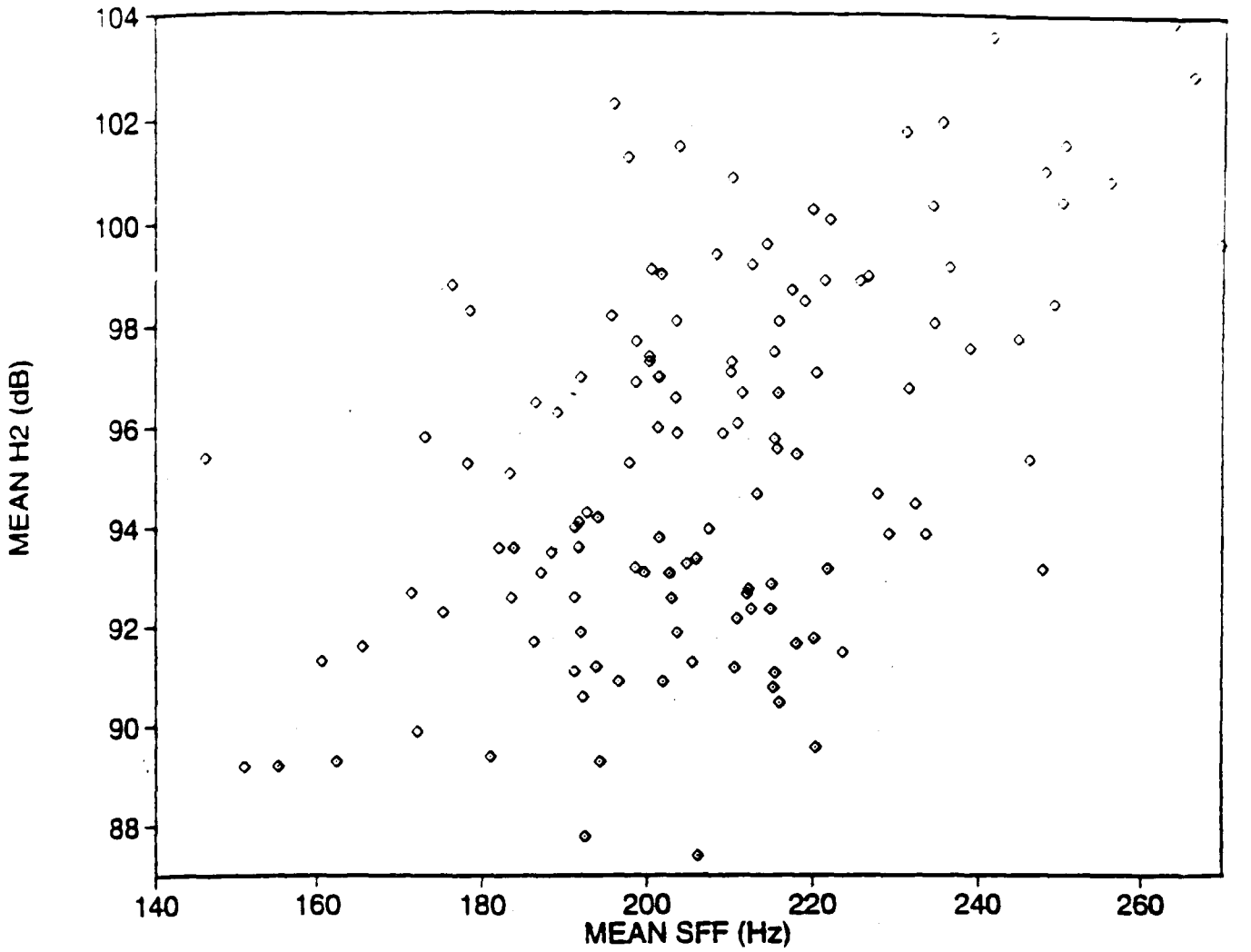


Figure 4.42: Scatter plot of F_0 SPEAKER MEANS versus H_1 SPEAKER MEANS for women (top) and men (bottom). The correlation coefficient was 0.567 for the women, and 0.645 for the men.



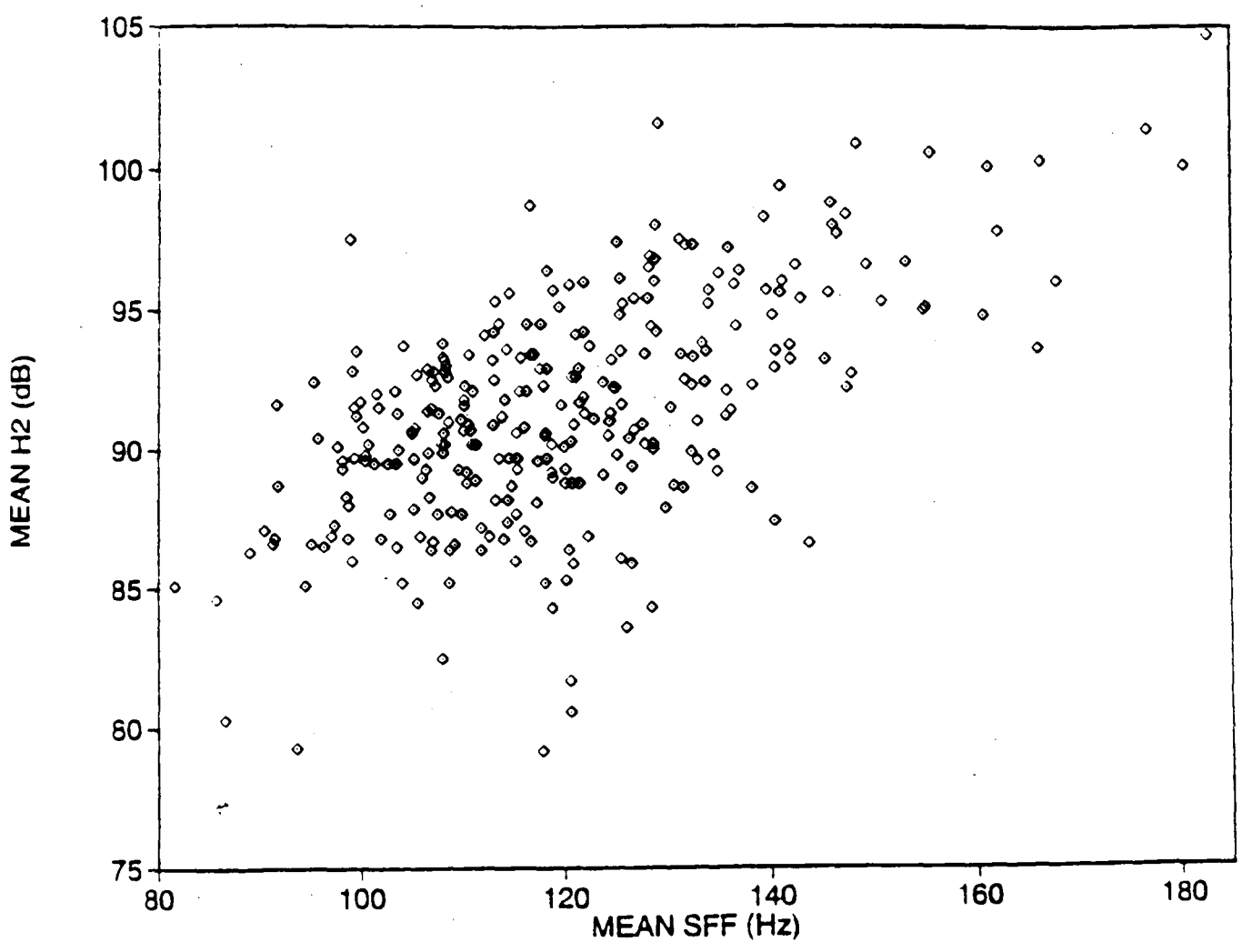


Figure 4.43: Scatter plot of F_0 SPEAKER MEANS versus H_2 SPEAKER MEANS for women (top) and men (bottom). The correlation coefficient was 0.492 for the women, and 0.589 for the men.

146-185Hz, and the 15% between 231-270Hz show just how different a woman's SFF can be to the expected value. Thus rather than citing a value for female SFF as 208Hz, a better description would be: The mean SFF of a female speaker will tend to lie within the range 185-231Hz, but may be as low as 146Hz and as high as 270Hz. Carrying out the same procedure for the male speakers, we arrive at a mean SFF in the range 103-137Hz, although it may be as low as 82Hz and as high as 183Hz. To gauge the accuracy of this procedure, the actual percentage of SPEAKER MEANS covered by the above ranges is 71% for the female data, and 72% of the male data. For the H_1–H_2 results (see below), the actual percentage covered was 72% for both sexes.

Another possible method for examining the extent of between-speaker variability involves the inter-quartile range (IQR) of the data. Considering again the SFF results, the IQRs for the female and male speakers were 193-220Hz and 108-129Hz respectively. Referring to the histogram of SPEAKER MEANS in Figure 4.9, we can see that the IQR tends to encompass the tallest intervals of the histogram (although the 100-110Hz interval for the male speakers is largely missed by the range). As these intervals are the tallest and are roughly equal in height, we can say that speakers are equally likely to have a mean SFF within this range. Furthermore, the IQR is not dependent upon the normal distribution assumption, representing as it does the actual range of the middle 50% of the data. However, the advantage of using the s.d. is that it is more reflective of the variability between speakers, as it encompasses a greater proportion of speakers.

Between-speaker variability -
Adopting the above methodology, the variability in SFF and H_1–H_2 between speakers can be summarised as follows:

SFF The mean SFF of a female speaker will tend to lie within the range 185-231Hz, but may be as low as 146Hz and as high as 270Hz. The mean SFF of a male speaker will tend to lie within the range 103-137Hz, but may be as low as 82Hz and as high as 183Hz.

H_1–H_2 The mean H_1–H_2 of a female speaker will tend to lie within the range -7.1dB to -2.5dB, but may be as low as -9.1dB and as high as 1.8dB. The mean H_1–H_2 of a male speaker will tend to lie within the range -8.2dB to -4.2dB, but may be as low as -11.7dB and as high as 0.8dB.

These figures clearly show the wide differences in mean SFF and H_1–H_2 produced by speakers.

Despite the huge variety of mean SFFs produced by female and male speakers, there was only a very small overlap between them. Note that the range of typical female means defined above does not include the male speaker with the highest mean SFF, and similarly for the male means. In contrast, there was a substantial overlap between the mean H_1–H_2s produced by individual women and men.

Within-speaker variability
The variability expressed by individual speakers in the values of the acoustic-phonetic measures can be explored by considering the range of values they attained when reading the TIMIT sentences. The mean range of SFF was 10.6st (with a s.d. of 4.5st) for the female speakers, and 9.7st (3.8st) for the male speakers; the Hertz equivalents were 122Hz (40Hz) and 66Hz (24Hz) respectively. The mean range of H_1–H_2 was 10.8dB (4.5dB) for the female speakers, and 7.0dB (2.9dB) for the male speakers. The procedure used to examine the extent of between-speaker variability is not as useful for the speaker ranges, as the distribution of ranges of both SFF and H_1–H_2 are highly skewed (although noticeably less

so for the female H_1–H_2 ranges). For both measures, the majority of ranges were clustered at the lower end of the scale, with a small percentage of speakers producing ranges that were relatively much larger (see the histograms in Figures 4.11 and 4.28, which show a long tail to the distributions on the right hand side). However, the variability expressed by individual speakers can be summarised as follows:

SFF The range of SFF produced by a female speaker will tend to be between 4.8-12.0st (or 58.3-150.0Hz), but may be as high as 24.7st (227.2Hz). The range of SFF produced by a male speaker will tend to lie within the range 5-12st (or 36-84Hz), but may be as low as 4.0st (25.8Hz) and as high as 24.7st (141.2Hz)[27].

H_1–H_2 The range of H_1–H_2 produced by a female speaker will tend to be between 4-13dB, but may be as low as 3.9dB and as high as 28.1dB. The range of H_1–H_2 produced by a male speaker will tend to be between 3.5-9.0dB, but may be as low as 2dB and as high as 24.1dB[28].

The above figures show just how different speakers can be with regard to their ranges of these acoustic-phonetic measures.

When expressed in semi-tones, the ranges of SFF produced by women and men tend to be similar, although there was a slight trend for men to have smaller ranges. Furthermore, women and men were equally likely to have a very large range of SFF (i.e. over 12st). In contrast, there was a sex-related difference in the range of H_1–H_2 produced by individual women and men. The vast majority of the men had a relatively restricted range, while the women were almost equally likely to have a relatively small or large range. The women also produced more extreme values of H_1–H_2, both positive and negative.

Variability due to different phone types

Effects on the values of SFF and H_1–H_2 were found to exist for different vowel phones. The difference between the phones with the lowest and highest group mean SFFs (/ao/ and /uw/) were 24Hz for the female speakers and 11Hz for the males. While being different in value, the distributions of individual mean SFFs for each phones were very similar in profile. Only the female speakers evidenced any differential behaviour in H_1–H_2 between phones. The greatest contrast between the sexes was for /ae/. Interestingly, the distribution of individual female and male mean H_1–H_2s for /ao/ were very similar in both range and value, although there was still a tendency for the women to have more positive values for H_1–H_2.

Variability due to different speaker types

One other source of speaker variability it is possible to investigate using the TIMIT data is the effect on the acoustic-phonetic measures of the extralinguistic speaker attributes described in the speaker information file. As will be seen, these can have a substantial effect of the values of the measures reported for each sex.

Probably the largest effect was found for speaker age. For speakers in their twenties and thirties, both SFF and relative first harmonic amplitude were very similar. However,

[27] The 4.8-12.0st range for the female speakers and 5-12st range for the male speakers each encompassed 77% of the speakers; in addition, 4.8st (25.8Hz) was the lowest female range. The 58.3-150.0Hz range for the female speakers encompassed 78% of the speakers, and 36-84Hz range for the male speakers encompassed 76% of the speakers.

[28] The figures for the female speakers were difficult to judge because of how spread out the ranges of individual women were (see Figure 4.28). The typical female range could easily have been defined as lying between 4-17dB, which encompasses 92% of the speakers. As it is, the 4-13dB encompassed 70% of the speakers. The 3.5-9.0dB range for the male speakers encompassed 81% of the speakers.

as speakers age past forty years, these results indicate that major changes occur in the values of the measures the speakers produce. Although no discernible trends in SFF were found for older men, older women appear to experience a drop in SFF of the order of 20-25Hz. While evidence from Stoicheff (1981) suggests that this is a direct result of women completing the menopause, the data presented here has most women completing the menopause in their forties, a decade earlier than Stoicheff found. The relative first harmonic amplitude fell (became more negative) for older women and rose for older men. to the extent that the group mean for men in their fifties was actually higher than the equivalent women's group mean. While no reliable trends were found linking speaker height and H_1–H_2, the SFF of both female and male speakers tended to decrease as the height of the speakers increased. Conversely, no significant differences were found between the SFFs of black and white speakers of either sex, while black speakers did not appear to use H_1–H_2 as a marker of speaker sex, unlike the white speakers. A speaker's dialect had a noticeable effect on both SFF and H_1–H_2. Excluding the dialect 'region' composed of speakers who moved around during childhood, the mean SFFs for the different dialects ranged over 20Hz for the women and 8Hz for the men, and the mean H_1–H_2s ranged over 1.4dB for the women and 1.1dB for the men.

However, the strength of any trends must be tempered by the knowledge that a number of speaker groups were under-represented on the TIMIT database, as any apparent trend may well be an artifact of having only a small sample of speakers. The vast majority of speakers were white, university-educated (and therefore probably from the middle classes) U.S. citizens aged between twenty and forty years. Thus, splitting the speakers into groups to analyse a particular speaker attribute was sometimes unsatisfactory. For example, only 8% (10) of the women and 4% (12) of the men were black, a fact which somewhat impoverishes the finding that black speakers possibly do not use H_1–H_2 as a marker of speaker sex. The vagaries of the population sampling process, combined with the extent of the variability in the acoustic-phonetic measures, means that this sample may not be truly representative of black people as a whole. Furthermore, although the effects of between- and within-speaker variability were clearly at work within the different speaker groups, it can be difficult to gauge the true extent of the variability from such small samples. This study is really the study of the acoustic-phonetic characteristics of young, white, middle class women and men from the U.S.A., and can only point to the *possibility* that trends exist for other speaker groups.

Summary

The results reported here show conclusively that the values of the acoustic-phonetic characteristics of speech are extremely varied. For speaker sex, the reporting of average values for different acoustic-phonetic measures has been shown to be, at best, inadequate, in that an average cannot hope to reflect the actual values attained by individual women and men. The results also indicate that sources of variability other than speaker sex can significantly effect expectations of the values of the measures. While an average value can indicate the possibility of the existence of a difference between certain speaker groups, on its own it disguises the true acoustic-phonetic behaviour of speakers.

However, the strength of the effect of some of the sources of variability is hard to estimate due to their under-representation on the TIMIT database. As well as the smallness of some of the speaker group samples noted above, the phone /uw/ and its allophone /ux/ were relatively poorly represented in relation to the other phones analysed.

4.4 Conclusions

4.4.1 Remarks on speaker characterisation

Models of speaker characteristics require a solid numerical foundation built on the analysis of real speech data. They must reflect not only the features which on average have a correlation with a particular speaker characteristic, but also the extent to which in real speech the values of these features vary between speakers and within the speech of individual speakers. To establish the true extent of the variability due to different speakers requires the investigation of large quantities of speech data. However, a review of the academic literature reveals that much of the investigation of speaker characteristics involves only small numbers of speakers in particular research conditions, while in the conclusions to these investigation it is often implied that these small samples are representative of the speaking population as a whole. Furthermore, when research studies are compared, they often prove to be inconclusive or contradictory, or only throw light on a particular group of speakers. The data is usually reported in insufficient detail to gauge the variability of a particular parameter.

From the literature review reported upon in Chapter 3, and from the research undertaken for this thesis, it is clear that speakers within any definable group exhibit great variation in whatever feature is being measured. While this thesis has concentrated on grouping speakers according to their sex, this variability was also shown to exist within speaker groups organised by age, height, dialect region and educational background. The variation occurs both in the mean values of speakers relative to the group means for that feature, and in the range of values attained by individual speakers even in the course of reading a small number of sentences. It is worth bearing in mind that the extensive variability shown by the TIMIT speakers occurred despite the fact that the data analysed for this study was read speech from a small number of sentences per speaker, recorded in an environment relatively free from noise and stress. There is a need then for researchers to analyse sufficient speakers to examine the extent of between-speaker variability, and sufficient speech from individuals to be able to account for within-speaker variability. Models of speaker characteristics also require the acknowledgement of the fact that in different situations, the features of the model could take on an entirely different range of values, for example due to changes in the speaker's affective state. Furthermore, speakers who are members of one group will also be members of many other speaker groups, each with their own attendant feature values.

4.4.2 Remarks on the acoustic-phonetic characterisation of speaker sex

There are three main points to be made regarding the acoustic-phonetic characteristics of speaker sex. These are dealt with below.

First, the acoustic-phonetic characteristics of the female voice are, in general, different from those of the male. From the analysis of the TIMIT data, the greatest degree of sexual dichotomy was found in SFF. The average female SFF was 73% greater than the male. Female formant frequencies were, on average, 15% greater than male formant frequencies. The smallest degree of sexual dichotomy was found in the relative amplitude of the first harmonic, for which the average female-male difference was 1.4dB. Furthermore, there is evidence from the literature of sex-related differences in such voice source features as jitter (e.g. Orlikoff & Baken 1990) and laryngealisation (e.g. Klatt & Klatt 1990. The combined influences of biology and acculturation instil a unique femaleness in women's voices, and

a unique maleness in men's voices. This ensures that the perception of speaker sex is a very easy, even trivial, task.

Secondly, there is no simple relationship between the acoustic-phonetic features of female and male voices. The female vocal apparatus is *not* a scaled down version of the male vocal apparatus. There appear to be fundamental differences in the anatomy of both the voice source and vocal tract, which are reflected in the acoustic speech signal. Furthermore, there is evidence that the vocal apparatus is used in different ways by women and men, principally to accentuate the acoustic-phonetic differences between them. Speaker sex perception studies have shown that sex is signalled in the acoustic-phonetic features of the speech signal in a number of ways, not just by fundamental frequency, with the result that the percept of speaker sex is extremely robust and able to survive substantial degradation.

Thirdly, the notion of an 'average' or 'ideal' speaker typifying each sex is at best inadequate. Individual speakers do *not* conform to group averages. This research has found that the extent of between- and within-speaker variability in the values of the acoustic-phonetic features is considerable. The citation of average values for sex is highly misleading, and results in an impoverished description of what constitutes speaker sex. Group averages disguise the wide range of values individuals utilise during everyday speech, the different ranges of values adopted by individuals in response to different situations, and the effect of extralinguistic speaker attributes such as age, height, ethnic group and socioeconomic background. To provide a richer description of the variability in the acoustic-phonetic features found between speakers, the average feature values are better expressed in a way which acknowledges the values of the bulk of the speakers. Thus:

- **Mean of speaking fundamental frequency**
  - The mean SFF of a female speaker will tend to lie within the range 185-231Hz, but may be as low as 146Hz and as high as 270Hz.
  - The mean SFF of a male speaker will tend to lie within the range 103-137Hz, but may be as low as 82Hz and as high as 183Hz.
- **Mean of relative first harmonic amplitude**
  - The mean H_1–H_2 of a female speaker will tend to lie within the range -7.1dB to -2.5dB, but may be as low as -9.1dB and as high as 1.8dB.
  - The mean H_1–H_2 of a male speaker will tend to lie within the range -8.2dB to -4.2dB, but may be as low as -11.7dB and as high as 0.8dB.

Similarly, within-speaker variability can be expressed in terms of the range of feature values a person is likely to attain during the course of their speech:

- **Range of speaking fundamental frequency**
  - The range of SFF produced by a female speaker will tend to be between 4.8-12.0st (or 58.3-150.0Hz), but may be as high as 24.7st (227.2Hz).
  - The range of SFF produced by a male speaker will tend to lie within the range 5-12st (or 36-84Hz), but may be as low as 4.0st (25.8Hz) and as high as 24.7st (141.2Hz).
- **Range of relative first harmonic amplitude**
  - The range of H_1–H_2 produced by a female speaker will tend to be between 4-13dB, but may be as low as 3.9dB and as high as 28.1dB.

- The range of H_1-H_2 produced by a male speaker will tend to be between 3.5-9.0dB, but may be as low as 2dB and as high as 24.1dB.

However, the results of any study into speech are inevitably constrained by the data upon which the analysis was based. Thus, while the need for a sound statistical basis to the description of the parameters of the voice has been acknowledged, the thoroughness with which the parameters can be described will almost inevitably be compromised by the sheer scale of analysis required to investigate all speaker types in all speaking conditions. One of the most important aspects to the investigation of speaker characteristics is therefore a description of the constitution of the speech data. The results of an investigation should include a description of what aspect of speaking behaviour and what speaker groups are represented by the data.

For the research reported here, the TIMIT database imposed two restrictions on the results, while the method of investigation imposed one further restriction. Firstly, the speakers represented on the database are in the main white, middle class, university-educated U.S. citizens aged between 20 and 40 years. Moreover, the male speakers outnumber the female speakers by more than two to one. Particularly in the light of the effects found due to different speaker types, it is insufficient to claim that the results of this study are representative of the population as a whole. Furthermore, its conclusions can only point towards the possible trends that exist for other speaker groups. Further research must be carried out to establish whether these results apply to other speaker types. Secondly, the database consists solely of read sentences. It is well-recognised that read speech is less dynamic than spontaneous speech, and therefore will not be truly representative of a person's normal speaking patterns. The use of less dynamic vocal characteristics will therefore affect the analysis and extent of within-speaker variability. Thirdly, the investigation reported upon in this thesis considered only a particular set of vowel types, and only three acoustic-phonetic features. It is possible, and indeed likely, that important sex-discriminating information is carried in other types of speech sounds and in other speech domains, while there is evidence available that sex is discriminated in other acoustic-phonetic features, such as jitter and laryngealisation.

It is therefore important to state that this thesis is representative only of a particular subset of the world's population, under particular speaking conditions. In reality, this research constitutes a study of some of the acoustic-phonetic characteristics of the read speech of young, white, middle class female and male U.S. citizens. That said, the size of the population sample analysed here provides for a fuller, more reliable description of some of the acoustic-phonetic markers of speaker sex, and the databases of feature values produced are a powerful tool for further analysis, focussing, for example, on the acoustic-phonetic consequences of phonetic context. Moreover, most of the research in speech science is conducted on speech read by white, middle class, university-educated (male) U.S. citizens, and as such provides a useful base from which to compare the results of this study.

4.4.3 Remarks on the automated analysis of speech databases

The four main stages of an automatic analysis

Four main stages have been identified in the automatic analysis of speech data, which can be summarised as follows:

1. **Preparation of a database of input speech.** The most important feature of the input database is that it be computer-readable. Thus the items of speech data (in the form of, for example, sentences) must be labelled so that they are uniquely identifiable by computer. Transcriptions (wide or narrow) of the speech data must be provided to enable the targetting of particular types of phone/phoneme or contexts. Again, these must be computer-readable.

2. **Establishment of structures to control the analysis of the data.** The size of the task involved in a large-scale automated analysis requires that the input and output of data be controlled and organised in a consistent manner. Protocols must be defined for the structures of files and directories to keep track of the output data, and to allow cross-referencing between the results of different analyses.

3. **Formation of a database of analysed speech.** Software must be written to handle the extraction of the specified data from the input database, the channeling of the data through parameter measurement tools, and the output of the analysed data into an organised form.

4. **Statistical analysis of the database of analysed speech.** The particular form of the statistical analysis of the output database is dependent upon the nature of the investigation, and the information available for the categorisation of the input data. This former point can take the form of an analysis of, for example, within- or between-speaker variability, of overall parameter behaviour in different contexts, or of the effect of different speaker attributes. Regarding the latter point, performing an analysis as a function of speaker attributes requires descriptions of, for example, the age, regional background or smoking habits of the sample of speakers.

One of the most striking lessons to be learnt from development of the analysis procedure was the amount of effort involved in enactment of each of these stages. Furthermore, the design of each stage depended upon the constraints imposed by other stages.

The preparation of the input database is probably the most time-consuming stage, most notably in the hand labelling of the phones/phonemes, but also in the initial collection of the recorded speech data. Obviously, the larger the database required, the greater the transcription load. Fortunately, the TIMIT database satisfied the requirement of a computer-readable data source, although the format of the database rendered the speech data less amenable to analysis by commercial database software. This necessitated the development of software capable of extracting data from the CD-ROM. To reduce some of the problems associated with the signal processing of speech sounds, and therefore reduce the level of sophistication required from the signal processing algorithms, only vowels were used in the analysis. Thus, the extraction software had to be able to target specific segments of speech with the TIMIT sentence files.

The structures to control the input, analysis and output of the data had to be designed to enable the complex cross-referencing required in the statistical analysis of the output. The intention to investigate the extent of between- and within-speaker variability required the establishment of two output databases of signal-processed speech: one containing the results of the processing of individual speech segments, and one containing a summary of the results for individual speakers. However, the most time-consuming part of the research proved to be the signal processing of the input speech data, and this is examined below.

Constraints on the analysis imposed by the variability of speech

Below, general points are made about the implications of speaker variability for automated analyses, followed by a description of how this was dealt with for this research.

The signal processing analysis of speech can be fraught with difficulties, even more so for the automatic analysis of acoustic-phonetic parameters on a large scale. The algorithms developed to compute the values of the parameters must be extremely robust in order to cope both with the wide range of values one must expect, and with acoustic speech data in forms other than the neatly-periodic waveforms to be found in the textbooks. Referring to the first point, most signal processing algorithms rely on the expectation that a parameter will lie within a certain range of values, i.e. within a specified search space. For small-scale analyses, investigating for example a single person's speech, the search space can be tailored to suit that speaker's range. Speech databases for use in a large-scale, automated analysis tend to contain either a relatively small amount of speech from a large number of speakers, raising the problem of between-speaker differences, or a large amount of speech from a relatively small number of speakers, raising the problem of within-speaker differences. The findings from this study suggest that the extent of between- and within-speaker variation, and therefore the wide range of values to be expected, imposes a large burden of reliability on the signal processing algorithms. The range of the search space must be such that it can cope with the variability to be expected in the parameters, while at the same time excluding (spectral) features which may confuse the algorithms.

Secondly, as most measurement of acoustic-phonetic characteristics involves the estimation of frequency values, steps are often taken to reduce the problems associated with the signal processing of speech. This is particularly important for an automatic analysis to remove some of the above-mentioned level of sophistication in the measurement algorithms. Even so, the effects of coarticulation between speech sounds and of such voice source characteristics as creak and diplophonia can play havoc with the expectations built into signal processing algorithms. While researchers often do their best to ignore such phenomena, because they are difficult to accommodate or fall outside of the range of expected values, or can be dealt with individually in a small-scale analysis, they are perfectly valid in terms of their effect on the perception of speaker sex, and ought to be considered in any characterisation.

In summary, the speech behaviour of the population as a whole can be remarkably varied, and so any automatic analysis of a large number of speakers must combine flexibility and reliability for there to be any confidence in the results it produces.

For the analysis of the TIMIT database, only vowel phones were used as input data, which ensured the signal processing algorithms need be designed only to cope with voiced speech. The advantage of using voiced speech is that a person's frequency characteristics are represented in the frequency domain by more prominent spectral features. These features, representing the fundamental frequency and its harmonics and formant frequencies, are therefore easier to locate and measure. Further steps were taken to reduce the problems associated with unexpected voice source characteristics (averaging the value of a parameter along the length of a phone rather than taking a single value from the centre) and coarticulation effects from neighbouring speech sounds (only averaging values from the middle 50% of the phone), to improve the accuracy of the estimated parameter values for each speech slice.

Despite the precautions taken, a substantial amount of time was invested in evaluating the performance of the measurement algorithms. While the fundamental frequency and harmonic amplitude measurement algorithms proved to be very robust, problems were

still encountered as a result of the wide range of parameter values and oddly-phonated phones, requiring manual adjustment of the algorithms. The formant frequency estimator could not be relied upon to produce a consistently accurate output, and as a result, the formant frequency data proved to be inadequate for an analysis of speaker variability. The problems associated with measuring the formant frequencies of female speakers, and more particularly with defining a search space for the estimator which encompassed all possible values for the first three formants while excluding F_1 and the spectral component of F_0, proved to be insurmountable given the time available.

Achievements of the analysis procedure

The software to carry out this analysis was designed to be flexible. Thus it is possible to target any type of phone for analysis, simply by changing the identifier of the phone(s) to be searched for in the TIMIT transcription files, and to perform any type of signal processing on the phones.

The software enables the creation of two types of output database, one containing the frequency characteristics of the speech slices in the analysis, and the other containing the average frequency characteristics of the speakers for each of the phones analysed. This provides a wealth of detail regarding the behaviour of the characteristics over a large volume of speech data, and for a large number of speakers.

The algorithms designed to measure the fundamental frequencies and relative first harmonic amplitudes of the vowel-phones, and their implementations on computer, were shown to be extremely robust. While they were unable to produce accurate results for every, single speech segment, further complexity could be built into the algorithms to allow them to cope with more unusual input.

The output databases are an extremely powerful source of information for statistical analysis. For this research, simple analyses were made possible for the investigation of between- and within-speaker variability, of the effects of extralinguistic speakers attributes such as age and dialect, and of the influence of phonetic context. The design of the software which processes the output databases allows for more complex statistical analyses to be incorporated.

Appendix A

Analysis of the TIMIT Database using UNIX

This appendix describes the UNIX tools used in the analysis of the TIMIT data. The UNIX commands are defined below in Section A.1, with particular reference to how they were employed in the analysis of the speech data used for this study. This comprises a brief summary of the UNIX knowledge required to understand the software written for the data analysis procedure, which is described in Appendix B. Finally, Section A.2 provides some examples of how the UNIX tools were used to analyse the TIMIT database.

The TIMIT CD-ROM does not contain a database in the usual sense. Rather it is a collection of speech data files (and related files) stored within a directory hierarchy. Although the database was intended to aid research into acoustic-phonetic phenomena, little such work has been reported to date, one of the reasons being that the organisation of the data does not lend itself easily to exploitation by commercial database software (Keating *et al.* 1992:823). However, the way the database has been structured for computer-readability does lend itself to analysis by the pattern-matching tools available for use within UNIX shell scripts. The **phn** files containing the phonetic transcriptions of the sentences (see Figure 4.5 for an example of the layout of a **phn** file) are set up in a way that is particularly suited to exploitation by the pattern scanning and processing language **awk** and the pattern matching command **grep**. Specific phones can be targetted using **grep**, which can then pass the line containing the phone's identifier and sample numbers to **awk**, which is able to form a unique label for the phone using the path to the sentence file and the sample numbers from the **phn** file. This particular piece of processing forms part of a shell script called **find_vowels**, and is used to establish a series of files listing labels pointing to all the data items (i.e. the instances of the phones uttered by the TIMIT speakers) to be used in the analysis (see Appendix B.2 for a more detailed description).

A.1 A description of the UNIX tools used in the analysis

A. awk – pattern scanning and processing language
awk scans each line in its input file(s) for matches to a defined pattern(s), and performs action(s) on matched lines. It can also be used to perform actions on every line in a file, for files whose lines are all in the same format, e.g. the **phn** files. To **awk**, input consists of a series of lines, or records, terminated by a newline. Each record, is made up of fields separated by white space (i.e. a space or tab), with the first field denoted by **$1**,

the second field by $2, etc., and the entire line denoted by $0. **awk** statements consist of a pattern-matching statement and an action statement, in the format:

```
awk 'pattern {action}' filename
```

where `filename` is the name of the input file. When there is no pattern, **awk** processes every line in the file; when there is no action, the entire matching line is printed out.

A pattern can be a simple relational expression, or a Boolean combination of expressions, and can be applied to one or more fields or the entire line. For example, the pattern `$2 == "fm:"` matches lines whose second fields contain the string '`fm:`'; while `$2<220.6 && $4!="samples"` triggers a match only if the second field has a numerical value less than 220.6 *and* the fourth field does not consist of the string 'samples'. An action is a sequence of one or more statements surrounded by braces. If there is more than one action statement, they are separated by semi-colons, newlines or right braces. Actions tend to involve either outputting a matched line in an different format (e.g. to edit input data, or to form command strings), or performing mathematical operations involving the data on the line. For this study, use was made of the following actions:

- **print** – simple output statement, without formatting. E.g.:
  `{print $2, $1}`
  outputs the second field of a matched line followed by the first, separated by a space.

- **printf** – formatted output statement, equivalent to the printf statement in C. E.g.:
  `{printf "%s\t%3d\n", $3,$5}`
  outputs the string in the third field followed by a three digit integer from the fifth field, separated by a tab. Note the use of \n to force the next output onto a new line.

- **if** – decision statement. E.g.:
  `{if ($2<10) {print $0}}`
  outputs the whole of a matched line only if the number in the second field is less than 10.

- **split(s,a,c)** – splits the string **s** into **a[1]** to **a[n]**, around character **c**. E.g.:
  `{split($2, a, "t")}`
  Here, if the second field consists of the string 'potato', a[1]='po', a[2]='a', a[3]='o'.

- **FS** – sets the field separator, i.e. the character separating fields, where the default is white space. E.g.:
  `{FS=":"}`
  sets the field separator to be a colon.

- The mathematical operators +, −, ∗, /, = and the C operators ++, − −, +=, −=, ∗=, /= can be used in expressions. E.g.:
  `{total+=$1; n++; average=total/n; print average}`
  outputs the mean of the numbers in the first field of a file.

- There are two special action statements: `BEGIN{action}` for performing actions before the first input from a file is scanned; `END{action}` for performing actions once all the input lines have been processed.

B. grep – pattern-matching command

grep scans each line in its input file(s) for matches to a defined pattern(s). Its output

consists of the whole of a line containing a matching pattern. While it is much more limited than **awk**, in that it cannot process matching lines, it is much faster in its operation, and is therefore time-saving when used to search large databases. **grep** searches for a given string in any position on a line, and regardless of any surrounding characters. **grep** statements take the form:

```
grep "pattern" filename
```

where `filename` is the name of the input file, and `pattern` can consist of any number of characters, including wildcards and whitespace. The double quotes around the pattern are optional, but are necessary when the pattern includes whitespace.

C. Input and output redirection

In UNIX, input to and output from commands is, by default, from the standard input and standard output. Input from and output to data files is handled in the following way (note that `command` can be a UNIX command or a specially-written shell script):

`command < datafile`	cause `command` to take its standard input from `datafile`
`command > datafile`	cause `command` to send its standard output to `datafile`
`command >> datafile`	cause `command` to add its output to the end of `datafile`

Output from one command can also be used as the input to another command using a 'pipe', |. In the following illustration, the output from `command1` is 'piped' to the input of `command2`:

```
command1 | command2
```

D. Shell variables

Shells allow you to assign values (strings or numbers) to variables:

```
variable1=value1
```

Note that no spaces are permitted around the equal sign. Retrieving the values of variables is achieved by preceding the name of the variable with a dollar sign, $:

```
variable2=$variable1
```

Variable names can also be assigned multiple values:

```
variable1="value1 value2 value3"
```

E. echo

The **echo** command is used to send data to the standard output, or redirected to file. For example:

```
echo this is output from echo
```

will display the string 'this is output from echo' on the screen.

F. Commenting

Comment lines within shell scripts are preceded by a hash sign, **#**, to stop them being interpreted by UNIX.

G. Command continuation

For neatness when writing shell scripts, very long commands can be safely broken up into two or more lines using a backslash, \.

H. Looping

Looping in UNIX allows repetition of a list of commands with different variable values:

```
for variable1 in value1 value2 value3
do
  list of commands
done
```

where `variable1` is assigned the values `value1`, `value2` and `value3` in turn. The same effect can also be achieved more flexibly by combining the use of a shell variable:

```
variable1_type="value1 value2 value3"

for variable1 in $variable1_type
do
  list of commands
done
```

I. Decision making

The format of the **if** command is as follows:

```
if decision1
then
  commands1
elif decision2
then
  commands2
else
  commands3
fi
```

The **case** command performs the same operation in a different way. It matches a given string with a number of possibe patterns, executing the appropriate set of commands:

```
case pattern in
  pattern1) commands ;;
  pattern2) commands ;;
  ...
esac
```

J. Command arguments

Command arguments provide a way of assigning values to a script's variables without

having to edit the script. For instance, when a shell script, `script1`, is run with a list of values:

```
script1 x1 x2
```

the arguments x1 and x2 can be accessed within the script as $1 and $2. As an example, if $1 was the name of a file, and $2 was a string, then the following line within `script1`:

```
grep $2 $1
```

would search the file whose name was given in $1, and output any lines containing the string $2.

K. Command substitution
This is a way of assigning the value of the output of a command to a variable name:

```
variable1=`command1`
```

For example, `variable1=`grep $2 $1`` would assign the output of the previous example to `variable`.

A.2 Using UNIX to analyse the TIMIT database

There now follows a series of examples to illustrate how the various UNIX facilities outlined above were used in the shell scripts written for the analysis of the TIMIT data. Note, the words in SMALL CAPITALS refer to particular types of data, file and directory names and structures, and are defined in Appendix B.1.

A. Accessing data files in turn
This had two major uses. The first was to use the structures for the TIMIT and analysis data directories to analyse large numbers of files containing the same type of data. In the following example, from the shell script **analyse_vowels**, the variable `target_file` is assigned the pathname of each of the SLICE NAMES FILES in a particular SLICE NAMES DIRECTORY. Within the body of the **for** loop, the same operations can then be performed on each of the SLICE NAMES FILES in turn:

```
file1=$phoneme_names_all.dialect/$sex.dr*.names

for target_file in $file1
do
  ...
done
```

The second major use was to access different variables in turn. In the following example, the variable `phone` is assigned a TIMIT phone identifier from a list of phones:

```
phone_types="aa ae ao iy uw ux"
```

```
for phoneme in $phone_types
do
  ...
done
```

B. Nesting of for loops

In the following example, from **do_variable_means**, the shell script **variable_means** is run a total of 120 times, each time with different combinations of the identifiers for speaker sex, **acoustic-phonetic variable**, **analysis variable directory** name, and phone:

```
sex_types="f m"
apv_types="f0 H1-H2 @H1 @H2"
variable_dir="age colour dialect education height"
phone_types="aa ae ao"

for sex in $sex_types
do
  for apv in $apv_types
  do
    for variable_directory in $variable_dir
    do
      for phone in $phone_types
      do
        variable_means $variable_directory $sex $apv $phone
      done
    done
  done
done
```

C. Extracting parameter names

In the following example, the pathname of the target file is split up to reach the identifier of an ANALYSIS VARIABLE. Thus, if the pathname of the target file was given as:

```
target_file=/timitdata/timitdata_f/aa_names_all.dialect/f.dr4.names
```

then the following command would assign the string `dr4` to the variable `name1`:

```
name1=`echo $target_file | \
  awk '{FS="/"; split($5,a,"."); print a[2]}'`
```

D. Using awk for counting

In the following example, **awk** increments the counter `n` by one every time it encounters the string 'samples' in the first field of the file:

```
awk '$1 == "samples" {n++}END{print n}' $file2
```

Appendix B

The Database Analysis Procedure

This appendix describes the procedures used in the extraction, signal processing and statistical analysis of the digitised speech data on the TIMIT database. This involved the development of file and directory structures to organise the data, and of protocols for the naming of speech slices, files and directories; the design of a suite of software (UNIX shell scripts and C programs) to handle the processing of the data; and an evaluation of the signal processing algorithms used to measure the acoustic-phonetic features of the input speech.

Described below is an overview of the main stages involved in the analysis procedure. The stages are described fully in the seven parts to this appendix, an outline of which follows the overview.

An overview of the analysis procedure

The analysis carried out for this thesis can best be thought of as consisting of three main stages: the setting up of the structures for the analysis; the extraction and signal processing of the speech data on the database; and the statistical analysis of the acoustic-phonetic data. A preliminary stage can also be identified, consisting of the preparation of the input data. With the TIMIT database, this has already been carried out, consisting of the phonetic transcription of the speech (allowing specific phones to be targetted), and the provision of the speech data and transcriptions in a computer-readable form on a CD-ROM. The three main stages will now be discussed in more detail.

For an analysis of the scale embarked upon for this thesis, it was necessary to define structures to facilitate the automatic input (in the form of the raw speech data) and output (in the form of signal processed data and statistical analyses) of large volumes of data. Thus protocols were designed to ensure the consistent naming of speech segments, files and directories, and the consistent organisation of the data held within them. The pattern-matching tools available for use within UNIX shell scripts proved to be particularly suited to the data manipulation task, and the protocols were developed with them in mind.

A core data set was established from six vowel phone types, through a series of files of labels pointing to the actual speech data on the TIMIT CD-ROM. Thus there was no need to store the raw speech data before signal processing, alleviating potential problems with lack of computer memory. The core data set consisted of almost 16.000 phones, or speech slices. The data extraction and signal processing procedures were governed by a single shell script. Using the files of phone labels (known as SLICE NAMES FILES, i.e. the files containing the names of the speech slices), the raw data was extracted from the CD-ROM and passed through the signal processing programs. The flexibility of the software design

allowed for any phone to be selected for analysis. and for any form of signal processing to be performed on it. The signal processed output for each slice (consisting of data on fundamental frequency, relative first harmonic amplitude and formant frequencies) was stored such that individual results could be accessed by referencing the labels in the SLICE NAMES FILES. In other words, a new database was created consisting of the frequency characteristics of every analysed slice.

At this stage, exhaustive checking of the results was carried out to evaluate the performance of the various signal processing algorithms, and to establish confidence in the accuracy of the output. The algorithms for the measurement of fundamental frequency and harmonic amplitude difference were shown to be very robust and accurate. However, doubt was raised over the consistency of the formant frequency estimator's accuracy. The process of evaluation also served to highlight the inherent variability of speech, even though the speech data from the TIMIT database consisted of read sentences recorded in a noise-free and unpressured environment.

Finally, for the statistical analysis, a further database was created, consisting of the average frequency characteristics of the speakers. Together with the database of speech slice frequency characteristics, this allowed for the statistical analysis of both between- and within-speaker differences. A further dimension was added to this analysis by the use of information provided on the CD-ROM about speaker attributes. Thus it was possible to categorise speakers by age, height, dialect, ethnic group and educational background, and to analyse their frequency characteristics accordingly.

An outline of this appendix

The analysis procedure is described in the following sections. Section B.1 describes the naming protocols, i.e. the protocols for the naming of speech slices, files and directories and the definition of data structures. Section B.2 describes the directory structure for the organisation of data input and output, and the establishment of the files of labels pointing to the speech slices. Section B.3 describes how the speech slices were extracted from the sentence files on the CD-ROM. Section B.4 describes the signal processing of the speech slices, or more specifically, the implementation of the signal processing algorithms. Section B.5 describes how the frame-by-frame results from the signal processing are passed through a simple statistical analysis to produce the data on the frequency characteristics of each slice. Section B.6 describes the exhaustive checking of the slice statistics, to ensure the signal processing programs performed as intended, and to seek out any unusual results (which may come from, for example, unusual articulations of the vowels). The end result is a database of the frequency characteristics of all the analysed slices. Finally, Section B.7 describes the procedures for the full statistical analysis of the databases of slice and speaker frequency characteristics. The procedures allowed for analysis of within- and between speaker variability, and of the effects of speaker attributes and phonetic context.

B.1 The naming protocols

For the exploitation of large numbers of data, it is too time-consuming to approach an analysis in an ad-hoc manner. Structures must be set up beforehand to organise the data exploitation. The analysis procedure for this study required the extraction of the speech waveform data from the database, the processing of this data into the required output, and a statistical analysis of the processed data. For this study, the raw data to be analysed consisted of almost 16,000 segments of speech, from which were produced many more thousands of output data items. It was therefore essential that the analysis procedure be automated. For this reason, and to take full advantage of the UNIX pattern-matching tools, it was also essential to establish a protocol for the layout and naming of data structures, files and directories.

For the names of files and directories, the naming protocols used on the TIMIT CD-ROM were maintained where possible. The advantage of this was compatibility with TIMIT file and directory names, making data extraction easier. For example, the path names of the TIMIT speech data files can be easily derived from the format of the SLICE NAMES. Elsewhere, the protocols relied on assigning identifiers to the various parameters, incorporating them into file and directory names in a consistent format.

The layout of the data structures within the files was designed with the UNIX pattern-matching tools in mind. The files of processed data (i.e. the values of the acoustic-phonetic parameters for each slice and speaker) had to be transparent to these tools to facilitate the easy extraction of particular data items for statistical analysis.

What follows is essentially a glossary explaining the terms, names and formats used in the analysis, and is to be used as a reference for the rest of Chapter 4. It also defines the identifiers used in the UNIX shell scripts and C programs to refer to them, and are given in **bold**. The names are given in SMALL CAPITALS, a convention which will be used for the rest of this appendix. How the protocols are incorporated into the directory structures used to organise the data analysis is described in the next section (B.2). The rest of this section is organised into four parts, defining the types of variable, data, file and directory established to facilitate the analysis.

A. Variable types

There are two distinct types of variable used at different stages of the analysis procedure. The first are the acoustic-phonetic parameters of fundamental frequency, relative first harmonic amplitude and the formant frequencies, which were produced by the signal processing of the speech data. The second are the variables which were examined for their influence on the acoustic-phonetic parameters, and around which the statistical analysis and characterisation of the acoustic-phonetic parameter data was organised.

The identifiers for both sets of variables were used in the UNIX shell scripts in the formation of the file and directory names.

- ACOUSTIC-PHONETIC VARIABLES These are the parameters from the acoustic-phonetic domain which were examined for their correlation with speaker sex. There are three classes of variable, each with its own identifier (or ACOUSTIC-PHONETIC VARIABLE class identifier). The variable classes, together with the variables within each class, are:

Identifier	Variable class	Variables
f0	Fundamental frequency	F_0
hd	Harmonic amplitude difference	H_1–H_2, H_1, H_2
fm	Formant frequency	F_1, F_2, F_3

The variables each have their own identifier (or ACOUSTIC-PHONETIC VARIABLE identifier). The variables and their identifiers are:

Identifier	Variable (units of measurement)
f0	Fundamental frequency, F_0 (Hz)
H1-H2	Relative amplitude of first harmonic, H_1–H_2 (dB)
@H1	Amplitude of first harmonic, H_1 (dB)
@H2	Amplitude of second harmonic, H_2 (dB)
F1	Frequency of first formant, F_1 (Hz)
F2	Frequency of second formant, F_2 (Hz)
F3	Frequency of third formant, F_3 (Hz)

The identifiers for H_1 and H_2 are preceded by the symbol '@' to distinguish them from the identifier for H_1–H_2. This is to avoid confusion when using the pattern-matching tool **grep**.

- ANALYSIS VARIABLES These are the extralinguistic and linguistic features which were examined for their effects on the values of the ACOUSTIC-PHONETIC VARIABLES. The extralinguistic features examined in this study (speaker age, ethnic group, dialect region, educational level, height, and sex) were chosen because the information was available on the TIMIT CD-ROM. It would be a simple matter to extend the number of features used in the analysis were the information available on them. The linguistic features examined were those of phone and phonetic context. Only some of the variables required an identifier (or ANALYSIS VARIABLE identifier). Each variable is subdivided into a number of groups, which are listed below:

Identifier	Variable	Groups
age	Age (years)	20-29, 30-39, 40-49, 50-59, 60 and over.
ethgrp	Ethnic group	Black, white.
dialect	Dialect region	New England, Northern, North Midland, South Midland, Southern, New York City, Western and Army Brat.
education	Highest educational level	High school, associate degree, bachelor's degree, master's degree, Ph.D.
height	Height (feet and inches)	5'1" and under, 5'2" to 5'3", 5'4" to 5'5", 5'6" to 5'7", 5'8" to 5'9", 5'10" to 5'11", 6'0" to 6'1", 6'2" to 6'3", 6'4" to 6'5", 6'6" and over.
–	Sex	Female, male.
–	Phone analysed	/aa/, /ae/, /ao/, /iy/, /uw/, /ux/.
–	Phonetic context of analysed phone	/CVC/, where C are the phones immediately before and after V, the analysis phone

The identifiers for each group are:

Variable	Identifiers of groups
Age	**20_29, 30_39, 40_49, 50_59, 60_**
Dialect	**dr1, dr2, dr3, dr4, dr5, dr6, dr7, dr8**
Education	**AS, BS, MS, HS, PHD**
Ethnicity	**black, white**
Height	**_51, 52_53, 54_55, 56_57, 58_59, 510_511, 60_61, 62_63, 64_65, 66_**
Sex	**f, m**
Phone	**aa, ae, ao, iy, uw, ux**

B. Data types

These are basically identifiers which are used to refer to specific items of data. The SLICE NAMES refer to particular phones on the TIMIT CD-ROM. and the SPEAKER NAMES refer to particular speakers.

- SLICE NAME A unique identifier for a particular phone, describing its speaker, the sentence it came from, and position it occurred in the sentence. As well as being a label for every phone used in the analysis, the SLICE NAME also gives the path to the **adc** file (the speech waveform file) from which the phone originated, as well as the sample numbers denoting the start and finish of that phone within the file. The format for the SLICE NAME is a four part character string, where the first part identifies the speaker's dialect (**dr1** to **dr8**); the second part consists of the SPEAKER NAME; the third part identifies the sentence spoken; and the final part, derived from the TIMIT **phn** file (the phonetic transcription file), gives the position of the phone in the speech waveform file, the numbers corresponding to the start and finish sample numbers of the phone. For example:

  dr1.faks0.sa1.14078to16158 This identifies a phone from the sentence **sa1** spoken by speaker **faks0**. The path name to the **adc** file is **dr1/faks0/sa1/sa1.adc**, and the phone starts at sample number **14078** and ends at sample number **16158**.

- SPEAKER NAME A unique identifier for a particular speaker. The format is the same as on the TIMIT database, and consists of a five character string. The first letter of the string identifies the speaker's sex (**f** or **m**), the second to fourth letters represent the speaker's initials, and the number on the end is to distinguish between speakers with the same initials (**0** for the first person, **1** for the second, etc.). For example:

  faks0 – female speaker A.K.S.. the first (or only) speaker with those initials.

  fjdm2 – female speaker J.D.M., the third person with those initials.

C. File types

These are the categories of data file created at various stages in the analysis. For each category, there are descriptions of the format of the file's name and contents. Also included here are descriptions of specific files, PROBLEM_SLICES and SLICES_TOO_SMALL, used to store information relevant to the analysis.

- ACOUSTIC-PHONETIC VARIABLE DATA FILE Contains the full, unabridged analysis data for a particular ACOUSTIC-PHONETIC VARIABLE. The files are organised by sex and dialect for each variable. The format for the file names is as follows:

Filename	Contents
SEX.DLT.**f0_data**	Fundamental frequency data.
SEX.DLT.**hd_data**	Harmonic amplitude difference data.
SEX.DLT.**fm_data**	Formant frequency data.

where SEX = speaker sex (**f** or **m**); DLT = dialect region (**dr1** to **dr8**).

Each file contains the results of the signal processing of each of the speech slices listed in the corresponding SLICE NAMES FILE. For each slice, the output consists of the frame-by-frame results of the signal processing, followed by the slice's identification number and the frame numbers used to compute the slice statistics, followed by the slice statistics themselves. An example of the file format, for a slice analysed for fundamental frequency, follows (for examples of the file format for the harmonic amplitude difference and formant frequency analyses, see Figures B.6 and B.7 respectively):

```
155.34  1927  103
155.34  1315  103
152.38  1100  105
152.38   772  105
161.62   790   99
:1: samples 2 to 4
f0: 153.4 1.71 ; 152.4 to 155.3
```

The first five lines show the fundamental frequency data for each analysis frame, each line consisting of F_0 and the position and amplitude of the cepstral F_0 peak. The last but one line consists of an identification number (which links it to the slice's SLICE NAME in the corresponding SLICE NAMES FILE), followed by the numbers of the frames used to compute the slice statistics. The final line displays a simple statistical analysis of the slice, and consists of an ACOUSTIC-PHONETIC VARIABLE identifier, followed by the mean and standard deviation, and the range of data.

- PROBLEM_SLICES — This is the name of a file set up during the checking of the signal processing results (see Section B.6) to keep a track of reanalysed slices, rejected slices and values considered to be atypical for the ACOUSTIC-PHONETIC VARIABLE under consideration. Each ACOUSTIC-PHONETIC VARIABLE DATA DIRECTORY contains a PROBLEM_SLICES file. The file entries were written in a consistent format – this allowed for the subsequent listing of, for example, all slices with a high F_0, or all slices that had been reduced in size. An example of the file format, showing the main categories of data (for the fundamental frequency analysis of the /aa/ phones), is as follows:

```
dr4.fedw0                          Low f0 (160.4 166.0 228.8)

dr4.fedw0.si1084.81160to83848      * Slice reduced from 81160to86245
                                   - little periodicity 2nd half of
                                   original slice. Low f0 (160.4).
                                   Context: /iy v aa v/) 'EVOLVE';
                                   utterance-final

dr4.feeh0.si471.28360to29533       ** Weird periodicity - 1st half
                                   (messy) has f0 110Hz, 2nd half
                                   220Hz. Context: /ix z aa r/
                                   'reflexES ARE'; 2nd last syll.

dr5.ftbw0.sx265.23930to25984       * -u option reset to 6.25msec
                                   (otherwise picker switches to 2nd
                                   rahmonic in 2nd half). Unstable
                                   periods give low f0 peaks
```

```
dr7.fisb0                    High (291.0) and low f0 (170.0)
                             for some slices
```

This tells us that speaker **fedw0** had a relatively low F_0 for her /aa/ phones (first line). One of her phones, in an utterance-final syllable, was reduced in length because there was little periodicity in the second half of the phone (second line). One of speaker **feeh0**'s phones, in the word 'are', was removed from the analysis because the frequency of the acoustic waveform jumped from 110Hz in the first half to 220Hz in the second (third line). The F_0 analysis program, known as 'picker', had to be rerun on a phone uttered by **ftbw0** because an unstable waveform periodicity had resulted in low cepstral F_0 peaks (fourth line). Finally, **fisb0** exhibited both high and low fundamental frequencies in two of her phones (fifth line).

- SLICE MEANS FILE Contains a list of slice statistics for a particular ACOUSTIC-PHONETIC VARIABLE. The files are organised by sex and dialect. The format for the file names is as follows:

Filename	Contents
SEX.DLT.**f0_means**	F_0 statistics.
SEX.DLT.**H1-H2_means**	H_1–H_2 statistics.
SEX.DLT.**@H1_means**	H_1 statistics.
SEX.DLT.**@H2_means**	H_2 statistics.
SEX.DLT.**F1_means**	F_1 statistics.
SEX.DLT.**F2_means**	F_2 statistics.
SEX.DLT.**F3_means**	F_3 statistics.

where SEX = speaker sex (**f** or **m**); DLT = dialect region (**dr1** to **dr8**).

The means are culled directly from the corresponding ACOUSTIC-PHONETIC VARIABLE DATA FILES, and constitute a summary of the statistics for each slice, or more importantly the slice means. An example of the data format, for a slice analysed for fundamental frequency (c.f. the example for the ACOUSTIC-PHONETIC VARIABLE DATA FILE), is as follows:

```
f0: 153.4 1.71 ; 152.4 to 155.3
```

This tells us that this slice had a mean F_0 of 153.4Hz, with a standard deviation of 1.71Hz, and that the F_0 within the slice ranged from 152.4Hz to 155.3Hz.

- SLICE MEANS SUPER-FILE Contains a list of all the slice means from the analysis of a particular ACOUSTIC-PHONETIC VARIABLE. In other words, it contains all the SLICE MEANS FILES grouped into a single file, and is used in the computation of the SPEAKER MEANS. An example of the data format, for a slice analysed for fundamental frequency (c.f. the example for the SLICE MEANS FILE), is as follows:

```
:23: 153.4 1.71 ; 152.4 to 155.3
```

Note that the difference between the file formats of the super-file and the means files is that a numerical identifier replaces the ACOUSTIC-PHONETIC VARIABLE identifier. This identifier corresponds to the identification number in the appropriate SLICE NAMES SUPER-FILE.

- SLICE NAMES FILE Contains a list of the SLICE NAMES for a particular phone. The files are organised by sex and dialect. The format for the file names is as follows:

SEX.DLT.names

where SEX = speaker sex, **f** or **m**; DLT = dialect region, **dr1** to **dr8**.

The file format consists of an identification number (which matches the identification number of the analysed slice in the corresponding ACOUSTIC-PHONETIC VARIABLE DATA FILE), followed by the SLICE NAME, followed by the phonetic context in which the phone was spoken. An example of the data format is as follows:

```
:1: dr1.faks0.sa1.14078to16158 / hv ae dcl /
```

Thus this entry represents an /ae/ phone from speaker **faks0**'s realisation of the word 'had' in the sentence **sa1**.

- SLICE NAMES SUPER-FILE Contains a list of all the SLICE NAMES. In other words, it contains all the SLICE NAMES FILE grouped into a single file. The file format is the same as for SLICE NAMES FILES. The identification number corresponds to the identification number in the appropriate SLICE MEANS SUPER-FILE.

- SLICES_TOO_SMALL A list of all the slices for a particular phone which were considered to be too small for analysis. This entailed all slices of less than 1000 samples (62.5msec). The file format consists of the SLICE NAME followed by the number of samples in the slice for reference purposes. An example of the format is as follows:

```
dr1.faks0.sa1.14078to15033    955
```

i.e. this slice contains only (15033 − 14078 =) 955 samples.

- SPEAKER MEANS FILE Contains a list of the statistics for a particular speaker from the analysis of a particular ACOUSTIC-PHONETIC VARIABLE. It is formed from a statistical analysis of all of that speaker's SLICE MEANS. The format for the file names is as follows:

Filename	Contents
f0.SPK	F_0 means.
H1-H2.SPK	H_1–H_2 means.
@H1.SPK	H_1 means.
@H2.SPK	H_2 means.
F1.SPK	F_1 means.
F2.SPK	F_2 means.
F3.SPK	F_3 means.

where SPK is the SPEAKER NAME.

Each file contains the means for each type of vowel phone spoken by the speaker, followed by the mean for all of that speaker's phones. The format consists of the identifier of the analysed phone (except for **all**, which indicates all the speaker's slices), followed by the mean and standard deviation, the number of slices involved in the analysis, and finally the range of data. An example of the data format, for a speaker's H_1–H_2 means, is as follows:

```
aa    -10.0   1.68   ;  7   ; -12.4  to  -7.9
ae    -10.1   1.28   ;  7   ; -11.9  to  -8.2
ao    -10.4   0.43   ; 11   ; -11.0  to  -9.4
all   -10.2   1.11   ; 25   ; -12.4  to  -7.9
```

Thus the 11 /ao/ vowels spoken by this speaker had a mean H_1–H_2 of -10.4dB (s.d. 0.43dB), and the SLICE MEANS (i.e. the mean H_1–H_2 for each slice) ranged from -11.0dB to -9.4dB.

- SPEAKER NAMES FILE Contains a list of the names of the speakers comprising a particular ANALYSIS VARIABLE group. The format for the file names is as follows:

Filename	Group identifier
SEX.AGE.**names**	AGE = **20_29, 30_39, 40_49, 50_59, 60_**
SEX.CLR.**names**	CLR = **black, white**
SEX.DLT.**names**	DLT = **dr1** to **dr8**
SEX.EDU.**names**	EDU = **AS, BS, MS, HS, PHD**
SEX.HGT.**names**	HGT = **_51, 52_53, 54_55, 56_57, 58_59, 510_511, 60_61, 62_63, 64_65, 66_**

where SEX = speaker sex, **f** or **m**.

The file format consists of simply a list of SPEAKER NAMES. For example, the file **m.40_49.names** contains a list of the male speakers aged between 40 and 49 years.

D. Directory types

These are the categories of directory created to store the files produced by the analysis. For each category, the directory's contents are listed, together with a description of the format of the directory's name.

- ACOUSTIC-PHONETIC VARIABLE DATA DIRECTORY Contains the ACOUSTIC-PHONETIC VARIABLE DATA FILES and SLICE MEANS FILES for each phone. Each directory also contains a PROBLEM_SLICES file. The directories are organised by vowel phone. The format for the directory names is as follows:

Filename	Contents
VWL_**f0_all.dialect**	Fundamental frequency data.
VWL_**hd_all.dialect**	Harmonic amplitude difference data.
VWL_**fm_all.dialect**	Formant frequency data.

where VWL = **aa**, **ae**, **ao**, **iy**, **uw**, **ux** for the fundamental and formant frequency data, and VWL = **aa**, **ae**, **ao** for the harmonic amplitude difference data

- SLICE NAMES DIRECTORY Contain the SLICE NAMES FILES for each phone. Each directory also contains a SLICES_TOO_SMALL file. The directories are organised by vowel phone. The format for the directory names is as follows:

  VWL_**names_all.dialect**

  where VWL = **aa**, **ae**, **ao**, **iy**, **uw**, **ux**.

- SPEAKER NAMES DIRECTORY Contains the SPEAKER NAMES FILES for the ANALYSIS VARIABLES age, ethnic group, dialect region, educational background and height. It was not necessary to assign directories for the other variables. The ANALYSIS VARIABLE identifiers are used for the directory names, i.e. **age**, **ethgrp**, **dialect**, **education**, **height**.

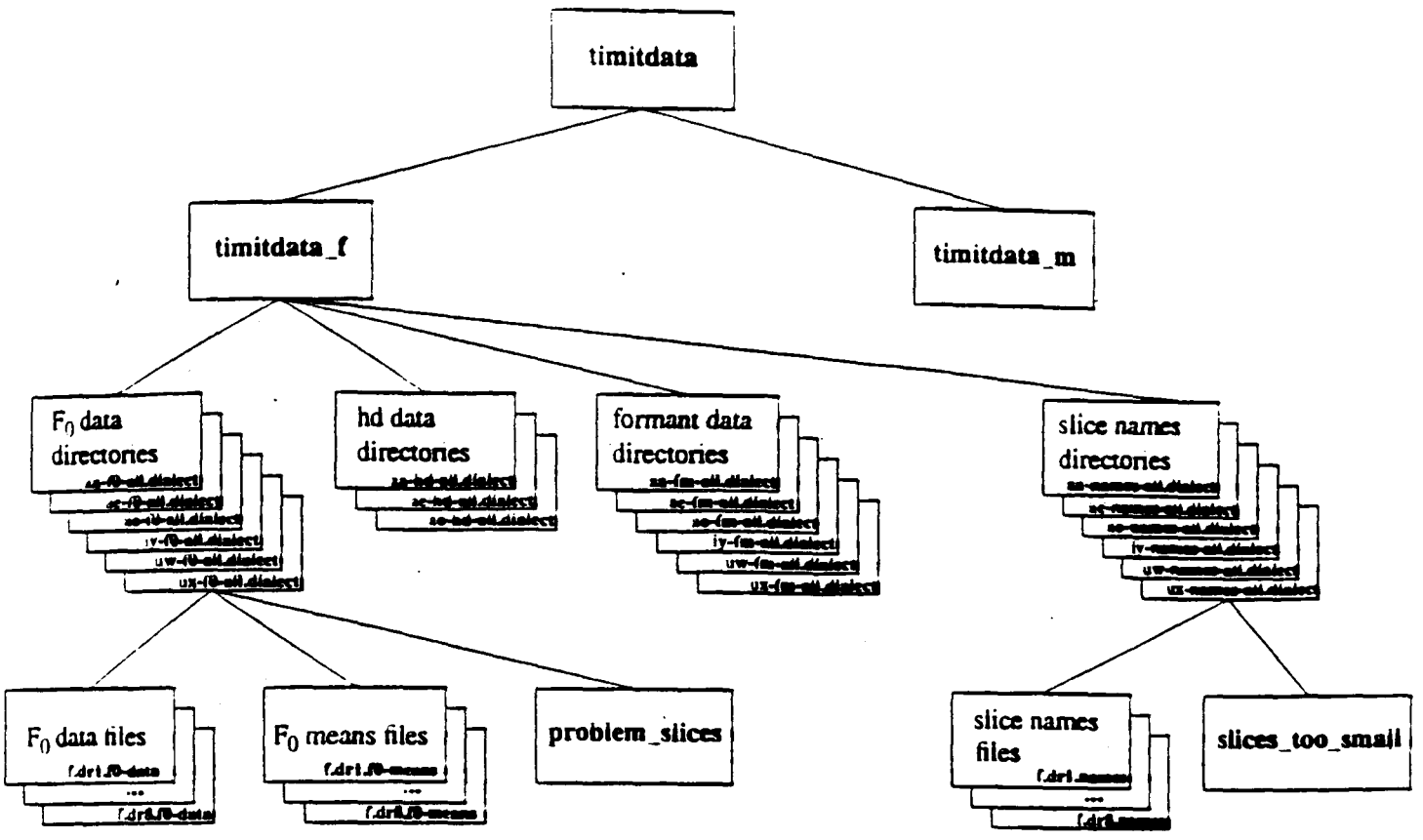


Figure B.1: The directory structure for the data analysis. Actual directory or file names are in bold.

B.2 Setting up the structures for the analysis

This section describes the pre-analysis work required to ensure the smooth running of the analysis procedure. First of all, this involved the design of directory structures to contain all the SLICE NAMES to be used as input data in the signal processing analysis, and to organise the storage of the results of the signal processing. Secondly, this involved the extraction of the SLICE NAMES from the TIMIT CD-ROM and their placement in the SLICE NAMES FILES.

The directory structure

The directory structure established for the automatic analysis of the speech slices is illustrated in Figure B.1. The directory structure is split by speaker sex into two halves: a directory for the analysis of the female speech called **timitdata_f**, and one for the male speech called **timitdata_m** (see row 2 of Figure B.1). Under each speaker sex directory are the SLICE NAMES DIRECTORIES (containing the names of the speech slices to be used as input data) and ACOUSTIC-PHONETIC VARIABLE DATA DIRECTORIES (where the output data, or results of the signal processing, is stored) – see row 3 of Figure B.1.

The ACOUSTIC-PHONETIC VARIABLE DATA DIRECTORIES contain all the signal processing data for the fundamental frequency, harmonic amplitude difference and formant frequency analyses. The results of the analyses for the three classes of ACOUSTIC-PHONETIC VARIABLE are stored by phone: thus there are six directories each for the fundamental and formant frequency data, and three for the harmonic amplitude difference data. The directory names are shown in row 3 of Figure B.1. Within each directory, the results of the signal processing are stored in the ACOUSTIC-PHONETIC VARIABLE DATA FILES and the SLICE MEANS FILES (see row 4 of Figure B.1). Both sets of files are divided by dialect,

yielding eight times two files per ACOUSTIC-PHONETIC VARIABLE DATA DIRECTORY. For example the female F_0 analysis data for the New York City dialect region is held in the file **f.dr6.f0_data**, and the SLICE MEANS are in **f.dr6.f0_means**. In addition, each directory contains a PROBLEM_SLICES file.

The SLICE NAMES DIRECTORIES contain the names of the slices corresponding to the analysis results stored in the ACOUSTIC-PHONETIC VARIABLE DATA DIRECTORIES. There is one directory for each of the six phones used in the analysis, and these are shown by name in row 3 of Figure B.1). Within each directory are the SLICE NAMES FILES (see row 4 of Figure B.1). These are divided by dialect, yielding eight files per SLICE NAMES DIRECTORY. In addition, each directory contains a SLICES_TOO_SMALL file.

Setting up the SLICE NAMES FILES

The establishment of the files containing the names of the speech slices to be used in the analysis is handled by the shell script **find_vowels**. This searches the TIMIT database for all instances of the specified phone types. It then forms the SLICE NAMES for each phone, and stores them by dialect in the SLICE NAMES DIRECTORIES.

To achieve this, the script makes use of the **phn** (phonetic transcription) files on the database (see Section 4.1.3). These files supply the position of each phone within a given sentence. Each line in the file consists of a phone's start and finish sample numbers and the TIMIT identifier for that phone (see Figure 4.5). It is therefore a simple matter to locate the target phones (and, through the sample numbers, their position in its **adc** (speech waveform) file) using the pattern-matching tool **grep**. The phone's SLICE NAME can then be formed using **awk**, thus:

```
# Match the phones in the transcription file with the target phone.
grep $phone $dialect/$spk_name/$sentence/$sentence.phn | \
  awk '{printf '$dialect'.'$spk_name'.'$sentence'.%dto%d\n", $1, $2}'
```

where $phone, $dialect, $spk_name and $sentence are the TIMIT identifiers for phone, dialect region, SPEAKER NAME and sentence respectively.

The task is made more difficult, however, because the format of the SLICE NAMES FILES also requires the phone's phonetic context. Using **awk** instead of **grep** to locate the target phones, it can remember the phones on previous lines in the **phn** file. Thus when **awk** locates a target phone it must wait until the following line (containing the next phone in the sentence) before outputting the SLICE NAME and phonetic context. This is coded as follows:

```
awk '{
  # If the phone on the previous line matches the target phone,
  # output the SLICE NAME.
  if(old == "'$phone'")
    {printf "'$dialect'.'$spk_name'.'$sentence'.%dto%d / %s \
             %s %s /\n", start, finish, old_old, old, $3}
  # Remember the current and previous phones.
  old_old=old; old=$3
  # Remember the start and finish samples of the current phone.
  start=$1; finish=$2
}' $dialect/$spk_name/$sentence/$sentence.phn
```

where $phone, $dialect, $spk_name and $sentence are the TIMIT identifiers for phone, dialect region, SPEAKER NAME and sentence respectively.

For example, for the two /iy/ phones in the transcription file reproduced in Figure 4.5, the output of **find_vowels** would be:

```
dr1.fdac1.si2104.14560to16120 / hh iy th /
dr1.fdac1.si2104.47480to51387 / s iy pau /
```

This would be stored in the SLICE NAMES FILE called:

timitdata/timitdata_f/iy_names_all.dialect/f.dr1.names

Obviously, using **awk** to locate the phones greatly increases the cost in computer processing time, but this is of no great importance as the operation to form the SLICE NAMES FILES need only be performed once.

The script **find_vowels** also constructs the file SLICES_TOO_SMALL, whereby slices consisting of less than one thousand samples are removed from the analysis. The procedure for this uses **awk** to split apart each SLICE NAME, retrieving the slice's sample numbers and hence the slice's length. In the following section of code from **find_vowels**, the comments refer to the example SLICE NAME **dr1.faks0.sa1.14078to15058**:

```
awk '{
  # Split the SLICE NAME into fields divided by '.'.
  # Thus a[4]=14078to15058.
  split($1, a, ".")
  # Split a[4] around the 't'. Thus b[1]=14078.
  split(a[4], b, "t")
  # Split a[4] around the 'o'. Thus c[2]=15058.
  split(a[4], c, "o")
  # Compute the difference between the sample numbers.
  diff = c[2] - b[1]
  # If slice length < 1000, output SLICE NAME to
  # the file slices_too_small.
  if (diff < 1000)
    {printf "%s   %d\n", $1, diff}
}' $names_file >> slices_too_small
```

where $names_file is the path to the target SLICE NAMES FILE.

The final action of **find_vowels** is to attach an identifier to each entry in the SLICE NAMES FILES. This is particularly important as it allows cross-referencing between a SLICE NAME and the results of the signal processing of the slice held in the ACOUSTIC-PHONETIC VARIABLE DATA FILES and SLICE MEANS FILES. In order to make the identifier transparent to the pattern-matching tools, the identifier consists of a number surrounded by colons, e.g. ':23:'. Using this convention, ordinary data values cannot be mistaken for identifiers. The numbering is achieved thus:

```
awk '{
  n++
  printf ":%d: %s\n", n, $0
```

```
}' $names_file >> $names_file.tmp

# Replace the unnumbered SLICE NAMES FILE with the numbered version.
mv $names_file.tmp $names_file
```

where $names_file is the path to the target SLICE NAMES FILE, and $names_file.tmp is a temporary file.

B.3 Extraction of the slices from the TIMIT CD-ROM

This section describes the extraction of the raw speech data from the TIMIT CD-ROM. Note, this does not involve any storage of the raw data, rather it is sent directly from the **adc** files to the programs which carry out the signal processing. Only the results of the signal processing are stored for further analysis. The shell script **analyse_vowels** carries out the extraction (as well as handling the signal processing of the speech data, described in Section B.4 below), using the SLICE NAMES held in the SLICE NAMES FILES to tell it which parts of the **adc** files are required. The final output from **analyse_vowels** is the ACOUSTIC-PHONETIC VARIABLE DATA FILES.

Each sentence on the database is written in digitally-sampled form in an **adc** file (see Section 4.1.3 for more details of the TIMIT files). The phone boundaries (the start and finish sample numbers delimiting the phone in the **adc** file) are extracted from the SLICE NAMES, and are used to cut out the appropriate portion of the speech signal from the **adc** file. The first job of **analyse_vowels** is to set up a temporary file of commands to perform this operation. The raw data representing the phone is then passed to the signal processing programs).

The file of commands to extract the speech data is set up using a procedure very similar to the one for checking the length of phones (see Section B.2 above). Thus:

```
awk '{
  # Split the SLICE NAME into fields divided by '.'.
  split($1, a, ".")
  # Split a[4] around the 't'.
  split(a[4], b, "t")
  # Split a[4] around the 'o'.
  split(a[4], c, "o")
  # Output the data extraction commands for the target phone.
  {printf "remove-timit-header < %s/%s/%s/%s.adc | swab | itt \
          | chop -s%s -f%s | '$operation'\n", \
     a[1], a[2], a[3], a[3], b[1], c[2]}
}' $names_file >> temporary_file
```

where $operation is the command to run the specified signal processing program; $names_file is the path to the target SLICE NAMES FILE.

This procedure is carried out for every target SLICE NAME. For example, for the SLICE NAME **dr1.faks0.sa1.14078to16158**, the command to remove the appropriate speech data from the **adc** file would be:

```
remove-timit-header < dr1/faks0/sa1/sa1.adc | swab | itt | \
  chop -s14078 -f16158 | signal_processing_program
```

where, the pathname **dr1/faks0/sa1/sa1.adc** is the path to the waveform (of the sentence **sa1** spoken by **faks0**) on the TIMIT database; **signal_processing_program** is the name of the program used to carry out either the fundamental frequency, harmonic amplitude difference or formant frequency analysis.

The C programs **remove-timit-header**, **swab** and **itt**[1] combine to convert the speech

[1]These programs were implemented at the Department of Computer Science, University of Sheffield by Dr. Martin Cooke.

data from the TIMIT data storage format into ASCII (see Section 4.1.3 for an explanation of the data formats). They pass the entire sentence waveform in ASCII format to the C program **chop**[2], which extracts the target phone from the **adc** file. Finally, this speech slice is passed to the appropriate signal processing program.

[2]This program was implemented at the Department of Computer Science, University of Sheffield by the author.

B.4 Signal processing of the slices

This section deals with the implementation of the signal processing algorithms used to analyse the input speech data. The software to perform the analysis was written in C. The analyses are run by the shell script **analyse_vowels**, as described at the end of the last section (B.3), whereby the speech slices are passed through the signal processing program relevant to the specified ACOUSTIC-PHONETIC VARIABLE. The analysis of fundamental frequency is achieved by performing a cepstrum on the speech data and locating the cepstral F_0 peak; the harmonic amplitude difference analysis consists of performing an FFT on the speech data and locating the first and second harmonics; the formant frequency analysis is achieved using the CSTR formant frequency estimator. Evaluations of each algorithm's performance are in Section B.6.

The signal processing of each class of ACOUSTIC PHONETIC VARIABLE is carried out by applying a succession of windows[3] along the length of a speech slice and performing a frequency analysis on each windowed portion of speech. Thus in effect, the values of the ACOUSTIC PHONETIC VARIABLES are sampled along the length of the slice, enabling an average of the intra-slice behaviour of each variable to be obtained (see Section 4.1.1 for a more detailed explanation of why this is necessary). For each analysis frame (i.e. the section of the speech slice defined by the application of the window), the results are output to the ACOUSTIC PHONETIC VARIABLE DATA FILES: thus each line of output for a slice contains the results of the processing for each frame.

A. Fundamental frequency analysis

Some issues in the application of cepstral analysis

As explained in Section 4.1.2, cepstra are computationally very expensive to perform, requiring two Fast Fourier Transforms (FFTs). However, it was possible to take some short cuts to reduce the computation time. Theoretical thoroughness requires the computation of the complex cepstrum rather than the cepstrum, in which the imaginery parts of the complex cepstrum contain phase information. As this study had no use for the phase information, only the cepstrum need be computed. Using a similar argument, we may replace the inverse FFT operation in the cepstrum calculation by a simple FFT.

The analysis window for the computation of the cepstrum must be neither too short nor too long. If the window is too short, the windowed portion of speech may contain insufficient information about its periodicity to produce a strong cepstral F_0 peak. For an F_0 peak of sufficient amplitude, at least two clearly defined glottal periods are required, bearing in mind the tapering caused by the windowing function (Rabiner & Schafer 1978). If the window is too long, too much variant speech information may be included. The cepstrum relies upon steady-state phonation to produce a strong F_0 peak in the quefrency domain. If the fundamental frequency is changing rapidly or jumps from one value to another – due to, for example, the transition from one phone or syllable to another – the amplitude of the F_0 peak can be severely reduced. The measures taken to limit the occurrence of such a situation were to use segments of speech which did not include cross-phone boundaries (i.e. consisted solely of a phone), to compute a mean value of F_0 for each slice; and to compute the mean from values obtained from the middle portion of the slice. For this latter measure, the analysis frames representing the first and last quarters of the slice were not included in the computation of the mean, effectively excluding much

[3]The Hamming windowing function was used to window the data. This function is most often recommended for the frequency analysis of speech data.

of the between-phone transitional information.

The length of the sampling window was set at 64msec (or 1024 samples, at the sampling rate used for the TIMIT data of 16000 samples/second), which for the average female F_0 of 200Hz encompasses 12.5 periods, or 7.5 periods for the male average of 120Hz. This may seem rather large, but the constraints of the FFT program required the number of data points in an analysis frame to be a power of 2. The next window size down was 32msec (512 samples), and it was feared that for some speakers this would prove to be too small. While it would be possible for the FFT program to be given variable window sizes, based on expectations of the speaker's F_0 or on previously computed values, the initial trials of the program did not indicate a need for this. Perhaps more importantly, the large window size smears over irregularities in the speaker's phonation caused by vocal perturbation (e.g. jitter, shimmer, diplophonia) or vocal pathology. A window size encompassing only two or three periods of the speech waveform results in the cepstral analysis being prone both to depressed F_0 peaks from a lack of periodicity in the waveform portion covered by the analysis window, and to 'phantom' F_0 peaks caused by, for example, situations where every second period in the waveform has a reduced amplitude. Trial analyses showed a 1024 sample window size to be a consistently accurate descriptor of both the vowel phone's mean F_0 and its internal F_0 dynamics.

Rabiner & Schafer (1978) listed three criteria for the optimal use of cepstral analysis, which are reproduced below in **bold**, together with comments related to the design of the cepstral peak-picking algorithm developed for this analysis:

- **Compute the cepstrum every 10-20msec, as the excitation parameters change relatively slowly in normal speech.** For this study the analysis window was shifted by 8msec (64 samples) after every computation of the cepstrum.

- **Search for the peak in the vicinity of the expected F_0 period.** Thus once a phone's F_0 has been identified in the first few analysis frames, it is generally valid to assume the F_0 of subsequent frames will be similar. The search can therefore be simplified by defining a search interval centred around the previously computed position of the cepstral F_0 peak. However, when seeking to define the initial search interval (i.e. to establish the fundamental frequency at the beginning of a slice), while the differences in average F_0 for each speaker sex indicate the use of different search intervals for female and male speakers, in practice the intonation patterns even within read sentences mean a person's SFF can encompass a huge range of values. For example, in speaker **frll0**'s production of the sentence **sa1**, between the first and second syllables the F_0 rose from 276Hz to 345Hz, and fell to 274Hz by the fourth syllable. At the end of the same sentence, between the ninth, tenth and twelfth syllables, the F_0 dropped from 250Hz to 193Hz, and then rose to 220Hz. Thus in just 3.7 seconds, this speaker's fundamental frequency ranged over more than 150Hz. One further limit on the range of the initial search interval is that the lower end must be clear of the large, low quefrency cepstral peaks caused by the vocal tract resonances. While Noll (1967) recommended an initial interval of 1-15msec[4]. However the lower end of the interval was too close to the cepstral vocal tract components, and so the search interval used here was the one recommended by Rabiner & Schafer (1978), namely 3-20msec (333-50Hz). This seemed a reasonable compromise between trying to avoid the vocal tract components and allowing for particularly high female fundamentals. Note that where high F_0s were expected,

[4]Note these are units of quefrency and represent the period of the fundamental. The search interval is equivalent to a frequency range of 1000–67Hz

the results were checked by hand. Furthermore, setting the upper limit to 20msec allowed for particularly low male fundamentals.

- **Check if the F_0 peak located exceeds some preset threshold.** For the analysis of clean vowel phones, a strong cepstral peak can be almost guaranteed. If care is taken to ensure the search interval minimum does not include the cepstral vocal tract components, the threshold can be set so that only the cepstral F_0 peak and the peaks of its rahmonics have sufficient amplitude to exceed it. Obviously this greatly simplifies the search.

The algorithm for the automatic tracking of the cepstral F_0 peaks is based on one designed by Noll (1967) for the more complex task of voiced-unvoiced segment detection. It is basically a peak-picking algorithm, with the addition of checks to ensure the correct peak has been located. As the algorithm was not required to cope with unvoiced segments, which would not produce an F_0 peak in the cepstral domain, it could be simplified to a search for the maximum peak within the search interval. Moreover, as the data examined for this study consisted of vowel phones uttered in a relatively noise-free environment, apart from the extreme edges of phones, where an adjecent phone can greatly influence the steady-state periodicity thereby depressing the F_0 peak, one could expecet the cepstral analysis to produce a strong F_0 peak. Thus an assumption could safely be built into the algorithm which assumed the presence of an F_0 peak in a given cepstrum. Furthermore, the tracking element of the algorithm is made fairly easy due to the relatively steady frame-by-frame fundamental frequency within phones. The algorithm can therefore begin its search of the next analysis frame at the approximate location of the previous frame's F_0 peak.

Description of the cepstral peak-picking algorithm

The following description of the cepstral peak-picking algorithm is made with reference to the flow diagram of the algorithm in Figure B.2 and the accompanying description of the flow diagram's boxes below. The algorithm proceeds in two stages, locating first the F_0 peak in the first analysis frame of a speech slice (see boxes 1-9), and then tracking the F_0 peaks through the analysis covering the rest of the slice (see boxes 10-17).

1. Read in the first frame.
2. Compute the cepstrum of the first frame.
3. Locate the maximum peak in the entire search interval.
4. If the search interval includes a quefrency of half the quefrency of the maximum peak, then the maximum peak may be the second rahmonic rather than the cepstral F_0 peak – look for peaks at half the quefrency of the maximum peak; else the quefrency value falls outside the search interval – assume the maximum peak is the cepstral F_0 peak and read in the next frame.
5. Set a temporary search interval of $\pm$0.5msec around half the quefrency of the maximum peak.
6. Locate the maximum peak in the temporary search interval.
7. Set a temporary threshold at half the original threshold value to make the maximum peak easier to locate.

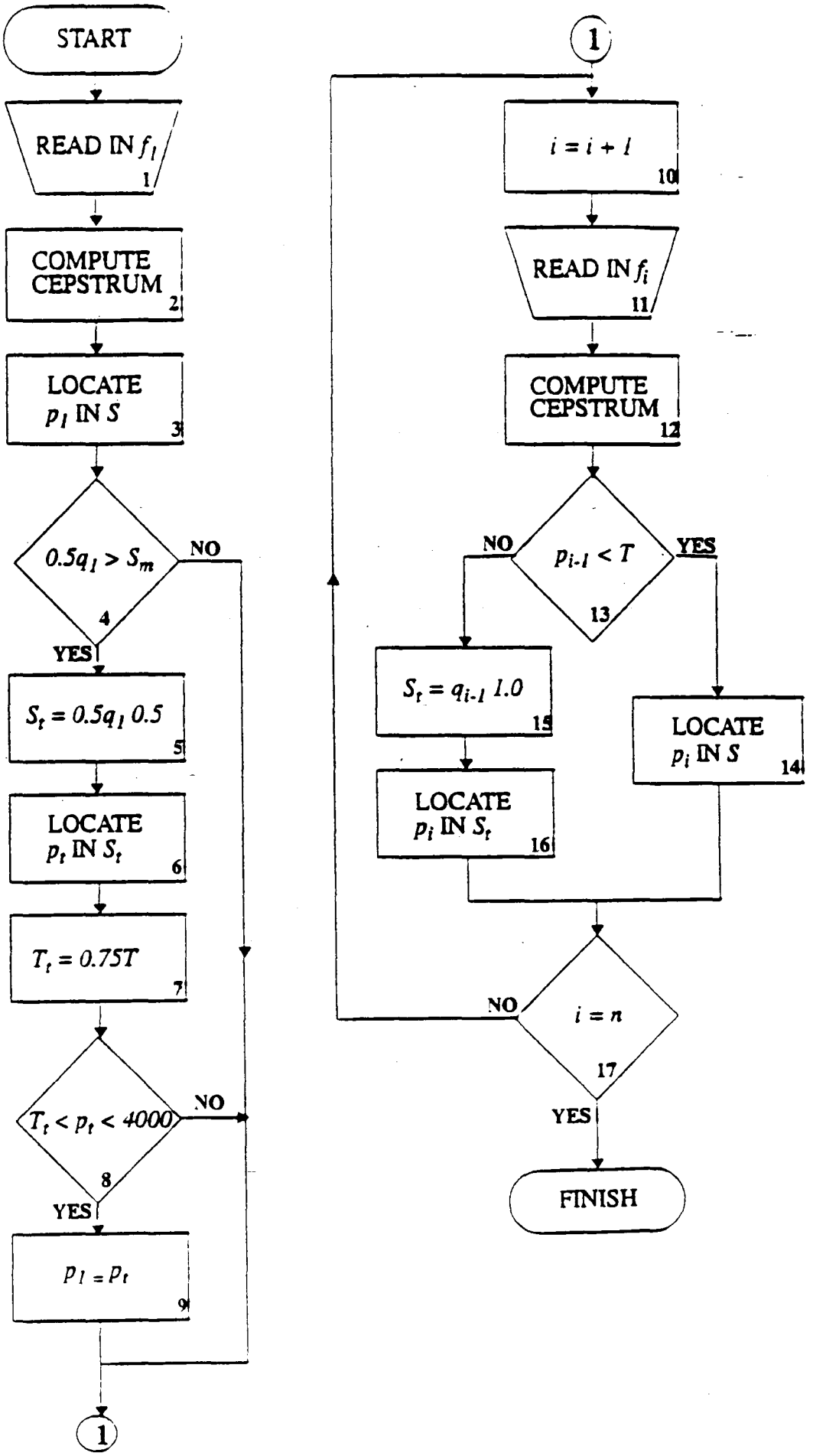


Figure B.2: Flow diagram of the cepstral peak-picking algorithm. Key to symbols: f_i is the i^{th} frame; i is the frame counter; n is the total number of frames; p_i is the maximum peak in the i^{th} frame; p_t is the maximum peak in the temporary search interval; q_i is the quefrency of p_i; S is the preset search interval; S_m is the search interval minimum; S_t is the temporary search interval; T is the preset threshold; T_t is the temporary threshold.

8. If the maximum peak in the temporary search interval exceeds the temporary threshold, but is less than 4000 (a value intended to avoid confusion with the high-amplitude cepstral vocal tract components), assume this maximum peak is the cepstral F_0 peak; else assume the original maximum peak is the cepstral F_0 peak.

9. Set the maximum peak for frame 1 to be the maximum peak from the temporary search interval.

10. Increment the frame counter.

11. Read in the next frame.

12. Compute this frame's cepstrum.

13. If the maximum peak from the previous frame exceeds the threshold, search for this frame's F_0 peak in the same place; else assume it may not have been an F_0 peak and search the entire search interval.

14. Locate the maximum peak in the entire search interval.

15. Set a temporary search interval of $\pm$1.0msec around the quefrency of the previous frame's maximum peak.

16. Locate the maximum peak in the temporary search interval.

17. If this is not the last frame, read in the next frame; else all the frames have been searched.

From the early trials of the algorithm it became clear that in the majority of cases, if the F_0 peak in the cepstrum of the first analysis frame was located accurately, then it was relatively easy to track the peaks in the remaining frames. Thus the first frame is analysed separately using an initial search interval of 3-20msec (see boxes 1-9 in Figure B.2). One other reason for this is that the preceding phone will tend to depress the F_0 peak's amplitude, sometimes causing the algorithm to pick out the first frame's second rahmonic as a legitimate F_0 peak. The algorithm locates first the maximum peak in the preset search interval (box 3). The algorithm then looks for a depressed peak at half the quefrency of this maximum peak (boxes 4-9), and a check is carried out to ensure this half quefrency value is within the search interval, to avoid confusion with the cepstral vocal tract components (box 4). If the peak at this half quefrency value exceeds a reduced threshold (box 8), then this peak is accepted as the first frame's F_0 peak (box 9).

The algorithm then tracks the rest of the peaks in the following way (boxes 10-17). It assumes the F_0 peak of the current frame will be at the approximate quefrency of the previous frame's F_0 peak, allowing for any between-frame change in fundamental frequency. Thus if the previous frame's F_0 peak exceeded the preset threshold (box 13), the maximum peak is located in a 2msec interval centred on the location of the previous frame's F_0 peak (boxes 15-16); and if the threshold was not exceeded, the previous frame's highest peak may have been a second rahmonic, and the entire search interval is investigated (box 14). It may seem worthwhile including many more checks to ensure only legitimate F_0 peaks are accepted, but in practice a combination of the strong cepstral peaks within the boundaries of vowel phones and the relatively large analysis window smoothing over irregularities in phonation render this by and large unnecessary.

The entire procedure is then repeated for all the phone's analysis frames. The frame-by-frame output for each slice consists of the fundamental frequency, the amplitude of

cepstral F_0 peak, and the sample number of the cepstral F_0 peak (related to the peak's quefrency), an example of which can be seen in Figure B.5.

B. Harmonic amplitude difference analysis

Some issues in the location of the harmonic peaks

The algorithm which locates the first two harmonics consists of a peak-picking strategy in the frequency domain. The FFT software used to compute the spectra of the input speech is the same as is used in the fundamental frequency analysis. The window size and shift are also the same as for the fundamental frequency analysis, and were chosen for the same reasons.

The algorithm for the automatic tracking of the first two harmonics uses a different peak-picking philosophy to that of the cepstral peak-picker. Whereas the cepstral peak-picker was limited to locating only a single prominent peak in the quefrency domain, the harmonic structure revealed by a wide-band spectrogram is represented by a succession of prominent peaks, necessitating a different search strategy. A threshold is set to limit the inclusion of spurious peaks during the search for the harmonic peaks. The amplitude value represented by the threshold was arrived at by a process of trial and error, and is such that only the (in general, prominent) harmonics are of sufficient amplitude to exceed it. The threshold was sex-dependent, reflecting the generally higher harmonic amplitudes of the female speakers. Thus, the threshold was set at 80dB for female speech data, and 75dB for male speech data.

The peak of the first harmonic, being the first prominent feature to be encountered in a spectrum, is located by advancing up its leading slope sample-by-sample, with the peak being reached when the slope starts falling. The harmonic peaks from the TIMIT vowel phone data tended to be clean as well as prominent, such that each successive sample was higher than the last until the apex was reached. Spurious peaks located away from the harmonic peak were almost always below the threshold, and were therefore rejected. The search for the second harmonic begins by leaping over the trailing edge of the first harmonic, and onto the leading edge of the second harmonic. Because the harmonics are multiples of the fundamental frequency, the second harmonic will lie at approximately twice the frequency of the first. However, to take into account the inaccuracies inherent in sampled data, the search is started at approximately two-thirds the distance between the first harmonic and the projected location of the second. The peak-picking then proceeds in the same way as for the first harmonic.

Description of the harmonic peak-picking algorithm

The following description of the harmonic peak-picking algorithm is made with reference to the flow diagram of the algorithm in Figure B.3 and the accompanying description of the flow diagram's boxes below. The algorithm proceeds in two stages following the computation of the frequency spectrum, finding first the amplitude of the first harmonic, H_1 (boxes 5-8), and then secondly the amplitude of the second harmonic, H_2 (boxes 9-12).

1. Set the frame counter to 0.
2. Increment the frame counter.
3. Read in the next frame.
4. Compute the log amplitude spectrum of the first frame.

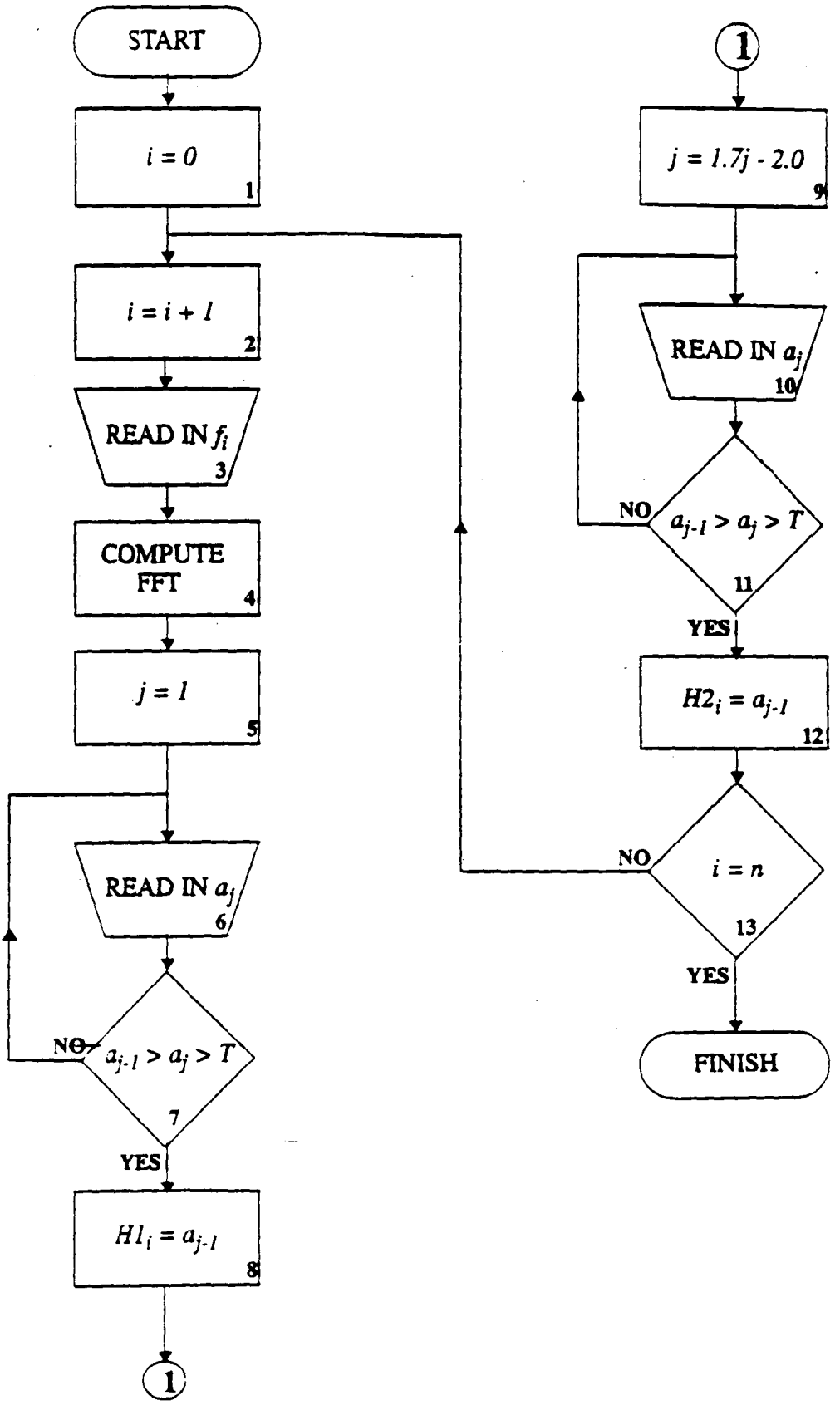


Figure B.3: Flow diagram of the harmonic peak-picking algorithm. Key to symbols: a_j is the amplitude of the i^{th} sample; $H1_i$ is the amplitude of the first harmonic in the i^{th} frame; $H2_i$ is the amplitude of the second harmonic in the i^{th} frame; f_i is the i^{th} frame; i is the frame counter; j is the sample counter; n is the total number of frames; T is the preset threshold.

5. Set the sample counter to 1.

6. Read in the next sample in the frame.

7. If the value of the sample is greater than the threshold, but less than the previous sample, then the previous sample represents the peak of the first harmonic; else read in the next sample.

8. Set H_1 for this frame to be the value of the previous sample.

9. Set the sample counter to jump over the trailing edge of the first harmonic's peak.

10. Read in the next sample in the frame.

11. If the value of the sample is greater than the threshold, but less than the previous sample, then the previous sample represents the peak of the second harmonic; else read in the next sample.

12. Set H_2 for this frame to be the value of the previous sample.

13. If this is not the last frame, read in the next frame; else all the frames have been searched.

To locate the first harmonic peak, the algorithm tests whether the amplitude of the current sample is less than the amplitude of the previous sample, whilst exceeding the threshold (box 7 in Figure B.3). If this is not the case, it reads in the next sample (box 6); but if it is the case, then it assumes it now on the downward slope of the harmonic peak, and sets the amplitude of the previous sample to be H_1 (box 8).

For the location of the second harmonic peak, the sample counter is set to start on the leading slope of the second harmonic (box 9). Again, the exact value for the beginning of the search for the second harmonic's peak was arrived at by trial and error. H_2 is then arrived at in the same manner as for H_1 (boxes 10-12).

The entire procedure is then repeated for all the phone's analysis frames. The frame-by-frame output for each slice consists of the amplitudes, sample numbers (related to the peak's frequency) and approximate frequencies of the two harmonic, and the difference between the amplitudes, H_1–H_2. An example of the output can be seen in Figure B.6.

C. Formant frequency analysis

The formant frequency analysis was performed using the CSTR formant frequency estimator. The principles behind the estimator were discussed in Section 4.1.2. The following description of the algorithm used to run the estimator is made with reference to the schematic view of the algorithm in Figure B.4 and the accompanying description of the Figure's boxes below.

1. Read in the next frame of input speech data from the TIMIT CD-ROM.

2. Convert the VAX binary format of the TIMIT data to audlab VOX format.

3. Compute the FFT of the data using the estimator's FFT program.

4. Run the formant frequency estimator.

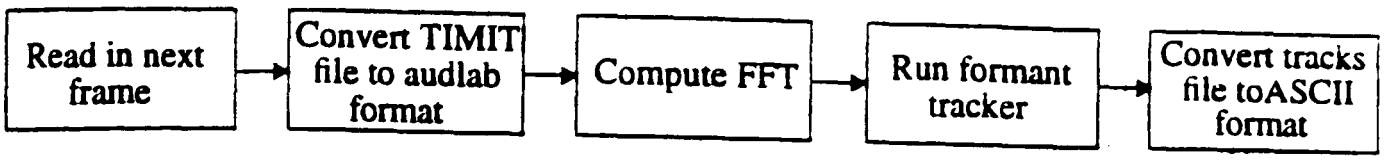


Figure B.4: Schematic view of the formant frequency tracking procedure.

5. Convert the audlab VOX format of the estimator's output to ASCII.

The estimator software uses a different data storage format to both the TIMIT database and the shell scripts, thus the passage of data between the various parts of the analysis procedure required transformations of the data format. Thus when the data is read off the TIMIT CD-ROM in VAX binary format (box 1), it must first be converted to the audlab VOX format required by the estimator (box 2). With the input speech data in the correct format, an FFT (the software for which is part of the estimator package) is performed (box 3), followed by the formant frequency estimation (box 4). The output of the estimator is finally converted from audlab VOX to ASCII format siutable for handling by the UNIX shell scripts[5].

The entire procedure is then repeated for all the phone's analysis frames. The frame-by-frame output for each slice consists of the frequencies of F_1, F_2 and F_3. An example of the output can be seen in Figure B.7. Note that the estimator also computes the energies and bandwidths of the first three formants, the total energy of the first three formants, and the overall energy in the spectrum. However, there was no way of evaluating the estimator's output for these measures, and so they were not used.

[5]The C program to carry out the conversion was implemented at the Department of Computer Science, University of Sheffield by Dave Abberley.

B.5 Performing the slice statistics

This section describes the production of the slice statistics as part of the output to the ACOUSTIC-PHONETIC VARIABLE DATAFILES. As with the signal processing analyses, the shell script **analyse_vowels** governs the progress of the data, although the actual statistical operations are carried out by programs written in C.

For each speech slice, the output from the signal processing programs, described in Section B.4 above, consists of the frame-by-frame results of the signal processing analysis. For example, the analysis of the fundamental frequency of the slice **dr3.fsjw0.si1333.30220to31916** produces the following output, the first number in each row representing the F_0 at successive points along the length of the slice. Thus during the production of this vowel sound (/aa/), the F_0 of speaker **fsjw0** fell from 203Hz to 184Hz:

```
202.53    2252    79
200.00    1986    80
197.53    1649    81
195.12    1747    82
188.24    2628    85
188.24    2891    85
183.91    2846    87
```

The frame-by-frame signal processing output for each slice is passed through a simple statistical analysis program to compute the mean, standard deviation and range of the values for the slice. Although the analyses of the different ACOUSTIC-PHONETIC VARIABLES required different statistical analysis programs (due to the different types of data output from each signal processing program), the procedure was essentially the same for each of them. As described in Section 4.1.1, the computation of the slice statistics did not include all the analysed frames, but used half the frames centred around the midpoint of the slice. Thus only the relatively steady-state middle portions of the vowel sounds were used in the analysis. Furthermore, the procedure of computing a mean parameter value for the phone from a number of frames, rather than from a single frame at the midpoint of the phone, produces a robust value relatively resistant to abnormal fluctuations in the parameter values.

The final action of **analyse_vowels** is to attach an identifier to each analysed slice in the ACOUSTIC-PHONETIC VARIABLE DATA FILES. This is particularly important as it allows cross-referencing with the SLICE NAMES FILES and SLICE MEANS FILES. In order to make the identifier transparent to the pattern-matching tools, the identifier consists of a number surrounded by colons, e.g. ':23:'. Thus when using **grep** for example, ordinary data values cannot be mistaken for the identifier. The numbering is achieved by searching for lines with the string 'samples' in them and reprinting those lines preceded by the identifier (incremented each time a search is successful). Thus:

```
awk '{
  if($1 == "samples")
    {n++
    printf ":%d: %s\n", n, $0}
  else
    {print $0}
}' $data_file >> $data_file.tmp
```

```
202.53    2252    79
200.00    1986    80
197.53    1649    81
195.12    1747    82
188.24    2628    85
188.24    2891    85
183.91    2846    87
:88: samples 2 to 5
f0: 195.2 5.07 ; 188.2 to 200.0
```

Figure B.5: The output of the fundamental frequency analysis of the slice **dr3.fsjw0.si1333.30220to31916**. The first column of the frame-by-frame results contains F_0, the second column the amplitude of the cepstral F_0 peak, and the third column the sample number related to the peak's quefrency. Frames 2–5 were used in the statistical analysis of the slice, giving a SLICE MEAN F_0 of 195.2Hz and a standard deviation of 5.1Hz. The range of data used to compute the mean was 188–200Hz.

```
# Replace the unnumbered ACOUSTIC-PHONETIC VARIABLE DATA FILE with
# the numbered version.
mv $data_file.tmp $data_file
```

where $data_file is the path to the target ACOUSTIC-PHONETIC VARIABLE DATA FILE, and $data_file.tmp is a temporary file.

Thus the final output of the analysis of each slice stored in the ACOUSTIC-PHONETIC VARIABLE DATA FILES consists of the frame-by-frame results of the signal processing, and the slice statistics, building up a database of the variable's intra- and inter-slice behaviour. Examples of the output for a single slice are shown in Figure B.5 for the fundamental frequency analysis, Figure B.6 for the harmonic amplitude difference analysis, and Figure B.7 for the formant frequency analysis. The SLICE NAME of the segment of speech used in the Figures is **dr3.fsjw0.si1333.30220to31916**. and the Figures represent analyses of an /aa/ vowel in the sentence **si1333** as spoken by speaker **fsjw0**. The 1696 samples in the phone produced seven analysis frames from the fundamental frequency and harmonic amplitude difference analyses; the formant frequency analysis produced 22 frames because the CSTR estimator required windows of 512 samples. Note that the identifier for each slice is the same in each case, enabling easy cross-referencing.

```
89.791   14   218   94.583   27   421   -4.792
88.916   14   218   93.960   27   421   -5.043
87.024   14   218   93.444   26   406   -6.420
87.525   13   203   93.258   26   406   -5.734
87.725   13   203   93.300   25   390   -5.576
87.533   13   203   93.396   25   390   -5.863
86.928   13   203   92.194   25   390   -5.266
:88: samples 2 to 5
@H1: 87.80 0.80 ; 87.02 to 88.92
@H2: 93.49 0.32 ; 93.26 to 93.96
H1-H2: -5.69 0.57 ; -6.42 to -5.04
```

Figure B.6: The output of the harmonic amplitude difference analysis of the slice **dr3.fsjw0.si1333.30220to31916**. The first column of the frame-by-frame results contains H_1, the second column the sample number related to the first harmonic's frequency, the third column the approximate frequency of the first harmonic, the fourth column H_2, the fifth column the sample number related to the second harmonic's frequency, the sixth column the approximate frequency of the second harmonic, and the final column H_1–H_2. Frames 2–5 were used in the statistical analysis of the slice, giving SLICE MEANS for the first two harmonic amplitudes of 87.8dB and 93.5dB respectively. The mean difference between the two harmonics is -5.7dB.

B.6 Checking the analysis data

This section describes the evaluation of the three signal processing programs, carried out to ensure the programs performed as intended. The evaluation procedure took the form of exhaustive checking of the slice results held in the ACOUSTIC-PHONETIC VARIABLE DATA FILES, and also served to seek out any unusual results (which may come from unusually articulated vowels or mislabelling of phones by the TIMIT phoneticians).

The section is split into five parts: a general description of the checking procedure as it was applied to the data for all three ACOUSTIC-PHONETIC VARIABLES; an evaluation of the performance of each of the three signal processing programs; and a description of the shell scripts used to check the updating of the ACOUSTIC-PHONETIC VARIABLE DATA FILES and SLICE NAMES FILES following any alteration to the results.

A. A description of the general checking procedure

The location of potentially erroneous slice results was carried out automatically by specially-written shell scripts. Particular use was made of the s.d. of the ACOUSTIC-PHONETIC VARIABLE values within a slice, of the frame-by-frame results for a slice, and of the expected ACOUSTIC-PHONETIC VARIABLE values for female and male speech. For slices whose results seemed suspect, visual inspection of the slice's acoustic waveform and its frequency or quefrency representation[6] was deemed particularly important to establish whether the

[6] A number of visual display tools written by members of the speech research group in the Department of Computer Science at the University of Sheffield enabled the inspection. Acoustic waveforms and individual spectra and cepstra were displayed using the tool **wave**. **spect** displays spectrograms as greyscale images, and can also be used to display cepstograms. However, because the resolution of its displays of small spectrograms is not very good, more use was made of **banktool**, which produces waterfall displays of spectrograms and cepstograms.

```
620.6   1934.3   2704.7
633.1   1989.0   2670.4
658.5   1951.3   2626.7
686.2   1890.1   2560.0
709.7   1832.7   2497.8
727.9   1786.3   2444.0
743.9   1724.9   2373.1
767.1   1694.4   2348.1
796.3   1675.5   2333.0
823.5   1654.7   2312.0
840.2   1637.2   2297.4
845.1   1625.4   2281.8
844.6   1614.6   2253.3
845.0   1603.1   2221.9
845.3   1571.2   2158.6
849.3   1558.5   2140.3
854.4   1554.8   2128.9
858.6   1557.5   2122.2
862.4   1561.2   2111.4
863.6   1563.8   2092.7
860.3   1555.1   2044.0
854.6   1559.8   2028.0
:88: samples 5 to 14
F1:  923.4  14.2 ; 906.8 to 948.7
F2: 1307.1   8.6 ; 1294.0 to 1319.4
F3: 2375.9  15.4 ; 2334.6 to 2387.2
```

Figure B.7: The output of the formant frequency analysis of the slice **dr3.fsjw0.si1333.30220to31916**. The first column of the frame-by-frame results contains F_1, the second column the F_2, and the third column F_3. Frames 5–14 were used in the statistical analysis of the slice, giving SLICE MEANS (s.d.s) for the first three formants of 923Hz (14Hz), 1307Hz (9Hz) and 2376Hz (15Hz) respectively.

results were acceptable or not, for two reasons. Firstly, because of the varied nature of the acoustics of speech sounds, the probable cause of an errant value was often not transparent, requiring inspection of the waveform and transformed representation to track down the problem. Secondly, the visual inspection of a large quantity of speech data allowed this researcher to build up a 'feel' for just how acoustically variable speech is, even when, as with the TIMIT data, the speech is read under very clean, unpressured conditions. While the location of suspect slices was carried out automatically, the decision on what actions to perform were taken manually. The options for the treatment of suspect slices located by the shell scripts were to:

1. Shorten the slice.
2. Rerun the signal processing program with different parameter settings.
3. Reject the slice.
4. Accept the slice.

Throughout the checking process, entries were made in the PROBLEM_SLICES files listing the actions carried out on suspect slices and any unusual features or variable values exhibited by the slices.

Finally, the ACOUSTIC-PHONETIC VARIABLE DATA FILES and the SLICE NAMES FILES had to be updated to take into account any changes made. If a slice was reduced in length, the signal processing programs had to be rerun on the shortened slice, and the following changes made: the old results for the slice in the appropriate ACOUSTIC-PHONETIC VARIABLE DATA FILES must be replaced by the new, and a new SLICE NAME, with altered sample numbers, must be entered in the SLICE NAMES FILE. If the signal processing program was rerun, the old results for the slice had to be replaced by the new. If a slice was removed altogether, its entries in the data and names files had to be removed. Finally, an entry of any alterations made must be entered in the appropriate PROBLEM_SLICES file for future reference. While the updating of the files was carried out as diligently as possible, human error inevitably creeps in, and so shell scripts were written to verify this tidying-up process.

B. Fundamental frequency data

The checking of the fundamental frequency data was based on the dual assumption that the SFF of a speaker of a particular sex, and therefore the measured F_0 of a slice, should fall within a certain range of values, and that the F_0 within a slice should not change by any great amount. These assumptions were incorporated into a shell script, **check_f0_means**, for the automatic checking of all the slices. While it would be possible to carry out the checks on either the ACOUSTIC-PHONETIC VARIABLE DATA FILES or the SLICE MEANS FILES, the advantage of using the SLICE MEANS FILES is that the checking can be carried out much faster, as these files contain only the statistical output for each slice. Using the ACOUSTIC-PHONETIC VARIABLE DATA FILES involves the additional step of searching for the slice statistics[7].

The range of F_0 considered acceptable was kept fairly small, the intention being that any major problems with the cepstral analysis and peak-picking algoritms would be highlighted. For an already tested system, this range could be much wider to reduce the

[7]This is fairly trivial to program, and is probably not much slower to perform as **grep** can be used to pattern match the lines containing the slice statistics.

time spent verifying the data. The ranges of acceptable F_0s for women and men were 170-260Hz and 94-149Hz respectively. As the initial trials of the programs indicated the cepstral analysis method was extremely robust, it was considered highly likely that any values outside of the ranges would be from speakers with atypically low or high fundamental frequencies, rather than as a result of the inadequate performance of the analysis programs.

If the slice showed a wide range of F_0, realised in a high standard deviation, the slice was examined further. The main problems a high s.d. would highlight would be phones adversely affected by phonetic context, and spurious F_0 values from one or more of the analysis frames. Slices with s.d.s greater than 3.9Hz were investigated.

check_f0_means works by examining each line in a SLICE MEANS FILE and marking with an asterisk those slices whose mean F_0 falls outside the acceptable range or whose s.d. is too high. Each suspect slice can then be matched with its corresponding entry in the SLICE NAMES FILE to find its SLICE NAME, and the slice investigated further. The script for **check_f0_means** is very simple, first setting the acceptable F_0 range:

```
# Test whether the variable 'sex' is f (female) or m (male).
if test "$sex" = f
then # FEMALE VERSION
  range_min=170.0
  range_max=260.0
else # MALE VERSION
  range_min=94.0
  range_max=149.0
fi
```

and then carrying out the checking as follows:

```
awk '{
  if ($2<'$range_min' || $2>'$range_max' || $2==NaN || $3>3.9)
    {printf "%s\t*\n", $0}
  else
    {print $0}
}' $data_file >> $data_file.tmp
```

where `$data_file` is the path to the target SLICE MEANS FILE, and `$data_file.tmp` is a temporary file.

The procedure for the examination of suspect slices depended upon the actual mean F_0 and s.d. of the slice in question:

- If the mean was outside the acceptable range, but the s.d. was relatively small (i.e. less than 3.9Hz), then a check on the frame-by-frame results for the slice in the fundamental frequency data file generally sufficed. If the amplitudes of the cepstral peaks (see the second column in Figure B.5) were sufficiently large, indicating strong periodicity, then the slice results were probably acceptable. In other words, the speaker probably used a relatively low or high F_0 for this phone, either because their SFF is relatively low or high for their sex, or because the sentence intonation required it.
- For female speakers, if the measured F_0 mean was low, this may have indicated a very high f0, one that fell outside the range of the search interval, and which was

not picked up by the cepstral analysis. Visual inspection of at least the acoustic waveform was required to establish the true F_0.

- Large s.d.s could indicate a number of things, including a very dynamic intra-slice F_0, spurious values in the frame-by-frame results, and corruption of the waveform due to the position of the phone in the sentence (particularly utterance-final positions) or its phonetic context (such as the influence of an adjecent fricative). Again, visual inspection of at least the acoustic waveform was required to establish the true F_0.

A total of approximately 1000 female and 1800 male slices were manually checked, involving at least an inspection of the appropriate fundamental frequency data file. This constituted approximately a fifth of core data set. The actions taken on slices, and the numbers of slices involved, are examined in detail below.

Rerunning the analysis program with a different search interval
In a number of cases, the peak-picking program had to be rerun with either the lower or upper limit of the search interval reset.

The program was rerun with a different upper limit to the search interval 45 times for the female speakers and 49 times for the male speakers. This was generally because in the initial portion of the slice the acoustic waveform had a double periodicity, resulting in two prominent peaks in the quefrency domain, one at double the quefrency of the other. Consider, for example, the middle of the acoustic waveform in Figure B.8: the F_0 appears to be 190Hz. But in the first third of the slice two different periods can be discerned, one of approximately 5msec (giving an F_0 of 190Hz), and one of 10msec (95Hz). (Indeed, in the first 20msec the F_0 is clearly 95Hz.) The consequences for the cepstral analysis can be seen in Figure B.9, where for the first few frames the peak at 10msec (representing an F_0 of 95Hz) is larger than the peak at 5msec (representing an F_0 of 190Hz). The peak-picking algorithm selected the higher amplitude peak, and thus reported the slice's F_0 as half its true value (of 190Hz). A double periodicity is even more evident in the acoustic waveform reproduced in Figure B.10, particularly so between 30-70msec. The most prominent peaks in the quefrency domain (see the cepstrum in Figure B.11) are clearly those at a quefrency of 13msec, equivalent to a fundamental frequency of 75Hz. However, careful examination of this slice, and the phones surrounding it, showed that this slice's F_0 was 150Hz. The peak-tracking program was not capable of dealing with such situations, and often ended up tracking the peak representing the lower F_0, particularly if this peak had the greater amplitude. Thus the F_0 results for the slice would tend to be abnormally low for either female or male speech. The remedy for this situation was to rerun the peak-tracking program, with the upper limit of the search interval set to exclude the rogue peak.

The program was rerun with a different lower limit to the search interval 8 times for the female speakers and 168 times for the male speakers. The main cause of the failure of the program was that the acoustic waveforms of the slices contained a false hint of dual periodicity, resulting in a peak at half the quefrency of the F_0 peak in the quefrency domain. This is illustrated for a female speaker in Figure B.12, and a male speaker in Figures B.13 and B.14. Further investigation of these slices revealed that when the two halves of the period in an acoustic wave are similar, this additional 'periodicity' is picked up by cepstral analysis. Consider the period beginning with the large positive peak at approximately 15msec in Figure B.13. The second half of the period is almost a scaled-down version of the first half. It is this quasi-periodicity that causes the rogue cepstral peak, visible at a quefrency of 4msec (equivalent to a F_0 of 260Hz) in Figure B.14, with the much larger F_0 peak at 8msec (120Hz). The remedy was to increase the lower limit of the search interval, to exclude the rogue cepstral peaks.

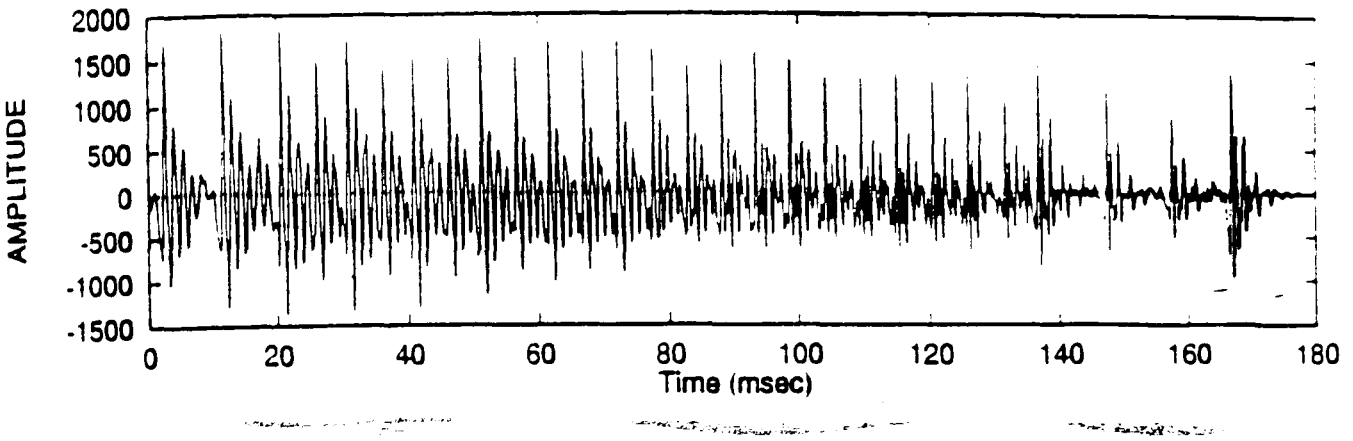


Figure B.8: Acoustic waveform of the slice **dr1.faks0.sa2.50010to52840**. This represents a production of the vowel phone /ae/, in an utterance-final position in the context /dh ae tcl/ ('THAT'). Note how the frequency halves in the final 40msec of the wave.

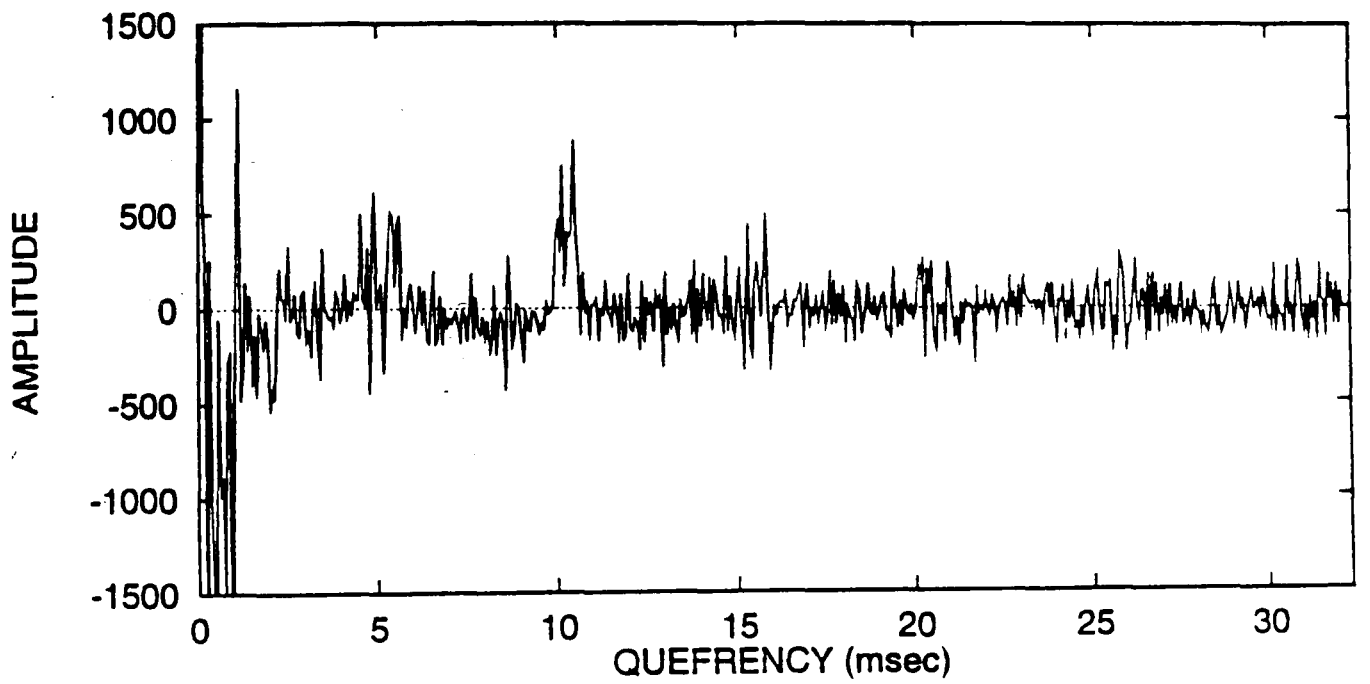


Figure B.9: Cepstrum from the middle of the slice **dr1.faks0.sa2.50010to52840**.

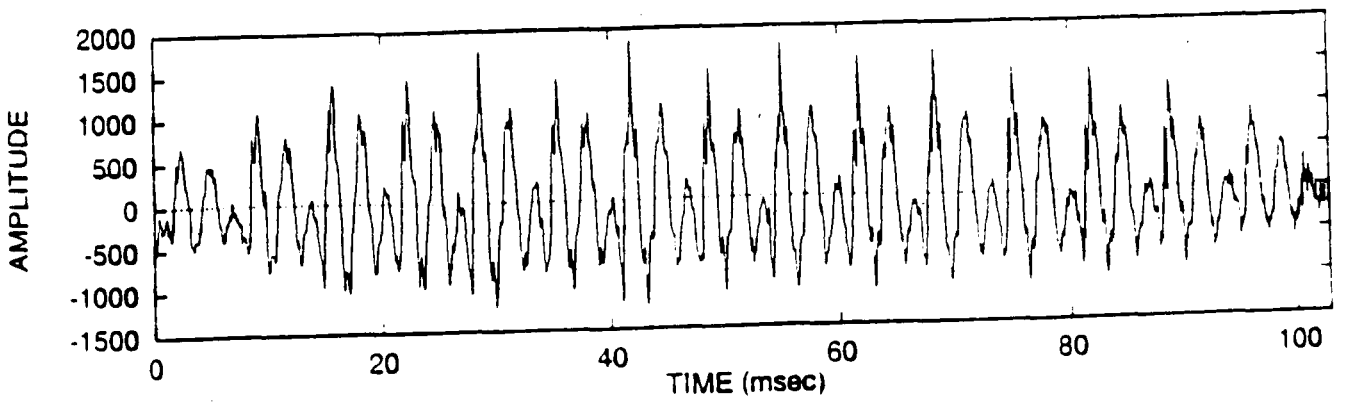


Figure B.10: Acoustic waveform of the slice **dr8.mbsb0.si723.17120to18750**. This represents a production of the vowel phone /iy/. in an utterance-final position in the context /dh iy z/ ('THESE').

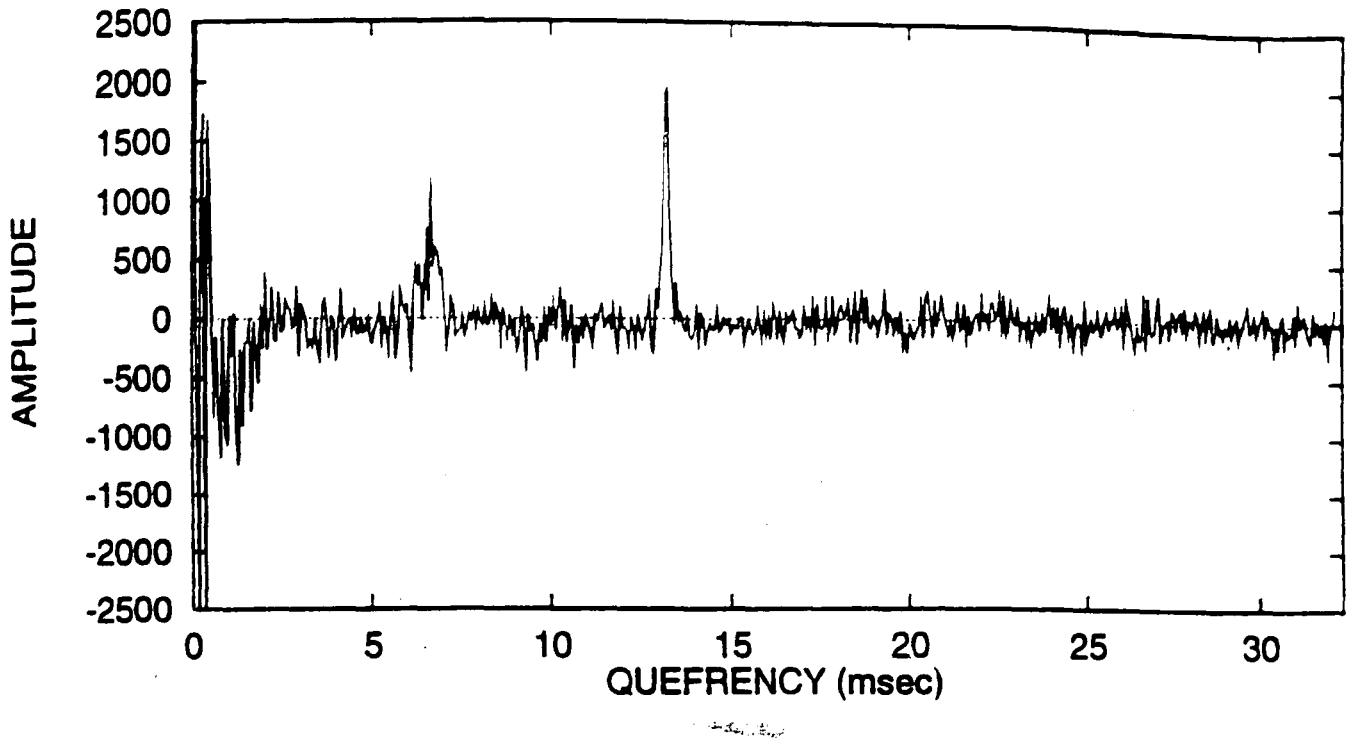


Figure B.11: Cepstrum from the middle of the slice **dr8.mbsb0.si723.17120to18750**.

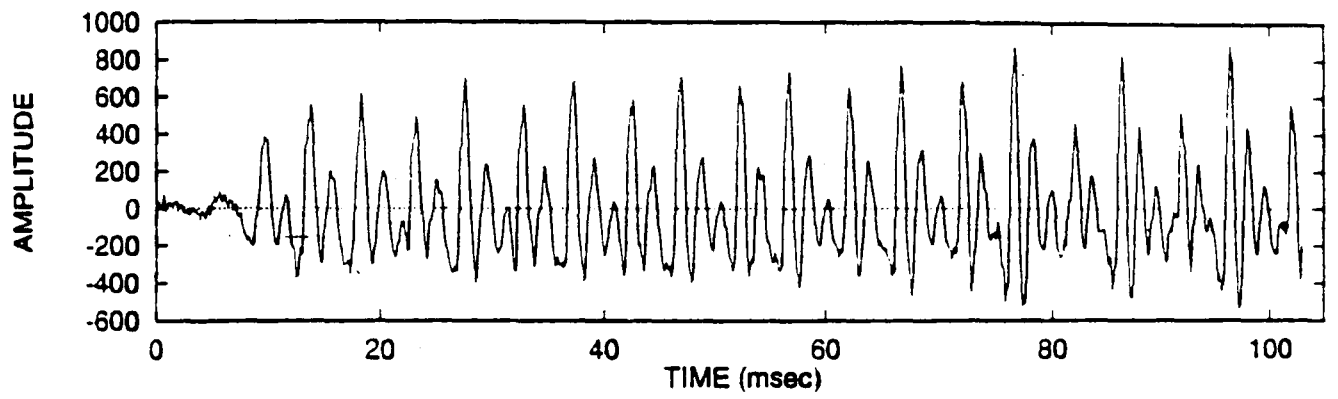


Figure B.12: Acoustic waveform of the slice **dr1.ftbr0.sx201.28770to30405**. This represents a production of the vowel phone /ux/, in the context /ch ae eh/ ('statuesque'). Although there is a hint of double periodicity, with F_0s of 100Hz and 200Hz, further investigation confirmed a low F_0 of 100Hz.

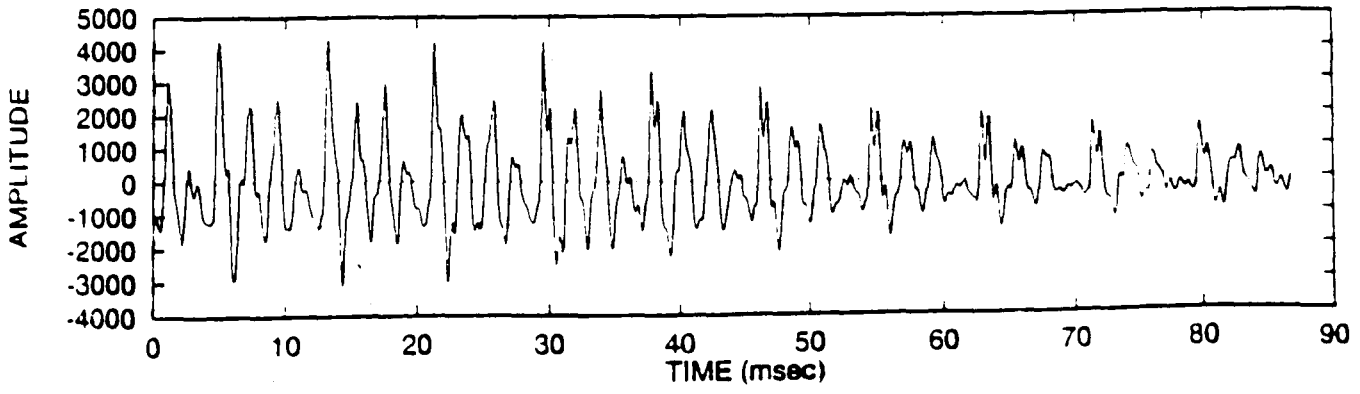


Figure B.13: Acoustic waveform of the slice **dr6.mrjs0.si1523.5849to7227**. This represents a production of the vowel phone /iy/, in the context /el iy w/ ('thoroughLY Wised').

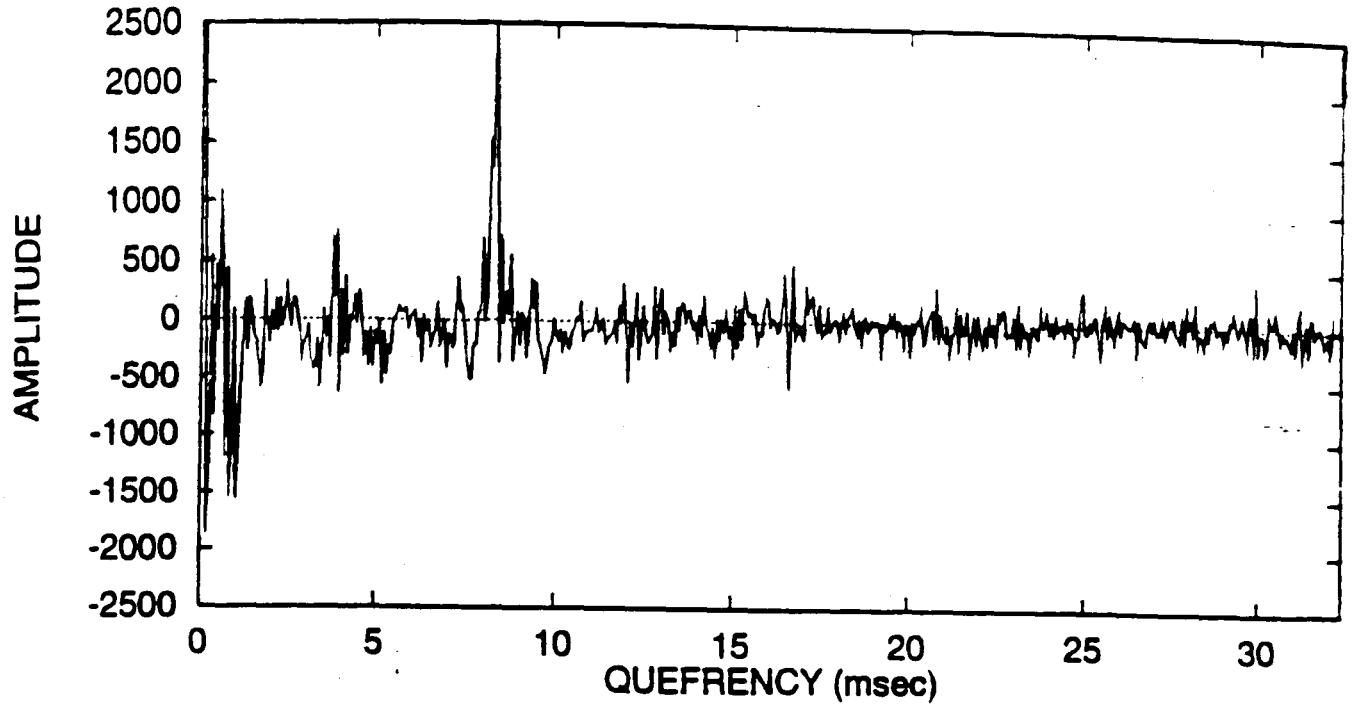


Figure B.14: Cepstrum from the slice **dr6.mrjs0.si1523.5849to7227**.

Another problem encountered was the few slices whose true F_0 lay outside the search interval. This entailed slices from female speakers with F_0s greater than 330Hz, and slices from male speakers with F_0s less than 50Hz. For four of the female slices, an exceptionally high slice F_0 meant cepstral F_0 peaks were not picked out by the peak-picking program, i.e. the slice's quefrency was less than the lower limit of the search interval (3msec, equivalent to 333Hz). The remedy was to reduce the lower limit of the search interval, while taking care not to include the cepstral vocal tract components. Similarly, in seven of the male slices, the fundamental frequency of the slice was too low for the program, as the slice's quefrency was greater than the upper limit of the search interval (20msec, equivalent to 50Hz). The remedy was to raise the upper limit of the search interval. The highest and lowest F_0s measured in this study were 384Hz (equivalent to a quefrency of 2.6msec) from a female speaker, and 40Hz (25msec) from a male speaker. It is possible that slices with exceptionally low or high F_0s were missed entirely: for example, if a female speaker's slice had a fundamental frequency of 400Hz, which produced a strong cepstral F_0 peak, the second rahmonic might be of sufficient strength to be picked out as the F_0 peak. The F_0 peak, being outside of the search interval, would be missed entirely, and the F_0 of the slice would be reported as an apparently legitimate 200Hz. Care was taken to examine all the slices for female and male speakers who evidenced a high or low SFF respectively. As the extremes of the search interval encompassed such a wide range of frequencies, it was unlikely that speakers with a exceptionally high or low SFF would be missed entirely, as a majority of their slices would be likely to fall within the bounds of the search interval.

Shortening the slice

For some slices it was felt the length of the slice had to be reduced. This was carried out on 129 female and 146 male slices. Despite the efforts to reduce the effects of coarticulation on the slice statistics, a number of the vowel phones were severely corrupted both by phonetic context and their position in a sentence. It is also possible that the bounds of some phones were wrongly placed in the original transcribing of the data. That said, the decision to reduce a slice was obviously a subjective one, but motivated by a desire to

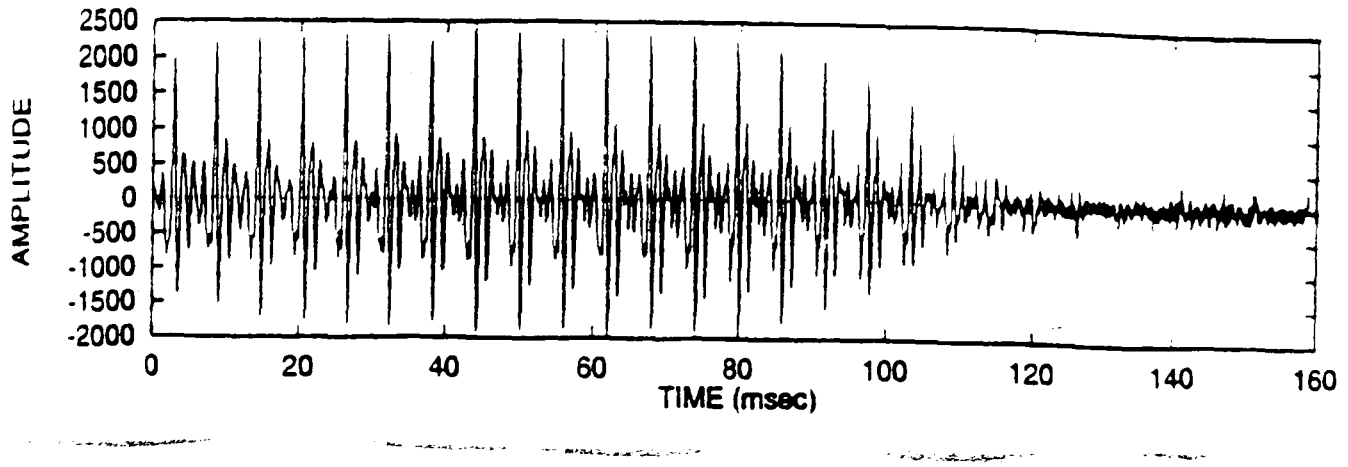


Figure B.15: Acoustic waveform of the slice **dr7.ftlh0.sx289.13657to16180**. This represents a production of the vowel phone /aa/, in the context /l aa sh/ ('gaLOSHes'). The last half of this slice was removed, the new slice name being dr7.ftlh0.sx289.13657to15350.

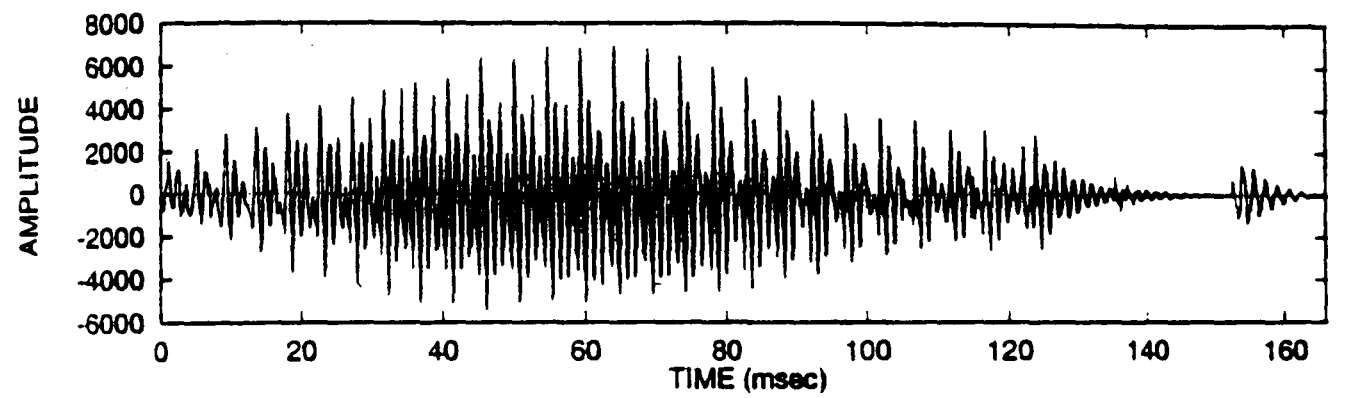


Figure B.16: Acoustic waveform of the slice **dr1.fjem0.sx184.5711to8339**. This represents a production of the vowel phone /ae/, in the context /kcl k ae tcl s/ ('CAT'S'). The last quarter of this slice was removed, the new slice name being dr1.fjem0.sx184.5711to7800.

have parameter values that were representative of both the rest of the slice in question and the speaker's other slices.

The biggest source of phonetic context problems were fricatives and consonantal closures. This is illustrated in Figures B.15, B.16 and B.17. The last third of the slice in Figure B.15, from 110msec onwards, shows how the vowel's phonation has been replaced by the noise of the following fricative. It is possible that the /aa/ ends and the /sh/ begins at around 110msec. The vowel phone in Figure B.16 loses its periodicity between 110msec and 130msec, followed by a period of silence. It is likely this was caused by the preparation for the /t/-closure. The slice in Figure B.17 suffers from fluctuations in amplitude in its first half, while the second half is largely unperiodic, although the vocal folds were obviously still phonating.

Approximately half the slices shortened occurred in the final syllable of the sentence. The consequences for a slice in an utterance-final syllable are illustrated in Figures B.18 and B.19, showing a reduction in waveform amplitude and lack of periodicity.

Rejecting the slice

For some slices, it was felt they were too corrupted to be left in the analysis. This decision was reached for 98 female and 197 male slices, and occurred for much the same reasons as

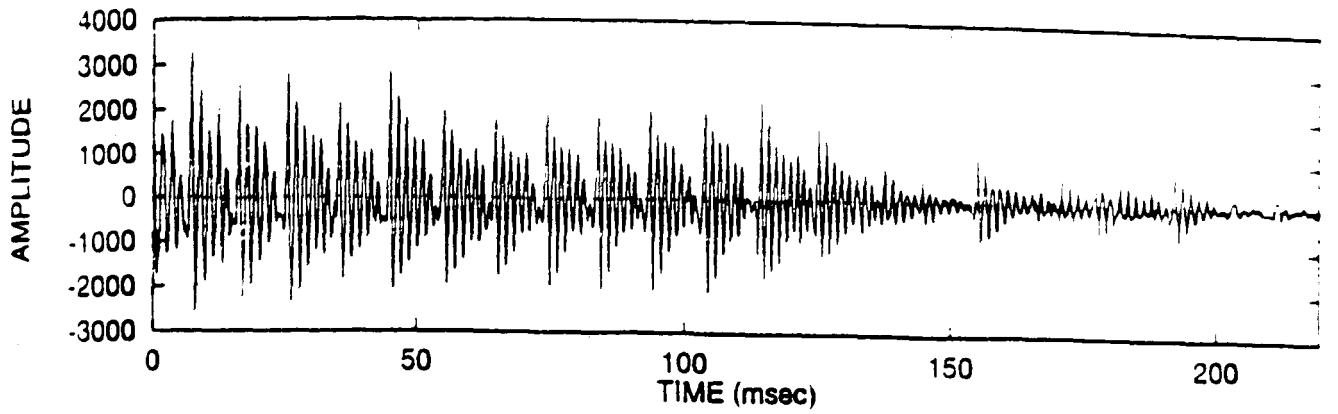


Figure B.17: Acoustic waveform of the slice **dr5.mvlo0.si517.22225to25720**. This represents a production of the vowel phone /aa/, in the context /kcl k ey aa s/ ('CHAOS'). The last third of this slice was removed, the new slice name being dr5.mvlo0.si517.22225to24570.

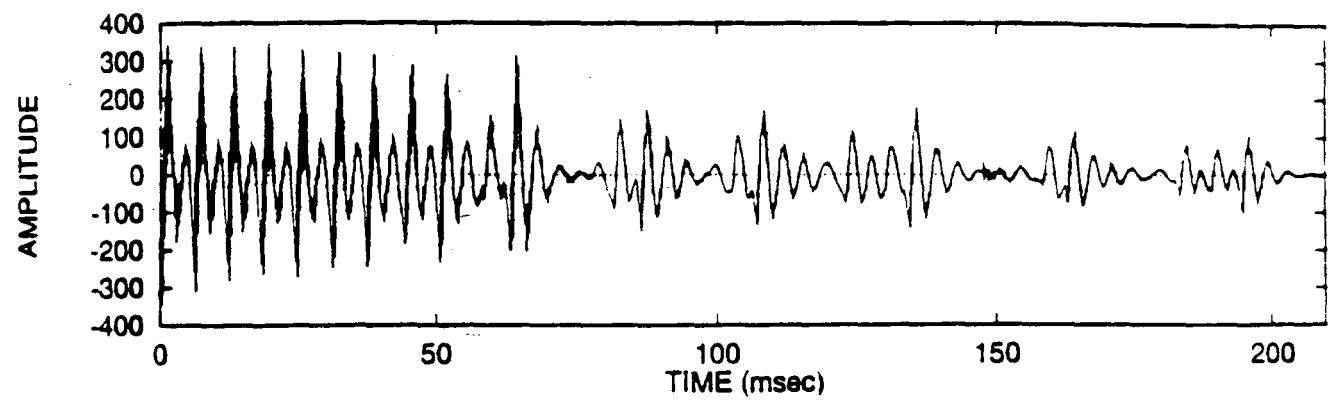


Figure B.18: Acoustic waveform of the slice **dr5.fdtd0.si2191.42310to46290**. This represents a production of the vowel phone /ux/, in the context /f y ux z/ ('reFUSE'). The last three-quarters of this slice was removed, the new slice name being dr5.fdtd0.si2191.42310to43330.

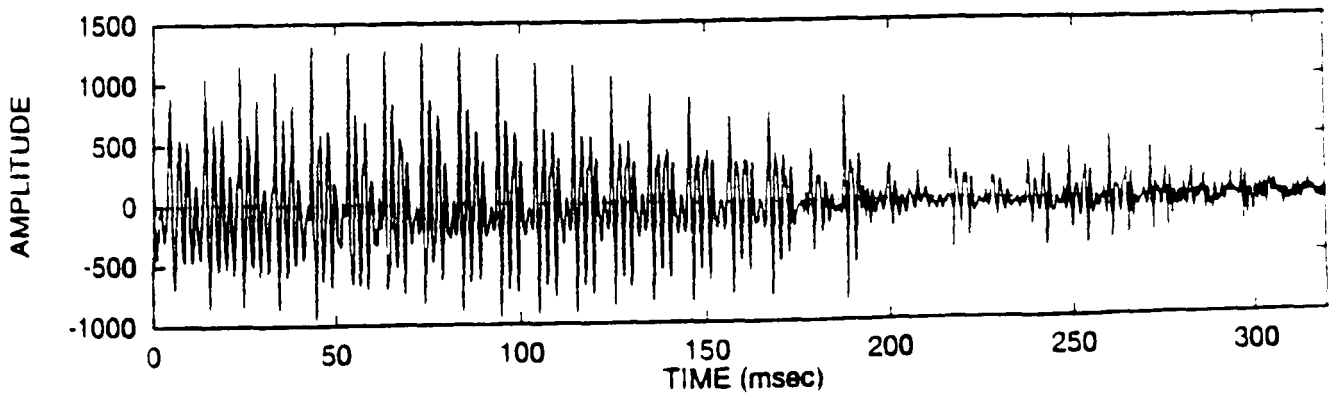


Figure B.19: Acoustic waveform of the slice **dr5.msfh1.sx190.33128to38200**. This represents a production of the vowel phone /ao/, in the context /dh iy ao h#/ ('THE OIL'). The last half of this slice was removed, the new slice name being dr5.msfh1.sx190.33128to35700.

the reduced slices. Interestingly, few speakers had more than two slices rejected. Even for the speakers with more rejected slices, this was only for a small proportion of their slices. For example, of the speakers with the most number of rejected slices, **ftbr0** had seven of her 41 slices rejected, **mgrt0** had six out of 45 rejected, and **mpab0** six out of 38.

As with the reduced slices, approximately half of the rejected slices occurred in utterance-final syllables, and involved 40 female and 96 male slices.

C. Harmonic amplitude difference data

The procedure for the checking of the harmonic amplitude difference data was based on the assumption that the slice parameters (i.e. H_1, H_2 and H_1–H_2, and an estimate of the fundamental frequency based on the location of the first harmonic) were not expected to vary much within a slice. As before, this assumption was incorporated into a shell script, **check_peak_diffs**, for the automatic checking of all the slices. This procedure involved checking the frame-by-frame slice data as well as the slice statistics, requiring the script to be run on the harmonic amplitude difference data files. The checks carried out will now be described.

The frame-by-frame results for a slice contain the sample number of the first harmonic (see column 2 in Figure B.6), from which a rough estimate of each frame's fundamental frequency can be estimated. This estimate proved to be 10-25Hz below the frame's true F_0. Taking into account the discrepancy between the estimated F_0 and the true F_0, a range of acceptable F_0s were defined for each sex: 203-250Hz (equivalent to sample numbers 13-16 for the 1024-sample window size) for female speech, and 109-141Hz for male speech (equivalent to sample numbers 7-9). Any frame with an F_0 outside of the acceptable range was investigated further.

The sample number of the second harmonic of each frame (see column 5 in Figure B.6) was checked to see if it was greater than twice the first harmonic's sample number. As the harmonics are multiples of the fundamental frequency, the frequency of the second harmonic should not be greater than twice that of the first harmonic. This was intended to guard against instances where the peak-picking program had failed to locate the second harmonic, and had found a higher harmonic instead.

The final series of checks were made on the intra-slice dynamics of H_1 and H_2. The first check involved seeing how the frame-by-frame harmonic frequency estimates changed. Thus, if a harmonic's sample numbers (i.e. the numbers in columns 2 and 5 in Figure B.6) differed by more than one between frames, then the slice was investigated further. This served to highlight instances where the peak-picking algorithm suddenly switched to a different harmonic. The other check was on the s.d.s of H_1–H_2, H_1 and H_2. Slices with a s.d. greater than 1.9dB were investigated further.

check_peak_diffs works by examining first the lines representing a slice's frame-by-frame results, followed by the slice statistics. The output differs from **check_f0_means** in that it takes the form of a report on each slice. Typical output is shown in Figure B.20. Thus slice 6 has problems with four frames having a relatively low F_0, and two of the slice parameters have s.d.s greater than 1.9dB. The line numbers refer to the line numbers of the output report file. The code for the checking is as follows:

```
# Set the slice counter and output first slice number.
awk 'BEGIN{slice_count=1;printf ":%d:\n", slice_count}
{
```

```
# This line ensures only the frame data is checked
# (and the stats are ignored).
if (NF==7 && $4!=";")
{
  ### Check the f0s. ###
  # For low f0s :
  if ($2 < '$l_sample' && $2 != old_field2)
    printf "LOW f0: Line %d, f0 = %d\n", NR, $3

  # For high f0s :
  if ($2 > '$u_sample' && $2 != old_field2)
    printf "HIGH f0: Line %d, f0 = %d\n", NR, $3

  ### Check the 2nd harmonic sample no. is <= twice ###
  ### the 1st harmonic sample no.                    ###
  if ($5 > (2*$2) && $5 != old_field5)
    printf "TOO BIG 2ND HARM: Line %d\n", NR
}

### Check the between-frame differences in sample numbers. ###
if (NF==7 && $4!=";" && old_field2>0)
{
  # If previous line differs from current line by more than one,
  # output the current line number and the difference.
  # FOR THE 1ST HARMONIC PEAK :
  sub1=$2-old_field2
  if (sub1 > 1 || sub1 < -1)
    printf "1ST HARM: Line %d, diff = %d\n", NR, sub1

  # FOR THE 2ND HARMONIC PEAK :
  sub1=$5-old_field5
  if (sub1 > 1 || sub1 < -1)
    printf "2ND HARM: Line %d, diff = %d\n", NR, sub1
}

### Check the s.d. ###
# Only the lines for the slice statistics contain semi-colons
if ($4==";" && $3 > 1.9)
  printf "SD: Line %d, s.d. = %.1f\n", NR, $3

# If the end of the slice's results have been reached,
# increment slice counter and output the next slice number.
if ($1 == "H1-H2:")
{
  slice_count++
  printf ":%d:\n", slice_count
}

# Remember sample no.s of the 1st and 2nd harmonic peak.
old_field2=$2
old_field5=$5
```

```
:1:
:2:
HIGH f0: Line 755, f0 = 265
HIGH f0: Line 757, f0 = 281
2ND HARM: Line 757, diff = 2
2ND HARM: Line 760, diff = -3
SD: Line 764, s.d. = 2.4
:3:
LOW f0: Line 1053, f0 = 187
:4:
:5:
:6:
LOW f0: Line 211, f0 = 187
LOW f0: Line 214, f0 = 171
LOW f0: Line 223, f0 = 156
LOW f0: Line 228, f0 = 140
SD: Line 238, s.d. = 5.9
SD: Line 239, s.d. = 3.3
:7:
:8:
```

Figure B.20: Example output from the shell script **check_peak_diffs**. For example, for slice 3, the frames reported on lines 755 and 757 have relatively high fundamental frequencies, the difference in second harmonic sample number between the frames on lines 756 and 757 and 759 and 760 are relatively high, and the slice parameter statistics on line 764 show a high s.d.

```
}' $data_file >> $file2
```

where `$data_file` is the path to the target ACOUSTIC-PHONETIC VARIABLE DATA FILE, `$file2` is the output file's pathname, and `$l_sample` and `$u_sample` are the boundaries of the acceptable fundamental frequencies.

The procedure for the examination of suspect slices depended upon what caused **check_peak_diffs** to signal the slice was suspect.

- If a high or low F_0 was signalled, a comparison between the reported F_0 and the true F_0 in the relevant fundamental frequency data file was usually a sufficient check. If there was a discrepancy, it was typically a falsely-reported 'HIGH F0' caused by the amplitude of the real first harmonic falling below the threshold set in the peak-picking program (80dB for female speech and 75dB for male speech). Thus what was being measured as the first harmonic was actually the second. To remedy this, the threshold had to be reset to, in general, 75dB for females and 70dB for males.

- If a difference between the sample numbers of the harmonics was signalled, it meant the harmonic had fallen below the threshold level, and the peak-picking program had switched to a higher harmonic. This usually occurred at the beginning or end of a phone, where coarticulation effects reduce the amplitudes of the harmonics. Thus it was usually outside of the interval used to calculate the slice statistics and was no cause for concern. If it happened within the statistics interval, the peak-picking program had to be rerun, with the threshold reset as above.

As the fundamental frequency analysis was carried out prior to the harmonic amplitude difference analysis, most of the slices causing problems had been identified. There were subsequently fewer suspect slices to deal with, although visual inspection of the frequency domain was still necessary for a number of slices. A total of 191 female and 318 male slices were manually checked. It was necessary to rerun the peak-picking program with the threshold reset a total of 88 times for the female speakers and 195 times for the male speakers.

D. Formant frequency data

While the initial trials of the fundamental frequency and harmonic amplitude difference analysis methods indicated a high degree of reliability, no such confidence was placed in the CSTR formant frequency estimator. This was partly because the automatic estimation of formant frequencies involves a much greater level of complexity, and partly because no formal evaluation of the estimator's performance had been carried out (see Section 4.1.2 for more details). Therefore this section is in two parts, the first dealing with estimator evaluation, the second with the checking of the slice results in the manner of the other two ACOUSTIC-PHONETIC VARIABLES.

Evaluation of the CSTR formant frequency estimator

As already discussed in Section 4.1.2, the evaluation of the CSTR formant frequency estimator had to answer two important questions: whether it was able to cope with female speech, and whether an automatic analysis could yield accurate results for different phones within a set search space. To answer these questions, three evaluation tasks were carried out. The first involved a random selection of the speech of both sexes to assess the estimator's general performance; the second compared the estimator's performance for different FFT window lengths; and the third examined female speech in more depth.

EVALUATION 1: Assessment of the estimator's overall performance

The evaluation of the estimator's overall performance was based on a comparison between the estimator's performance using an upper limit, or u-setting, of 3100Hz, and a hand measurement of the formant frequencies. The u-setting defines the upper limit of the search space for the estimator, the default for which is 2700Hz. This was considered to be too low to encompass the majority of female third formants. The lower limit, or l-setting, was left at 250Hz, the estimator's default. The hand measurements were conducted using a spectrogram display tool[8] on a NeXT computer, with a spectrogram window length of 128 samples for the female speakers, and 256 samples for the male speakers[9]. The measurements were made in the centre of the slice, although any intra-slice drift in the formants was taken into account. However, due to the resolution of the display tool, it proved difficult to estimate the frequency of F_1, particularly for the female speakers, and it was difficult to ascertain the degree of confidence that could be placed in the estimates.

For each phone, eight slices were (fairly) randomly selected for test data (in fact, two dialect regions were randomly selected for each phone, with four slices taken from each dialect). The word 'fairly' is used to reflect the fact that most of the slices had a fundamental frequency close to the mean, while the rest were representative of more extreme F_0s.

[8] Designed by Malcolm Crawford at the Department of Computer Science, University of Sheffield.

[9] These values reflect the window lengths required to give the best spectrogram resolution. Using a window length of 128 samples for the male speakers, the opening and closing of the vocal folds in the form of vertical lines is clearly visible, interfering with the spectrogram's resolution. Moreover, the 256-sample window gave the best resolution for F_1, the hand estimation of which tended to be a major problem. Using a 256-sample window for female speech created very marked harmonics, resulting in ill-defined formants.

Where the output of the estimator and the hand estimates differed, the estimator was rerun with a different u-setting. The results of the evaluation are reproduced in Table B.1 for the female speakers, and Table B.2 for the male speakers. From the evaluation exercise, a number of points can be made regarding the estimator's general performance:

- When F_3 lies close to the u-setting, its true value is often underestimated. This could be because the bandwidth of F_3 straddles the upper limit of the search space, with the result that the estimator is not measuring the entire formant. Occasionally, this will cause the estimator to ignore F_3, picking out a prominent feature with a lower frequency instead.

- When the phone is poorly articulated, resulting in ill-defined formants, the estimator can jump in value for some of the analysis frames (usually to another formant) as it seeks other candidates for the formants. This can occur in cases where the amplitude of the phone is decreasing, where there is alot of coarticulation from the adjacent phones, and where there is some aperiodicity in the phonation of the vowel. However, there were also a number of examples where the estimator coped well with such phones, illustrated in Figures B.21 and B.22, showing an acoustic waveform and its spectrogram.

- For a number of the slices, the formants were not steady-valued, changing by as much as 1000Hz or more within the slice. The estimator generally coped well with such input, the drift in formant frequency being reflected in large s.d.s, except where the drift took F_3 close to or above the u-setting.

- When the u-setting is raised or lowered, the estimator typically raises or lowers its frequency estimates by small but significant amounts. For example, when the u-setting is raised from 3100Hz to 3500Hz the estimator generally adds 30-40Hz to its estimate of F_3. Obviously this has implications for the accuracy of the estimator's output.

- The estimator appears to perform well on female speech, interpolating successfully between the relatively widely-spaced harmonics to produce accurate formant frequency values. See Figure B.23 for an example of this.

- Based on its performance with relatively high-valued first formants (i.e. in the vowels /aa/, /ae/ and /ao/), the estimator performs well with an l-setting of 250Hz, even for the female speakers (see evaluation exercises using speakers with high F0s). As already stated, it was not possible to accurately hand-measure the frequencies of the first formants, particularly for low-valued first formants. Thus the evaluation was unable to comment on the estimator's output for the first formants of /iy/, /uw/ and /ux/.

EVALUATION 2: Assessment of the estimator's performance using different FFT window lengths

This evaluation compared the estimator's performance on the phone /iy/ using FFT window sizes of 256 and 512 samples. Two female speakers were chosen, one with a relatively high SFF (speaker **faks0**, whose mean SFF was measured at 264Hz) and one with a relatively low SFF (**fdac1**, with a mean SFF of 161Hz). The second and third formants of /iy/ for female speakers are generally high in frequency and fairly close together (2800Hz and 3300Hz respectively according to Peterson & Barney (1952)). This creates a considerable test for the estimator, especially when using female speech, for a number of reasons:

	Slice name	Context	F_0	F_1	F_2	F_3	Best u
	dr4.fbmj0.si1776.35000to39320	/jh aa bcl/	140	900	1300	2500	3100
	dr4.fcft0.si548.21932to25560	/l aa f/	190	900	1250	3000	3100
/	dr4.fgjd0.sx369.2381to6600	/b aa bcl/	231	900	1400	2700	3100
a	dr4.fmcm0.sx280.20094to24200	/r aa zh/	190	900	1500	2500	3100
a	dr6.fjdm2.sx232.53369to58122	/y aa bcl/	180	900	1400	2900	3100
/	dr6.fmgd0.si2194.26653to31560	/jh aa bcl/	140	800	1300	2850	3300
	dr6.fsgf0.si2187.25147to28440	/r aa kcl/	200	900	1300	2900	3100
	dr6.ftaj0.sx429.30401to33160	/b aa dx/	210	950	1300	2600	3100
	dr1.faks0.si1573.70910to74200	/sh ae bcl/	240	880	1850	2600	3100
	dr1.fecd0.si2048.12080to14730	/hh ae z/	210	850	2100	3000	3500
/	dr1.fjsp0.sx174.33720to36178	/t ae n/	260	600	2800	3200	3500
a	dr1.fkfb0.sa2.32439to35720	/r ae gcl/	200	800	2200	2900	3500
e	dr6.fjdm2.sa2.48010to52422	/dh ae tcl/	250	1000	2000	3100	3500
/	dr6.flag0.sx294.18632to21890	/r ae f/	230	950	1750	2700	3100
	dr6.fmju0.sx129.17295to19885	/k ae tcl/	220	900	2000	2850	3500
	dr6.fsgf0.sa2.5960to9120	/n ae s/	230	800	2400	2950	3100
	dr2.fdnc0.si2287.6920to8439	/k ao r/	290	600	900	2400	3100
	dr2.fdxw0.sa1.26600to29440	/w ao sh/	190	650	950	2500	3100
/	dr2.frll0.sx434.30600to34590	/f ao n/	220	800	1150	2500	3100
a	dr2.fslb1.si644.2760to5366	/q ao l/	230	850	1400	2500	
o	dr6.fdrw0.si1423.9680to13360	/d ao n/	210	800	1200	3500	3700
/	dr6.fjdm2.sx142.14732to17617	/l ao ng/	310	800	1200	3250	3500
	dr6.fmgd0.sa1.41401to44950	/q ao l/	200	600	1100	3000	3500
	dr6.fsdj0.sa1.28073to30600	/w ao sh/	190	750	1100	2700	3100
	dr4.fcft0.sx278.16280to19175	/t iy dcl/	220	450	2650	3050	3500
	dr4.fedw0.si1084.7480to12440	/th iy m/	240	450	2500	3250	3500
/	dr4.feeh0.si471.9613to12760	/l iy pau/	290	600	2500	3000	3500
i	dr4.fgjd0.si818.38433to42064	/l iy pau/	180	550	2550	3000	3500
y	dr7.fisb0.sx319.28845to31644	/l iy q/	270	550	2400	3100	3500
/	dr7.fjrp1.sx262.28210to31123	/s iy q/	180	450	2600	3050	3500
	dr7.fpab1.si1471.15940to19010	/z iy s/	300	600	2400	3000	3500
	dr7.ftlh0.si1639.7880to10933	/v iy pau/	250	500	2700	3300	3700
	dr1.fdac1.sx214.15666to18214	/ux uw q/	165	500	850	2950	3500
	dr1.fetb0.sx338.18440to22392	/g uw n/	230	600	1150	3200	3500
/	dr1.fkfb0.sx258.39812to42510	/y uw q/	200	500	1600	2700	3100
u	dr1.fsma0.sa1.21849to24440	/s uw tcl/	240	500	1300	2700	3100
w	dr5.fcdr1.sx106.8606to11560	/y uw zh/	220	450	1200	2400	3100
/	dr5.fexm0.sa1.30200to32794	/s uw q/	220	450	1100	2750	3100
	dr5.fjxm0.sx131.20440to23000	/n uw bcl/	220	450	1100	3000	3500
	dr5.futb0.sx34.10110to13080	/d uw tcl/	200	400	1300	2900	3500
	dr1.fdac1.sa1.28310to32520	/s ux tcl/	160	450	950	2600	2900
	dr1.fecd0.sa1.31260to34200	/s ux tcl/	220	550	2450	3050	3500
/	dr1.fjem0.sx274.14520to18840	/s ux dh/	230	500	1750	2500	3100
u	dr1.fsah0.sa1.19960to22600	/s ux tcl/	180	400	1800	2600	3100
x	dr7.fisb0.sa1.32672to35960	/s ux tcl/	220	450	2000	2650	3100
/	dr7.fkde0.sx241.26600to29480	/hh ux dcl/	220	450	2400	3000	3500
	dr7.flet0.si507.14287to16961	/n ux w/	200	600	2000	2950	3500
	dr7.fsxa0.sa1.19560to22225	/s ux tcl/	180	400	1750	2450	3100

Table B.1: Results of the evaluation of the CSTR formant frequency estimator for female speech. Notes: (1) The F_0 values are the means for the slice, as computed for the fundamental frequency analysis (to the nearest 10Hz); (2) The formant frequency values are from the hand estimates (to the nearest 100Hz); (3) The value reported in the column 'Best u' is the u-setting (in Hz) which produced the most accurate estimate of the three formant frequencies.

	Slice name	Context	F_0	F_1	F_2	F_3	Best u
	dr3.mcal0.sx328.13160to16520	/t aa m/	130	700	1100	2800	3100
	dr3.mdtb0.si1200.11080to14120	/m aa n/	90	700	1000	3000	3300
/	dr3.mhpg0.sx280.18381to23080	/r aa zh/	100	700	1100	2200	3100
a	dr3.mlns0.sx417.8230to11840	/l aa dcl/	150	700	1100	2200	3100
a	dr7.mbth0.si505.18172to22768	/b aa m/	130	600	1000	2800	3300
/	dr7.mdpb0.sa1.30413to33170	/w aa sh/	140	750	1100	2500	3100
	dr7.mkdb0.sx422.45080to48130	/w aa sh/	120	700	1100	2400	3100
	dr7.mpfu0.sa1.34131to37264	/w aa sh/	100	750	1100	2500	3100
	dr3.madc0.sa2.6459to9525	/q ae s/	130	800	2000	2500	3100
	dr3.mgaf0.sa2.30440to34040	/dh ae tcl/	80	650	1500	2500	3100
/	dr3.mmar0.sa1.6360to8731	/hv ae dcl/	140	750	1850	2500	3100
a	dr3.mwdk0.si806.7840to11280	/dx ae n/	100	600	2000	2600	3100
e	dr8.mcxm0.sa2.33810to36389	/dh ae tcl/	80	700	1400	2300	3100
/	dr8.mjln0.sa2.34280to38360	/dh ae tcl/	100	700	1600	2600	3100
	dr8.mkdd0.sa2.33546to36760	/r ae gcl/	100	700	1700	2500	3100
	dr8.mpam0.sx379.3087to5160	/p ae m/	150	700	1700	2400	3100
	dr5.mahh0.sx214.5042to8200	/m ao n/	130	550	950	2500	2900
	dr5.mjxa0.si1507.4512to7770	/w ao l/	130	650	850	2900	3100
/	dr5.mrjm3.si2078.21640to25413	/d ao gcl/	100	600	900	2700	3100
a	dr5.mwem0.sa1.47645to50827	/ax ao l/	120	700	1000	2500	2900
o	dr6.mbma1.sx324.22512to25933	/ch ao zh/	120	550	950	2300	3100
/	dr6.mkes0.si1883.40898to44200	/s ao dcl/	140	650	1000	2500	3100
	dr6.msjk0.sx426.5932to9065	/m ao th/	140	650	950	2500	2900
	dr6.msvs0.sa1.49155to52556	/q ao l/	130	600	850	2700	3100
	dr2.mdem0.sx428.19863to22684	/m iy pau/	120	550	2000	2500	3100
	dr2.mjar0.sx278.17550to21347	/t iy dcl/	160	400	2450	3050	3500
/	dr2.mprb0.sx305.12449to15560	/ch iy z/	110	350	2400	2900	3100
i	dr2.mzmb0.sx266.33960to36760	/n iy dcl/	90	500	2200	2800	3100
y	dr8.mbsb0.si1353.14240to16720	/n iy dcl/	140	450	2200	2600	3100
/	dr8.mjth0.si1296.19927to22475	/l iy v/	100	400	2600	3000	3100
	dr8.mrre0.sx164.15630to18280	/t iy pau/	100	400	2400	2850	3100
	dr8.mtcs0.sx172.29117to32209	/ay iy v/	120	450	2400	3100	3500
	dr1.mcpm0.sx204.29585to33170	/l uw z/	130	400	1250	2600	3100
	dr1.mpgh0.sx294.24856to27948	/z uw z/	120	450	1350	2600	3100
/	dr1.mpgr0.sx330.17030to19802	/p uw n/	100	500	1100	2600	2900
u	dr1.mwbt0.sx383.30080to34280	/p uw dcl/	120	400	1000	2400	2700
w	dr2.mdlb0.sa1.24280to26520	/s uw tcl/	130	400	1550	2400	3100
/	dr2.mjar0.si1988.14538to17640	/ux uw f/	130	400	850	2400	3100
	dr2.mjhi0.sx338.18381to21765	/g uw n/	90	500	1000	2400	3100
	dr2.mrjm0.sx418.19800to23066	/z uw hh/	150	450	1550	2400	3100
	dr1.mjeb1.si837.20467to24240	/y ux z/	120	400	1600	2400	3100
	dr1.mkls0.sa1.20880to23720	/s ux tcl/	130	400	1600	2200	3100
/	dr1.mreb0.sx295.45430to48353	/t ux h#/	100	400	1650	2200	3100
u	dr1.mtrr0.sx378.26920to29911	/n ux aa/	130	500	1350	2350	3100
x	dr4.mdma0.sx350.19630to23603	/k ux n/	150	500	1550	2200	3100
/	dr4.mjls0.sx106.5160to9080	/hv ux dcl/	200	400	2000	2500	3100
	dr4.mjxl0.sx182.10862to14914	/l ux pau/	110	450	1700	2500	3100
	dr4.msfh0.sa1.29820to34280	/s ux tcl/	150	450	1750	2600	3100

Table B.2: Results of the evaluation of the CSTR formant frequency estimator for male speech. See Table B.1 for an explanation of the table headings.

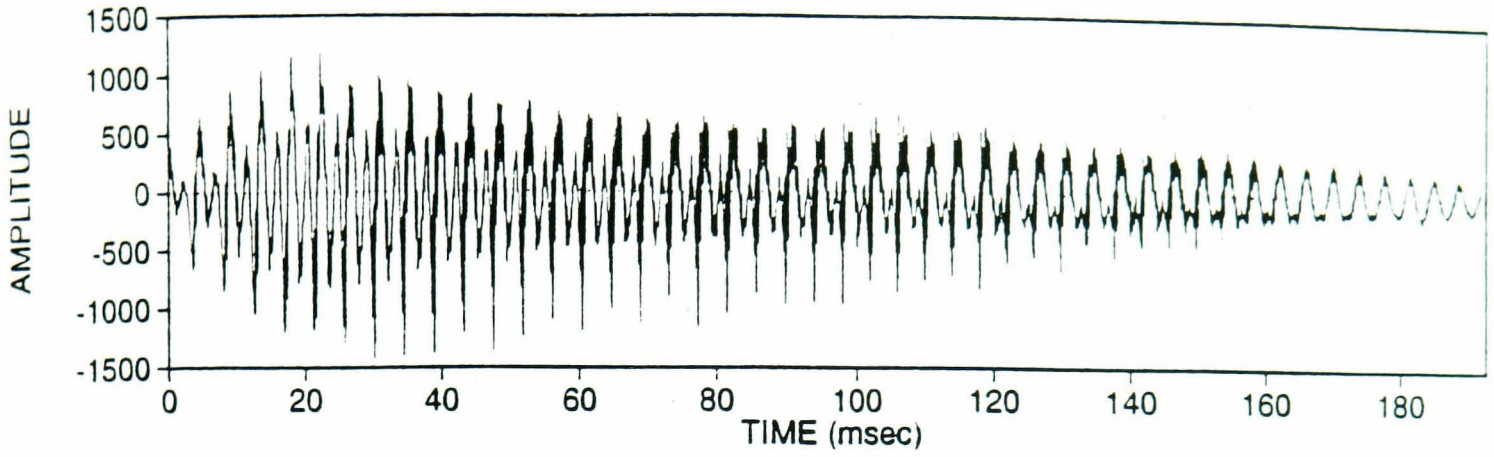


Figure B.21: Acoustic waveform of dr7.ftlh0.si1639.7880to10933. an example of a slice in which the amplitude decreases throughout the slice. See Figure B.22 for the spectrogram of this slice.

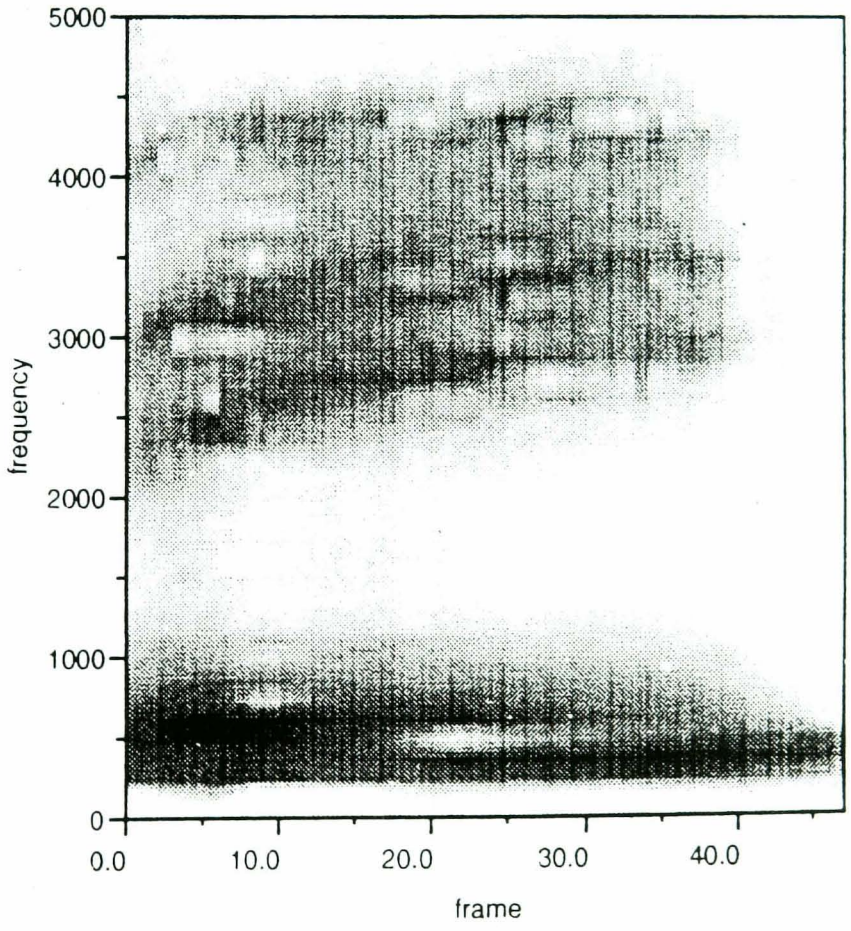


Figure B.22: Spectrogram of dr7.ftlh0.si1639.7880to10933. an example of falling formant energy, mirroring the falling amplitude of the acoustic waveform. There are a number of examples of where the formants fade through a slice, and yet the estimator is able to estimate the frequency values accurately. Note also that the estimater was able to following the rising F_2 and F_3 accurately, although the u-setting had to be raised to 3700Hz to take in to account the high value of F_3 in the second half of the slice.

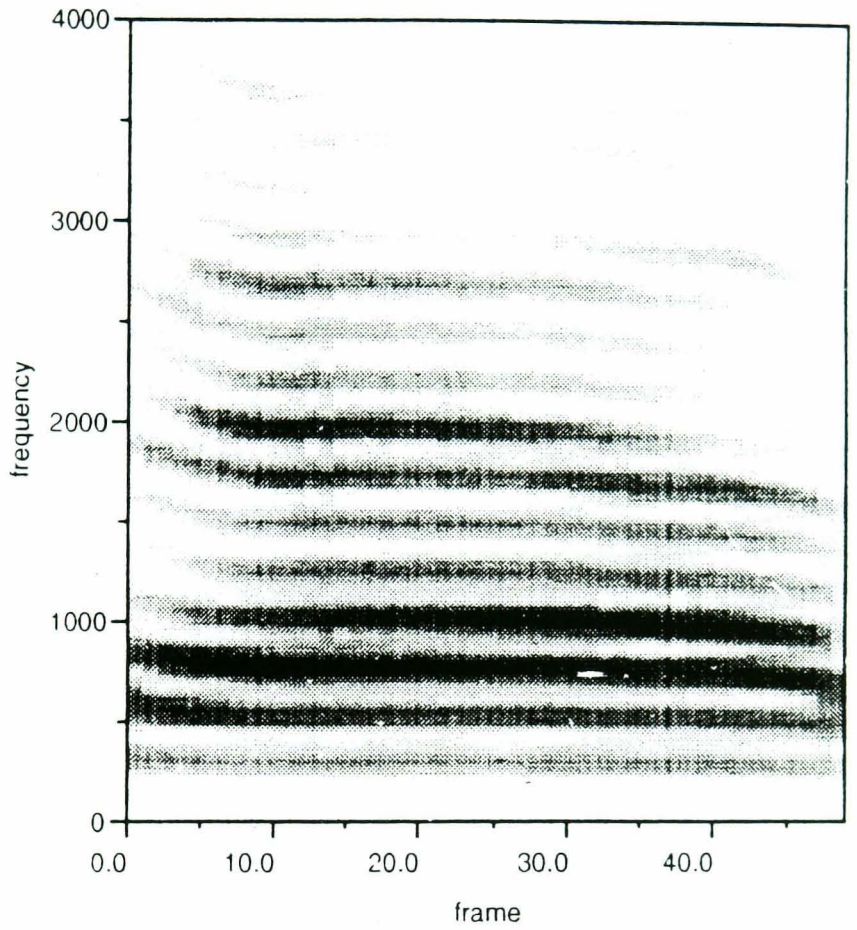


Figure B.23: Spectrogram of dr1.faks0.si1573.70910to74200, an example of how the formant frequencies of speech with a high F_0 can be spread over two harmonics. Consider F_2: for the first two thirds of the slice the formant consists of a dark, broad harmonic above a lighter one, the darker harmonic gradually fading over the second half as the formant falls in frequency. The estimator correctly picks a point in between the two harmonics, tracking F_2 as it falls from 1890Hz at the beginning, to 1850Hz in the middle, to 1600Hz at the end. Similarly, F_1 rises from 750Hz at the start of the slice, to 900Hz at the centre, to 800Hz at the end.

1. F_3 will be close to, if not in excess of, the upper limit of the search space. Often, if F_3 is within 100-200Hz of the u-setting, the estimator will underestimate or even miss the formant altogether.

2. With the energy of formants being defined over the harmonic structure, the formants of female speakers have less definition because their harmonics are more widely spaced. Consequently the estimator has greater difficulty in locating female formants.

3. Because F_2 and F_3 are relatively close together, the estimator sometimes mistakes them for a second formant spread across two (or more) harmonics. This becomes particularly so for cases where F_4 has crept into the search space, and is mistaken for F_3.

It was hoped that the smaller window size would make the formants more visible to the estimator: when using a window size of 512 samples, more of the slice's periodic information is captured, producing spectrograms in which the harmonics are more prominent than the formants. However, the results for the two window sizes were very similar, lending weight to Crowe's assertion that the estimator is reasonably insensitive to adjustment of the parameters of the FFT. In fact, when the estimator was run with an FFT window size of 128 samples, there was a marked deterioration in performance.

EVALUATION 3: Assessment of the estimator's performance with female speech

Two assessments of the u-setting were carried out using examples of female speech, on the phones /iy/ and /ae/. In both cases, the data was checked for speakers with relatively low and high F_0s.

The first assessment used the speakers **faks0** and **fdac1** again. What became clear was that, while the estimator could be made to peform accurately, different slices required different u-settings, and that this problem existed for both speakers. If the u-setting was too low, then F_3 could be left out of the search space; on the other hand, if it was set too high, then F_3 could be overestimated, or, even worse, F_4 could be included in the search space. In the latter case, the estimator often merged F_2 and F_3 into one formant, leaving F_4 to be treated as F_3. Regarding the problem of accurate measurment of low first formants, the estimator appeared to perform well for both speakers, and in agreement with the hand estimates. However, with mean frequencies for /iy/ of 480Hz and 450Hz for **faks0** and **fdac1** respectively, this is substantially in excess of the mean F_1 of 310Hz for /iy/ reported by Peterson & Barney (1952).

In the second assessment, the speakers **fsem0** and **fcdr1** were used, with mean SFFs of 250Hz and 184Hz respectively. This assessment highlighted the problem of an automatic formant frequency analysis of different speakers. Hand measurements showed **fsem0** had mean formant values of 850Hz, 2000Hz and 3200Hz, while **fcdr1**'s formants were at 700Hz, 1800Hz and 2300Hz – there is a 900Hz difference in mean F_3. While the estimator could be made to produce accurate results for speakers with high or low F_0s by adjusting the u-setting, for half of **fsem0**'s slices the upper search space limit had to be reset to 3700Hz. This would cause a substantial overestimation of F_3 for speakers such as **fcdr1** who have a far lower F_3.

Checking the formant frequency data

The checking of the results in the formant frequency data files was carried out using the

shell script **check_formant_diffs**. This shell worked in a similar way to **check_peak_diffs** in that it looked at the formant frequencies of successive analysis frames for each slice, and output those that differed by more than 100Hz to a report file. Typical output from **check_formant_diffs** for a slice is as follows:

```
:24:
 F2: Line 846, diff = -143
 F3: Line 847, diff = -104
 F2: s.d. = 205
```

Thus for slice number 24, the estimates of F_2 (from the frame on line 846 of the relevant formant frequency data file) and F_3 (on line 847) differ from the estimates of the previous frame by more than 100Hz. Furthermore, the s.d. of F_2 for the slice is greater than 100Hz.

A total 74 slices were checked for the female productions of /ux/, 55 slices for the male /ux/, and 33 slices for the male /ae/. Unfortunately, time constraints meant it was not possible to examine the other phones in more detail. However, based on the evaluations and the experience gained from the checking described above, it was possible to make the following final remarks on the performance of the CSTR formant frequency estimator.

One confusing aspect of these evaluation exercises is the lack of agreement between the hand-measured formant frequencies reported here and the values reported in the literature. While the results for /aa/, /ae/ are similar to those reported by Peterson & Barney (1952 – see Table 3.10) and Fant (1959 – see Table 3.11), those for /iy/ for females, the first formant of /iy/ for males, and the first and second formants of /uw/ and /ux/ for both sexes are very different. Some of the discrepancy between the F_1 results can probably be explained by the difficulty in obtaining accurate hand measurements from the spectrogram viewing tool. The resolution of the spectrograms made it difficult to distinguish between the voice bar (the dark line across the bottom of a spectrogram representing the fundamental frequency component) and the first formant. However, particularly for the male speakers, the hand measurements of F_1 appeared to be a fair reflection of the true values. Yet it is difficult to square the hand measured results for, say, /iy/ (400-450Hz for the males, 450-600Hz for the females), with the Peterson & Barney means (270Hz for the males, 310Hz for the females). There is even less reason for the discrepancies in the higher formants, as it was much easier to estimate their values. What is interesting in the light of these discrepancies is the fact that the estimator's output reflected the hand measured values rather than the previously reported values. It is therefore possible that the estimator does in fact produce accurate results.

Although the estimator could be made to produce results that reflected the hand measurements, this was often at the expense of rerunning the estimator with a different u-setting. While this is to some extent to be expected, given the variation in mean F_3 for different phones reported by Peterson & Barney and Fant, it does create the problem of finding optimum u-settings for an automatic analysis involving different phone types. To further complicate matters, between-speaker differences in formant values can make a nonsense of any attempt to do this. As an illustration, consider the hand measured formant values in Tables B.1 and B.2. Five of the values for F_3 of /ux/ for the females range from 2450-2650Hz, while the other three range from 2950-3050Hz. Similarly, for the male productions of /aa/, five of the F_3 measurements range from 2200-2500Hz, and the rest fall in the range 2800-3000Hz. How does one reconcile the desire to include all possible F_1 and F_3 values, whilst still excluding F_0 and F_4? And how does one avoid an overestimation of, in particular, F_3 due to an over-large search range?

Using the results of the evaluations, and in particular the results reported in Tables B.1

and B.2, an attempt was made to define optimum u-settings for the estimator. These were as follows:

- /aa/ – For the female speakers, u=3100Hz seemed adequete, with F_3 hand-measured at 2600-2900Hz. For the males, while F_3 is generally in the region 2200-2500Hz, it can rise to 3000Hz; therefore, u=3100Hz should be used.

- /ae/ – With F_3 about 2800-3200Hz, a u-setting of 3500Hz is suggested for the females. For the male speakers, although 25 of the slices checked showed accurate results for u=3100Hz, 6 had u reset to 2900Hz to avoid a low F_4; with F_3 being between 2400-2700Hz, a u-setting of 2900Hz is recommended.

- /ao/ – The female F_3 is either around 2500Hz, or 3000-3200Hz, indicating a u-setting of 3300Hz (set to avoid the influence of F_4). A low F_4 caused problems for some of the male slices, suggesting a u-setting of 2900Hz to accommodate F_3 values of 2500-2700Hz.

- /iy/ – F_3 is between 3000-3300Hz for the females, therefore u=3500Hz should be used. For the males, F_3 is in the region 2600-3000Hz, indicating a u-setting of 3300Hz (to avoid under-estimation of the third formant when it is close to 3000Hz).

- /uw/ – F_3 is around 2700-3000Hz, suggesting a u-setting of 3300Hz for the female speakers. For the males, F_3 is around 2400-2600Hz, therefore u=2900 should be used.

- /ux/ – Although two thirds of the slices checked showed accurate results for the female speakers using u=3100Hz, a number of slices had a third formant at 2900-3100Hz, suggesting a u-setting of 3300Hz. With a low F_4 in evidence for a number of the male slices, and with F_3 generally falling between 2200-2500Hz, a u-setting of 2900Hz is recommended.

To summarise, the most striking conclusion to be drawn from this evaluation is that an automatic analysis of formant frequencies is very difficult to carry out accurately. There are two main reasons for this. One is that the mean formant frequencies of different phones are so varied, implying that different search spaces would be optimal for different phones. Even for male speakers it is not possible to define a single search space to contain only the first three formants, while excluding the fundamental frequency component at the lower end of the search space and F_4 at the upper end. The second problem is that individual speakers do not readily conform to group means. In other words, while the majority of speakers may be close to the mean for a particular phone, some speakers will have values that are substantially different. Thus even when a search space is defined to include most of the expected values, the formants of some speakers will inevitably be wrongly estimated. Whilst this may have little effect on the mean formant frequency values, especially when analysing the speech from a large number of speakers, the results are impoverished by not being able to describe the complete range of expected values for the formants, and analyses of between- and within-speaker variation in the formant frequencies are rendered less meaningful.

E. Description of the shell scripts used to check the updating

There now follows a description of the shell scripts used to check the updating of the ACOUSTIC-PHONETIC VARIABLE DATA FILES and SLICE NAMES FILES following any alteration to the results.

check_problem_slices ascertains whether the SLICE NAMES FILES have been properly updated, for both reduced and removed slices. Care was taken to report these two actions accurately and in a consistent format in the PROBLEM_SLICES file, enabling **check_problem_slices**'s updating to be carried out by pattern-matching. The script performs two checks, the first of which involves seeing if the corrected SLICE NAME entered in PROBLEM_SLICES appears in the appropriate SLICE NAMES FILE:

```
# Match the string "* Slice reduced" in the PROBLEM_SLICES file;
# if a match occurs, set slice_name to be the name of this slice
# (which appears in the first field of PROBLEM_SLICES).
for slice_name in `grep "\* Slice reduced" problem_slices | \
  awk '{print $1}'`
do
  # Set the slice's dialect, so the correct SLICE NAMES FILE
  # is accessed.
  dialect=`echo $slice_name | awk '{FS=".";print $1}'`

  # If slice_name is found in the SLICE NAMES FILE, set a flag.
  # If at the end of the search, no flag has been set,
  # output slice_name.
  awk '$2 == "'$slice_name'" {n=99}
    END{if(n!=99) {print "'$slice_name'"}
  }' $path2/$sex.$dialect.names
done
```

where $path2 is the path to the SLICE NAMES DIRECTORY; $sex is the sex of the speaker.

The second check involves seeing if the removed SLICE NAME reported in PROBLEM_SLICES appears in the SLICE NAMES FILE:

```
# Match the string " **" in the PROBLEM_SLICES file; if a match
# occurs, set slice_name to be the name of this slice (which appears
# in the first field of PROBLEM_SLICES).
for slice_name in `grep " \*\*" problem_slices | awk '{print $1}'`
do
  # Set the slice's dialect, so the correct SLICE NAMES FILE
  # is accessed.
  dialect=`echo $slice_name | awk '{FS=".";print $1}'`

  # If slice_name is found in the SLICE NAMES FILE,
  # output slice_name.
  awk '{
    if($2 == "'$slice_name'") {print "    '$slice_name'"}
  }' $path2/$sex.$dialect.names
done
```

where $path2 is the path to the SLICE NAMES DIRECTORY; $sex is the sex of the speaker.

check_data_files ascertains whether the ACOUSTIC-PHONETIC VARIABLE DATA FILES have been properly updated and are in the correct format. When new slice results need to be

placed in a data file, it is possible to forget to update the slice's identification number, or even sometimes to remove the old results. As the example ACOUSTIC-PHONETIC VARIABLE DATA FILE output in Figures B.5 to B.7 show, the line containing the slice's identification number should have the word 'samples' in the second field; if this is not the case, then the file has not been updated properly:

```
# Find all instances of the string 'samples' in the
# ACOUSTIC-PHONETIC VARIABLE DATA FILE.
grep "samples" $data_file | \
awk '{
  # Keep a count of the no. of occurrences of "samples".
  n++
  # Output lines NOT containing 'samples' in the second field.
  if ($2 != "samples") {print $0}
# Output the filename and the number of occurrences of "samples".
}END{printf "Number of slices in '$name_of_file'\t\t%4d\n", n}'
```

where **$data_file** is the path to the target ACOUSTIC-PHONETIC VARIABLE DATA FILE, and **$name_of_file** is the name of that file.

The script also outputs the number of lines in the corresponding SLICE NAMES FILE for comparison with the number of slices found in the ACOUSTIC-PHONETIC VARIABLE DATA FILE.

Lastly, the SLICE NAMES FILES and ACOUSTIC-PHONETIC VARIABLE DATA FILES need to have their identification numbers updated (to take into account the slices that have been removed), and the SLICE MEANS FILES need to be reformed. This is achieved in the scripts **renumber_names_files** and **renumber_data_files** by reprinting all lines containing the identification numbers, discarding the number in the first field and replacing it, incrementally.

B.7 Statistical analysis of the slices and speakers

Statistical analyses can be performed using the SLICE MEANS or the SPEAKER MEANS. The procedures for carrying out the analysis are different, and are described separately below. For the SLICE MEANS, statistics can be performed on all of the slices (to obtain overall speaker sex characteristics for a particular ACOUSTIC-PHONETIC VARIABLE), as a function of an ANALYSIS VARIABLE (i.e. as a function of age, ethnic group, dialect, education, height or phone), and as a function of phonetic context (to see how a vowel phone behaves in different phonetic contexts, for example, how /aa/ vowels are influenced by immediately following closures). For the SPEAKER MEANS, statistics can be performed for all of the speakers, and as a function of an ANALYSIS VARIABLE.

A. Performing statistical analyses on the SLICE MEANS

The computation of statistics based on dividing the analysed data by dialect region and vowel phone is relatively easy to perform, the reason being that the output database is already organised by dialect and phone, in the form of the SLICE MEANS FILES. The computation of statistics for the ANALYSIS VARIABLES age, ethnic group, education and height, and for phonetic context requires some preliminary organisation of the SLICE MEANS, and is described below. Simple statistics can then be produced using the shell script **full_stats**, and is dealt with later.

Analysis as a function of ANALYSIS VARIABLE

The procedure for obtaining age, ethnic group, education and height statistics involves forming the SLICE MEANS and SLICE NAMES SUPER-FILES (complementary files containing all the results for a particular ACOUSTIC-PHONETIC VARIABLE, and all the labels required to identify their owners), and using them to set up files of SLICE MEANS for the specified ANALYSIS VARIABLE. The SLICE MEANS SUPER-FILE is formed using the shell script **list_all_slice_means**, which essentially concatenates all the SLICE MEANS FILES into a single file, preceded by an incremental identifier. The relevant code is:

```
awk '{n++; printf ":%d: %s %s %s %s %s %s\n",n,$2,$3,$4,$5,$6,$7}'
  *_f0_all.dialect/$sex.dr*.f0_means >> all_f0_means
```

where **$sex** is the speaker sex.

The wild cards (i.e. *) in the path to the input file ensure every SLICE MEAN is sent to the super-file, with the files being accessed in alphabetical order. The SLICE NAMES SUPER-FILE is formed using the shell script **list_all_slice_names**, and is used to identify the owner of each entry in the SLICE MEANS SUPER-FILE, via its identifier. The UNIX code is essentially the same as for **list_all_slice_means**, the only difference being the input and output pathnames:

```
awk '{n++; printf ":%d: %s %s %s %s %s %s\n",n,$2,$3,$4,$5,$6,$7}'
  *_names_all.dialect/$sex.dr*.names >> all_f0_names
```

The shell script **extractor** is used to set up the SLICE MEANS directories for a particular ANALYSIS VARIABLE. The directory structure is in the same format as the structure for dialect region used to perform the signal processing analysis (see Section B.2, and Figure B.1). Each ANALYSIS VARIABLE group is given its own file. Thus for ethnic group, the female speakers' slice results for fundamental frequency for the phone /ae/

are stored in the directory **ae_f0_all.ethgrp** under the filenames **f.black.f0_means** and **f.white.f0_means**. For each group, **extractor** uses the SPEAKER NAMES FILE for the group to search the SLICE NAMES SUPER-FILE for slices uttered by the speakers comprising the group. The corresponding entries in the SLICE MEANS SUPER-FILE, located using the incremental identifiers, are output to files for each group. The relevant code is as follows:

```
for target_file in $anal_var_dir/$sex.*.names
do
  # Extract identifier of group being analysed from the pathname
  # (to help form part of output filename).
  name1=`echo $target_file | awk '{FS="/";print $8}' | \
    awk '{FS=".";print $2}'`

  # This awk gets each target SPEAKER NAME in turn, and assigns it
  # to the variable TARGET by command substitution.
  for target_speaker in `awk '{print $1}' $target_file`
  do
    # This awk matches the target SPEAKER NAME with its counterpart
    # in the SLICE NAMES SUPER-FILE, and assigns the counterpart's
    # number to the variable TARGET_NUMBER by command substitution.
    for target_number in `grep "$target_speaker" all_names | \
      awk '{print $1}'`
    do
      # This awk matches the target SLICE NAME's number with its
      # counterpart in the SLICE MEANS SUPER-FILE and outputs
      # that SLICE MEAN.
      grep "$target_number" all_"$apv"_means >>
        "$phone"_"$apv"_all.$anal_var_dir/$sex.$name1."$apv"_means
    done
  done
done
```

where `$anal_var_dir` is the name of the ANALYSIS VARIABLE DIRECTORY, `$sex` is the speaker sex, `$apv` is the ACOUSTIC-PHONETIC VARIABLE identifier, `all_"$avp"_means` is the filename of the SLICE MEANS SUPER-FILE, `all_names` is the file name of the SLICE NAMES SUPER-FILE. Note how the output file name is unique for each ANALYSIS VARIABLE group.

Analysis as a function of phonetic context

The procedure for the analysis of phonetic context is very similar to the above, also utilising the slice super-files. Again the procedure involves running an extraction shell script, called **extractor_context**, to set up SLICE MEANS FILES, which can subsequently have statistical analyses performed upon them. **extractor_context** differs from **extractor** only in one line in the main body of the script, the line which finds the identifier of the target SLICE NAME in the SLICE NAMES SUPER-FILE:

```
for target_number in `grep "$target_speaker" all_names | \
  grep "$context" | awk '{print $1}'`
```

where `$target_speaker` is the target SPEAKER NAME from the SPEAKER NAMES FILES, `$context` is the name of the phonetic context to be matched, `all_names` is the file name of the SLICE NAMES SUPER-FILE.

The `context` argument, i.e. the phone's phonetic context, is defined at the beginning of the script. For example, to set up SLICE MEANS FILES for the context /d aa r/, the line `context="/ d aa r /"` should be used. A specific sentence on the TIMIT database can be targetted by replacing the pattern for the **grep** by `"$target_speaker.$sentence"`. Thus by including the line `sentence=sa1` at the beginning of the script, only the calibration sentence **sa1** is targetted by the script.

extractor_context works by using a **grep** to locate, in the SLICE NAMES SUPER-FILE, all the target sentences uttered by the target speaker. Each SLICE NAME contains phonetic context information (i.e. the phonetic context of the phone represented in the slice), so that when the SLICE NAMES are passed through the second **grep** it locates only those phones uttered in the specified context. The rest of the script works in the same way as **extractor**. Note that by using the SPEAKER NAMES FILES to provide the list of speakers, an analysis of phonetic context can be carried out as a function of one of the ANALYSIS VARIABLES.

Performing simple statistics

The new SLICE MEANS FILES, produced in the same format as the original SLICE MEANS FILES ordered by dialect, can now have statistical analyses performed on them using the same, or very similar, shell scripts. The computation of simple statistics is now described.

full_stats computes the means and s.d.s of the specified data. It also outputs the number and range of data. The statistics are computed for each ANALYSIS VARIABLE group, and for the total sample. For example, for the age ANALYSIS VARIABLE, statistics are computed for the six age groups and then for all the groups taken together. Firstly, the data for each group (i.e. the mean of each slice in the group) is extracted, and their sum and sum of squares are stored in a temporary file along with the number of data summed:

```
awk '{sum+=$2; sum_squares+=$2*$2}
  END{printf "%f  %f %d '$sex' '$apv'\n", sum, sum_squares, NR}
' $data_file >> $tf1
```

where `$sex` is the speaker sex, `$apv` is the ANALYSIS VARIABLE group identifier, `$datafile` is the path name of the SLICE MEANS FILE, and `$tf1` is the name of a temporary file.

The computation of the statistics is handled by a single **awk** statement, the main body computing the means for each group, the END statement computing the overall mean:

```
awk '
{       ### First carry out the analysis for each ANALYSIS
        ### VARIABLE group ...
  # Compute running totals for the whole sample.
  sum_total+=$1
  sum_squares_total+=$2
  n_total+=$3
  # Compute mean and s.d. for the group.
  av=$1 / $3
  if ($3 == 1) # If there is only one datum, force s.d. to be zero.
    sd=0.0
  else
    sd=sqrt(($2-$3*av*av)/($3-1))
  # Output "sex.apv number_of_data mean ( sd )"
```

```
  printf "%s.%s\t  n= %4d   %.2f ( %.2f )\n", $4, $5, $3, av, sd}
END {     ### Now carry out the analysis for the whole sample ...
  # Compute mean and s.d.
  av=sum_total/n_total
  sd=sqrt((sum_squares_total-n_total*av*av)/(n_total-1))
  # Output "Totals: number_of_data mean ( sd )"
  printf "  Totals: n= %4d   %.2f ( %.2f )\n\n", n_total, av, sd}
' $tf1 >> $stats_file
```

where **$tf1** is the name of a temporary file, and **$stats_file** is the path name of the output file containing the statistics.

An example of the output, for a fundamental frequency analysis of female /uw/ phones as a function of dialect region, follows:

```
f.dr1    n=   35   210.11 ( 38.27 )
f.dr2    n=   22   217.82 ( 27.59 )
f.dr3    n=   13   206.50 ( 29.58 )
f.dr4    n=   16   221.58 ( 45.54 )
f.dr5    n=   28   215.99 ( 32.02 )
f.dr6    n=   12   215.92 ( 40.84 )
f.dr7    n=   22   228.38 ( 39.36 )
f.dr8    n=    3   229.77 ( 23.78 )
  Totals: n=  151   216.74 ( 35.95 )
```

B. Performing statistical analyses on the SPEAKER MEANS

For the statistical analysis of the SPEAKER MEANS, certain structures must be in place, namely the SPEAKER NAMES FILES and the SPEAKER MEANS FILES. When formed, the SPEAKER NAMES FILES are organised into directories as illustrated in Figure B.24. The files are accessed by the shell scripts **do_variable_analysis** and **extractor** to perform statistical operations on the SPEAKER MEANS and SLICE MEANS respectively, as a function of an ANALYSIS VARIABLE. The SPEAKER MEANS FILES form the core data for the statistical analysis of the speakers, and as such constitute a database of the means of each ACOUSTIC-PHONETIC VARIABLE for each speaker. Simple statistics can then be performed on the SPEAKER MEANS. The shell scripts **do_phone_analysis** and **do_variable_analysis** carry out analyses as functions of phone and ANALYSIS VARIABLE respectively.

Forming the SPEAKER NAMES FILES

Information about the attributes of the TIMIT speakers is contained in the file **trnspkr.log**. The data for each ANALYSIS VARIABLE is collated into groups automatically by shell scripts. The different representations of data in **trnspkr.log** require processing by different scripts, which are described separately below. The group data is used to form the SPEAKER NAMES FILES.

For the ANALYSIS VARIABLES ethgrp, dialect and education this is relatively easy, as the relevant information in **trnspkr.log** is simply represented. Referring to the extract from **trnspkr.log** in Figure 4.8, the identifiers for speaker ethnic group, dialect region and educational level are listed in columns 9, 5 and 10 respectively, while the speaker's initials are in column 2. **awk** is used to get the speaker initials and speaker attribute information from the relevant column (or field, to use **awk** notation), which are then used to form the SPEAKER NAMES FILES. For example, consider the formation of the dialect SPEAKER

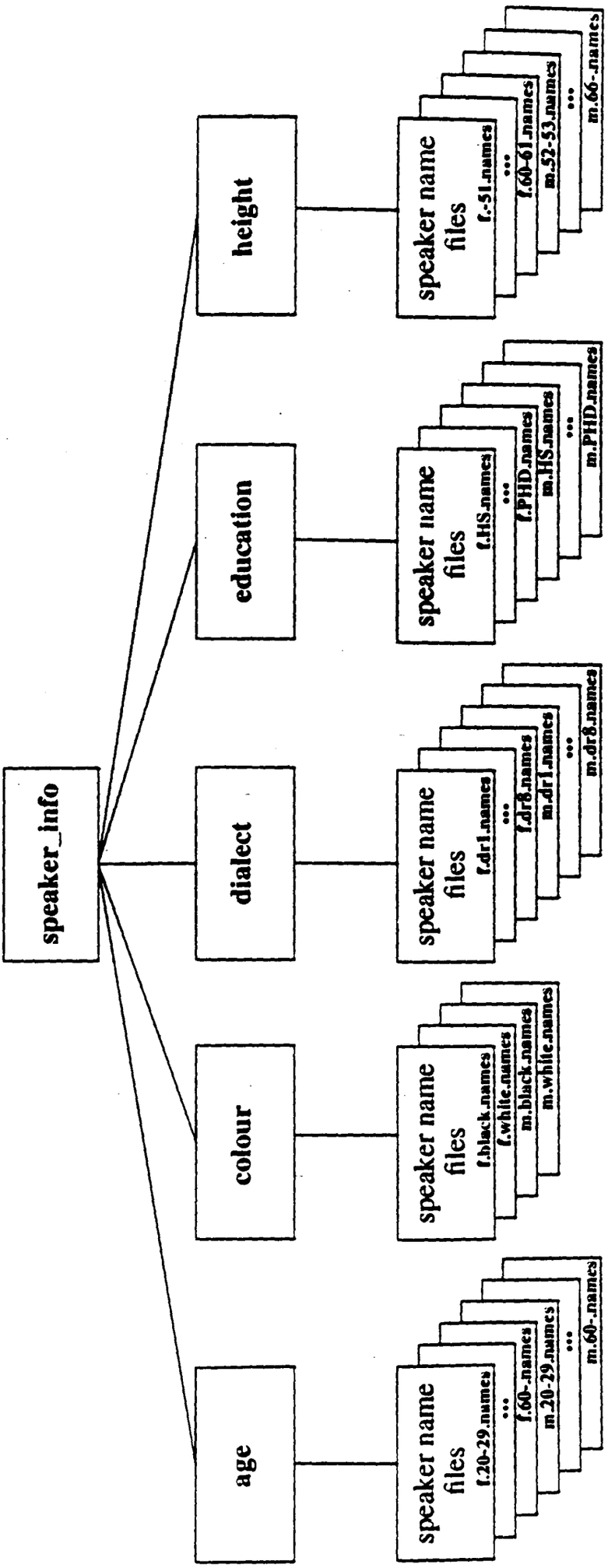


Figure B.24: The directory structure for the ANALYSIS VARIABLES age, ethgrp, dialect, education and height. Actual directory or file names are in bold.

NAMES FILES, using the shell script **form_speaker_names.dialect**. In the following piece of UNIX code, the **awk** first locates every line in **trnspkr.log** containing the identifiers of both the relevant sex and the dialect. Note that in **trnspkr.log**, the speaker initials are in uppercase letters, which must be converted to lowercase to form the SPEAKER NAMES. First, an uppercase version of the SPEAKER NAME is produced. This is not entirely trivial as some speaker initials have no numbers after them, while others do. For example, CMM is converted to **FCMM0**, while JDM1 is converted to **MJDM1**. The **awk** command **length** is used to test the length of the speaker initials string, then the UNIX command **dd** is used to convert the output from the **awk** from upper to lowercase:

```
# Define the parameter lists.
dialect_type="1 2 3 4 5 6 7 8"
sex_type="f m"

# For each sex in turn ...
for sex in $sex_type
do
  # Set the capitalised version of the sex parameter
  # (as TRNSPK.LOG uses capital letters).
  if test "$sex" = f
  then
    SEX=F
  else
    SEX=M
  fi
  # For each dialect region in turn ...
  for dialect in $dialect_type
  do
    awk '$4 == "'$SEX'" && $5 == "'$dialect'" {
      if (length ($2) == 3) # If string length is 3, add a zero ...
        printf "'$sex'%s0\n", $2
      else # else, initials already suffixed ...
        printf "'$sex'%s\n", $2
    }' trnspkr.log | \
    dd conv=lcase >> speaker_info/dialect/$sex.dr$dialect.names
  done
done
```

The output, files of SPEAKER NAMES for each dialect, are stored in the dialect ANALYSIS VARIABLE DIRECTORY (see Figure B.24).

For the ANALYSIS VARIABLES age and height, forming the SPEAKER NAMES FILES is more difficult because of the way this information is represented in **trnspkr.log**. The speaker's age must be calculated from the speaker's birthdate and the recording date of the speech material, held in columns 3 and 7 of **trnspkr.log**, and the speakers must then be sorted into age groups. Note that the format of the dates follows the U.S. convention (i.e. month/day/year). In the following piece of code, the first **awk** locates all the speakers of one sex and outputs the recording and birth dates and the speaker's initials, such that all the data is separated by backslashes. For example, for the speaker PGL (see speaker 17 in Figure 4.8), the output would be `1/20/86/6/29/61/PGL`. This is piped to the second **awk** which computes the speaker's age from the years in the recording and birth dates. Note that the script takes in to account cases where the speaker had not had a birthday

by the time of the recording date. For example, PGL's age at recording was 24, not the 25 one gets from subtracting the birth year from the recording year:

```
# AWK1: reorder the data in trnspkr.log, so that standard output
# contains lines in the form:
#   recording-date/birth-date/speaker-initials
awk '$4 == "'$SEX'" {
  printf "%s/%s/%s\n",$3,$7,$2
}' trnspkr.log | \
# AWK2: Compute each speaker's age, and send output to the
# temporary file.
awk '{
  FS="/" # Set file separator to a backslash
  if ($1 < $4)
    # If recording month < birth month,
    # speaker age = recording year - birth year - 1 ...
    print $7, $3-$6-1
  else
    # else, speaker age = recording year - birth year.
    print $7, $3-$6
}' | dd conv=lcase > temp_file1
```

where $SEX is the speaker sex in capitalised form; temp_file1 is the name of a temporary file.

After conversion from uppercase to lowercase, the output from the second **awk**, comprising a SPEAKER NAME and the speaker's age, is sent to a temporary file from which the speakers are sorted into age groups as follows:

```
path1=speaker_info/age

awk '$2 < 30 {print $1}' temp_file1 > $path1/$sex.20_29.names
awk '$2 > 29 && $2 < 40 {print $1}' temp_file1 > $path1/$sex.30_39.names
awk '$2 > 39 && $2 < 50 {print $1}' temp_file1 > $path1/$sex.40_49.names
awk '$2 > 49 && $2 < 60 {print $1}' temp_file1 > $path1/$sex.50_59.names
awk '$2 > 59 {print $1}' temp_file1 > $path1/$sex.60_.names
```

where $sex is the speaker sex.

To form the height SPEAKER NAMES FILES, a special version of **trnspkr.log** is used, where the notation used in column 8 for feet (a single quote, ') and inches (a double quote, ") is replaced by colons (:). The reason for this is that **awk** does not like to pattern match against quotes, which it uses as special characters. Thus, for example, the line for speaker 12 in Figure 4.8 becomes:

```
12   CMJ   1/16/86  M    6    1X   2/06/58  5:7:  WHT   MS
```

The categorisation then proceeds in much the same way as for the speaker ages, in the following code, the first **awk** locates all the speakers of one sex and prints out the speaker's height and initials, such that all the data is separated by colons. For example, for the speaker CMJ, the output would be 5:7:CMJ. This is piped to the second **awk** which removes the colons, and sends the output to a temporary file:

```
# AWK1: reorder the data in trnspkr.log, so that the standard output
# contains lines in the form:
#   height-in-feet:height-in-inches:speaker-initials
awk '$4 == "'$SEX'" {
  print $8$2
}' trnspkr.log | \
# AWK2: Output the data to the temporary file in the form:
#   speaker-initials height-in-feet height-in-inches
awk '{FS=":";printf "%s %s %s\n", $3, $1, $2}' | \
dd conv=lcase > temp_file1
```

where $SEX is the speaker sex in capitalised form; temp_file1 is the name of a temporary file.

The speakers are sorted into height groups as follows:

```
path1=speaker_info/height

awk '$2==5 && $3<2 {print $1}' temp_file1 > $path1/$sex._51.names
awk '$2==5 && $3>1 && $3<4 {print $1}' temp_file1 > $path1/$sex.52_53.names
awk '$2==5 && $3>3 && $3<6 {print $1}' temp_file1 > $path1/$sex.54_55.names
awk '$2==5 && $3>5 && $3<8 {print $1}' temp_file1 > $path1/$sex.56_57.names
awk '$2==5 && $3>7 && $3<10 {print $1}' temp_file1 > $path1/$sex.58_59.names
awk '$2==5 && $3>9 {print $1}' temp_file1 > $path1/$sex.510_511.names
awk '$2==6 && $3<2 {print $1}' temp_file1 > $path1/$sex.60_61.names
awk '$2==6 && $3>1 && $3<4 {print $1}' temp_file1 > $path1/$sex.62_63.names
awk '$2==6 && $3>3 && $3<6 {print $1}' temp_file1 > $path1/$sex.64_65.names
awk '$2==6 && $3>5 {print $1}' temp_file1 > $path1/$sex.66_.names
```

where $sex is the speaker sex.

Forming the SPEAKER MEANS FILES

The SPEAKER MEANS FILES are formed from the SLICE MEANS FILES using the shell **speaker_means**. The shell searches the SLICE MEANS FILES for all the slices uttered by a particular speaker and computes a mean value for all the slices of a particular phone and for all the slices. These means are stored in individual files for each speaker. These SPEAKER MEANS FILES form the core data for the statistical analysis of the speakers. and essentially comprise a database of the frequency characteristics of each speaker.

In the following piece of code from **speaker_means**, the simple statistical analysis is carried out on each phone in turn (the code for the analysis is essentially the same as that for the simple statistics in **full_stats**). The script uses the name of the speaker to **grep** all of that speaker's slices from the appropriate SLICE NAMES FILE, extracting each slice's identifier from the SLICE NAME. This serves as the line number of the SLICE MEANS FILE where the appropriate mean will be found. The speaker's SLICE MEANS for the phone are stored in a temporary file for subsequent statistical analysis.

```
phone_type="aa ae ao iy uw ux"
# For each phone in turn ...
for phone in $phone_type
do
  # Set pathname for SLICE NAMES FILE.
```

```
  names_file="$phone"_names_all.dialect/$sex.$dialect."$apv"_means

  # Set pathname for SLICE MEANS FILE.
  means_file="$phone"_"$apv"_all.dialect/$sex.$dialect.names

  # Match the speaker's name in the names file; if a match occurs,
  # set slice_number to the slice's line number.
  for slice_number in `grep "$speaker_name" $names_file | \
    awk '{FS=":";print $2}'`
  do
    # Match the slice number from the names file with the relevant
    # line number from the means file, and output that line.
    awk 'NR == "'$slice_number'" {
      print $2
      }' $means_file >> $temp_file1
  done

  [ ... Carry out the statistical analysis on the phone ... ]
  [ ...       using the data in the temporary file      ... ]

  # Put ALL the vowel means into one temporary file.
  cat $temp_file1 >> $temp_file2
  rm $temp_file1
done
```

Once the means for each phone have been computed, the mean for all slices is computed using the data in the second temporary file. The shell variables `sex`, `dialect`, `apv` and `speaker_name` can be set using another shell script, as in the following piece of code from **do_speaker_analysis**:

```
apv_types="H1-H2 @H1 @H2"
sex_types="f m"
# For each ACOUSTIC-PHONETIC VARIABLE ...
for apv in $apv_types
do
  # For each sex ...
  for sex in $sex_types
  do
    # For each dialect directory of speaker names ...
    for data_file in speaker_info/dialect/$sex.dr*.names
    do
      # Extract the dialect's identifier from the pathname
      dialect=`echo $data_file | awk '{FS="/";print $8}' | \
        awk '{FS=".";print $2}'`

      # For each speaker ...
      for speaker_name in `awk '{print $0}' $data_file`
      do
        speaker_means $speaker_name $dialect $sex $apv
      done
    done
  done
```

```
done
```

An **awk** is used to extract each SPEAKER NAME in turn from each dialect SPEAKER NAMES FILE. The shell variable **dialect**, extracted from the pathname to the SPEAKER NAMES FILE, is used in the formation of the pathname to the relevant SLICE NAMES FILE.

Simple statistical analysis as a function of phone

Referring to the format of the SPEAKER MEANS FILES (see Section B.1), the phone identifier in the first field of each line allows the statistics for a particular phone to be extracted from every speaker's file of means. Thus an analysis of the frequency characteristics of a particular phone can be performed. Alternatively, using the SPEAKER NAMES FILES established for one of the ANALYSIS VARIABLES, statistics can be performed for a phone as a function of an ANALYSIS VARIABLE (see below).

Simple statistical analysis as a function of ANALYSIS VARIABLE

The shell script to perform statistical analyses of the ANALYSIS VARIABLES, **variable_means**, is very similar to **speaker_means**. Using the SPEAKER NAMES FILES for the relevant ANALYSIS VARIABLE, each speaker's file of frequency characteristics is searched for the line containing the specified phone identifier, which is stored in a temporary file. The statistical analysis is performed in the same way as for **speaker_means**.

```
# Select the analysis variable groups from the
# correct ANALYSIS VARIABLE DIRECTORY.
if test "$variable_directory" = dialect
then
  variable="dr1 dr2 dr3 dr4 dr5 dr6 dr7 dr8"
elif test "$variable_directory" = age
then
  variable="20_29 30_39 40_49 50_59 60_"
elif test "$variable_directory" = height
then
  variable="_51 52_53 54_55 56_57 58_59 510_511 60_61 62_63 64_65 66_"
elif test "$variable_directory" = education
then
  variable="AS BS MS HS PHD"
elif test "$variable_directory" = ethgrp
then
  variable="black white"
fi

# For each analysis variable group in turn ...
for variable_group in $variable
do
  # Set file name for the files of speaker names.
  names_file=speaker_info/"$variable_directory"/$sex.$variable_group.names

  # For each speaker ...
  for speaker_name in `awk '{print $0}' $names_file`
  do
    # Put the stats for the specified vowel into a temporary file.
```

```
    grep "$phone" $path1/$apv.$speaker_name | \
      awk '{print $'$field'}' >> $temp_file1
  done

  [ ... Carry out the statistical analysis on the phone ... ]
  [ ...       using the data in the temporary file      ... ]

  # Put ALL the vowel means into one temporary file.
  cat $temp_file1 >> $temp_file2
  rm $temp_file1
done
```

where path1 is the path to the SPEAKER MEANS FILES.

Again, the shell variables **sex**, **variable_directory**, apv, phone and **field** are set using another shell script, as in the following piece of code from **do_variable_analysis**:

```
apv_types="F1 F2 F3"
phone_types="aa ae ao iy uw ux all"
variable_dir="age ethgrp dialect height"
sex_types="f m"
field_types="2"

# For each sex ...
for sex in $sex_types
do
  # For each ACOUSTIC-PHONETIC VARIABLE ...
  for apv in $apv_types
  do
    # For each directory of ANALYSIS VARIABLE groups ...
    for variable_directory in $variable_dir_types
    do
      # For each field of the SPEAKER MEANS FILES ...
      for field in $field_types
      do
        # For each phone
        for phone in $phone_types
        do
          variable_means $variable_directory $sex $apv $phone $field
        done
      done
    done
  done
done
```

Note the use of the shell variable **field**, which allows a specfic datum to be extracted from each SPEAKER MEANS FILE. Thus by setting the correct field number, an analysis can be carried out on: each speaker's mean (field=2) and s.d. (field=3), the number of slices spoken by each speaker (field=5), and the minimum (field=7) and maximum (field=9) of each speaker's range.

References

Abercrombie D (1967) *Elements of general phonetics.* Edinburgh Univ. Press.

Ainsworth WA (1975) "Intrinsic and extrinsic factors in vowel judgements." In Fant & Tatham (1975:103-13).

Alexander AB (1971) "Aspects of symmetry in male and female laryngeal function." Folia Phoniatrica **23**, 247-51.

Bailey PJ, Summerfield Q (1980) "Information in speech: Observations on the perception of [s]-stop clusters." Journal of Experimental Psychology: Human Perception & Performance **6**, 536-63.

Barry MC (1986) "Synthesising female voice quality: Parameters and test methods." Cambridge Papers In Phonetics and Experimental Linguistics **5**.

Barry WJ, Hoequist CE, Nolan FJ (1989) "An approach to the problem of regional accent in automatic speech recognition." Computer Speech & Language **3**, 355-66.

Beasley DS, Davis GA (ed.s) (1981) *Aging: Communication Processes and Disorders.* Grune & Stratton: New York.

Benjamin BJ (1986) "Dimensions of the older female voice." Language and Communication **6**, 35-45.

Bennett S (1983) "A 3-year longitudinal study of school-aged children's fundamental frequencies." Journal of Speech and Hearing Research **26**, 137-42.

Bennett S, Weinberg B (1979) "Sexual characteristics of preadolescent children's voices." Journal of the Acoustical Society of America **65**, 179-89.

Bickley C (1982) "Acoustic analysis and perception of breathy vowels." MIT R.L.E. Speech Communications Group: Working Papers **1**, 71-82. or MIT Working Papers in Speech Communication **1**, 73-83.

Biever D, Bless D (1989) "Vibratory characteristics of the vocal folds in young adults and geriatric women." Journal of the Voice **3**, 120-31.

Bladon RAW (1982) "Arguments against formants in the auditory representation of speech." In Carlson & Granström (1982:95-102).

Bladon RAW (1985) "Acoustic phonetics, auditory phonetics, speaker sex and speech recognition: a thread." In Fallside & Woods (1985:29-38).

Bladon RAW, Henton CG, Pickering JB (1983) "Testing an auditory theory of speaker normalisation." Preprints of the 10th International Congress of Phonetic Sciences (Utrecht). Or "Outline of an auditory theory of speaker normalisation" Proceedings of the 10th International Congress of Phonetic Sciences (Utrecht).

Bladon RAW, Lindblom B (1981) "Modeling the judgement of vowel quality differences." Journal of the Acoustical Society of America **69**, 1414-22.

Bless DM, Abbs JH (ed.s) (1983) *Vocal fold physiology: Contemporary research and clinical issues.* San Diego: College Hill Press.

Bless D, Biever D, Shaik A (1986) "Comparisons of vibratory characteristics of young adult males and females." Proceedings of the International Congress of the Voice **2**, 46-54.

Boerman IE (1967) "Vertical facial development from ages five to seventeen." MS thesis, Univ. of Michigan.

Bogert BP, Healy MJR, Tukey JW (1963) "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross cepstrum and saphe cracking." In Rosenblatt (1963:209-43).

Bonaventura M (1935) "Ausdruck der Persönlichkeit in der Sprechstimme und im Phonogramm." Archives ges. Psychol. **94**, 501-70.

Borden GJ, Harris KS (1984) *Speech science primer: Physiology, acoustics and perception of speech.* 2nd ed. Williams & Wilkins.

Broad DJ (1976) "Toward defining acoustic phonetic equivalence for vowels." Phonetica **33**, 401-24.

Brown P, Levinson S (1979) "Social structure, groups and interaction." In Scherer & Giles (1979:291-341).

Carlson R, Granström B (ed.s) (1982) *The representation of speech in the peripheral auditory system.* Elsevier Biomedical Press: Amsterdam.

Carlson R, Granström B, Karlsson I (1991) "Experiments with voice modelling in speech synthesis." Speech Communication **10**, 481-9.

Chasaide A (1987) "Glottal control of aspiration and of voicelessness." Proceedings of the 11th International Congress of Phonetic Sciences (Tallinn, Estonia) *6*, 28-31.

Chasaide A, Gobl C (1987) "Cross-language study of the effects of voiced/voiceless consonants on the vowel voice source characteristics." Journal of the Acoustical Society of America Supplement 1 **82**, S116.

Charlip WS (1968) "The aging female voice: Selected fundamental frequency characteristics and listener judgements." Unpublished doctoral disseration, Purdue University.

Chiba T, Kajiyama M (1941) *The vowel - Its nature and structure.* Tokyo. OR (1958) Phonetic Society of Japan: Tokyo.

Cohen JR, Crystal TA, House AS, Neuberg EP (1980) "Weighty voices and shaky evidence: A critique." Journal of the Acoustical Society of America **68**, 1884-6.

Coleman RF (1971) "Effect of waveform changes upon roughness perception." Folia Phoniatrica **23**, 314-22.

Coleman RF, Mabis JH, Hinson JK (1977) "Fundamental frequency-sound pressure level profiles of adult male and female voices." Journal of Speech and Hearing Research **20**, 205-11.

Coleman RO (1971) "Male and female voice quality and its relationship to vowel formant frequencies." Journal of Speech and Hearing Research **14**, 565-77.

Coleman RO (1976) "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice." Journal of Speech and Hearing Research **19**, 168-80.

Coleman RO (1983) "Acoustic correlates of speaker sex identification: implications for the transsexual voice." Journal of Sex Research **19**, 293-306.

Cooley JW, Tukey JW (1965) "An algorithm for the machine calculation of complex Fourier series." Mathematics of Computation **19**, 297-301.

Cooper FS, Peterson GE, Fahringer GS (1957) "Some sources of characteristic vocoder quality." Journal of the Acoustical Society of America **29**, 183.

Cooper M (1987) "Human Factor aspects of voice input/output." Speech Technology **3**(4), 82-6.

Cornut G, Riou-Bourret V, Louis MH (1971) "Contribution à l'etude de la voix parlee et chantee de l'enfant normal de 5 à 9 ans." Folia Phoniatrica **23**, 381-9.

Crelin ES (1973) *Functional anatomy of the newborn.* Yale Univ. Press: New Haven.

Crowe AS (1988) "Generalised centroids: A new perspective on peak picking and formant estimation." Proceedings of the 7th FASE Symposium (Edinburgh, U.K.) *2*, 683-9.

Curry ET (1940) "The pitch characteristics of the adolescent male voice." Speech Monographs **7**, 48-62.

Daniloff RG, Schuckers G, Feth L (1980) *The physiology of speech and hearing.* Prentice-Hall: New Jersey.

Darby JK (ed.) (1981) *Speech evaluation in psychiatry.* Grune & Stratton: NY.

Deem JF, Manning WH, Knack JV, Matesich JS (1991) "Comparison of pitch perturbation extraction procedures with aduolt male and female speakers." Folia Phoniatrica **43**, 234-45.

Delattre P (1967) "Acoustic or articulatory invariance?" Glossa **1**, 3-25.

Delattre P, Liberman AM, Cooper FS (1955) "Acoustic loci and transitional cues for consonants." Journal of the Acoustical Society of America **27**, 769-73.

de Pinto O, Hollien H (1982) "Speaking fundamental frequency characteristics of Australian women: then and now." Journal of Phonetics **10**, 367-75.

Dew BAH, Hardwick PJ, Roach PJ, Shirt MA, Kirby HF (1986) "Voice degradation problems in using automatic speech recognisers." Proceedings of the International Conference on Speech Input/Output; Techniques and Applications (London), 319-23.

Doddington GR, Schalk TB (1981) "Speech recognition: turning theory into practice." IEEE Spectrum **18**(9), 26-32.

Doherty ET, Hollien H (1978) "Multiple-factor speaker identification of normal and distorted speech." Journal of Phonetics **6**, 1-8.

Duffy RJ (1970) "Fundamental frequency characteristics of adolescent females." Language and Speech **13**, 14-24.

Edwards JR (1979) "Social class and the identification of sex in children's speech." Journal of Child Language **6**, 121-7.

Eguchi S, Hirsh I (1969) "Development of speech sounds in children." Acta Oto-laryngologica **257** (supplement), 5-51.

Emanuel FW, Lively MA, McCoy JF (1973) "Spectral noise levels and roughness ratings for vowels produced by males and females." Folia Phoniatrica **25**, 110-20.

Endres W, Bambach W, Flosser G (1971) "Voice spectrograms as a function of age, voice disguise, and voice imitation." Journal of the Acoustical Society of America **49**, 1842-8.

Fairbanks G, Herbert EL, Hammond JM (1949) "An acoustical study of vocal pitch in seven and eight year-old girls." Child Development **20**, 71-8.

Fairbanks G, Wiley JH, Lassman FM (1949) "An acoustical study of vocal pitch in seven and eight year-old boys." Child Development **20**, 63-9.

Fallside F (1985) "Frequency-domain analysis of speech." In Fallside & Woods (1985:41-80).

Fallside F, Woods WA (ed.s) (1985) *Computer Speech Processing.* Prentice-Hall.

Fant CG (1959) "Acoustic analysis and synthesis of speech with applications to Swedish." Ericsson Technics **15**, 3-108; reprinted in Fant (1973:32-83).

Fant CG (1960) *Acoustic theory of speech production.* Mouton: The Hague.

Fant CG (1966) "A note on vocal tract size factors and non-uniform F-pattern scalings." Speech Transmission Laboratory, Quarterly Progress and Status Report (KTH, Stockholm), no. 4, 22-35; reprinted in Fant (1973:84-93).

Fant CG (1973) *Speech sounds and features.* MIT Press: Cambridge, Mass.

Fant CG (1975) "Non-uniform vowel normalisation." Speech Transmission Laboratory, Quarterly Progress and Status Report (KTH, Stockholm), no. 2-3, 1-19.

Fant CG (1985) "The voice source: Theory and acoustic modeling." In Titze & Scherer (1985:453-464).

Fant CG, Tatham MAA (ed.s) (1975) *Auditory analysis and perception of speech.* Academic Press: London.

Fay PJ, Middleton WC (1940) "Judgement of Kretschmerian body types from the voice as transmitted over a public address system." Journal of Social Psychology **12**, 151-62.

Firchow ES *et al.* (ed.s) (1972) *Studies for Einar Haugen.*

Fischer-Jorgensen E (1967) "Phonetic analysis of breathy (murmured) vowels in Gujarati." Indian Linguistics **28**, 71-139.

Fischer-Jørgensen (1972) "Formant frequencies of long and short Danish vowels." In Firchow *et al.* (1972).

Fry DB (1979) *The physics of speech.* Cambridge: Cambridge Univ. Press.

Fujimura O (ed.) (1988) *Vocal Physiology: Voice Production, Mechanisms and Functions.* New York: Raven Press.

Fujisaki H, Kawashima T (1968) "The roles of pitch and the higher formants in the perception of vowels." IEEE Transactions on Audio and Electroacoustics **AU-16**, 73-7.

Gilbert HR, Weismer GG (1974) "The effects of smoking on the speaking fundamental frequency of adult women." Journal of Psycholinguistic Research **3**, 225-31.

Giles H, Scherer KR, Taylor DM (1979) "Speech markers in social interaction." In Scherer & Giles (1979:343-81).

Goldstein UG (1976) "Some speaker-identifying features based on formant tracks." Journal of the Acoustical Society of America **59**, 176-82.

Goldstein UG (1980) "An articulatory model for the vocal tracts of young children." Sc.D. dissertation, MIT.

Graddol D (1986) "Discourse specific pitch behaviour." In Johns-Lewis (1986:221-37).

Graddol D, Swann J (1983) "Speaking fundamental frequency: Some physical and social correlates." Language and Speech **26**, 351-66.

Gunter CD, Manning WH (1982) "Listener estimations of speaker height and weight in unfiltered and filtered conditions." Journal of Phonetics **10**, 251-7.

Günzburger D (1989) "Voice adaptation by transsexuals." Forthcoming in Clinical Linguistics and Phonetics.

Günzburger D (1991) "Breathiness in male and female speakers." Proceedings of the International Congress of Phonetic Sciences (Aix-en-Provence) **3**, 62-5.

Günzburger D, Bresser A, ter Keurs M (1987) "Voice identification of prepubertal boys and girls by normally sighted and visually handicapped subjects." Language and Speech **30**, 47-58.

Günzburger D, de Vries M (1989) "How do minor acoustic cues affect male and female voice quality?" Proceedings of the European Conference on Speech Communication and Technology **2**, 143-5.

Hammond JM (1947) "An objective study of the pitch characteristics of eight-year-old girls during oral reading." M.A. thesis, State University of Iowa.

Hanley, Snidecor (1967) "Some acoustic similarites between languages." Phonetica **17**, 141-8.

Hanson W, Emanuel FW (1979) "Spectral noise and vocal roughness relationships in adults with laryngeal pathology." Journal of Communication Disorders **12**, 113-24.

Harshman R, Ladefoged P, Goldstein L (1977) "Factor analysis of tongue shapes." Journal of the Acoustical Society of America **62**, 693-707.

Hasek C, Singh S, Murry T (1980) "Acoustic attributes of preadolescent voices." Journal of the

Acoustical Society of America **68**, 1262-5.

Hecker MH, Stevens KN, von Bismark G, Williams CE (1968) "Manifestations of task-induced stress in the acoustical speech signal." Journal of the Acoustical Society of America **44**, 993-1001.

Henton CG (1989) "Fact and fiction in the description of female and male pitch." Language and Communication **9**, 299-311.

Henton CG, Bladon RAW (1985) "Breathiness in normal female speech: Inefficiency versus desirability." Language and Communication **5**, 221-7.

Herbert EL (1942) "An objective study of the pitch characteristics of seven-year-old girls during oral reading." M.A. thesis, State University of Iowa.

Hermes DJ (1991) "Synthesis of breathy vowels: Some research methods." Speech Communication **10**, 497-502.

Hermes DJ (1992) "Pitch analysis." European Speech Communication Association Tutorial: Comparing Speech Signal Representations (Sheffield, UK), 1-29.

Hicks D (1989) "Anatomical basis of the glottal chink." Paper presented at the American Speech-Language-Hearing Association meeting, St. Louis, USA.

Higgins MB, Saxman JH (1989) "A comparison of intrasubject variation across sessions of three vocal frequency perturbations indices." Journal of the Acoustical Society of America **86**, 911-6.

Higgins MB, Saxman JH (1991) "A comparison of selected phonatory behaviours of healthy aged and young adults." Journal of Speech and Hearing Research **34**, 1000-10.

Hillenbrand J (1987) "A methodological study of perturbation and additive noise in synthetically generated voice signals." Journal of Speech and Hearing Research **30**, 448-61.

Hirano M, Kurita S, Nakashima T (1983) "Growth, development and aging of human vocal folds." In Bless & Abbs (1983:22-43).

Holbrook A, Fairbanks G (1962) "Diphthongs, formants and their movements." Journal of Speech and Hearing Research **5**, 38-58.

Hollien H, Dew D, Phillips P (1971) "Phonational frequency ranges of adults." Journal of Speech and Hearing Research **14**, 755-60.

Hollien H, Jackson B (1973) "Normative data on the speaking fundamental frequency characteristics of young adult males." Journal of Phonetics **1**, 117-20.

Hollien H, Majewski M (1977) "Speaker identification by long-term spectra under normal and distorted speaking conditions." Journal of the Acoustical Society of America **62**, 975-80.

Hollien H, Malcik E (1962) "Adolescent voice change in southern Negro males." Speech Monographs **29**, 53-8.

Hollien H, Malcik E (1967) "Evaluation of cross-sectional studies of adolescent voice change in males." Speech Monographs **34**, 80-4.

Hollien H, Malcik E, Hollien B (1965) "Adolescent voice change in southern white males." Speech Monographs **32**, 87-90.

Hollien H, Paul P (1969) "A second evaluation of the speaking fundamental frequency characteristics of post-adolescent girls." Language and Speech **12**, 119-24.

Hollien H, Shipp T (1972 or 1971) "Speaking fundamental frequency and chronological age in males." Journal of Speech and Hearing Research **15**, 155-9.

Hollien H, Tolhurst GC, McGlone RE (1982) "Speaking fundamental frequency as a function of vocal intensity." Unpublished representation to the Dreyfus Foundation.

Honjo I, Isshiki N (1980) "Laryngoscopic and voice characteristics of aged persons." Archives of Otolaryngology **106**, 149-50.

Horii Y (1980) "Vocal shimmer in sustained phonation." Journal of Speech and Hearing Research **23**, 202-9.

Horii Y (1982) "Jitter and shimmer differences among sustained vowel phonations." Journal of Speech and Hearing Research **25**, 12-4.

Huber D (1989) "Voice characteristics of female speech and their representation in computer speech synthesis and recognition." Proceedings of the European Conference on Speech Communication and Technology **2**, 477-80.

Hudson AI, Holbrook A (1981) "A study of the reading fundamental vocal frequency of young Black adults." Journal of Speech and Hearing Research **24**, 197-201.

Huffman MK (1987) "Measures of phonation type in Hmong." Journal of the Acoustical Society of America **81**, 495-504.

Hunter W, Garn S (1972) "Disproportionate sexual dimorphism in the human face." American Journal of Physical Anthropology **36**, 133-8.

Ingeman F (1968) "Identification of the speaker's sex from voiceless fricatives." Journal of the Acoustical Society of America **44**, 1142-4.

Ingrisano D, Weismer G, Schuckers GH (1980) "Sex identification of preschool children's voices." Folia Phoniatrica **32**, 61-9.

Javkin H, Hanson B, Kaun A (1991) "The effects of breathy voice on intelligibility." Speech Communication **10**, 539-43.

Johns-Lewis C (ed.) (1986) *Intonation in discourse.* Croom Helm: London.

Johnson K (1990) "The role of perceived speaker identity in F_o normalisation of vowels." Journal of the Acoustical Society of America **88**, 642-54.

Jørgensen (1969) In German. Phonetica **19**, 217-54.

Kahane J (1975) *The developmental anatomy of the human prepubertal and pubertal larynx.* Doctoral dissertation, University of Pittsburgh.

Kahane J (1978) "A morphological study of the human prepubertal and pubertal larynx." J. Am. Anat. **151**, 11-20.

Kahane J (1981) "Anatomic and physiologic changes in the aging peripheral speech mechanism." In Beasley & Davis (1981).

Kallail KJ, Emanuel FW (1985) "The identifiability of isolated whispered and phonated vowel samples." Journal of Phonetics **13**, 11-7.

Karlsson I (1985) "Glottal waveforms for normal female speakers." Speech Transmission Laboratory, Quarterly Progress and Status Report (KTH, Stockholm), no. 1, 31-6.

Karlsson I (1991) "Female voices in speech synthesis." Journal of Phonetics **19**, 111-20.

Keating P, Blankenship B, Byrd D, Flemming E, Todaka Y (1992) "Phonetic analyses of the TIMIT corpus of American English." Proceedings of the International Conference on Spoken Language, 823-6.

Kelley A (1977) "Fundamental frequency measurements of female voices with aging." Paper presented at the American Speech and Hearing Association Convention, Chicago.

Kent RD (1976) "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies." Journal of Speech and Hearing Research **19**, 421-47.

Kent RD, Burkart R (1981) "Changes in the acoustic correlates of speech production." In Beasley & Davis (1981).

Kersta LG, Bricker PD, David EE (1960) "Human or machine?" Journal of the Acoustical Society of America **32**, 1502.

King E (1952) "A roentgenographic study of pharyngeal growth." Angle Orthod. **22**, 23-37.

Kirchner JA (1970) *Pressman and Keleman's physiology of the larynx.* rev.ed. American Academy of Opthamology and Otolaryngology: Washington DC.

Klatt DH (1986) "Detailed spectral analysis of a female voice." Journal of the Acoustical Society of America **80**, supplement 1, S97.

Klatt DH (1987) "Review of text-to-speech conversion for English." Journal of the Acoustical Society of America **82**, 737-93.

Klatt DH, LC Klatt (1990) "Analysis, synthesis and perception of voice quality variations among female and male talkers." Journal of the Acoustical Society of America **87**, 820-57.

Koopmans-van Beinum FJ, van Bergem DR (1989) "The role of 'given' and 'new' in the production and perception of vowel constrasts in read text and in spontaneous speech." Proceedings of the European Conference on Speech Communication and Technology **2**, 113-6.

Krogman WM (1962) "The human skeleton in forensic medicine." CC Thomas: Springfield.

Krook MIP (1988) "Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis." Folia Phoniatrica **40**, 82-90.

Künzel HJ (1989) "How well does average fundamenal frequency correlate with speaker height and weight." Phonetica **46**(1-3), 117-25.

Labov W (1966) *The social stratification of English in New York City.* Washington DC: Center for Applied Linguistic.

Labov W (1972) *Sociolinguistic patterns.* Univ. of Philidelphia Press: Philidelphia.

Ladefoged P (1975) *A course in phonetics.* Harcourt Brace Jovanovich: New York.

Ladefoged P (1983) "The linguistic use of different phonation types." In Bless & Abbs (1983:351-60).

Ladefoged P, Antoñanzas-Barroso N (1985) "Computer measures of breathy voice quality." UCLA Working Papers In Phonetics **61**, 79-86.

Ladefoged P, Maddieson I, Jackson M (1988) "Investigating phonation types in different languages." In Fujimura (1988:297-317).

Lass NJ (1981) A reply to Cohen *et al.* (1980). Journal of the Acoustical Society of America **69**, 1204-6.

Lass NJ, Almerino CA, Jordan LF, Walsh JN (1980a) "The effect of filtered speech on speaker race and sex identifications." Journal of Phonetics **8**, 101-12.

Lass NJ, Barry PJ, Reed RA, Walsh JM, Amuso TA (1979a) "The effect of temporal speech alterations on speaker height and weight identification." Language and Speech **22**, 163-71.

Lass NJ, Beverley AS, Nicosia DK, Simpson LA (1978a) "An investigation of speaker height and weight identification by means of direct estimates." Journal of Phonetics **6**, 69-76.

Lass NJ, Brong GW, Ciccolella SA, Walters SC, Maxwell EL (1980b) "An investigation of speaker height and weight discriminations by means of paired comparison judgements." Journal of Phonetics **8**, 205-12.

Lass NJ, Brown WS (1978) "Correlational study of speakers heights, weights, body surface areas, and speaking fundamental frequencies." Journal of the Acoustical Society of America **63**, 1218-20.

Lass NJ, Davis M (1976) "An investigation of speaker height and weight identification." Journal of the Acoustical Society of America **60**, 700-3.

Lass NJ, Dicola GA, Beverley AS, Barbera C, Henry KG, Badali MK (1979b) "The effect of phonetic complexity on speaker height and weight identification." Language and Speech **22**, 297-309.

Lass NJ, Hughes KR, Bowyer MD, Waters LT, Bourne VT (1976) "Speaker sex identification from voiced, whispered and filtered isolated vowels." Journal of the Acoustical Society of America **59**, 675-8.

Lass NJ, Kelley DT, Cunningham CM, Sheridan KJ (1980c) "A comparative study of speaker height and weight identification from voiced and whispered speech." Journal of Phonetics **8**, 195-204.

Lass NJ, Mertz PJ, Kimmel KL (1978b) "The effect of temporal speech alterations on speaker race and sex identifications." Language and Speech **21**, 279-91.

Lass NJ, Philips JK, Bruchey CA (1980d) "The effect of filtered speech on speaker height and weight identification." Journal of Phonetics **8**, 91-100.

Lass NJ, Tecca JE, Mancuso RA, Black WI (1979c) "The effect of phonetic complexity on speaker race and sex identifications." Journal of Phonetics **7**, 105-18.

Laver J (1975) "Individual features in voice quality." PhD dissertation, Univ. of Edinburgh.

Laver J (1976) "The semiotic nature of phonetic data." York Papers In Linguistics **6**, 55-62.

Laver J (1988) "Cognitive science and speech - a framework for research." Work in Progress, Dept. of Linguistics, Univ of Edinburgh, no. 21, 83-114.

Laver J, Trudgill P (1979) "Phonetic and linguistic markers in speech." In Scherer & Giles (1979:1-32).

Liberman AM (1970) "Some characteristics of perception in the speech mode." Perception & its Disorders **48**, 238-54.

Liberman AM, Cooper FS (1972) "In search of the acoustic cues." In Valdman (1972:329-38).

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) "Perception of the speech code." Psychological Review **74**, 431-61.

Lieberman P (1967) *Intonation, perception, and language.* MIT Press: Cam., Mass.

Lieberman P, Blumstein SE (1988) *Speech physiology, speech perception, and acoustic phonetics.* Cambridge Univ. Press: Cambridge.

Lindblom B, Öhman S (ed.s) (1979) *Frontiers of speech communication research.* Academic Press: London.

Linke CE (1973) "A study of pitch characteristics of female voices and their relations to vocal effectiveness." Folia Phoniatrica **25**, 173-85.

Luchsinger R, Arnold G (1965) *Voice-Speech-Language.* Constable: London.

Lyons J (1977) *Semantics.* (2 vol.s) Cambridge Univ. Press: Cambridge.

Majewski W, Hollien H, Zalewski J (1972) "Speaking fundamental frequency characteristics of Polish adult males." Phonetica **25**, 119-25.

Margulies MK (1979) "Male-female differences in speaker intelligibility: Normal versus hearing impaired listeners." In Wolf & Klatt (1979:363-6).

Mattingly IG (1966) "Speaker variation and vocal tract size." Journal of the Acoustical Society of America **39**, 1219.

McGlone RE, Hollien H (1963) "Vocal pitch characteristics of aged women." Journal of Speech and Hearing Research **6**, 167-72.

McGlone R, McGlone J (1972) "Speaking fundamental frequency of eight-year-old girls." Folia Phoniatrica **24**, 313-7.

Meditch A (1975) "The development of sex-specific speech patterns in young children." Anthropoligical Linguistics **17**, 421-33.

Michel JF, Hollien H, Moore P (1966) "Speaking fundamental frequency characteristics of 15, 16 and 17 year-old girls." Language and Speech **9**, 46-51.

Milenkovic P (1987) "Least mean square measures of voice perturbation." Journal of Speech and Hearing Research **30**, 529-38.

Miller JL, Grosjean F, Lomanto C (1984) "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications." Phonetica **41**, 215-25.

Miller JL, Volaitis LE (1989) "Effect of speaking rate on the perceptual structure of a phonetic category." Perception and Psychophysics **46**, 505-12.

Miller RL (1953) "Auditory tests with synthetic vowels." Journal of the Acoustical Society of America **25**, 114-21.

Monsen RB, Engebretson AM (1977) "Study of variations in the male and female glottal wave." Journal of the Acoustical Society of America **62**, 981-93.

Moses JP (1940) "Is medical phonetics an essential part of otorhinolaryngology?" Archives of Oto-Rhino-Laryngology **31**, 444-51.

Moses JP (1941) "Theories regarding the relation of constitution and character through the voice." Psychological Bulletin **38**, 746.

Murry T, Hollien H, Müller E (1975) "Perceptual responses to infant crying: Maternal recognition and sex judgements." Journal of Child Language **2**, 199-204.

Mysak ED (1959) "Pitch and duration characteristics of older males." Journal of Speech and Hearing Research **2**, 46-54.

Negus VE (1962) *The comparative anatomy and physiology of the larynx.* Hafner: New York.

Nittrouer S, McGowan RS, Milenkovic PH, Beehler D (1990) "Acoustic measurements of men's and women's voices: A study of context effects and covariation." Journal of Speech and Hearing Research **33**, 761-75.

Nolan FJ (1983) *The phonetic bases of speaker recognition.* Cambridge Univ. Press: Cambridge.

Nolan FJ (1987) "Linguistic versus personal variation in speech recognition." EuroTech87 **2**, 476-9.

Noll AM (1964) "Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection." Journal of the Acoustical Society of America **36**, 458-65 or 296-302.

Noll AM (1967) "Cepstrum pitch determination." Journal of the Acoustical Society of America **41**, 293-309.

Nord L, Ananthapadmanabha TV, Fant G (1986) "Perceptual tests using an interactive source filter model and considerations for synthesis strategies." Journal of Phonetics **14**, 401-3.

Noyes JM, Frankish CR (1989) "Gender differences in speech recogniser performances." Proceedings of the European Conference on Speech Communication and Technology **1**, 110-2.

Oppenheim AV (1967) "Generalized superposition." Information & Control **11**, 528-36.

Oppenheim AV, Schafer RV (1968) "Homomorphic analysis of speech." IEEE Transactions on Audio and Electroacoustics **AU-16**, 221-6.

Oppenheim AV, Schafer RV, Stockham TG (1968) "Nonlinear filtering of multiplied and convolved signals." IEEE Transactions on Audio and Electroacoustics **AU-16**, 437-65.

Orlikoff RF, Baken RJ (1990) "Consideration of the relationship between the fundamental frequency of phonation and jitter." Folia Phoniatrica **42**, 31-40.

Pallett DS (1985) "Performance assessment of automatic speech recognisers." Journal of Research of the National Bureau of Standards **90**(5), 371-87.

Pandit PB (1957) "Nasalisation, aspiration and murmur in Gujarati." Indian Linguistics **17**, 165-

Parsons T (1986) *Voice and Speech Processing.* McGraw-Hill.

Perkell JS (1979) "On the nature of distinctive features: Implications of a preliminary vowel production study." In Lindblom & Öhman (1979).

Perkins WH (1971) "Vocal function." In Travis (1971:481-503).

Peterson GE, Barney HL (1952) "Control methods used in a study of vowels." Journal of the Acoustical Society of America **24**, 175-84.

Pols, Tromp, Plomp (1973) "Frequency analysis of Dutch vowels from 50 male speakers." Journal of the Acoustical Society of America **53**, 1093-101.

Rabiner LR, Schafer RW (1978) *Digital processing of speech signals.* Prentice-Hall.

Reed L (1985) "Military applications of voice technology." Speech Technology **2**(4), 52-50?.

Remez RE, Rubin PE (1990) "On the perception of speech from time-varying acoustic information: Contributions of amplitude variation." Perception & Psychophysics **48**, 313-25.

Riordan CJ (1977) "Control of vocal tract length in speech." Journal of the Acoustical Society of America **62**, 998-1002.

Robb MP, Saxman JH (1985) "Developmental trends in vocal fundamental frequency of young children." Journal of Speech and Hearing Research **28**, 421-7.

Rollins AM (1985) "Speech recognition and manner of speaking in noise and quiet." Proceedings of the Conference on Human Factors in Computing Systems (ACM New York) 197-9.

Rosenblatt M (ed.) (1963) *Proceedings of the Symposium on Time Series Analysis.* John Wiley: New York.

Sachs J (1975) "Cues to the identification of sex in children's speech." In Thorne & Henley (1975).

Sachs J, Lieberman P, Erikson D (1973) "Anatomical and cultural determinants of male and female speech." In Shuy & Fasold (1973:74-84).

Sambur MR (1975) "Selection of acoustic features for speaker identification." IEEE Transactions on Acoustics, Speech and Signal Processing **ASSP-23**, 176-82.

Saxman JH, Burke KW (1967) "Speaking fundamental frequency characteristics of middle-aged females." Folia Phoniatrica **19**, 167-72.

Scherer KR (1981a) "Vocal indicators of stress." In Darby (1981:171-87).

Scherer KR (1981b) "Speech and emotional states." In Darby (1981:189-220).

Scherer KR, Giles H (ed.s) (1979) *Social Markers in Speech.* Cambridge Univ. Press: Cambridge; Editions de la maison des sciences de l'homme: Paris.

Schwartz MF (1968) "Identification of speaker sex from isolated voiceless fricatives." Journal of the Acoustical Society of America **43**, 1178-9.

Schwartz MF, Rine HE (1968) "Identification of speaker sex from isolated, whispered vowels." Journal of the Acoustical Society of America **44**, 1736-7.

Shevchenko K (1989) "What's in a voice: A system of regional and social acoustic characteristics based on the analysis of 100 British English voices." Proceedings of the European Conference on Speech Communication and Technology **2**. 131-4.

Shuy RW, Fasold RW (ed.s) (1973) *Language attitudes: Current trends and prospects.* Georgetown Univ. Series in Language and Linguistics: Washington DC.

Slawson AW (1968) "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency." Journal of the Acoustical Society of America **43**, 87-101.

Smith PM (1979) "Sex markers in speech." In Scherer & Giles (1979:106-46).

Smith PM (1985) *Language, the sexes and society.* Blackwell: Oxford.

Snidecor JC (1951) "The pitch and duration characteristics of superior female speakers during oral reading." Journal of Speech and Hearing Disorders **16**, 44-52.

Söderston M, Hammarberg B (1992) "Vocal fold closure, perceptual and fundamental frequency characteristics in normally speaking females before and after voice training." Karolinska Inst., Huddinge Hospital, Phoniatric and Logopedic Progress Report 8, 23-9. Presented at the International Congress on the Voice (Besançon, 1991).

Söderston M, Lindestad P-Å(1990) "Glottal closure and perceived breathiness during phonation in normally speaking adults." Journal of Speech and Hearing Research **33**, 601-11.

Söderston M, Lindestad P-Å, Hammarberg B (1989) "Vocal fold closure, perceived breathiness and acoustic characteristics in normal-speaking adults." Paper presented at the Vocal Fold Physiology Conference (Stockholm, Sweden).

Spencer LE (1988) "Speech characteristics of male-to-female transsexuals: A perceptual and acoustic study." Folia Phoniatrica **40**, 31-42.

Stoicheff ML (1981) "Speaking fundamental frequency characteristics of nonsmoking female adults." Journal of Speech and Hearing Research **24**, 437-41.

Strange W, Verbrugge RR, Shankweiler DB, Edman TR (1976) "Consonant enviroment specifies vowel identity." Journal of the Acoustical Society of America **60**, 213-24.

Tanner JM (1978) *Foetus into man.* Open Books: London.

Tartter VC (1980) "Happy talk: perceptual and acoustic affects of smiling on speech." Perception & Psychophysics **27**, 24-7.

Tartter VC (1989) "What's in a whisper?" Journal of the Acoustical Society of America **86**, 1678-83.

Thomas T, Peckham J, Frangoulis E, Cove J (1989) "The sensitivity of speech recognisers to speaker variability and speaker variation." Proceedings of the European Conference on Speech Communication and Technology **1**, 408-11.

Thorne B, Henley N (ed.s) (1975) *Language and sex: Difference and dominance.* Newbury House: Rowley, Mass.

Tielen MTJ (1989) "Intelligibility of male and female voices under a few noise conditions." Proceedings of the European Conference on Speech Communication and Technology **2**, 127-30.

Titze IR (1989a) "Physiologic and acoustic differences between male and female voices." Journal of the Acoustical Society of America **85**, 1699-707.

Titze IR (1989b) "On the relation between subglottal pressure and fundamental frequency in phonation." Journal of the Acoustical Society of America **85**, 901-6.

Titze IR, Scherer RC (ed.s) (1985) *Vocal fold physiology: Biomechanics, acoustics and phonatory control.* The Denver Center For Performing Arts: Denver, Colorado.

Titze IR, Horii Y, Scherer RC (1987) "Some technical considerations in voice perturbation experiments." Journal of Speech and Hearing Research **30**, 252-60.

Traunmüller H (1981) "Perceptual dimension of openness in vowels." Journal of the Acoustical Society of America **69**, 1465-75.

Travis LE (ed.) (1971) *Handbook of Speech Pathology and Audiology.* New York: Appleton.

Trudgill P (1974a) *The social differentiation of English in Norwich.* Cambridge University Press: Cambridge.

Trudgill P (1974b) *Sociolinguistics: An introduction.* Harmondsworth: Penguin Books.

Trudgill P (1975) "Sex, covert prestige and linguistic change in the urban. British English of Norwich." In Thorne & Henley (1975).

Vaissière J (1985) "Speech recognition: a tutorial." In Fallside (1985:191-242).

Valdman A (ed.) (1972) *Papers in linguistics and phonetics in the memory of Pierre Delattre.* Mouton: The Hague.

van Bergem DR, Koopmans-van Beinum FJ (1989) "Vowel reduction in natural speech." Proceedings of the European Conference on Speech Communication and Technology **2**, 285-8.

Visick D, Johnson P, Long J (1984) "The use of simple speech recognisers in industrial applications." Proceedings of Interact'84: 1st IFIP Conference on Human-Computer Interaction (London).

Vuorenkoski V, Lenko H, Tjernlund P, Vuorenkoski L, Perheentupa J (1978) "Fundamental voice frequency during normal and abnormal growth, and after androgen treatment." Archives of Disease in Childhood **53**, 201-9.

Walker G, Kowalski C (1972) "On the growth of the mandible." American Journal of Physical Anthropology **36**, 111-8.

Waterworth JA (1984) "Interaction with machines by voice: human factors issues." British Telecom Technology Journal, **2**(4), 8.

Weinberg B, Bennett S (1971a) "Recognition of the sex of esophageal talkers." Journal of the Acoustical Society of America **50**, 147.

Weinberg B, Bennett S (1971b) "Speaker sex recognition of 5 and 6-year-old children's speech." Journal of the Acoustical Society of America **50**, 1210-3.

Wells JC (1982) *Accents of English.* Cambridge Univ. Press: Cambridge.

Wilcox KA, Horii Y (1980) "Age and changes in vocal jitter." Journal of Gerontology **35**, 194-8.

Williams CE, Stevens KN (1972) "Emotions and speech: some acoustic correlates." Journal of the Acoustical Society of America **52**, 1238-50.

Williams CE, Stevens KN (1981) "Vocal correlates of emotional states." In Darby (1981:221-40).

Wilpon JG, Roberts LA (1986) "The effects of instructions and feedback on speaker consistency for automatic speech recognition." Proceedings of the International Conference on Speech Input/Output; Techniques and Applications (London), 319-23.

Wolf JJ, Klatt DH (ed.s) (1979) *Speech communication papers presented at the 97th meeting of the Acoustical Society of America.* Acoustical Society of America: New York.

Yumoto E, Gould W, Baer T (1982) "Harmonics-to-noise ratio as an index of the degree of hoarseness." Journal of the Acoustical Society of America **71**, 1544-9.

analysis of the data by speaker variable (i.e. by age, dialect, etc.). Particular attention is paid to the distribution of the mean and the range of values produced by each speaker to facilitate the analysis of between- and within-speaker variability in Section 4.3. Also discussed in this section are the sex-differentiating potentials of the three acoustic-phonetic measures. The conclusions are presented in Section 4.4, which includes remarks on speaker characterisation in general, on the characterisation of speaker sex, and on the automatic analysis of speech databases.

Note that throughout this chapter reference is made to SPEAKER MEANS and SLICE MEANS, where a SPEAKER MEAN is the mean value produced by a particular speaker for a particular acoustic-phonetic measure, and a SLICE MEAN is the mean value of an individual speech slice for a particular acoustic-phonetic measure.

4.2.3 Formant frequency

Reported below are the overall group means for F_1, F_2 and F_3 for both sexes, and an analysis of the formants by phone. These results are compared with the formant frequency data reported in the literature.

The analysis of the formant frequency data is limited in scope due to the lack of confidence in the output of the formant frequency estimator, and due to insufficient time to carry out a complete investigation. The reasons for this are discussed.

Analysis of overall data

The overall group means for first, second and third formant frequencies are given in Table 4.40, and are illustrated in Figure 4.38. The scaling factors ranged from 9-20% for F_1, 12-20% for F_2 and 5-19% for F_3, with overall factors for the formants of 16%, 18% and 11% respectively. This produces a general scaling factor of approximately 15%, which means that on average the female speakers' formant frequencies were 15% greater than the male speakers'. This figure is comparable with the female-male differences reported in the literature. Now due to the different types of phones measured, the overall formant scaling factors given in this study are to some extent incompatible with those given in Tables 3.10 and 3.11 for the Peterson & Barney (1952) and Fant (1959) studies. However, we can directly compare the phones common to each study, namely /aa/, /ae/, /iy/ and /uw/ (and /ux/, an allophone of /uw/, from the TIMIT data), and are given below. By recalculating overall formant frequency means using only the figures for these four phones[21], we can also compare the overall formant scaling factors for the three studies[22]:

		all	PB	F	T
F_1	f	600	570	620
	m	490	470	540
% diff.		22	21	15
F_2	f	1750	1560	1820
	m	1490	1350	1570
% diff.		17	16	16
F_3	f	2910	3050	2730
	m	2530	2570	2420
% diff.		15	19	13

The results from the TIMIT study are most similar to Peterson & Barneys' results, especially for F_1 and F_2, although the female-male difference is somewhat less for F_1. Fant's results are noticeably different, especially for F_2 and F_3. However, Fant's results are also rather different to Peterson & Barneys' results, in particular for F_2 and the female F_3. The main reason for this is probably that Fant's subjects were Swedish, uttering Swedish phonemes (see also Section 3.1.3). He derived his vowel phoneme names by finding the nearest equivalent U.S. English phoneme through F-pattern matching. This indicates that these Swedish phonemes are not always a good match for the U.S. phonemes. The implication is that in the direct vowel comparisons, more weight should be attached to the comparison with the Peterson & Barney data. There is however further cause for caution. Peterson & Barney measured their formant frequencies from vowels spoken in a simple

[21] The Peterson & Barney and Fant means were computed by averaging the means for the four phones under consideration. The TIMIT overall means also include the figures for /ux/.

[22] In the following tables, unless stated otherwise the column headed 'PB' refers to Peterson & Barneys' (1952) study, 'F' refers to Fant's (1959) study, and 'T' refers to this study of the TIMIT data. The formant frequencies are rounded to the nearest 10Hz.

of the between-phone transitional information.

The length of the sampling window was set at 64msec (or 1024 samples, at the sampling rate used for the TIMIT data of 16000 samples/second), which for the average female F_0 of 200Hz encompasses 12.5 periods, or 7.5 periods for the male average of 120Hz. This may seem rather large, but the constraints of the FFT program required the number of data points in an analysis frame to be a power of 2. The next window size down was 32msec (512 samples), and it was feared that for some speakers this would prove to be too small. While it would be possible for the FFT program to be given variable window sizes, based on expectations of the speaker's F_0 or on previously computed values, the initial trials of the program did not indicate a need for this. Perhaps more importantly, the large window size smears over irregularities in the speaker's phonation caused by vocal perturbation (e.g. jitter, shimmer, diplophonia) or vocal pathology. A window size encompassing only two or three periods of the speech waveform results in the cepstral analysis being prone both to depressed F_0 peaks from a lack of periodicity in the waveform portion covered by the analysis window, and to 'phantom' F_0 peaks caused by, for example, situations where every second period in the waveform has a reduced amplitude. Trial analyses showed a 1024 sample window size to be a consistently accurate descriptor of both the vowel phone's mean F_0 and its internal F_0 dynamics.

Rabiner & Schafer (1978) listed three criteria for the optimal use of cepstral analysis, which are reproduced below in **bold**, together with comments related to the design of the cepstral peak-picking algorithm developed for this analysis:

- **Compute the cepstrum every 10-20msec, as the excitation parameters change relatively slowly in normal speech.** For this study the analysis window was shifted by 8msec (64 samples) after every computation of the cepstrum.

- **Search for the peak in the vicinity of the expected F_0 period.** Thus once a phone's F_0 has been identified in the first few analysis frames, it is generally valid to assume the F_0 of subsequent frames will be similar. The search can therefore be simplified by defining a search interval centred around the previously computed position of the cepstral F_0 peak. However, when seeking to define the initial search interval (i.e. to establish the fundamental frequency at the beginning of a slice), while the differences in average F_0 for each speaker sex indicate the use of different search intervals for female and male speakers, in practice the intonation patterns even within read sentences mean a person's SFF can encompass a huge range of values. For example, in speaker **frll0**'s production of the sentence **sa1**, between the first and second syllables the F_0 rose from 276Hz to 345Hz, and fell to 274Hz by the fourth syllable. At the end of the same sentence, between the ninth, tenth and twelfth syllables, the F_0 dropped from 250Hz to 193Hz, and then rose to 220Hz. Thus in just 3.7 seconds, this speaker's fundamental frequency ranged over more than 150Hz. One further limit on the range of the initial search interval is that the lower end must be clear of the large, low quefrency cepstral peaks caused by the vocal tract resonances. While Noll (1967) recommended an initial interval of 1-15msec[4]. However the lower end of the interval was too close to the cepstral vocal tract components, and so the search interval used here was the one recommended by Rabiner & Schafer (1978), namely 3-20msec (333-50Hz). This seemed a reasonable compromise between trying to avoid the vocal tract components and allowing for particularly high female fundamentals. Note that where high F_0s were expected,

[4]Note these are units of quefrency and represent the period of the fundamental. The search interval is equivalent to a frequency range of 1000–67Hz

Age	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
20 – 29	82	-4.8	(2.2)	-9.1	-6.1	-5.1	-3.5	0.5
30 – 39	32	-4.8	(2.7)	-8.5	-7.0	-4.8	-3.5	1.8
40 – 49	10	-5.0	(1.8)	-7.3	-7.0	-4.5	-4.1	-1.7
50 – 59	4	-5.7	(0.9)	-6.3	-6.2	-6.1	-5.2	-4.3
≥ 60	2	-5.2	(0.1)	-5.2	–	-5.2	–	-5.1

Table 4.31: Female mean H_1–H_2 data (to nearest 0.1dB) by age.

Age	n	Mean	(s.d.)	Min	Q1	Q2	Q3	Max
20 – 29	183	-6.3	(1.9)	-10.2	-7.5	-6.4	-5.2	0.8
30 – 39	86	-6.2	(2.1)	-11.7	-7.6	-6.6	-4.9	0.3
40 – 49	13	-5.1	(2.2)	-8.3	-6.7	-5.4	-3.3	-0.9
50 – 59	7	-4.7	(2.0)	-7.0	-6.7	-4.1	-2.9	-2.3
≥ 60	1	-3.4	(–)	–	–	-3.4	–	–

Table 4.32: Male mean H_1–H_2 data (to nearest 0.1dB) by age.

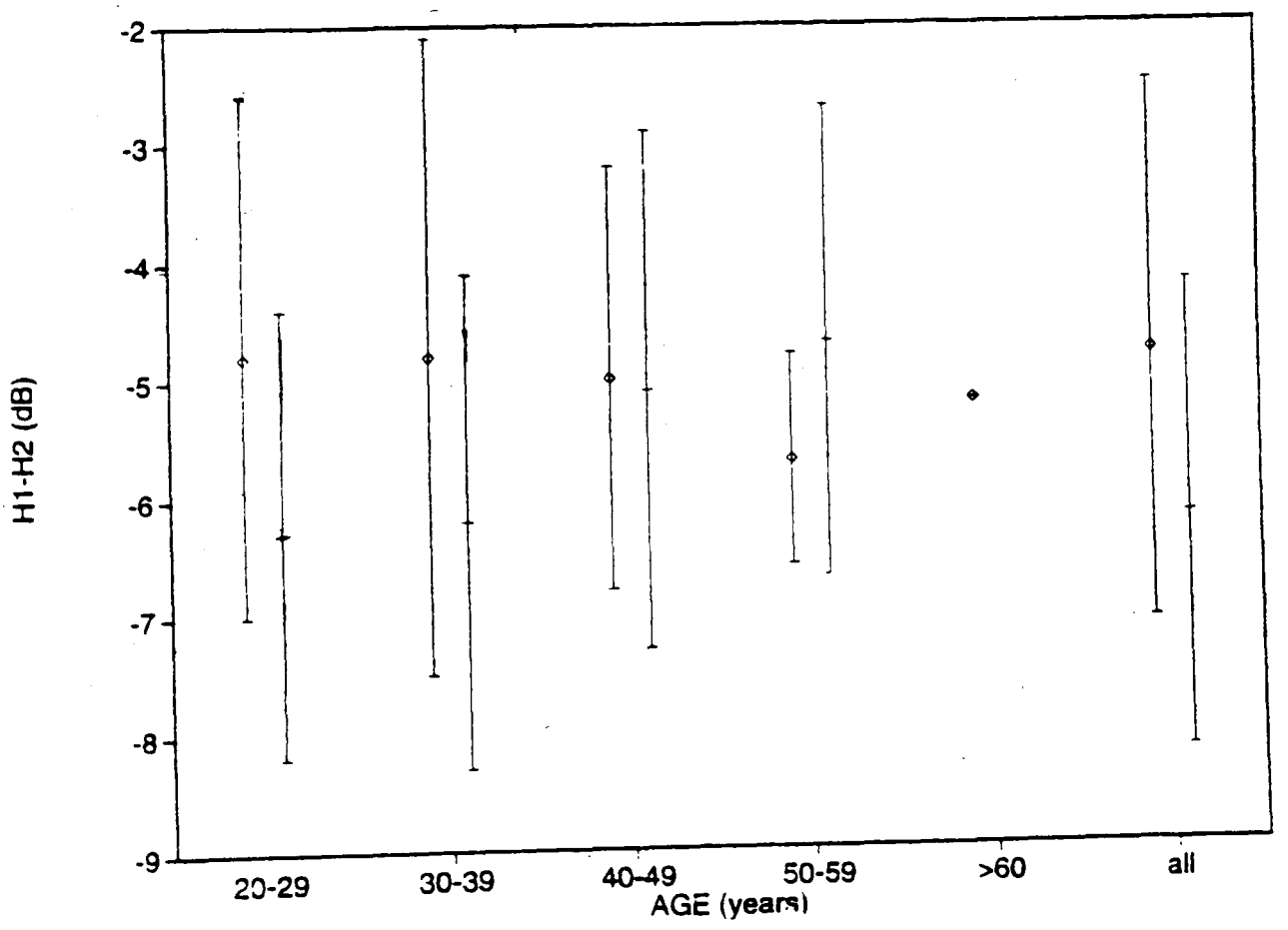


Figure 4.32: Mean H_1–H_2 (dB) by age for female (diamond) and male (cross) speakers. The s.d.s are represented by vertical lines around the mean (note, no s.d. intervals are given for the age group '60 and over' as there were so few speakers). The mean for all speakers is on the right.

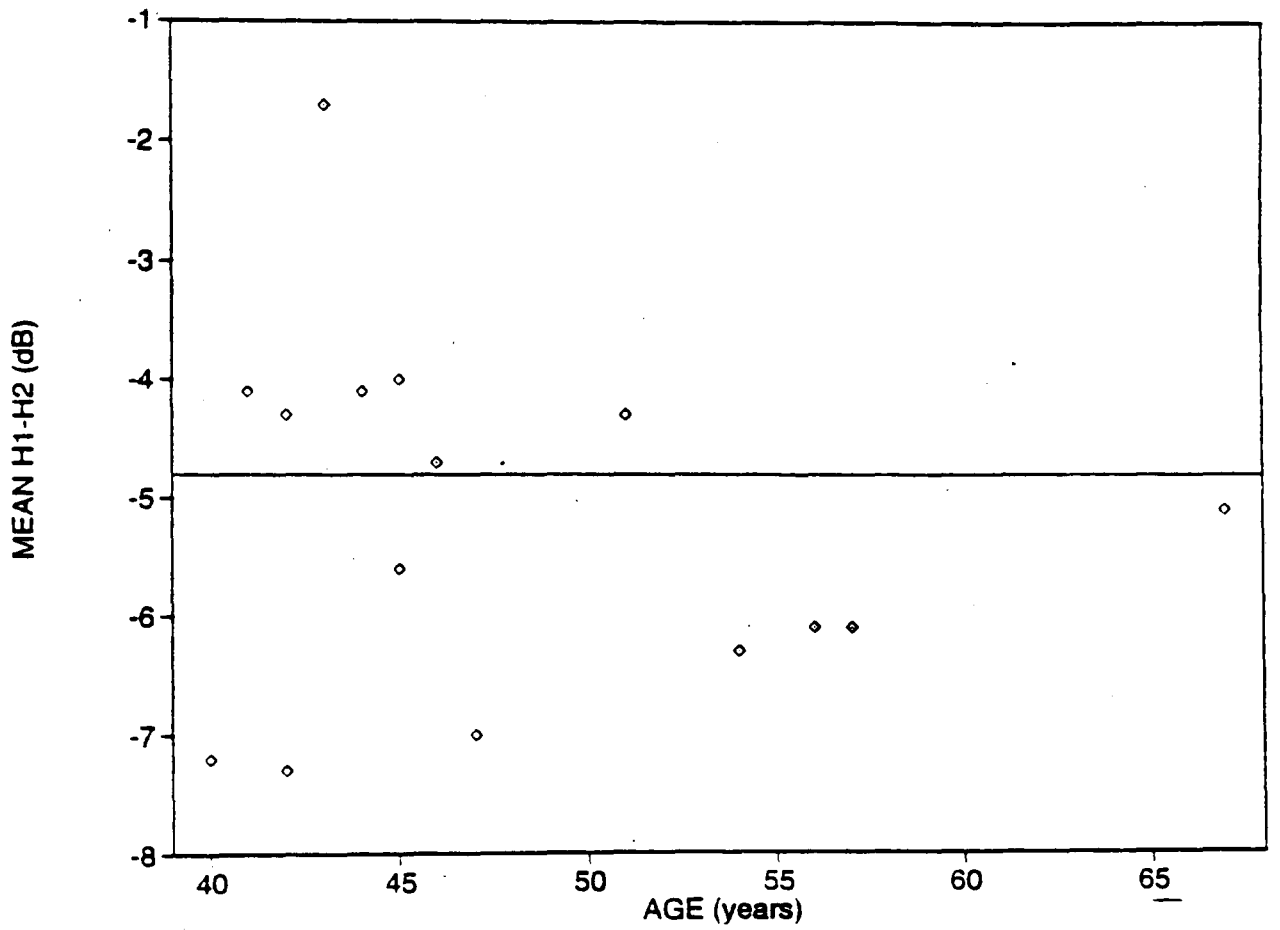


Figure 4.33: Mean H_1–H_2s (dB) of the female speakers over 40 years old. The solid line at -4.8dB represents the mean H_1–H_2 of all the female speakers. Note: The 85 year-old speaker fkfb0 has been left off the graph – her mean H_1–H_2 was -5.2dB.

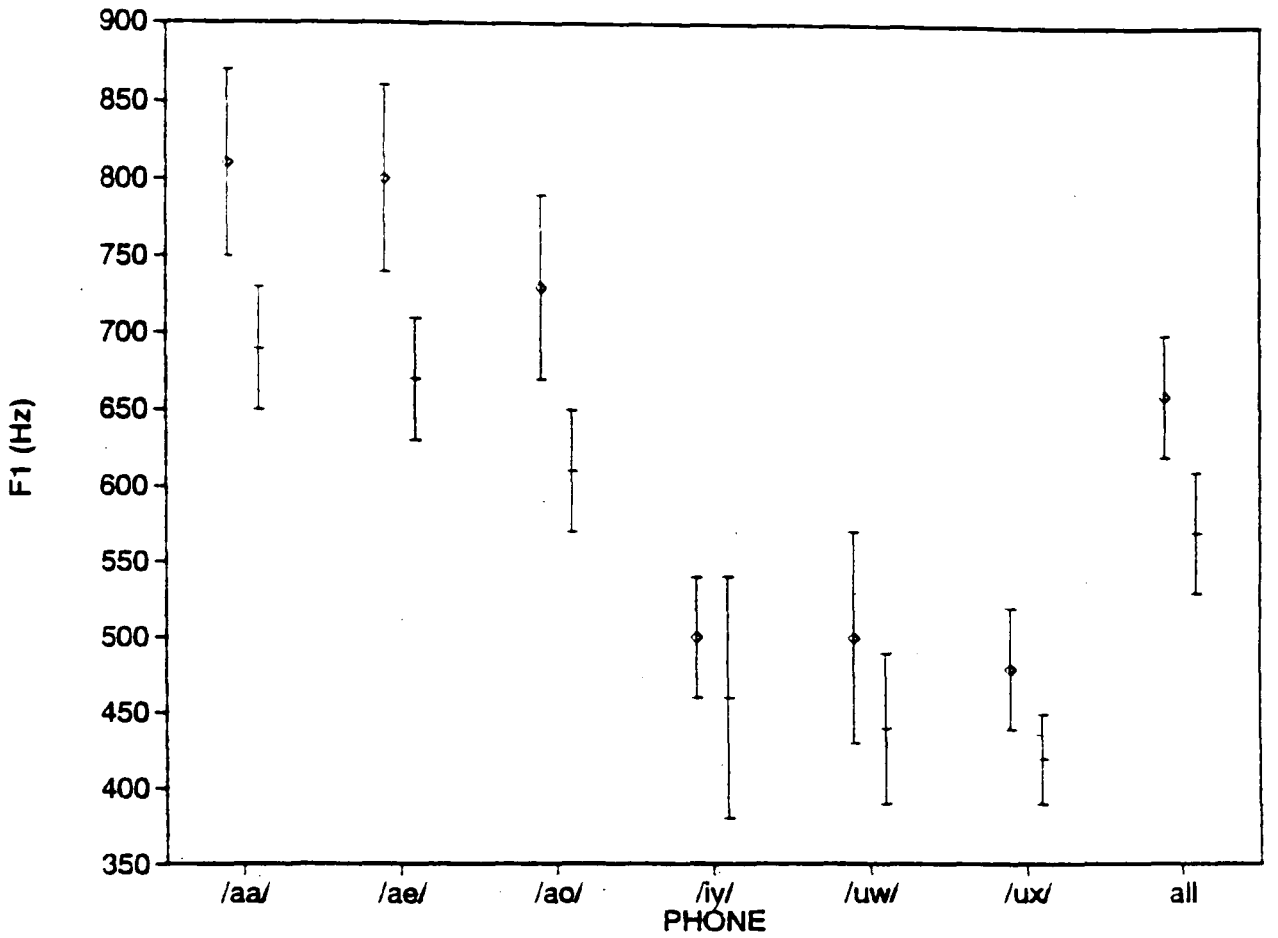


Figure 4.39: Mean and s.d. of F_1 (Hz) by phone for female (diamond) and male (cross) speakers. The means for all the phones are on the right.

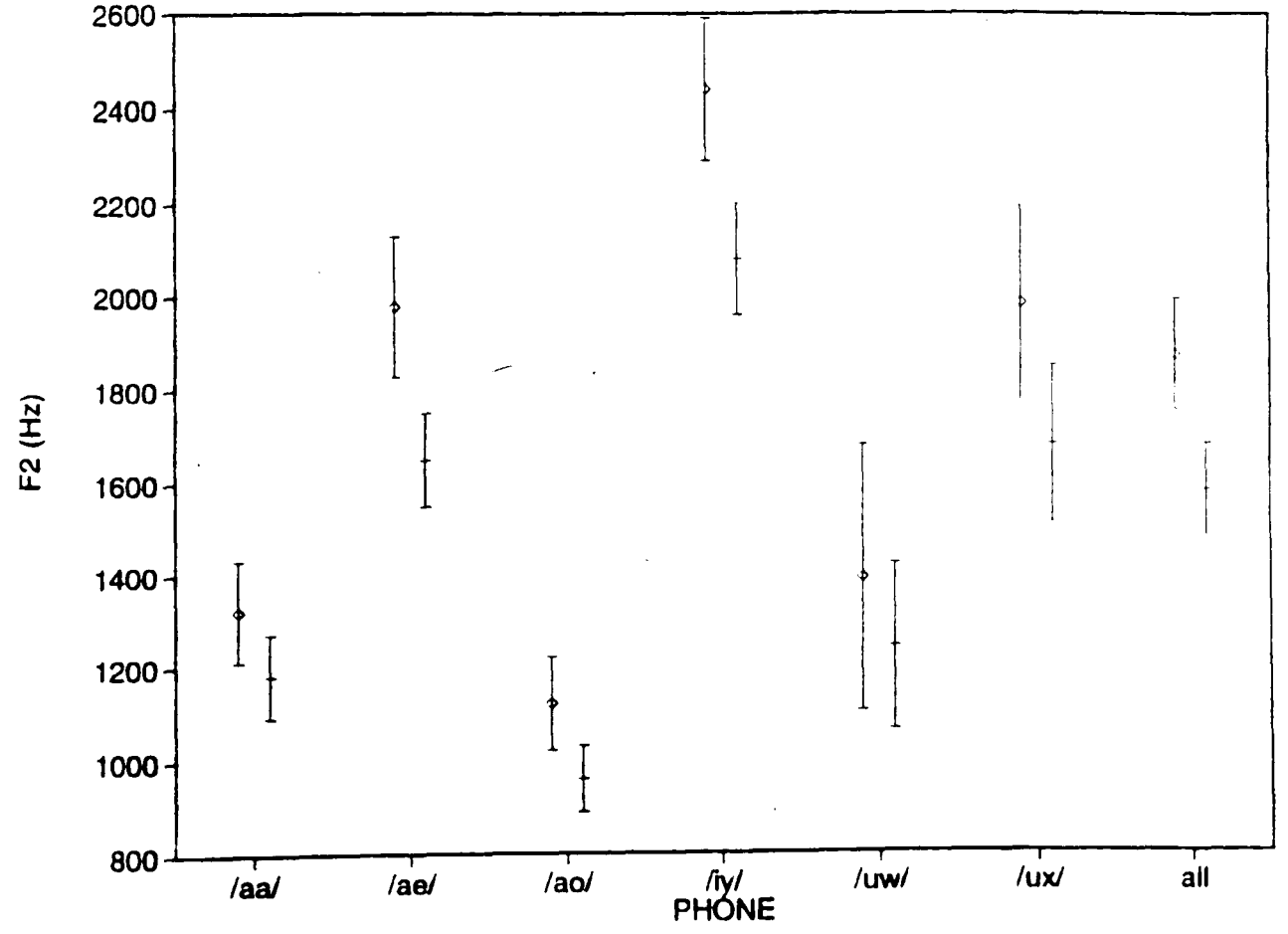


Figure 4.40: Mean and s.d. of F_2 (Hz) by phone for female (diamond) and male (cross) speakers. The means for all the phones are on the right.

As the correlation coefficients and the scatter plots in Figures 4.42 to 4.43 show, for both sexes, increasing SFF indicates increasing harmonic amplitudes, although there was a larger correlation for the male speakers. Furthermore, there appears to be a *general* tendency for speakers with higher fundamentals to have higher harmonic amplitudes. This has already been indicated by the tendency found for female speakers to have higher first and second harmonic amplitudes. Further evidence comes from correlation analyses performed on *all* the SPEAKER MEANS, which produced coefficients of 0.721 using F_0 and H_1 as variables, and 0.595 using F_0 and H_2 as variables.

The presence of a correlation between SFF and any of the formant frequencies was also investigated. Thus, correlation analyses were performed using the SFF and the formant frequency SPEAKER MEANS as variables. As the coefficients reproduced in the table below show, there were no strong correlations between SFF and any of the formants, although the formant frequencies of some of the phones showed a degree of correlation.

Analysis	Sex	/aa/	/ae/	/ao/	/iy/	all
F_0 vs. F_1	f	0.177	0.258	0.155	0.384	0.310
	m	0.224	0.133	0.079	-0.098	-0.018
F_0 vs. F_2	f	0.035	0.084	0.076	0.126	0.208
	m	0.068	0.104	0.025	0.102	0.100
F_0 vs. F_3	f	-0.017	0.219	-0.008	0.258	0.213
	m	0.057	0.102	0.135	0.043	0.118

In summary, the SFF appears to be the most powerful and most consistent acoustic-phonetic measure for the differentiation of speakers of a different sex. However, it is highly unlikely that we base our perceptions of speaker sex on SFF alone, given the high degree of sexual dichotomy to be found in other measures. The formant frequencies appear to be a further powerful source of dichotomy, although more investigation is needed to establish to which formants and to which phones this applies. Finally, the amplitudes of the first two harmonics also appeared to be capable of sex-differentiation.

3. Is there such thing as an average speaker for each sex?

Assessing the extent of speaker variability

Figures for acoustic-phonetic measures are often reported in the literature as if they apply to all speakers of a particular type, for example, the fundamental frequency of men *is* 120Hz. The research reported in this thesis has shown that within any identifiable group of speakers there is generally a substantial amount of both within- and between speaker variation in the value of the measure

The extent of between-speaker variability for a particular acoustic-phonetic measure can be investigated by examining the distribution of SPEAKER MEANS, and by using the standard deviation of the SPEAKER MEANS as a rough guide. For a normally-distributed sample, the interval encompassed by the s.d. of the sample contains 68.3% of the data. While the distributions examined in this thesis are unlikely to ever be *exactly* normal, they are generally close enough to allow the s.d. to serve as a rough guide to the range containing the greatest density of data. Consider, for example, the SFF results. The overall mean female SFF was measured at 208Hz, with a s.d. of 23Hz. From this we can reasonably state that the majority (i.e. approximately 70%) of female SPEAKER MEANS fell within the range 185-231Hz. The remainder of the means fall outside this range (between 146Hz and 270Hz), and may be considered as outliers. However, for the thorough characterisation of female SFF characteristics, it is essential that they are included in any description of female SFF. The approximately 15% of the women who had a mean SFF between