SYNTHETIC VOICE DESIGN AND IMPLEMENTATION:

A HUMAN FACTORS INVESTIGATION

Christopher K. Cowley, B.A. Hons.

A thesis submitted in fulfilment of the requirements of
Bournemouth University for the degree of
Doctor of Philosophy.

1999

Abstract

The limitations of speech output technology emphasise the need for exploratory psychological research to maximise the effectiveness of speech as a display medium in human-computer interaction.

Stage 1 of this study reviewed speech implementation research, focusing on general issues for tasks, users and environments. An analysis of design issues was conducted, related to the differing methodologies for synthesised and digitised message production. A selection of ergonomic guidelines were developed to enhance effective speech interface design.

Stage 2 addressed the negative reactions of users to synthetic speech in spite of elegant dialogue structure and appropriate functional assignment. Synthetic speech interfaces have been consistently rejected by their users in a wide variety of application domains because of their poor quality. Indeed the literature repeatedly emphasises quality as being the most important contributor to implementation acceptance. In order to investigate this, a converging operations approach was adopted. This consisted of a series of five experiments (and associated pilot studies) which homed in on the specific characteristics of synthetic speech that determine the listeners varying perceptions of its qualities, and how these might be manipulated to improve its aesthetics.

A flexible and reliable ratings interface was designed to display DECtalk speech variations and record listeners perceptions. In experiment one, 40 participants used this to evaluate synthetic speech variations on a wide range of perceptual scales. Factor analysis revealed two main factors: "listenability" accounting for 44.7% of the variance and correlating with the DECtalk "smoothness" parameter to .57 ($p<0.005$) and "richness" to .53 ($p<0.005$); "assurance" accounting for 12.6% of the variance and correlating with "average pitch" to .42 ($p<0.005$) and "head size" to .42 ($p<0.005$).

Complimentary experiments were then required in order to address appropriate voice design for enhanced listenability and assurance perceptions. With a standard male voice set, 20 participants rated enhanced smoothness and attenuated richness as contributing significantly to speech listenability ($p<0.001$). Experiment three using a female voice set yielded comparable results, suggesting that further refinements of the technique were necessary in order to develop an effective methodology for speech quality optimization.

At this stage it became essential to focus directly on the parameter modifications that are associated with the the aesthetically pleasing characteristics of synthetic speech. If a reliable technique could be developed to enhance perceived speech quality, then synthesis systems based on the commonly used DECtalk model might assume some of their considerable yet unfulfilled potential.

In experiment four, 20 subjects rated a wide range of voices modified across the two main parameters associated with perceived listenability, smoothness and richness. The results clearly revealed a linear relationship between enhanced smoothness and attenuated richness and significant improvements in perceived listenability ($p<0.001$ in both cases).. Planned comparisons were conducted between the different levels of the parameters and revealed significant listenability enhancements as smoothness was increased, and a similar pattern as richness decreased. Statistical analysis also revealed a significant interaction between the two parameters ($p<0.001$) and a more comprehensive picture was constructed.

In order to expand the focus of and enhance the generality of the research, it was now necessary to assess the effects of synthetic speech modifications whilst subjects were undertaking a more realistic task. Passively rating the voices independent of processing for meaning is arguably an artificial task which rarely, if ever, would occur in 'real-world' settings. In order to investigate perceived

quality in a more realistic task scenario, experiment five introduced two levels of information processing load. The purpose of this experiment was firstly to see if a comprehension load modified the pattern of listenability enhancements, and secondly to see if that pattern differed between high and and low load.

Techniques for introducing cognitive load were investigated and comprehension load was selected as the most appropriate method in this case. A pilot study distinguished two levels of comprehension load from a set of 150 true/false sentences and these were recorded across the full range of parameter modifications. Twenty subjects then rated the voices using the established listenability scales as before but also performing the additional task of processing each spoken stimuli for meaning and determining the authenticity of the statements.

Results indicated that listenability enhancements did indeed occur at both levels of processing although at the higher level variations in the pattern occured. A significant difference was revealed between optimal parameter modifications for conditions of high and low cognitive load ($p<0.05$).

The results showed that subjects perceived the synthetic voices in the high cognitive load condition to be significantly less listenable than those same voices in the low cognitive load condition. The analysis also revealed that this effect was independent of the number of errors made. This result may be of general value because conclusions drawn from this findings are independent of any particular parameter modifications that may be exclusively available to DECtalk users.

Overall, the study presents a detailed analysis of the research domain combined with a systematic experimental program of synthetic speech quality assessment. The experiments reported establish a reliable and replicable procedure for optimising the aesthetically

pleasing characteristics of DECtalk speech, but the implications of the research extend beyond the boundaries of a particular synthesiser. Results from the experimental program lead to a number of conclusions, the most salient being that not only does the synthetic speech designer have to overcome the general rejection of synthetic voices based on their poor quality by sophisticated customisation of synthetic voice parameters, but that he or she needs to take into account the cognitive load of the task being undertaken. The interaction between cognitive load and optimal settings for synthesis requires direct consideration if synthetic speech systems are going to realise and maximise their potential in human computer interaction.

Contents

Chapter eight: Major findings and final conclusions

Acknowledgements

I would like to thank the following people for their help and encouragement:

Dr David Heathcote,
Dr Linda Hole,
Professor Dylan Jones,
Dr Alison Murray,
Dr Philip Kreeger,
Sonya Cowley,
and my parents

The contents of this thesis are the results of the candidates own investigation, except where stated to the contrary. The work has not been submitted, nor is it being currently presented for any other degree.

Refereed publications and resources associated with this thesis

Cowley, C.K., Miles, C. and Jones, D.M. (1990) The incorporation of synthetic speech into the human-computer interface. in Lovesey, E.J. (Ed) Contemporary Ergonomics, London: Taylor and Francis. (Presented at the 1990 Ergonomics society conference at Leeds University).

Cowley, C.K. and Jones, D.M. (1992) Synthesized or digitized? A guide to the use of computer speech. Applied Ergonomics, 23, 3, pp 172-177.

Cowley, C.K. and Jones, D.M. (1992) More than meets the eye: issues relating to the application of speech displays in human-computer interaction. Displays: Technology and Applications, 13, 2, pp 69-75.

Cowley, C.K. and Jones, D.M. (1992) A Human Factors guide to computer speech: Design criteria for the effective implementation of speech technology into computer applications. SIGCHI, 24, 2, ACM Press.

Cowley, C.K. and Jones, D.M. (1993) Assessing the quality of synthetic speech. in Baber, C. and Noyes, J (Eds) Interactive Speech Technology London: Taylor and Francis. Presented at Interactive Speech Technology at NEC Birmingham, 15th May 1993

The major findings from the first three chapters were also presented in two research films written and directed by the author and screened at CHI'90 (Colorado U.S.A.) and CHI'95 (Boston U.S.A.):

A human factors guide to computer speech *ACM SGVR* (1990)
Voice systems: An inventor's guide *ACM SGVR* (1995)

Section One:

Human factors and computer speech technology: An ergonomic overview

Foreword to section one:

Aims and objectives

Research initiatives with the domain of computer speech input and output technology attempt, at varying levels of detail, to duplicate human capacities for speech interpretation and production. As yet however, many such implementations fall short of fully realising this aim and this is especially so in the case of speech synthesis. The limitations of the technology emphasise the need for thorough psychological research to maximise the effectiveness of systems at their current and potential levels of development. In order to achieve this aim it is vital to assess the requirements, expectations and reactions of users as they are confronted with technology which utilises the auditory modality to replace or supplement the visual-manual interfaces with which most are familiar.

In the case of synthetic speech output, it would seem that the all too common dissatisfaction with synthetic voices (discussed in Chapters 2 and 3) may result in a lowering of motivation to accept them within interfaces. It has been established that more cognitive effort is required to understand synthetic speech-and the additional effort is unlikely to be applied by an unmotivated listener.

Text-to-speech synthesis is too much of a passive, mechanical activity. At the time of writing, text is automatically converted to speech independently of meaning and, to a large extent, grammar. It is hardly surprising that such speech is perceived as unpleasant, unnatural, and repetitive, even when it is highly intelligible. Yet, improvements in technological sophistication may not solve this problem, no matter how remarkably advanced the syntactic and semantic parsing, no matter how elegant and sophisticated the algorithms, applications are not likely to be accepted by users who are either unmotivated to comprehend synthetic voices, or who find the use of speech an irritant due to inappropriate

implementation strategies. A truly conversational computer is essentially redundant if no one wants to converse with it.

In order to address these issues, section one of the study consists of an extensive analysis of the use of computer speech synthesis and digitisation technology in human-computer interaction. The aim is initially to adopt a wide scope of inquiry in order to produce an extensive ergonomic overview of the research domain. This will involve the evaluation and refining of the foundations and development of ergonomic speech output research initiatives. This part of the study addresses the requirements and expectations of users of speech output devices within the context of both task and environment. Ergonomic guidelines for the effective implementation of text-to-speech synthesis and speech systems in general are developed and updated.

The overall objective of section one then is to determine the position of speech output research within the general context of human-computer interaction, to determine the various strengths and weaknesses of the spoken modality and to develop implementation criteria which will facilitate effective use of the technology.

The foreword to section two provides a statement of aims and objectives for the second, and major, stage of the study.

Chapter one:

The incorporation of computer speech into human-computer interfaces

1.1. Summary

The role of speech output technology in human-computer interaction requires a functional analysis of the attributes of the speech, the task, the environment, and most importantly, the needs of the users, their requirements, preferences and expectations.

In the process of prototyping and testing a range of prototype speech synthesis systems (including a 'teleshopping' system and a number of speech presentation systems for the experimental procedures in section two of this thesis), it became apparent that human factors considerations are of paramount importance to speech system design. Furthermore, analysis of the literature in the area revealed that many computer speech systems are built with inadequate reference to them.

Thorough examination of the literature clearly revealed the lack of systematic psychological research in this area. Human factors research appears to have been undertaken almost as an afterthought, rather than as a central priority. The overall impression was that speech system design has been almost exclusively technologically driven and that the enormous potential of speech output applications has been consistently undermined. Over and over again speech output systems and products have been introduced with a fanfare. These products are usually short-lived and are discontinued by manufacturers when users show indifference or active rejection. At the time of writing, except for a very few select (and dwindling) application domains, speech output technology remains marginalized and has the status of a gimmick or novelty.

In order to address this, the initial requirement of this research was an ergonomic overview of speech output literature and research to uncover appropriate and inappropriate design strategies or principles. The aim being to allow the development of a series of informed ergonomic recommendations highlighting some fundamental costs and benefits when implementing speech output systems, especially in cases where speech is to be selected to replace or enhance existing textual or graphical displays (a strategy which is prone to potential usability problems).

Based mainly on psychological and ergonomics literature concerning computer speech technology, and influenced by practical experience gained during design and testing of various speech systems, the first set of guidelines are concerned with the general application of speech output within the human computer interface. This aims to clarify where and how to use speech output to enhance human computer interactions from a psychological perspective whilst at the same time attempting to make the most of the technology.

Also covered in this chapter are some general concepts of relevance to effective speech system design. These include dialogue flexibility, the attribution of personality to synthetic (and indeed digitised/human) voices, and the intelligibility of computer speech.

1.2. Introduction

The degree to which the incorporation of speech into computer systems either enhances or degrades human-computer interaction depends on a number of inter-related factors concerning the functional appropriateness of speech to the task or sub task, environmental suitability, and the needs of particular user populations. Unless all three factors are adequately considered during the developmental cycle, a speech output implementation can easily fail. Varying levels of analysis conducted in these areas will determine system success. This is because even for closely similar applications, the addition of speech output can have a variety of effects on the perceived quality of the system and the effectiveness of the interaction.

Such complexities mean that an effective methodology for speech system development must necessarily be dynamic and flexible. It is not possible to invent design strategies for all possible system implementations and enhancements, therefore development initiatives are best presented in the form of ergonomic guidelines derived from on-going research and usability studies. These guidelines or recommendations could then be applied with flexibility across the potential application range.

1.3. The development of speech facilities for computer interfaces:

Historical context

Up until the 1970's speech input and output facilities both presented technical and cost effectiveness difficulties which discouraged their use in many computer applications. More recently, although technical problems are still delaying widespread implementation of speech recognition systems, for synthesis these problems have now

been greatly reduced. Consequently, the considered use of computer generated speech in system design is generally accepted to be a valid alternative to standard text-only media (Pisoni, Nusbaum and Greene, 1985, Oberteuffer, 1995, Cole et al. 1995, Liberman, 1995).

Although Wolfgang Ritter Von Kempelen constructed a talking machine in 1791 (made from a compressible leather tube, bellows and a reed), the earliest electronic synthesis of speech was achieved by Dudley in 1939. In the nineteen fifties and sixties various attempts to analyse and synthesis natural speech were developed and a wide range of synthesis systems were produced (Westall, Johnston and Lewis, 1996). Most of these were based on the source filter model of speech production. This technique entails an electronic simulation of the human vocal tract (see chapter 2). Formant synthesis, is an example of a source filter technique and is arguably the most successful technique to date. The DECtalk synthesiser used in this study is a well known example of a formant synthesis device.

With the advent of reliable and highly intelligible electronic synthesis, speech output functions could be incorporated easily and cheaply into virtually any system. Unfortunately, in many cases this led to speech output being implemented with little consideration as to whether it is particularly desirable or advantageous from the user's viewpoint.

In many cases, the human factors problems which resulted from the inclusion of speech into unsuitable systems could only be solved by removing the speech facility entirely (which defeated the whole object of the exercise). Peacock (1984) cites examples of such early implementation failures.

Coupled with the 'problem' of widespread accessibility and low cost was a prevailing inclination to include speech in computer systems

wherever possible based on a flawed conception of the capabilities of the technology and inaccurate assumptions about human nature (Jones, 1989). This tendency arose from a common belief that the introduction of speech would *automatically* enhance the performance of a system and that users would be thrilled to interact with the new output medium, enthusiastically embracing the latest technological 'breakthrough'. Unfortunately, this was rarely the case and a more realistic and common scenario was a pronounced decrement of task performance and efficiency and a growing irritation with the mechanical voice of synthetic speech and/or the repetitive nature of the dialogue. The irritating nature of repetitive speech is still reported in the literature (Vries, G. and Johnson, G.I. 1997).

"..voice synthesis promises to become as important and ubiquitous as the video terminal." (Gutcho ,1985)

Predictions that speech technology which would change the nature of human computer interaction fundamentally and irrevocably have clearly not been realised. When we consider computer speech technology as a whole, we can see that speech recognition systems appear to have finally reached a level of maturity where implementation is becoming a practical reality. However, with speech synthesis, implementation is not really making an impact at all except in highly specialised areas (see Chapter 3 for further discussion of this topic). This is despite the fact that highly intelligible speech is available cheaply and has been for some time.

To summarise, opportunities for widespread implementation of synthetic speech devices have, to a certain extent, been undermined by optimistic predictions based on false assumptions. Users have not rushed to embrace the technology with anywhere near the enthusiasm that the developers have anticipated. Before we address

the reasons for this, it is necessary to evaluate the functional potential of the spoken modality within human-computer interaction.

1.4. Advantages of the use of speech

Human speech production and reception (and the associated mental processes) have evolved to the point where verbal communication is perhaps the most effective and natural method of communication between people. Using highly sophisticated, multi-level code, we transmit and receive extremely detailed information. In theory, language processing skills could be especially beneficial if an individual is engaged in tasks which require active use of the hands or eyes in demanding conditions, for example, whilst driving a car or operating a keyboard and screen.

Computer speech technology has now advanced to the point where practical commercially viable output devices can be utilised for a variety of applications which will hopefully take advantage of some of these unique attributes of language as a mode of communication, yet widespread implementation has failed to occur. In order to understand why, it is important to determine where speech output may be really an appropriate facility to implement and where it most certainly is not.

The overall potential of speech technology lies perhaps in reducing or reallocating operator workload by providing an alternative input/output channel to the traditional visual-manual screen/keyboard arrangement. In some applications this may significantly reduce the user's workload and enhance efficiency.

There are a number of distinct and unique advantages of the spoken mode of human-computer interaction over traditional screen and keyboard control. The fact that messages can be conveyed in speech

without reliance on the visual channel is of particular relevance when the nature of the task does not require constant visual monitoring, such as a task where new information is only presented occasionally or irregularly. In some situations it may only be convenient for the user to look at visual displays at certain stages of the interaction (for example, whilst flying an aircraft, driving a car or inspecting a production line). For tasks such as these, audio output could be advantageous as a method of information presentation that allows unrestricted physical movement and visual attention. Distance and orientation to the source of sound is, up to a point, completely arbitrary and multi-tasking is facilitated.

As increased computational power facilitates the amount and rate of information exchange, the already cluttered visual landscape can become overloaded and inefficient. The ubiquitous multiple window environments common to many contemporary computer users can conceal or confuse relevant information and cause navigational problems for the user. Speech may help to relieve the congestion by providing a separable serial information source to either compliment the visual field or to transmit information which is unrelated to the visual task(s) in an appropriately distinct modality.

Speech is also well suited to bringing information such as status messages, feedback and especially, warnings or alarms to the attention of an operator, again this may be done regardless of the direction of the current visual focus (Baber, 1993). Stanton and Baber (1997) compare speech and text for alarm handling and conclude that while some speech alarms may be unsuitable for tasks that include a memory component, synthetic speech messages are useful for tasks where an immediate response is required, or where the operator is away from the interface.

Spoken messages certainly seem to have more of an impact on the user than messages presented in text. Research into speech

recognition, dialogue design and error correction has highlighted the failure of users to notice, respond to and act on visual feedback. This is in striking comparison to their reaction and response to spoken feedback. In one experiment, Frankish and Noyes (1990) report that subjects failed to correct seventeen per cent of recognition errors when visual feedback was used, but less than one per cent when spoken feedback was used. If errors are rare, such as when accuracy is high in a recognition task, then feedback monitoring failures are more likely to occur (Frankish and Noyes, 1993). In such cases the attention-grabbing properties of auditory messages can offer a significant advantage, although this alerting effect can be diminished by repeated use (Burns, 1979).

There are many different types and styles of speech output and many different tasks where speech can be utilized. Yet there has been little research conducted into matching specific types of voices to specific types of tasks. Furthermore, the research that has been done is limited to the selection of a suitable voice from a set of voices rather than the development of a principled methodology for researching the reasons for such a selection. It seems likely that certain types of voices would be more or less appropriate for various tasks, yet this has not been thoroughly studied. Experiment five in this thesis contributes to our understanding in this area.

(Section 1.9 provides a summary of some of these points in the form of design guidelines).

1.5. Disadvantages of the use of speech

In addition to a reduction of the alerting attributes of speech, excessive use of the audio channel often results in a negative evaluation of the system. In an experiment utilising a teleshopping simulation (Tucker, 1989), subjects were required to complete a

number of simple transactions in a purchasing task where information was presented both visually and auditorily (in the form of synthetic speech) to varying degrees. They were then required to rate the interaction on a number of scales (e.g. boring-interesting). Results showed that the quantity of speech output used was negatively correlated with the user's judgment of the acceptability of the system. The more speech was used, the less the users liked the system. Such results are by no means uncommon. Peacock (1984) reported incidents where drivers have disabled voice-warning systems in their cars because they resented the machines "telling them what to do", and that a speaking elevator quickly became irritating to regular users. System designers therefore need to consider the possibility that speech may rapidly lose any alerting, informative or aesthetically pleasing attributes and may become annoying to users exposed to it, leading to irritation and, in some cases, outright hostility. This problem is still reported. Vries and Johnson (1997) conducted an exploratory study investigating the potential for the implementation of spoken help into a car stereo system. They conclude that although the use of spoken help messages enhances the performance of (and would be acceptable to) novice users, "subjects became annoyed when they had to listen repeatedly to the same messages".

When examining the major drawbacks to the implementation of speech output, it is useful to distinguish between negative reactions based on the functional inappropriateness of spoken *information* and those based on a dislike of the characteristics of a particular *voice* used in the application. The former is concerned with whether or not the adoption of speech for particular messages enhances the functionality or the application and the latter is concerned with the personal preferences of the user and the potential attribution of personality based on the specific characteristics of the voice. Bearing in mind that a combination of poor design practice in both of these areas will most likely result in extremely negative perceptions of the system (indeed, an interface which presents useless messages spoken by an unpleasant voice is certain to totally alienate the user), these

two aspects of speech system design will be discussed separately.

1.6. Spoken information: dialogue characteristics

If speech messages are to be included in an interface, there should be sufficient flexibility in the dialogue to allow it to be adapted to accommodate a wide range of user experience. The number of speech messages needed for the user to successfully complete a task or sub task may well vary as expertise develops. Clearly, it may be useful in some systems to allow the user the option to switch between audio and visual modalities on occasions. As a general rule, future systems incorporating speech should be flexible in this regard. The user should feel empowered and be able to decide whether or not he/she wants more or less spoken or visual information during interaction.

A consideration of the differing expectations of users further highlights the importance of flexibility in speech system design. Most users of computers have preconceptions about their nature, capabilities and 'intelligence' which in turn determines the user's judgment of the level of skill required to use them. People tend to attribute far more intelligence to a system than it actually possesses as soon as the system communicates in natural language phrases (Boden 1984). Research by Simpson (1986) featured a hotel-booking simulation with speech input and output. Results showed that as the content of speech messages was made more natural by the inclusion of redundant words and phrases typical of human-human dialogues (e.g. "Please", "thank you"), users' responses echoed this redundancy, under the assumption that the machine could understand a much wider vocabulary than before. This was not observed for the same task with textual messages and feedback replacing the speech. So it appears that the use of speech, and the addition of dialogue conventions imported from human-human conversations, can increase the naturalness of the interaction. Consequently, the user may import expectations derived from human-human communication

and may then hold a greater over-estimation of the system's intelligence or capabilities than held with text-only systems. This will increase the likelihood of the user attempting to interact with the system at a level of sophistication that exceeds the capabilities of the programming and hardware. When the system then fails to live up to their exaggerated expectations, they are likely to find this especially frustrating. Jones (1989) refered to this tendency as "speech-induced anthropomorphism" and warns of the negative repercussions for speech interface design, especially when the system also has speech recognition facilities which require precise commands within a confined vocabulary set.

Technological advances over the last ten years have not removed the unrealistic expectations of users. Conversing with a computer that can successfully deal with spoken dialogue in a completely human-like way is still a long way off (Wolf et al, 1997). Hone and Baber (1999) recognise this problem and have developed a dialogue constraint methodology to try and model effective dialogues in order to enhance recognition accuracy in speech applications. However, unrealistic assumptions about what the technology can do remain (perhaps as a result of the cinematic portrayal of fictional speech systems as being error-free, unconstrained, natural language systems).

As the user's preconceptions are continually modified through moment-by-moment interpretations of evidence based on the success of their dialogue with the device, it is important to make sure that they have a clear idea of where they are within the dialogue and what input the computer can, and cannot deal with at particular stages. This may not be the easiest design requirement to implement, especially if the system has varied adaptive qualities. For example: if the system uses a mixture of textual and spoken output, the user will approach the system with certain expectations with respect to the system's responses to input. Such responses are often new instructions which represent implicit confirmation of the adequacy of the user's previous action. If the system does not respond (or is not

perceived to respond), the user may interpret this as an inadequacy in their previous response, or, a fault of the system. If a message in a particular modality is anticipated, but does not occur (due perhaps to the adaption process having streamlined the spoken instructions in line with variations in expertise), there may be an impasse where both user and system do nothing on the assumption that their 'turn' is complete and that their next action is dependent upon a response from the other. This contrasts with human-human interaction in which implicit confirmation is used as the basis for reception of a message (Hayes and Reddy, 1983).

Dybkjaer and Bernsen (1998) recognise the importance of diagnostic evaluation when developing high quality spoken language systems and state that no rigourous methodology exists for the systematic and exhaustitive diagnostic evaluation of all aspects of spoken language interaction. In order to address this they develop a set of guidelines which they recommend should be applied as a "design guide" prior to implementation of dialogue systems.

Dialogue styles are closely related to the usability of speech systems. Lauretta and Deffner (1996) investigated dialogue styles for telephone information interfaces and concluded that ease of use is determined by both cognitive load and the time required to complete a task. Consequently the design of sophisticated dialogue requires considerable ingenuity, and an exhaustive analysis of the fine details of the interaction is required.

1.7. Personality attribution

Users readily ascribe personality to speech synthesis systems on the basis of the device's speech style. For example, Michaelis and Wiggins (1982) reported that some listeners tended to consider unfamiliar synthetic voices as "evil and sinister". Unfortunately, attempts to improve the naturalness of synthesised speech by adding rule-

generated stress can result in decreased intelligibility (McPeters and Tharp, 1984). In this case there appeared to be a negative correlation between the intelligibility of the synthesised speech and its aesthetically pleasing characteristics.

Even speech of a very high quality can be regarded as unacceptable by users; e.g. a study which used digitised human speech (Edman and Metz, 1983) showed that even highly intelligible speech may be regarded as "machine-like", "harsh" or "flat". Assertions about users' adaptability to a synthesised voice, for example, that "people would probably learn to tolerate the verbal inadequacies of mechanical speech and would soon become accustomed to its quaint style" (Frude, 1983) - were therefore overly optimistic and unrealistic.

The rejection of synthetic speech on the basis of its lack of quality forms the basis of the experiments featured in the second part of the thesis and is discussed in detail in chapter 3.

1.8. Intelligibility

User-system dialogue problems may also occur because of difficulties with the intelligibility of speech. Synthetic speech in particular has been shown to be generally less easy to understand than natural speech and in some settings it may not be possible to identify a synthetic speech message correctly on its first presentation (Pisoni, Nusbaum, Luce and Schwab 1983). In such cases the dialogue may need to incorporate message repetition, message summaries, or an alternative information source for important messages, such as a visual display (which rather defeats the object of the exercise). Another tactic is to allow the user time to become practised in listening to the speech by providing an introductory message. A common observation is that people learn to understand synthetic speech (Pisoni, Nusbaum and Greene, 1985). New listeners usually

have far more difficulty in understanding a speech synthesiser than people accustomed to hearing the speech. It has been shown (Schwab, Nusbaum and Pisoni, 1985) that practice in listening to a synthesiser can rapidly improve a listener's ability to understand the output of a speech device.

Problems of intelligibility of spoken messages maybe due to the serial nature of speech and this may be exaggerated by both the complexity and quantity of the information to be conveyed. Instead of reviewing and correlating information by simply looking at different places on a screen, a user of a speech display has to both maintain and update information in his/her memory. For instance, Kidd (1982) has shown that the additional processing required by users of speech displays can reduce their ability to make accurate selections from acoustically presented menus. Speech menus with clear, non-overlapping categories of menu items posed few difficulties for users but menus with 'fuzzy' categories or unrelated items proved to be very difficult to understand and remember.

Having covered a number of general issues which require consideration during speech system development, the salient points in this chapter will be presented within a set of ergonomic guidelines. Firstly though, it is necessary to examine the methodological validity of ergonomic guidelines as a practical research and development tool.

1.9. Guidelines: An appropriate methodology?

Ergonomics has a significant contribution to make in the design of the human-computer interface. Speech technology has the potential to revolutionise the way we interact with computers and as systems are finally emerging into the public domain, ergonomics can provide us with valuable design strategies and detailed and specific guidelines

and standards. An emphasis on ergonomic analysis during system design is an important first step towards a quantification of the requirements and expectations of users, and their relationship to various tasks and environments.

Whilst by no means being a panacea for the development and evaluation of interfaces, the use of ergonomic criteria (including design guidelines) during the developmental cycle can certainly be beneficial. Ergonomic criteria can help designers to avoid common design errors, to identify and diagnose potential usability problems and to streamline the developmental cycle during iterative design (Bastien and Scapin, 1993).

The research of Williges and Williges (1982) summarised general characteristics that distinguish visual/manual interfaces from speech interfaces. These include rate of information presentation, availability of scanning, directionality of display, user memory requirements, and environmental influences on display/control usability. Whilst these points remain valid today, their suggested approaches to guideline development are a product of their historical context. These approaches consisted of:

(i) The reinterpretation of data from studies involving non-speech auditory displays (due to the fundamental differences between textual and speech displays, existing guidelines for visual displays and keyed-entry may not be appropriate for speech input and output).

(ii) The prediction of speech system requirements for effective communication by simulation of speech hardware capabilities.

(iii) The development of empirical and theoretical models of human-computer communication with speech displays.

It can be argued that since 1982, a considerable number of speech studies and research initiatives have been conducted and therefore it

is no longer so important to rely on reinterpretation of data from non speech research. Also, the need to simulate hardware capabilities has been largely mitigated by technological advances (less so in the case of speech recognition where 'Wizard of Oz' simulation may still hold value as a productive research technique).

Finally, whilst modelling and hypothesis-testing have their place, the development of effective and flexible guidelines must necessarily be a dynamic, iterative process incorporating an examination of the changing *functions* of speech and vision as they relate to particular applications. Any interactions between guidelines also need to be considered as do the extent to which design goals are being achieved throughout the development cycle. Although the use of guidelines will not eliminate the need for both iterative design of the interface and testing with an appropriate set of users, it should substantially improve the initial design of the dialogue and reduce the number of iterations required.

Guidelines for determining whether to adopt the audio or the visual channel for the display of information in computer systems were outlined as far back as Deatherage (1972). The general principles and recommendations from this early research have remained largely consistent within the literature. Michaelis and Wiggins (1982) published guidelines which cover similar points which are still being cited today (see Shneiderman, 1998). Having established the appropriateness of the choice of the speech modality for specific functions in the computing process, these generally accepted recommendations can now be updated and expanded.

To conclude, a guideline approach in this area is only likely to be effective if the guidelines are considered flexible recommendations rather than rigid rules. Changes and development in speech application demands and technological refinement may make individual guidelines more or less useful. New guidelines may need to be constructed and existing ones modified if we are to keep pace with

and enhance system evolution with the use of speech.

1.10. Summary: Guidelines for the development of speech-output displays

Having discussed many of the areas of concern to the speech system designer, it is now possible to summarise this chapter by making some initial recommendations and observations. These relate to the use of the spoken and textual modalities within computer applications, and may be particularly relevant to decisions to replace textual messages with spoken ones. They may also be of use during the construction of systems which utilise both modalities. These guidelines offer an extension and elaboration of Deatherage's original list. Whilst some of these may appear simplistic or obvious, it is necessary to quantify the basic components of an interface in terms of functionality. The assignation of speech or text to an inappropriate function is likely to result in serious usability problems and this should clearly be avoided, at the risk of presenting an oversimplification of design criteria.

In order to encompass the varying aspects of potential speech interfaces, they have been subdivided into three sections relating to user, task and environment.

The user:

1. Use speech if the message is simple. Use text if it is complex. A textual message can be easily reviewed.

2. Use speech if the message is short. Use text if is long. A long spoken message can be tiring to listen to and the transient nature of speech may result in some of the information being forgotten.

3. Use speech or text in order to maintain modality consistency within defined stages of the interaction. i.e. don't use one modality if the other is anticipated.

4. Do not use speech and text modalities simultaneously. The user may miss an auditory message whilst concentrating on reading a textual one, or the reverse.

5. Do not present important information in speech until the user has had an opportunity to listen and attune to the speech style and quality. People *learn* to understand synthetic speech through practice and exposure.

<u>The task</u>:

6. Use speech if the user's job requires them to move about continuously or to be frequently away from a terminal. Use text if the user's job allows them to remain in one position.

7. Use speech if the message will not need to be referred to later. Use text if the message will need to be referred to later.

8. Use speech if the message calls for immediate action. Use text if the message does not call for immediate action. Speech demands attention and is particularly suited to feedback and urgent warnings.

<u>The environment:</u>

9. Use speech if the auditory system of the user is not overburdened. Use text if the visual system of the user is not overburdened. For example, a spoken message may be lost amongst the variety of sounds in a noisy office environment, a textual message may go unnoticed if it appears on a particularly cluttered screen.

10. Use speech if the receiving location is too bright, or dark adaptation is necessary. Use text if the receiving location is too noisy.

11. Use text rather than speech if the message contains information that requires security. i.e. passwords. Unless headphones are used, speech is a public event and people in the vicinity may overhear sensitive or confidential information. Even if the information is not confidential, speech output devices can also be very irritating to other people in a public setting.

Chapter Two

Speech synthesis and digitisation: Technological limitations, implementation and user preferences.

2.1 Summary

This chapter discusses a number of issues relevant to the implementation and use of the various types of computer speech displays. Both high-quality, computer-generated synthesised speech and digitised human speech are now available for implementation in a wide range of applications. Some devices can produce either types of output. Consequently, the designer may have to determine the appropriateness of the different technologies for particular applications. In addition to technical and cost determinants, a range of human factors issues govern the appropriateness of synthesised or digitised speech for a particular system. The choice to implement a particular technology should therefore be an informed decision which takes into account the skills and preferences of potential users, the demands of the task and operational setting, the limitations of the technology, and the psychological issues relating to speech production, perception and understanding.

Digitised messages are created auditorily and synthesised messages textually. There are a number of differences between spoken and written language, some of which may cause receivers to interpret the intended meaning of the same message in different ways. Consequently, the different techniques require consideration at points of message creation, editing, transmission and reception.

Synthesised speech is easy to edit in terms of lexical content, but the production of sophisticated prosodic features is not always straightforward. In contrast, unedited digitised speech will normally have appropriate prosodic cues and therefore may be the favoured medium for many applications. Although primitive, the rule-based prosodic features applied to synthesised speech remain consistent as messages are edited or concatenated. Whereas, for digitised utterances, editing or concatenation often results in a disruption of prosodic cues and a unintelligible message. This chapter concerns a number of issues relating to the design, implementation, and use of speech output. It gives special prominence to the cognitive processing of computer speech and provides illustrations of both good and poor design practice. Topics covered include a consideration of users' preferences, the importance of prosody, usability issues, and the psychological implications of textual and spoken input. It concludes with guidelines for the development of synthesised and digitised speech displays.

2.2 Introduction

Until recently, computers almost exclusively used one input and one output medium: textual input generated by a keyboard or keypad and visual output displayed on a VDU screen or printer. Although developments in direct manipulation interfaces (utilising windowing environments and analogue input devices such as mice) have enabled users to organise and process greater quantities of information, in many cases the amount of information available to the user has increased proportionately. This is often a result of modern, high-speed networking interfaces which allow users access to a number of applications simultaneously on a single terminal. Consequently, for many tasks, the visual channel has become overloaded.

Use of spoken output offers potential solutions to this problem. A range of devices is available which generate speech mechanically or output stored speech messages. These are both inexpensive and relatively easy to use. It is clear however, that many people have expectations of speech technology that go far beyond what is currently available and possible.

Two major types of technology are available: synthesised speech, in which textual information serves as a basis for a process by which speech is generated using a set of rules, and digitised speech, in which a speaker's utterances are stored for reproduction in digital form and may be called up in their original or edited form at will.

Human factors research can help to determine when, where and how synthesised and digitised speech technology should be used to enhance human-computer interaction. Effective use of the different technologies depends upon variables which include the characteristics of the user's preferences, and the constraints imposed by the task and physical environment. It is unlikely that any particular type of speech device or individual voice will be suitable for speech output in all situations. Rather, the combination of task and user characteristics associated with a particular interface environment will dictate the applicability of the different speech technologies.

2.3 Technological limitations

Although some systems have utilised non-speech audio to provide information and feedback (eg, Apple Macintosh computers which utilise a limited selection of basic auditory cues/alarms), the use of speech has been frequently overlooked because of technological

limitations. For example: in the case of digitised speech, usage has often been precluded by computer memory constraints. In the past, if an extensive vocabulary was required, few systems would have had sufficient memory to make the implementation a practical option, especially if memory-intensive, high-fidelity digitised speech was required. Technological developments such as data compression techniques and the availability of cheaper computer memory have reduced the extent of this problem and may ultimately eliminate it.

In the case of synthesised speech, the availability of computer memory is not an issue since messages are usually created when they are required rather than stored in memory. Despite this, and a number of other advantages, speech synthesisers have not achieved the widespread implementation predicted for them in the early 1980s. This is because the quality of output from prototype systems has frequently provoked very negative evaluations from users. This has been observed in a number of studies and has resulted in very few systems utilising synthetic speech output (this is further discussed in Chapter 3).

However, the technology is available to allow speech displays to *potentially* assume an equal footing with other more traditional interfaces. Some contemporary devices can utilise both machine-generated synthesised speech and digitised human speech to either supplement visual displays or to provide alternative speech-only interfaces. Applications which may benefit from the inclusion of computer speech are widespread and include document annotation, message store-and-forward systems and computer-aided instruction packages. Furthermore, there is great potential for speech as an aid to navigation through complex interfaces. It can be combined with visual cues to inform users of changing states in the system or the task at hand. This may be especially beneficial in multi-tasking environments where a number of tasks proceed concurrently and

information pertaining to their status exceeds the constraints of visual representation.

2.4 The Importance of Human factors

While the choice of the right device for the task may be influenced by technical considerations, the success of the system in terms of acceptance and usability is mainly determined by the emphasis placed on human factors issues during interface design and usability testing. Informed decisions on when, where and how to use the devices are particularly dependent on an understanding of the psychological processes involved in the editing and cognitive processing of messages.

An appropriate human factors research focus for speech technology is not achieved by the provision of generic rules, but by applying flexible recommendations and informed insights derived from research to each unique scenario. The costs and benefits of employing speech devices are dependent upon a number of variables, including the preferences of potential users and the constraints imposed by the task and environment. It is unlikely that any particular type of speech device will be suitable for speech output in all situations. Rather, the combination of task and user characteristics associated with a particular interface will determine the applicability of the different speech technologies.

2.5 Techniques for composing and editing computer-speech messages:

Synthesised speech

Speech synthesis refers to a machine's conversion of written language into auditory form. A synthesiser is a sound-generating mechanism which is functionally analogous to the human vocal tract. Synthesisers electronically model the physical structures of the anatomy such as the oral and nasal cavities and the vocal cords. The human vocal tract is physically complex and a clear understanding of the functional relationships between the various components has, and is likely to, remain an active research topic (Klatt and Klatt, 1990).

2.6 Text-to-speech

Text-to-speech is the conversion of unrestricted text to an acoustic speech signal. This is not easily achieved and involves a series of complex processes. At the beginning of this decade, the understanding of the dynamic behaviour of the human system permitted the synthesis of intelligible machine speech, but not speech that was indistinguishable from that of a human being (White, 1990). Progress over the last few years in this area has not led to any major significant improvements.

For many devices, synthetic speech is mechanically generated using rules of correspondence between written words and acoustic sound signals. This is usually accomplished without reference to actual human speech although the sequence of events involved in text-to-speech synthesis are a reflection of the human cognitive capability for reading aloud (Allen, 1980).

Text is entered into a keyboard, this is then 'normalised'. Most texts contain a great deal of symbolic material such as numbers and abbreviations which require transformation. In the normalisation stage, these are converted into their full alphabetical forms. "Dr" becomes "doctor", "misc" becomes "miscellaneous", "100" becomes "one hundred" etc. This is by no means straightforward because the pronunciation of various abbreviations is frequently determined by context, eg. "Dr." could refer to both "Doctor" and "Drive" (in an address). Punctuation is also subject to variability due to context.

At the next stage an exceptions dictionary is employed to identify words which are exceptions and differ from standard pronunciation conventions (for example: *pint*, when compared with *mint, stint, tint, lint, flint* etc). Most modern devices contain both in-built and user-definable exception dictionaries for words that deviate from standard pronunciation conventions (i.e. DECtalk, which has a six thousand word exception dictionary), but such dictionaries cannot realistically hope to cover more than a limited sub-section of the myriad of alternative pronunciations in language.

In the event that the word is missing from the dictionaries, then letter-to-speech rules are used. Instead of using entire words, 'morphs' are used, these are the units of speech that go to make up parts of words. The rules needed to create natural sounding speech this way are highly complex. English has by far the most complex relationship between spelling and word pronunciation of any alphabetic language (O'Malley, 1990) and the development of a rule-based system which can tackle the process of text-to-speech conversion with anywhere near the precision of a human speaker is certainly a formidable task.

White (1990) stresses that speech is remarkable for both the variety of rules it follows and also for the number of rules it violates. Take the letter combination: *"gh"* This can be pronounced in a wide

variety of ways depending on the word it appears in. For example, in *"ghost", "through", "thorough", "bough",* and *rough"* the letters are pronounced in completely different ways and in the word *"thought"* they are not pronounced at all. It may be possible for a meticulous designer to identify a significant proportion of exceptions like these and compile them into a dictionary but there are many exceptions that arise in day-to-day communication which cannot easily be predicted. For example, many surnames have idiosyncratic pronunciations which vary between regions, dialects and cultures. Street and place names are also subject to wide variability. Unpredictable exceptions are almost certain to feature regularly in many implementations and users may have neither the time nor the inclination to do the necessary modifications needed to resolve them. As can be seen, the development of a rule-based system to accurately convert text into speech is by no means a simple process.

At the next stage, the text is subjected to comprehensive stress placement rules. For many commercial synthesisers, stress assignment is carried out on individual words rather than complete sentences. This can lead to inadequate pronunciation because, although the words may sound adequate when spoken individually, stress assignment is often very different within the context of a sentence and the result may be stylised and unnatural sounding speech (Waterworth and Talbot, 1987). It may be necessary to further adjust the phonemic representation according to context by applying morphophonemic rules. Additional phonological rules are used to determine appropriate allophonics, parameters are specified to modify features such as pitch and speaking rate (these can be set by the user), and a waveform is created to produce the final auditory output.

2.7 Formant speech synthesis

A formant speech synthesiser generates speech using appropriate sets of time-varying parameters. These consist of the formants of speech and their corresponding bandwidths and amplitudes. This technique attempts to copy the formants in human speech (the peaks in the energy spectrum of the natural speech wave). It is achieved using a set of rules which vary a set of parameters over time. The Formants are physically produced using a set of filters which resonate at varying frequencies.

Two types of formant synthesizers exist, parallel and cascade (see figure 2.1). In parallel, an excitation filter is applied to all the formants and their outputs are combined, whereas in cascade (or serial) synthesis, the output of each formant forms the input for the next. Both techniques have strengths and weaknesses. Parallel synthesis is generally better for producing nasal and fricative sounds while cascade synthesis produces better non-nasal voiced sounds (Linggard, 1985).

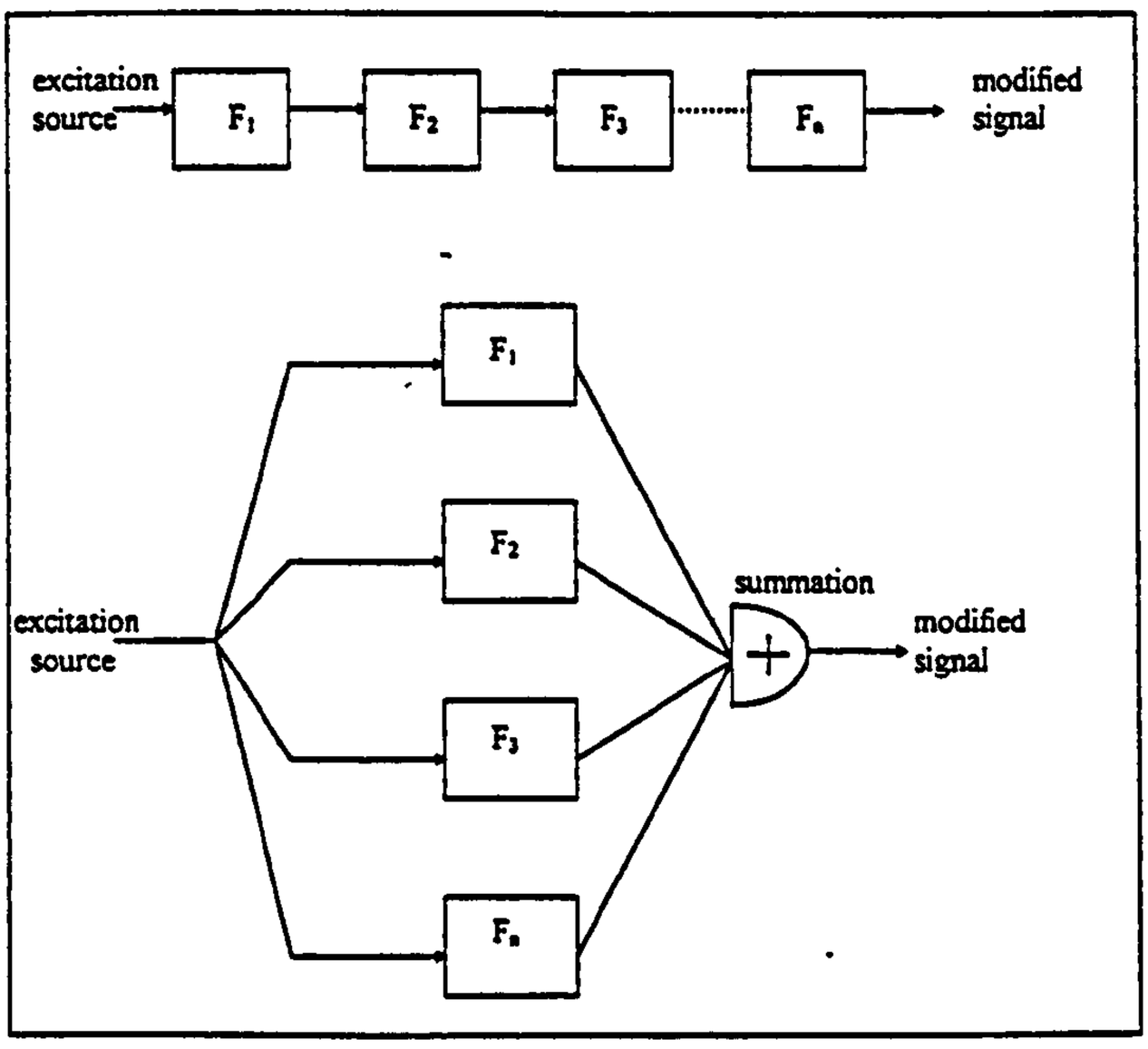


Figure 2.1. A simple diagram of speech synthesis models.. Cascade/serial is shown at the top and parallel at the bottom (Adapted from Vine, 1999)

Klatt, in the early 1980s, capitalised on the strengths of these two techniques by combining them and introducing some sophisticated refinements to produce the generic synthesis by rule system that was incorporated into many commercial systems and is still widely used today.

DECtalk, the synthesiser used in this research, uses rule-based or formant speech synthesis, (although it also has a text-to-speech capability). DECtalk produces speech by using a combination of cascade and parallel synthesis. Although there are other methods of producing mechanical speech and various refinements have been made to Klatt's original model, DECtalk produces the style of speech which most people associate with computer speech synthesis and the DECtalk technology is still being implemented in various applications today. For example, Toby Churchill communication aids for people who cannot speak (winner of the 1996 dti export award).

2.8 Digitised speech

Digitised speech is recorded human speech which is filtered, sampled and usually stored electronically in a compressed form. The samples can then be retrieved and- reconverted to analogue form using a digital-to-analog-converter. This process reproduces the speech that was originally recorded with little or no degradation or effects on intelligibility. Although there may be some loss of speech quality due to the sampling rate and the number of bits used to code the speech wave-form, the resulting speech quality is usually acceptable. When listeners hear such speech they typically have little difficulty perceiving or understanding it (Edman and Metz, 1983). Whilst in digital form, the samples can be edited using one of a range of software applications designed for the manipulation of digitised sounds. Words or phrases recorded separately can be concatenated and in other ways modified, or complete sentences can be recorded

and played back. Such devices are clearly limited to the vocabulary elements available and bear little or no relation to the ways humans produce speech (Waterworth and Talbot, 1987). The technique of concatenation is also highly likely to result in a confusing-sounding message which can require the allocation of additional cognitive resources on the part of the listener to ensure accurate interpretation. In 'normal' speech, features such as pitch and intonation normally change smoothly, stressed segments are equally spaced in time, and pronunciation and intonation depend strongly on context. This is essential in order to achieve fluency and naturalness. Consequently, if single stored versions of each word are joined together in the arbitrary orders that might be needed for a wide variety of messages, the concatenation process will usually disrupt natural intonation and rhythm. Concatenated messages can be extremely difficult to listen to, even when the quality of the individual words is very high (Holmes, 1988).

This problem can be addressed with the process of perceptual centring. Put simply, this technique involves pin-pointing the psychological moment when a listener perceives that a word has been spoken. Using this as a reference, the gaps between the words can be adjusted and the concatenated message has a much more natural delivery. (For detailed discussion of the technical issues involved in this process, see Morton et al, 1976, Marcus, 1981, and Cooper et al, 1986).

The different techniques involved in the creation of digitised and synthesised speech displays can determine the usefulness of these devices when they are incorporated in particular applications. Operational settings and user characteristics can indicate the potential benefits or problems of employing either technology. For speech digitisation, a quiet setting and a practised reader and clear speaker are needed to establish a good standard of clarity. A message may be unintelligible simply because the voice is fatigued or the

speaker is in a dusty or otherwise inhospitable environment. Problems may occur if the reader possesses a strong or unusual accent which others may find confusing or if they are simply unpracticed or self-conscious (this is likely to be particularly relevant when other people are present). For synthesised messages, a different set of skills are required. Although machines can take any ASCII text and convert it to speech, messages may benefit from the user's programming skills, technical knowledge of the potential of the device and literary competence may be necessary to produce natural sounding results. A user adept at creating quality digitised messages may lack the skills necessary to program a synthesiser to produce sophisticated output.

The reverse can also occur. The fundamental differences between the two processes suggest that the choice of either technology for applications must take into account the skills of the user and how these may interact with the task at hand.

Users of speech output devices require facilities that offer a high degree of direct manipulation of the output message (Gould and Boies, 1984). With currently available systems, users can create and manipulate synthesised speech simply and easily, using text-editing skills which most computer users will have already developed. With digitisation, although generation of an original message is simply a case of voice recording, any manipulation of the message will involve the use of a specially designed interface and may be a laborious, exacting and time-consuming process.

Furthermore, editing of digitised speech usually involves listening to the message. In contrast, synthesised speech can be created and edited by manipulating the textual basis, without having to listen to the message. One side-effect of this difference is that since the presentation of speech is inherently a more public event than the use of text alone, the construction of private or confidential digitised messages may, in some situations, be problematic. (Stern, 1984).

Once a particular speaker has been selected for a digitised speech application, the system, in order to appear consistent, is dependent on that same human speaker for new vocabulary. It is not possible to add even a single new word without making a new recording. Subsequently, revised digitised recordings can be unsatisfactory in the absence of the original speaker as the change in voice may prove disorientating to the user and may fragment the flow of the message. In contrast, messages for synthesis can be edited, adapted and revised by a number of individuals in a number of environments without becoming degraded or suffering disruptions of continuity.

Although digitised waveforms can be displayed visually, the semantic content of digitised messages can only be determined by the message in its auditory form. Auditory information is evanescent: once uttered, the information is not readily available for inspection. The information may be misheard, misinterpreted or forgotten and the user must repeat a procedure to display the information again. Synthesised speech messages are not confined to auditory display and can be composed and viewed in textual form. This enables messages to be attended to, reviewed, printed out and cross-referenced at leisure. With digitisation, the message would need to be listened to and transcribed if a written version was required.

2.9 User requirements and preferences

How and why people select particular communications modalities for certain tasks is an important issue for both behavioural researchers and software product developers, yet little is known about the criteria for such decisions (El-Shinnawy and Markus, 1997).
User preferences for digitised or synthesised speech systems may be partly determined by the different input modalities of the two devices (i.e. speech and text), the nature of the application and the basic acceptability of the voices that can be produced by either

system.

As far as the task is concerned, in 'benign' conditions (i.e. when the task is either short, easy or undertaken infrequently), users perceive both vocal and textual input modalities as equally useful.

With more stressful conditions, research has shown that the spoken modality is preferred unless the task is a spatial one (Robert, Fiset and Bergeron, 1989). If the user is undertaking a task in which a great number of voice annotations are required, it is likely that the additional effort required to compose and type messages for synthesis (in contrast to the simple and quick recording of digitised messages) will make the task more arduous. Generally, consistency between input mode and output mode is preferred by users. This may determine preferences for digitised speech (where the audio modality is used for both input and output) over synthesised speech (where different modalities are used).

Subjective acceptability of machine speech, as indicated by preference ratings, does not necessarily correlate with objective performance (Rosson and Mellon, 1985). i.e. out of a range of voices, a voice that yields the highest performance when it is used to undertake a task, will not necessarily be rated as the most acceptable of the set. Despite this, desirable voice characteristics are, to some extent, application dependent. In some situations, users may prefer a machine-style synthesised voice, in others, a natural-sounding digitised voice may be preferred. Such preferences are determined by the extent to which the application simulates human communications. A machine-style voice may be preferred for conveying system messages whereas a human-style voice may be preferred for messages from another user (Simpson, McCauley, Roland, Ruth, and Williges, 1987). Although most synthesised speech devices offer a number of different voices, and careful tuning of the source text can add a degree of dialect or accent, the variety of possible digitised voices is potentially unlimited. Accordingly, the

task of finding an acceptable voice for a particular application may be facilitated by digitised speech to a greater extent than by synthesised speech.

2.10 The importance of Prosody:

Prosody refers to paralinguistic features which modify the sense of a sentence by changing intonation, stress or timing of the speech. It is usually inherent in digitised speech and can be beneficial in conveying the essential meaning of messages, and prosodic features can help the user to grasp the most important information in messages of an instructional nature (Van Ness, 1986, Negroponte, 1995). A major weakness with available text-to-speech synthesisers is the restricted intonational repertoire. This may partially account for the tedious repetitiveness which most users experience when listening to more than a few sentences produced by a particular system. Indeed Silverman (1985) blamed the combination of restricted repertoire and phonetically incorrect output for the generally perceived unpleasantness of synthesised speech.

Prosody is a natural part of speech production; seldom is speech without it. With synthesised speech, messages are normally typed and consequently may be devoid of the idiomatic cues that enable the listener's interpretation of intended meaning. Furthermore, most synthesisers can apply rule-based prosodic features to messages which may be inappropriate and may then obscure the meaning. The achievement of a degree of prosody comparable to digitised speech will usually entail a considerable knowledge of the capabilities and subtleties of the speech device and much time-consuming manipulation of source code and voice parameters. Inadequate prosody may demand increased cognitive effort from the receiver of synthesised spoken output as he/she attempts to make sense of the message. It can also result in a misunderstanding of the intended meaning of the message (Pisoni, 1981). For example, consider the

message "console operators to contract changes". In textual form, the meaning is ambiguous because of the different ways the words "console" and "contract" can be pronounced. With a sentence such as this, rule based synthesisers are only likely to produce the intended pronunciation by chance. By way of contrast, an unedited digitised version would normally have the appropriate emphasis automatically supplied by the speaker. Considering the complexity of English pronunciation, it is likely that messages for synthesis will frequently require adjustments at the text-input stage.

Synthetic speech increases the load on information processing capacities, especially short-term memory. Luce, Feustel, and Pisoni (1983) discovered that the perception of synthetic speech imposes greater demands on processing capacities than those imposed by the perception of human speech. Extra effort is required to encode phonetic segments of speech which have inadequately specified acoustic cues. As a consequence, fewer resources will be available for any additional tasks that listeners are asked to perform simultaneously and which draw on the same resources. As the memory capacity is limited, extra allocation to encoding may well affect the rehearsing of earlier items. Luce et al's study examined free recall of a list of both synthesised and naturally spoken words for a number of different presentation conditions designed to vary in difficulty (e.g., subjects were required to perform additional short-term memory recall tasks). Results showed that as the difficulty of the tasks increased, there was a more rapid decrement in performance for the synthetic than for natural speech (Figure 2.2). They concluded that there is an overall decrement in performance with synthetic as opposed to natural speech. These findings have been duplicated (Waterworth and Holmes, 1986). One method of improving the recall of synthetic speech is to use larger than natural, but grammatically correct, pauses in the sentence structure. Such exaggerated pauses facilitate rehearsal (Nooteboom, 1983), but still do not overcome the need for greater processing capacity in comparison to human speech processing.

It appears possible that the negative consequences of employing synthesised speech rather than digitised speech are apparent, not from measures of intelligibility and comprehension (Jusczyk, 1986), but rather from the extent that synthesised speech imposes a greater load on processing capacities. Very often the extra effort required to understand synthetic speech results in a reduction of spare information processing capacity, even though no reduction in intelligibility is shown. This effort will only become apparent if the person undertakes another task or if something has to be held in memory whilst listening.

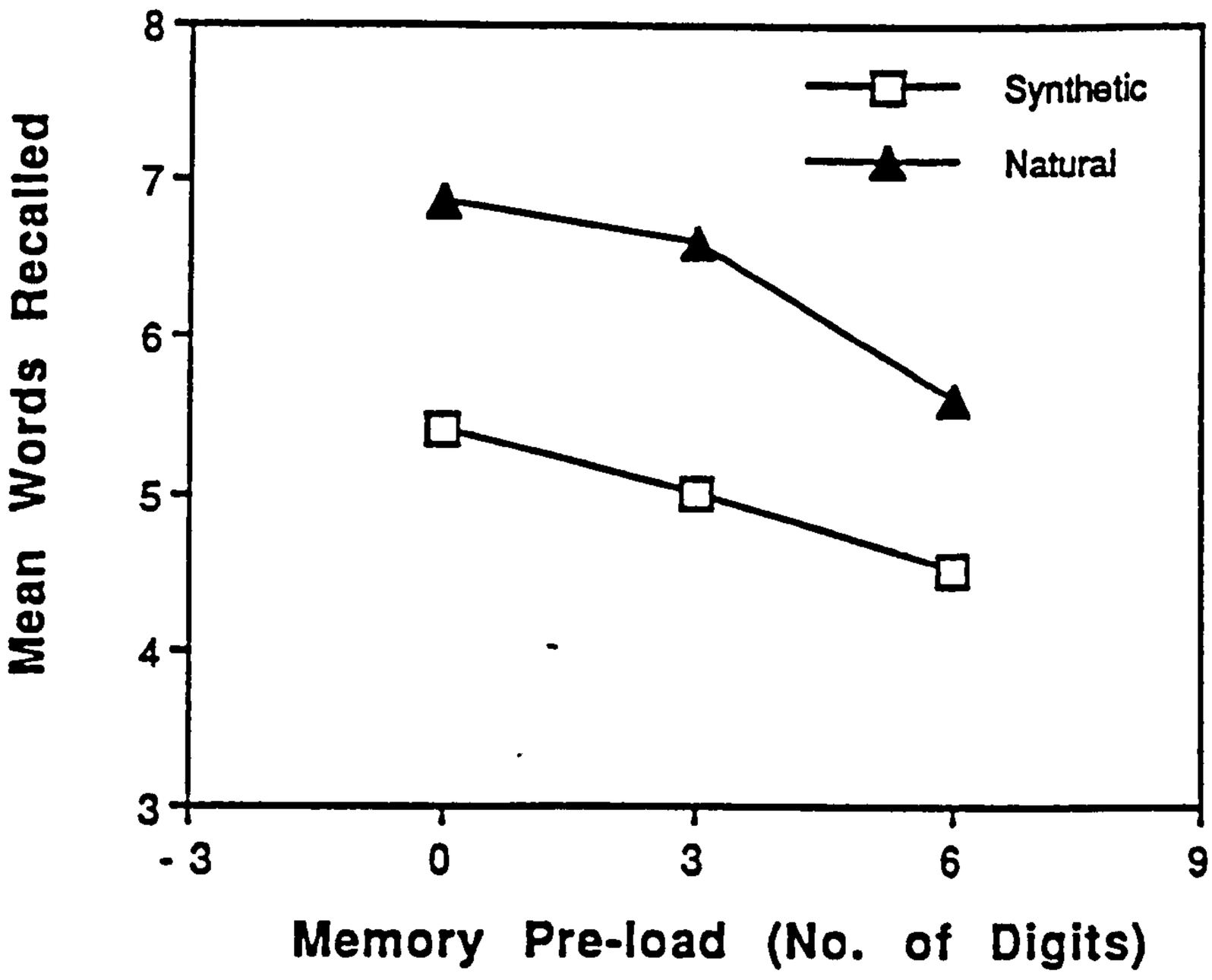


Figure: 2.2 Mean number of natural and synthetic words recalled as a function of memory pre-load. Adapted from Luce, Fuestel and Pisoni (1983)

In a restricted range of settings however, the distinctive features inherent in a voice with a mechanical style can increase efficiency and, in comparison to digitised speech, may be preferred by users. This is especially true when synthesised speech and concatenated digitised speech are compared. For example, in one study pilots evaluated an electronic warning system with messages generated in a variety of ways using synthesised and concatenated digitised speech. They expressed a preference for a distinctive, slightly mechanical sounding synthesised voice, one that was easily discriminable from the jumble of speech usually heard over an aircraft's radio. They reported extreme dissatisfaction with the slow speaking rate of all digitised voices, caused by the artificial pauses that were introduced by the word concatenation method used to generate the messages (Simpson, Marchionda-Frost, and Navarro, 1984). Synthesised speech here offers an advantage in that most synthesisers enable the speaking rate to be dramatically altered without affecting the prosody of the speech. Depending on the context, and possibly on the amount of natural speech present, synthesised speech may be distinctive and can therefore serve an alerting function (Simpson et al 1984). However, the physical correlates of this distinctiveness remain to be determined experimentally.

Prosodic limitations are a function of the state of technology of synthesised speech production and *may* eventually become insignificant. Some contemporary speech synthesisers can be programmed to produce speech which includes approximations of the hesitations, variations and imperfections of natural speech. However admirable this may be as a technological achievement, it may prove problematic for users in certain situations if the programming becomes so sophisticated that the machine speech is indistinguishable from human speech. There is an accumulation of reports from pilots who have served in speech display flight simulation studies that the voice ought not to sound too human in case it is confused with human speech such as radio or intercom

communications (Simpson, 1981).

The limitations of synthesised speech may gradually diminish as contemporary devices produce output which is less and less distinguishable from natural speech. One possible side-effect of this is that it may exacerbate the anthropomorphic projection of 'human' qualities onto machine devices. Users may be tempted to treat the device as another person. They may attempt natural conversation when confronted with high-quality speech displays on a telephone system, or overestimate the capabilities of the system. This could have unfortunate side-effects: users may become less disciplined in their use of commands and assume that computers can accept natural language when they cannot and this may lead to an increase in the likelihood of dialogues breaking down. This was reported by Jones (1989) and any improvements in the realism of synthetic speech over the last ten years may result in this problem occuring more frequently.

2.11 Implementation Guidelines

The scarcity of experimental evidence on the ergonomic aspects of speech output was noted by Van Ness in 1986 and it can be argued that the situation has yet to be significantly improved. This is perhaps due to an over-emphasis on technologically-driven research initiatives at the expense of psychological and ergonomic enquiry. Hence, general prescriptive recommendations for speech system design are currently derived as much from observations and insights during developmental cycles as from direct experimental scrutiny.

The following guidelines act as a conclusion to this chapter and contrast the two different techniques of speech output which have

been discussed. They are intended to facilitate interface design decisions when, as is sometimes the case with speech facilities, the developer has the option to include both synthesised and digitised and may have to determine which technology is more appropriate for particular users, environments and functions. They consist of points which, when taken into consideration during speech system design, are likely to aid designers in the choice of synthesised or digitised speech devices for particular application scenarios. As with the guidelines in chapter 1, they should be regarded as general and flexible recommendations rather than rigid rules.

1. For speech digitisation, a quiet setting and a good reader are required for clarity of constructed messages. In contrast, synthesised messages will be unaffected by auditory interference or the speaking capabilities of an orator.

2. If digitised speech messages will need to be edited or upgraded at various stages during the general use of the system, they will need the original speaker to be present so that consistency may be maintained. In contrast, synthesised messages can be edited at will by anyone conversant with the system and without special settings.

3. If the messages featuring in the system are likely to contain words and phrases with untypical pronunciation, such as would be found in an average database of peoples names and addresses, synthesised speech devices may frequently apply inappropriate phonetic rules and may produce peculiar or unintelligible output. Digitisation is here recommended. Of course, people can and do apply inappropriate pronunciation rules, but human text-to-speech conversion is still far more sophisticated than any computer approximation when dealing with the overwhelming complexities of human language.

4. Compared with digitised speech, storage space of synthesised data is very modest. Therefore, synthesised speech may be advantageous

where computer memory is restricted. This consideration is diminishing in importance as storage technology becomes more and more refined and inexpensive, although the fact that synthetic speech is comparatively free of all memory constraints will mean that storage issues will always require consideration.

5. Synthesised speech is more intelligible and more attention-getting in a noisy environment than digitised or normal human speech. This factor can be utilised if messages are to function as warnings or alarms.

6. The perception of synthesised speech imposes greater demands on processing capacities than digitised speech. This may interfere with other tasks carried out simultaneously. Subsequently, synthesised speech may be more problematic than digitised speech in situations where users are overloaded with information.

7. At sentence level, synthesisers have difficulty in producing speech with natural-sounding prosody. Prosodic features can be programmed in by the user, but only with considerable effort. Therefore, for messages where prosodic factors are essential to convey intended meaning, digitisation may be a more effective strategy.

9. With messages of considerable complexity or which need to be transcribed with perfect accuracy, synthesised speech may offer an advantage because messages can be studied visually, at the source text level, as well as auditorily. They can also be printed out in a semantically intelligible form.

10. Digitised speech systems can utilise a potentially unlimited variety of different voices. Without considerable extra work, synthesisers have a limited number of voices. Therefore, it may be easier with digitised speech to find an acceptable and appropriate voice for a particular user and application. Although that fact that a

user can *modify* a synthetic voice in line with their own preferences may offset this advantage somewhat.

11. The speaker of digitised messages can be identified with ease. Synthesised messages can be made or changed by anyone. If the speaker can be identified the message may signify more authority or convey more confidentiality. Para-linguistic cues (such as voice tone and stress) in digitised messages can also convey the mood of the speaker which may be essential for the highlighting of such factors as urgency, gravity or underlying meaning.

2.12 Conclusions to section one

To conclude, section one of this thesis has addressed and examined the major issues concerning the use of computer speech output in human-computer interaction, both in terms of (briefly) historical context, and in relation to the requirements of contemporary users. The guidelines presented in conclusion to chapters one and two can be applied across a wide range of speech system applications. Whilst the entire range of guidelines will not always be relevant in every situation, they should at least encourage an informed human factors perspective during the development of speech systems.

In answer to any charge that many of the guidelines are simply common sense and state the obvious, one need only consider the continuing adoption of speech output facilities for inappropriate tasks and settings. For example, the supermarket chain 'Safeway' has announced the introduction of talking shopping trolleys. Each trolley will now come equipped with a synthetic speech device which will welcome the customer and accompany him/her around the shop announcing special-offers and making supposedly helpful remarks and suggestions, presumably based on a consumer profile derived from previous purchasing data. The author is hard pressed to think

of a more inappropriate application. The individuals responsible for this 'innovation' (and indeed the customers this will be inflicted on) would surely benefit from some common-sense advice on human-factors issues. The application of guideline 11 (page 34) concerning privacy and the use of speech in public settings might, in this case, have saved the developers time and trouble as well as the sanity of the general public.

The guidelines from chapters 1 and 2 have featured in a number of publications (see page 8), they were also presented in two research films directed by the author which were included in the formal video programmes for CHI'90 and CHI'93. They are also cited in the literature by other authors (eg Stanton and Baber 1997, Baecker at al 1995). Furthermore, some of the basic principles underlying the work date back nearly 30 years (Deatherage, 1972). The material is out there for those who wish to find it, yet speech is consistently employed inappropriately. It seems likely that in the cases that have been discussed, the developers may have been been overwhelmed by their enthusiasm for the technology, concentrating on what is *possible*, rather than what is *desirable*. Some consideration of published design principles for speech systems would almost certainly have helped to avoid the implementation disasters that have frequently occurred..

Even if applied effectively, an ergonomic research strategy can only go so far. In the current study this approach has illustrated and defined design criteria for successful system implementation, but, as stated previously, no matter how advanced the functional analysis of the interaction and how elegant and principled the dialogue design, applications are not likely to be accepted if the users simply find the computer speech unpleasant to listen to.

Section one has taken an exploratory/ergonomic paradigm to the point where further inquiry would be inexpedient and a shift in emphasis is now required. The next stage of the research will

therefore adopt an experimental paradigm in order to investigate synthetic speech perception and the determinants of user acceptance of speech applications, in particular synthetic voice analysis and design.

Section Two:

Synthetic speech design:

An evaluation of speech quality

Foreword to section two:
Aims and objectives

"There is an evident necessity, then, for some means of assessing objectively the quality of synthetic speech." (Barry , 1990)

"The improvement of the quality of synthetic speech is now the major concern" (Helender , 1993)

Whilst there has been a significant amount of work which has sought to improve the intonation of synthetic speech in terms of its intelligibility and thus make it more acceptable to listeners, there is little work that has attempted to improve the aesthetic qualities of the voices a synthesiser can produce. Although a number of studies have reported adverse reactions to synthetic speech (as discussed in chapters 1 and 2), and chapter 3 will demonstrate that substandard speech quality has been consistently cited as a major cause for such reactions, very little research has focussed specifically on this problem.

Ergonomic strategies and guideline development have their place, but the most important part of designing speech-based systems to be used *by* people has to involve the testing of prototypes *on* people. It is only through analysing the reactions of people during exposure to auditory output and applying the results systematically to synthetic speech design that we will reach a stage where we can build systems that truly take advantage of the flexibility of computer speech technology.

Consequently, the aims and objectives of section two of this study will be addressed within an extensive empirical investigation of synthetic speech design. A laboratory-based experimental methodology will be used in order to investigate the relationships

between voice parameter modification and the perceptions and reactions of listeners. This will initially require the development of an efficient and reliable ratings procedure and an associated assessment methodology for the investigation and quantification of synthetic speech perception within a laboratory based environment.

The next stage will involve a comprehensive factor analytic study and a progressive experimental program in order to investigate the relationships between specific synthetic speech characteristics and the perceptions and reactions of listeners. Assuming such relationships can be usefully specified, complimentary experiments will be conducted which will incorporate the application of calibrated modifications of speech parameters to both male and female synthetic voices. The aim of this part of the inquiry will be to determine the extent to which the precise control of perceptual attributes is possible across a balanced range of DECtalk synthetic voices. Finally, a major experiment will be conducted that will investigate the effects of independently verified cognitive load on perceptual evaluations of synthetic speech quality.

The overall aim of the second stage of the research is therefore to determine the precise characteristics of synthetic voices which influence usability and acceptability within the framework of a valid and reliable laboratory-based ratings environment. Throughout stage two, the empirical work will be supported by an extensive review of related literature in order to provide a firm academic foundation for the research. It is hoped that the research conducted here will enable the many speech system designers to implement computer speech with greater confidence concerning the overall usability and acceptability of such systems. This in turn may enable speech output technology to finally demonstrate some of the enormous potential that has been anticipated for so long. Section one of this thesis has investigated the process of improving the appropriate implementation of computer speech, section two will investigate the possibilities for improving the speech itself.

Chapter Three

Experiment 1:

Speech quality assessment:

A factor analytic approach

3.1 Summary

"A number of alternative directions of current research aim at the ultimate goal of fully natural synthetic speech. One especially promising trend is the systematic optimisation of large synthesis systems with respect to the formal criteria of evaluation." Liberman (1995)

Human speech conveys far more than just semantic information. The quality, style and specific paralinguistic characteristics of the speech signal can convey a wealth of information to the listener. This, with or without additional visual cues, is used to build up a complex profile of the speaker which may include age, sex, mood, race, personality and a host of other perceived characteristics. This tendency occurs automatically and in reaction to exposure to any type of identifiable speech signal, either human or machine generated.

Consequently, caution should be applied when implementing speech devices in the human-computer interface. As discussed in the previous chapter, "speech-induced anthropomorphism" (Jones 1989) can result in an overestimation of a system's capabilities. However this tendency can have other consequences, the most common being

that synthetic voices are perceived as having a distinct personality and this perception is usually negative. However, the picture is more complex than this, even if listeners of synthetic voices do not attribute personality characteristics (negative or otherwise), there appears to be something unique about mechanical speech which elicits a negative reaction in listeners. Put crudely, synthetic speech gets on people's nerves. Unless this adverse reaction to machine generated speech can be investigated, understood and (hopefully) alleviated, the promise of widespread implementation will not be fulfilled. The series of experiments reported in this thesis will attempt to address these research issues.

The first experiment is an attempt to determine the relationships between voice qualities and listeners' perceptions and rated reactions. In order to achieve this, twenty four distinct synthetic voices produced by a DECtalk speech synthesiser were rated on thirty perceptual dimensions by forty judges. Emphasis was on objective evaluation of voice quality rather than speaker characteristics, although these two aspects are not always easy to separate (discussed later). Factor analysis of the perceptual ratings recovered two strong factors characterised as "*Listenability*", associated with the voice parameters 'Richness' and 'Smoothness', and *"Assurance"*, associated with 'Average Pitch' and 'Head Size'. Whilst these voice parameters have been shown to have an association with perceptual factors in a previous study (Rosson and Cecela 1986), the factors extracted in this study are qualitatively different. Subsequently, it is important to determine the extent to which extracted factors are constrained by the stimuli and response measures offered.

3.2 Introduction:

"*Over the past ten or fifteen years there was a widespread view that text-to-speech synthesis has been largely solved... ...What those who judged that the task of building synthesisers was all but finished failed to get right was the lay user's reaction to what was available: it turned out that users were not impressed at all by text-to-speech synthesis - they want intelligibility of course but they also want naturalness...*" *(Tatham, 1993)*

An abundance of various types of voice response devices have been developed for integration into human-computer communication systems. Most of these devices, such as the Digital Equipment Corporation's DECtalk and DECvoice, produce speech which is sufficiently intelligible and dynamic to allow implementation into a variety of commercial, industrial, military, and educational applications.

Intelligible synthetic speech has been available for use within the human-computer interface for more than two decades, yet the implementation of speech synthesisers has not exactly been widespread. Furthermore, even in the rare cases where synthesisers *have* been implemented commercially, they have usually failed to achieve long-term acceptance by users. Researchers working in this area have discovered that protracted exposure to even the most sophisticated and intelligible synthetic speech can lead to negative reactions from listeners. These can range from less than favourable impressions of the technology up to outright hostility illustrated by examples of exasperated users physically disconnecting the devices or selecting alternative communication methods.

Tatham (1993) argues that the production of highly intelligible speech appears to have caused a premature shift in emphasis for speech technology research. The production of intelligible speech

seems to have been perceived as indicative of "all the toothpaste having been squeezed out of the tube" and a cue to switch emphasis, in at least some cases, towards the minefield of challenges in speech recognition research, or "squeezing the toothpaste back in", to use Bristow's analogy (1986). However, speech synthesis has not been 'solved' as a research topic, there is much more to do if the potential of the technology is to be realised.

3.3 Voice quality, a major obstacle to acceptable implementation

Speech synthesis has been rejected because of its lack of quality in a number of application domains. Gray (1984) discussed the implementation of speech output in an educational context. He stated the necessity for the speech to be of high quality as the number one criteria to be observed, and, on the basis of this, rejected synthetic speech in favour of recorded human speech.

When describing the surprising delay in the adoption and use of speech synthesis in the telecommunications industry, Roe (1984) stated that the blame for this lay with the need to find a technique to quantify the speech quality obtainable from a synthesis system.

Speech synthesis has not revolutionised human-computer interaction in any especially noticeable way. The main exception to this general implementation failure is the use of synthesis devices as prosthetics for the disabled, either as text-reading devices for those with visual impairment, or as communication prosthetics for speech-impaired individuals. Success in such areas though is highly influenced by the fact that the technology creates opportunities for the user that would not otherwise exist. Here, it should be remembered that such people are rather a captive market for synthesisers and their adoption of the technology and constant usage are often matters of necessity rather than any particular fondness

for synthetic voices. Indeed, a significant proportion of the research that is addressing improving voice design identifies these population subsets as, at least initially, the main source of demand and the immediate beneficiaries of any improvements to voice design (Murray et al 1991, Wilson, 1996). This *should* encourage high standards in voice design. Clearly the demands of users who are *identified* and *personified* in terms of the characteristics of their synthetic voice are likely to be much higher than those of the casual listener.

This is indeed the case, for speech impaired individuals, intelligibility is by no means all that is required, and while such people are likely to report any improvement to their situation favourably, using computer generated speech that sounds mechanical may still be embarrassing or awkward, and for the listener poor quality speech may encourage inappropriate or prejudicial attitudes (Hunnicutt, 1995).

For people dependent on synthesis in order to talk, poor quality speech remains a major problem. Edwards (1991) discussed the use of synthesis by speech-impaired individuals and stated that even the best current synthesizers were "*of such low quality that they inevitably detract from a person's persona*". He continued by stating it was no surprise that some users choose to reject speech synthesizers *on the basis of their poor quality* in favour of another communication method. Edwards stated that many gross qualitative features had still not been successfully addressed and that it was unfortunate that developers did not seem to realize how important they are to this particular group of users.

Unfortunately, research intended to directly benefit such users has not been considered especially high-priority, arguably due to the limited market and lack of commercial opportunities for voice prosthetics, compared with, for example, business/office or telecommunication applications. Not that these application domains

have had much success with synthesis either, as we have already seen.

There remains some room for refinement of the technical aspects of rule-based text-to-speech systems. The DECtalk synthesiser has now been commercially available for many years, and its output is still considered by most to be the best example of state-of-the-art synthetic speech. Although there are other synthesisers available, the DECtalk model remains standard in many systems that utilize synthetic speech, for example Toby Churchill Ltd, a leading European manufacturer of communication aids for the speech impaired currently use DECtalk technology in their state of the art speech prosthetics.

The fact that DECtalk is still considered to be the definitive synthesiser seems a little curious on consideration of speech technology in the context of computer technology in general where design standards are constantly improved and superseded. When compared with the dramatic improvements in voice recognition technology over the last decade, synthesis technology has barely moved.

Some potentially fruitful research avenues for text-to-speech synthesis have now been identified. These include the development of sophisticated strategies for syntactic and, especially, semantic parsing. These may enhance variability in the articulation of the speech as well as improve subtle pragmatic effects which convey to the listener such things as mood and attitude (Tatham 1993). Such research may well eventually result in more human-style voices which speak more naturally and reduce the incidence of misapplied prosodic cues. However it is important to stress that technological sophistication in text-to-speech algorithms alone will not suddenly remove the predictable scenario of listener irritation and subsequent rejection that has been apparent through thirty years of indifference to synthetic speech in computer systems. As of the time

of writing, technological refinements have failed to make much of an impact.

To conclude, the literature that has been discussed clearly indicates that poor quality synthetic speech has been frequently responsible for the rejection of such systems in the classroom, in telecommunications, and in applications for the disabled. It seems probably that other instances of adverse reactions reported in previous chapters may have also been the result of poor quality. Furthermore, the problem remains today. Bernson et al (1998) state that the drawback to successful speech synthetic speech systems is that "*parametric (synthetic) speech quality is still low for many languages*" and they include the evaluation of user satisfaction through questionnaire/multiple scaling amongst a list of research issues which remain, to date *"unsolved"*.

Therefore, what is needed if research is going to attempt to break the deadlock is a consistent analysis of what makes synthetic speech unlistenable and grating, what makes it sound unconfident or unreliable. If the specific speech characteristics responsible for consistent perceptions of negativity could be uncovered, it should be possible to develop speech with an improved chance of user acceptance and successful implementation in a wide range of application domains. The need for an effective technique to quantify and improve speech quality has been identified repeatedly in the literature. This is the aim of the current research.

3.4 Voice quality: Towards an empirical solution

Human voice quality and style are major factors which allow us not only to differentiate between speakers who would be considered to have the same accent, but also to distinguish between a single speaker's attitudes towards us and their various moods and idiosyncrasies. The paralinguistic function of voice quality may also impart to the listener specific cues concerning the content of the message. For example; a whispery voice tends to impart an impression of confidentiality whereas a sustained creaky voice tends to signal bored resignation (Gobl and Chasaide 1992).

While there have been a number of explorations of human voice quality which have investigated between-subject differences (for example, Fritzell et al 1986), there has been virtually no work on the types of quality variations available to the single speaker. Furthermore, research into listeners' perceptions of the attitudes associated with particular voice qualities is even more scarce.

When considering listeners perceptions of voice qualities:

"..... relatively little is known about this: we have little to go by other than impressionistic observations." (Gobl and Chaisaide, 1992).

Research has shown consistently that the attributions and predictions that listeners make about human speakers based on voice characteristics are similarly made about synthetic speech systems. Personality is sometimes ascribed on the basis of the device's speech style and quality, even when a voice is clearly perceived as machine generated (see section 1.7). Indeed, during the experimental program undertaken in the current study, the experimenter noticed that some subjects tended to refer to the synthesiser as "He/Him" or "She/Her" and also tended to make comments concerning their perceptions of

the machine's "personality" or "mood", even though the DECtalk is a beige plastic box with no distinguishing features or obvious 'character' of any kind.

This tendency appears to happen automatically (although the scales selected for use within a ratings experiment may influence this) and can be problematic if an analysis of quality independent of personality attributions is sought. A small number of studies have attempted to investigate the underlying characteristics of speech quality and style which influence listener's perceptions, some have been more successful than others. The standard methodology has been to adopt a factorial analysis and scaling technique in an attempt to uncover the presence of the underlying factors in speech signals which influence users perceptions and judgments. In many cases, researchers have adopted perceptual scales derived from personality research to assess human and machine-generated voices in terms of appropriateness for tasks and acceptability for users. Rarely have objective measures of speech quality been attempted.

Cox and Cooper (1981) used a paired comparison technique to contrast the appropriateness of seven different recorded human voices for a telephone announcement task. The study attempted to assess the relevance of speech style as an indicator of various speaker attributes. A scaling methodology was used to determine the "features of the speech that were considered by the subjects important in selecting a preferred speaker". Their predictive measures were factors derived from a set of personality ratings originally employed in a study of speech style and social evaluation (Giles and Powesland, 1975). These were used to assess judges ratings of the speaker's personality and social characteristics (i.e. hard working/lazy, honest/dishonest, generous/ungenerous, wealthy/poor) rather than specific qualities of their voices. Thus, although they found that two factors, 'agreeableness' and 'assertiveness', appeared to be related to judged appropriateness of

voices for the task, it is not clear what perceptual qualities or features of the speech were responsible for these personality attributions. It could be argued that there is no obvious relationship between speech quality and wealth/honesty/diligence and that quantification of these concepts within speech could never be achieved with any degree of reliability.

Furthermore, the paired comparison methodology used in the study is limited in that such a technique tells us little more than how one voice compares with another on a set of personality scales. This may be useful if a practical decision needs to be made regarding choice amongst a limited voice set for a particular application, but explains little about the qualities of the voices which result in such a rating except in terms of their relationship to each other. A more objective approach is required if we wish understand the perception of speech quality beyond preferences from a limited voice set.

To sum up, Cox and Cooper highlight criteria used in the identification and comparison of voices but provides little information about the specific qualities of speech which listeners were perceiving and reacting to. This is because such studies tend to overlook the distinction between the personality listeners ascribe to a speaker and their perception of the voice's more objective features.

Of more relevance to the present research is work directed at uncovering such objective features and their perceptual implications. A more exacting and replicable example of the application of factor analysis in speech perception scaling was conducted in 1986 by Rosson and Cecela. This research was a direct attempt to assess the perception of specific objective qualities of voices and to understand the implications for the judgment of appropriateness for different application scenarios.

Rosson and Cecela make a simplifying distinction between voice *quality* and voice *style*. Voice quality is conceptualised as its

inherent timbre, its fundamental frequency and associated harmonic structure. Voice style is conceptualised as a function of the manner in which words are produced, their intonational and durational characteristics which are determined by intra-speaker stylistic variation (physiologically determined individual differences) and are influenced by learned socio-culturally determined articulatory habits. In short, voice *style* is concerned with the articulation of the speech, whereas voice *quality* is concerned with acoustic properties.

Whilst both sets of characteristics will influence listener's reactions to a voice, an attempt to assess quality independently of style is likely to lead to more objective and practical findings. This is because it is possible to make a quantitative analysis of voice quality, whilst assessment of voice style is dependent more on qualitative analysis. The use of a synthesiser allows a precise assessment because the characteristics associated with speech quality can be manipulated precisely and dramatically with minimal influence on speaking style.

Their study assessed eight listener's perceptual and appropriateness evaluations of sixteen variations on one synthetic voice modified across four voice quality parameters, Richness, Head size, Average Pitch and Smoothness (see procedure for descriptions of these parameters). Factor analytic and regression techniques were used to map the relationship between voice quality, perceptual evaluations and appropriateness measures.

As with Cox and Cooper (1981), Rosson and Cecela used varimax orthogonal rotation of the data matrices to isolate the principal component factors (see section 3.5 for discussion of this technique). The analysis of perceptual ratings yielded a two factor solution accounting for 76.7% of the total variance. Factor one was characterised as a "bigness" or "fullness" dimension with the strongest contributing scales being: big/small, low/high pitch, rich/thin and heavy/light. Factor two was characterised as "Clarity"

with the strongest contributors being: clear/muffled, polished/sloppy and smooth/rough. It should be noted that the scales that made strong contributions prompted ratings of objective speech characteristics rather than ratings of the personality of the speaker.

Regression analyses using the voice quality parameters as predictors indicated that three effects had unique contributions for the "fullness" factor. They found that the strongest contribution was from the richness variable with voices high in richness having higher scores on "fullness". Head size and pitch were also associated, and effects were especially strong when high head size was associated with a low pitch. A similar analysis on the "clarity" factor indicated the largest effect was associated with smoothness with voices high in smoothness having high "clarity" scores.

Rosson and Cecela also rated the voices for appropriateness for a wide range of voice-output scenarios. Judgments yielded a three factor solution accounting for 73.3% of the total variance. Factor one was characterised as "information provision" with scenarios such as 'phone-machine', 'tutor' and 'catalogue' being the strongest contributors. Factor two and three, characterised as "entertainment" and "feedback" respectively, accounted for a relatively small proportion of the variance. The analysis assessed the degree to which appropriateness factor scores were predictable by scores on the two perceptual factors. One strong relationship was discovered, the "information provision" factor appeared to be associated with both "fullness" and "clarity" ratings with best voices having high scores on both factors. The other two appropriateness factors were "entertainment" and "feedback" these were marginally related to "fullness". Regression analyses using the voice quality variables as predictors on the "information provision" factor revealed a large effect of richness and a marginal interaction between richness and pitch. Voices high in richness being judged more appropriate for this factor, especially when combined with a low pitch.

Factor analytic studies such as these are inherently constrained by the stimuli and response measures offered to subjects. Rosson and Cecela stated that there may well have been important voice quality manipulations and/or perceptual variables that they failed to examine. The current study seeks to address such limitations by the use of a much wider selection of voices, scales and judges so that a much more comprehensive evaluation of the issues of speech perception scaling is achieved.

In expanding the voice set, its seemed appropriate to capitalise on the strong effects reported by Rosson and Cecala. So the same parameters were used (richness, smoothness, average pitch, and head size), but this time modifications were applied beyond the single 'standard male' voice, and to the full range of voices that can be produced by the DECtalk synthesiser. The stimuli were expanded from Rosson and Cecala's use of one simple sentence for all presentations, to a large set of phonetically diverse sentences designed to demonstrate a much more comprehensive selection of DECtalk's capabilities. Also, the validity of the ratings obtained was enhanced by the use of a comparatively much greater number of judges.

Furthermore, a larger selection of perceptual scales was used, both unique to this study, and documented in previous research. These were derived from a variety of sources (see procedure). The selection and use of the perceptual scales requires special consideration. As has been noted, many previous studies in this area have failed to adequately differentiate between objective voice quality and attributions of personality characteristics, mood and various other predictions about the speaker. Specifically quantifying exactly which characteristics of the speech are being rated by all subjects is not an easy task to accomplish. Furthermore, scaling adjectives can be ambiguous and may mean different things to different people. Rosson and Cecela's study is perhaps the only one discussed where scales which contribute to the main factor suggest listeners were

rating objective quality with reasonable precision.

As we have seen, listeners can attribute many different features and characteristics to synthetic speech. Consequently, it was decided that for the initial experiment, subjects would be provided with a wide range of scales, covering both objective voice quality and various speaker characteristics. This was done deliberately in order to see whether listeners were capable of differentiating between the assessment of perceived characteristics of the speaker and the rating of the objective characteristics of the voices. An examination of the contributing scales for any factors that emerged should help to clarify this question.

To sum up, the methodology used in early studies such as Cox and Cooper (1981) is only really appropriate for testing subjective preferences for one voice over another and fails to provides any information about the specific qualities of speech which listeners perceive and react to. Early studies such as this tended to overlook the distinction between the personality listeners ascribe to a speaker and their perception of the voice's more objective features. Rosson and Cecela (1986) addressed this problem with a useful quality/style distinction and made some initial progress into speech quality assessment. Their experiment produced strong effects and established factorial analysis as an appropriate (although controversial) technique for speech quality assessment data. However they themselves state their research was limited in scope and was only a partial mapping of the area. The current study seeks to build on previous research and significantly expand the state of knowledge in this research domain. In order to achieve this an experimental program is necessary starting with a closer look at the theory and method of factorial analysis, common practice in this line of research and an effective and widely used, experimental technique.

3.5 Factor analysis: Theory and technique

Factor analysis refers to a variety of statistical techniques whose common objective is to represent a set of variables in terms of a smaller number of hypothetical variables. The technique of factor analysis assumes that there is a system of underlying factors and a system of observed variables. The observed variables are linear combinations of underlying source variables or factors. The underlying factors are responsible for the covariation among the observed variables. The logical (mathematical) properties of the correspondence are such that one causal system of factors always leads to a unique correlation system of observed variables but not vice versa. Only under very limited conditions can one unequivocally determine the underlying causal structure among the factors from the correlations among the observed variables. The most that can be achieved is the conclusion that the structure of any observed data is either consistent or inconsistent with a particular factor model based on such a postulate. Having said this, the validity and reliability of a particular factor model can be determined by subsequent experiments which either confirm or refute the conclusions from the initial analysis. This strategy was adopted for this study with further experiments being undertaken to assess the validity of the factor model that emerged. These are described in chapters four and five.

There are two basic types of procedure: 'confirmatory factor analysis' and 'exploratory factor analysis'. The first technique posits that there are some underlying factors for a set of variables and then seeks to test specific hypotheses. As this experiment is of an exploratory rather than confirmatory nature, the second technique was used. This technique attempts to reduce a set of variables into a number of underlying factors.

The technique involves firstly, preparing a covariance data matrix from which a number of initial factors are extracted. The number of these is determined by the Kaiser or Eigenvalue criterion (Kim and Meuller, 1978) which is a 'rule-of-thumb' for determining the number of initial factors. Eigenvalue of greater or equal to one indicates the factors to be extracted. The underlying factors are orthogonal, independent of each other, they do not interact with each other. The first factor accounts for as much variance as possible, the second factor accounts for as much of the residual variance left unexplained by the first factor, etc.

In the rotation stage, simpler and more readily interpretable results are obtained. No method of rotation improves the degree of fit between the data and the factor structure. Any rotated factor solution explains exactly as much covariation in the data as the initial solution. What is attempted through rotation is a possible simplification not a twisting of the data to get a different result. The Varimax orthogonal rotation method is commonly used in analyses of this type. The factor loadings obtained in the analysis are the correlations between the variables and the hypothetical factors and show the extent of any relationships that are uncovered.

Having briefly examined the theory and methodology behind this type of analysis, the technique is accepted as an appropriate tool for the purposes of the current study and the first experiment can proceed (a critique of this form of analysis is included later in section 3.10).

3.6 Method

3.6.1 Subjects

Forty subjects whose first language was English were used in the experiment, twenty males and twenty females with ages ranging from seventeen to forty years. Each subject was paid six pounds for their participation. All were familiar with computers but none had ever regularly used speech synthesis systems or had any prior knowledge of the aims of the experiment.

3.6.2 Design

The experiment was a between subjects design conducted in order to collect a large amount of data suitable for multi variate factorial analysis. As such, all forty subjects followed the same procedure rating a wide range of synthetic voice modifications on quality assessment scales. The predictor variables were the various modifications made to the range of DECtalk voices throughout the presentation of the stimuli sentences during the experiment (see section 3.6.4).

3.6.3 Equipment: The DECtalk synthesiser

A DECtalk formant speech synthesiser was used for the experiment. This device was originally based on the MITalk-79 system but new letter-to-phoneme rules developed by Hunnicutt (1980) were added to produce a new system called Klattalk (Klatt, 1982). In 1982, Klattalk was licensed to DEC for commercial use (Logan et al 1989).

This device was made commercially available in 1983, it converts ASCII code into high-quality synthetic speech and is often reported to be the most superior synthesiser available in terms of intelligibility (Greene et al, 1986) and naturalness (Nusbaum et al, 1984). DECtalk has managed to retain its reputation for a number of years although some recent research has demonstrated that the formant synthesis technique is possibly no longer clearly ahead of other techniques (see Klaus et al, 1997, discussed in chapter 8, section 8. 3). The device was controlled by a microvax and a standard VT340 keyboard and monitor was used for the presentation of the scales and the collection of data from the subjects.

The DECtalk speech synthesiser produces 7 default voices. One voice, the 'light female' voice, was not used because it has virtually identical parameters to the child's voice and is perceptually indistinguishable. The six most distinct were chosen:-

1. Standard Male
2. Standard Female
3. Deep Male
4. Deep Female
5. Older Male
6. Child (non-specific gender)

3.6.4 Voice parameter modification

In order to produce a varied and comprehensive voice set, the six default voices were modified using four voice-design parameters. Each default voice was used to create four members of the set resulting in twenty four different and distinctive synthetic voices. The four parameters were each modified as far as was possible without exceeding their ranges (see Appendix 2D).

Smoothness:

Smoothness is caused by a decrease in voicing energy at higher frequencies and is the opposite of brilliance which comes from an increase in voicing energy. Professional singing voices that are trained to be able to sing above an orchestra are usually high in brilliance. The smoothness parameter is appropriate because we would intuitively expect modification to produce changes in voice quality but to have marginal influence on the perceived identity of the speaker. For each default voice, smoothness was increased by 50% producing six modified voices for the stimuli set.

Richness:

Voice richness or forte is associated with the appearance of a low amplitude nasal formant. Rich voices carry well and are more intelligible in noisy environments, while smooth voices can sound more 'agreeable' to the ear. Richness quality was modified by either + or - 50%. For example, if the default pitch was too close to the parameter's limits to allow a 50Hz increase, a 50Hz reduction was chosen. As the criterion for adaptation of the speech was to create a wide range of voices and determine the influence of the parameters in general, rather than in a particular direction, a unidirectional modification across the voice set was not essential.

Average pitch:

The fundamental frequency (indicated in Hertz) of voiced speech, which determines the perceived pitch, is widely used in all languages to convey information that supplements the sequence of phonemes. In the English language, pitch changes provide additional information about a sentence, such as whether it is a question, statement or a command. For example, to indicate a question, a speaker may raise the pitch at the end of the sentence. Pitch changes can also convey the mood of the speaker and have been suggested for indicating the urgency of a message with raises in pitch corresponding with greater perceived urgency (Simpson et al, 1984).

Every voice has a different average pitch and pitch range (which is expressed as a percentage change relative to the current average). In the experiment, average pitch was either increased or decreased by 50Hz depending on the default setting. Pitch range was not modified as this can result in markedly distinct changes in the character or style of the speech, a high pitch range resulting in a 'sing-song' style whilst a low range produces a 'Dalek' monotone.

Head size

The head size variable is literally, a computerised simulation of the type of acoustic changes that would be apparent if the vocal tract cavities were somehow enlarged or shrunk. The head size variable is realised acoustically through changes in formant positioning and amplitude. Human head size has a strong influence on a person's normal speaking voice. Larger musical instruments tend to produce lower notes, and humans with larger heads tend to have lower, more resonant voices. Decreasing head size produces a higher voice, such

as in a child or adolescent. Voices with enhanced resonance were created by increasing head size by 15% for the six default voices.

3.6.5 The control program

A ratings program was written and compiled in the Pascal programming language. The program allowed a brief example of the ratings procedure for the experiment and then, using a randomisation algorithm, scrambled the order of presentation for the voices, the scales and the sentences. The program then collected the full 720 ratings from each subject (24 voices rated on 30 scales each). Six breaks were included where subjects could pause and rest for a few minutes if they desired. Various strategies were used to ensure that the ratings obtained were accurate and reliable. All keys on the keyboard were disabled apart from the ones of use in the experiment to avoid input of inappropriate data. If the subjects were to accidentally hit the wrong key, a 'bleep' would sound and the program would prompt for appropriate input. On completion of the task, the program unscrambled the ratings and created an intelligible data matrix for each subject which was suitable for statistical analysis.

Appendices one and two contain a detailed discussion of the control program, including an overview of development and evolution, a description of the various human-factors problems encountered during preliminary testing, and a summary of the architecture and procedures. The code itself used for this experiment is listed in Appendix 2E.

3.6.6 Stimuli

Rosson and Cecela (1986) provided subjects with only one short sentence ("some new information has just become available") as a stimulus for all perceptual and appropriateness ratings. Linguists have identified 17 vowel phonemes and 24 consonant phonemes in English, all of which can be simulated by the DECtalk synthesiser. It seems likely that the use of such a simple, limited and repetitive stimulus may perhaps have influenced the ratings obtained. In the current study, 30 sentences were constructed using, as far as was practical, all of these phonemes in equal frequencies (Appendix 2B). Many of the sentences were transcribed phonetically to remove any pronunciation errors and to give the best possible output. As far as possible, the American inflections that the DECtalk produces for certain utterances were minimised by phonetic transcription. Marginal modifications in speech rate were made for some of the words in the sentences to compensate for the fact that the DECtalk tends to pronounce certain letter combinations faster than others. This is due to the logistics of the text to speech algorithms and can affect the naturalness of the delivery. Listeners may make perceptual judgments about voices both from the qualities of the speech and the content. In general, the annoyance of sound is very much dependent on the information that the sound brings with it (Galer, 1974). In order to minimise the possibility of ratings being influenced by the semantic content of the messages, the sentences were designed to be as unemotive and dispassionate as possible.

3.6.7 Scales

The scales were constructed with the intention of obtaining a wide-ranging and comprehensive selection of perceptual ratings. Thirty bipolar, Likert-style 5 point rating scales were constructed (Appendix 2C). These were chosen by drawing on appropriate research (eg. Nusbaum et al 1984 and mood adjective checklists). The intention was to provide subjects with as wide a range of scales as possible so that those responsible for rating any specific perceptual qualities of the speech might emerge in sufficient number to allow the development of an effective ratings tool. The scales included both objective voice quality adjectives as well as some that refer to speaker characteristics. If voice quality ratings were seen to be quantifying perceptual factors independently of ratings of speaker personality, the resultant adjective set could be extremely useful for further experimental work.

3.6.8 Procedure

The experiment took place in a sound-proof laboratory at the University of Wales Human Factors Research unit within the School of Psychology. The subjects, who participated in the experiments individually, were seated in front of a terminal and the DECtalk and given instructions about the use of the ratings procedure. They were then given a brief example of the scaling procedure (four trials) in order to familiarise them with the scale presentation style, the use of the keyboard, and to familiarise them with the experience of hearing synthetic speech for the first time. The subjects were told to attend to, and make judgments about, the quality of the speech only, and to try to disregard the semantic content of the sentences (this was not considered to be an especially hard request due to the bland and repetitive nature of the stimuli). DECtalk occasionally sounds

rather American, so all subjects were instructed to interpret the 'British - Foreign' scale as being 'English as first language' vs 'English not first language'.

The subjects were asked to imagine how they might feel if exposed to the various voices on a daily, routine basis and to make their judgments accordingly. For each trial, subjects were first presented with a 5 point ratings scale on the monitor. Two seconds later the synthesiser output one of the sentences at a clear but comfortable volume. Headphones were not necessary because the laboratory was soundproofed and is was important to allow as much comfort as possible during the length of the experiment (approximately 2 hours). The delay between scale and speech presentation was to allow time for them to read and understand the particular rating required before hearing the speech. Then the screen prompted for a rating. The subject would then enter their choice by pressing one of the keys, 1 to 5. Their choice was echoed visually via a highlight on the scale calibration in reverse video. Finally, the subject was offered the choice to either change and re-submit their rating or confirm their choice by pressing return which would initiate the next presentation. When all the speech samples had been presented and rated, the subjects were thanked, debriefed and paid a small sum for their participation. No subjects reported difficulty or confusion with the rating procedure but a number of them reported boredom with the repetitive nature of the experiment.

On completion of the experiment, the scores were compiled electronically and subjected to analysis. This process included re-scoring of the data so that positive adjectives were always scored high and negative always low.

3.7 Results

The data matrices were subjected to factorial analysis using the statistical package SPSS. Factors were retained if their Eigenvalue was at least one (so that a factor accounted for at least as much of the variance as a single variable). Varimax orthogonal rotation was used to render the principal component factors more interpretable. The analysis yielded a 3 factor solution accounting for 62.5% of the total variance (see figure 3.1 below for a simple representation of the variance). The correlation cut-off point was set at +/- 0.5 (see appendix 3 for the entire set of extracted factors and their associated scales). This type of analysis was considered appropriate due to its consistent and effective use in analysis of comparable data in previous speech perception scaling experiments (Cox and Cooper 1981, Rosson and Cecela 1986).

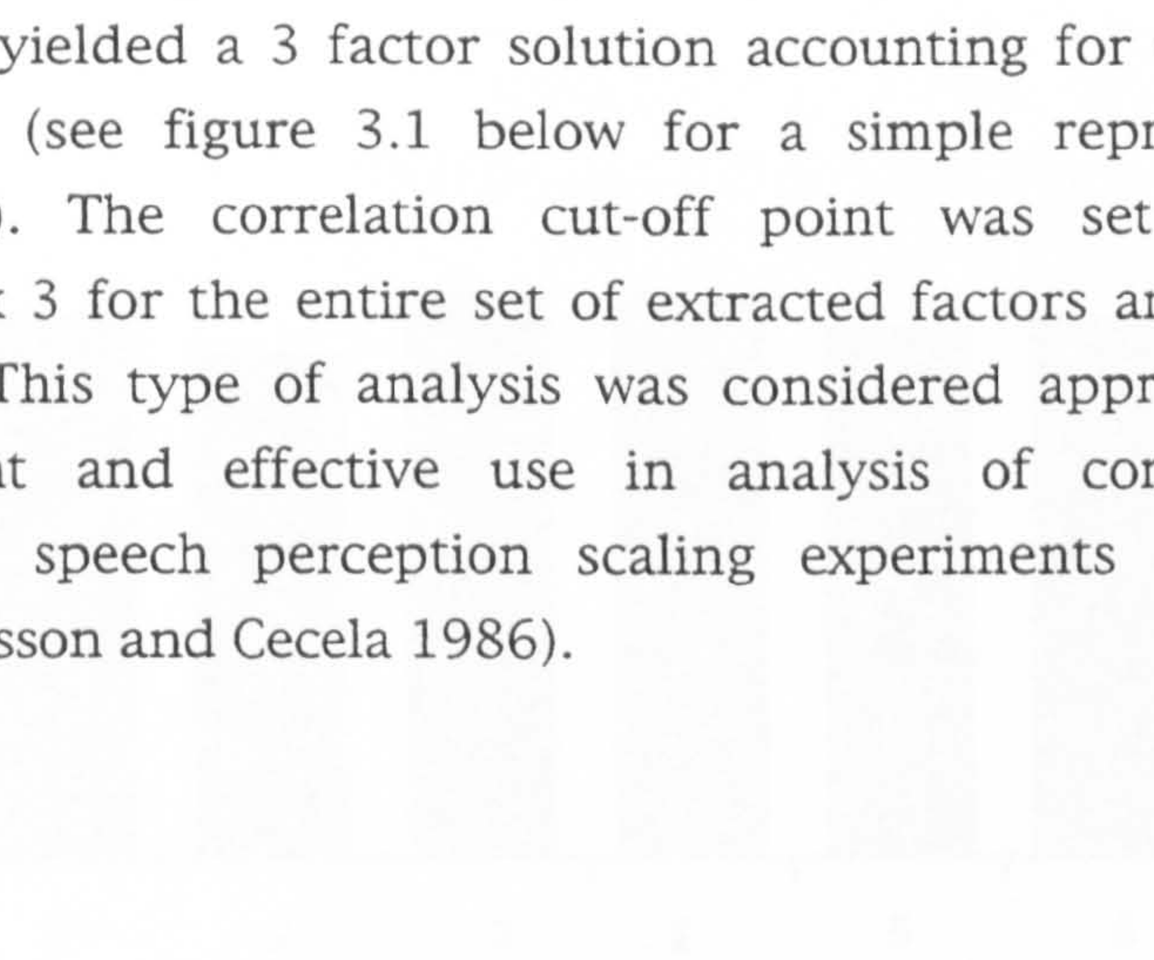
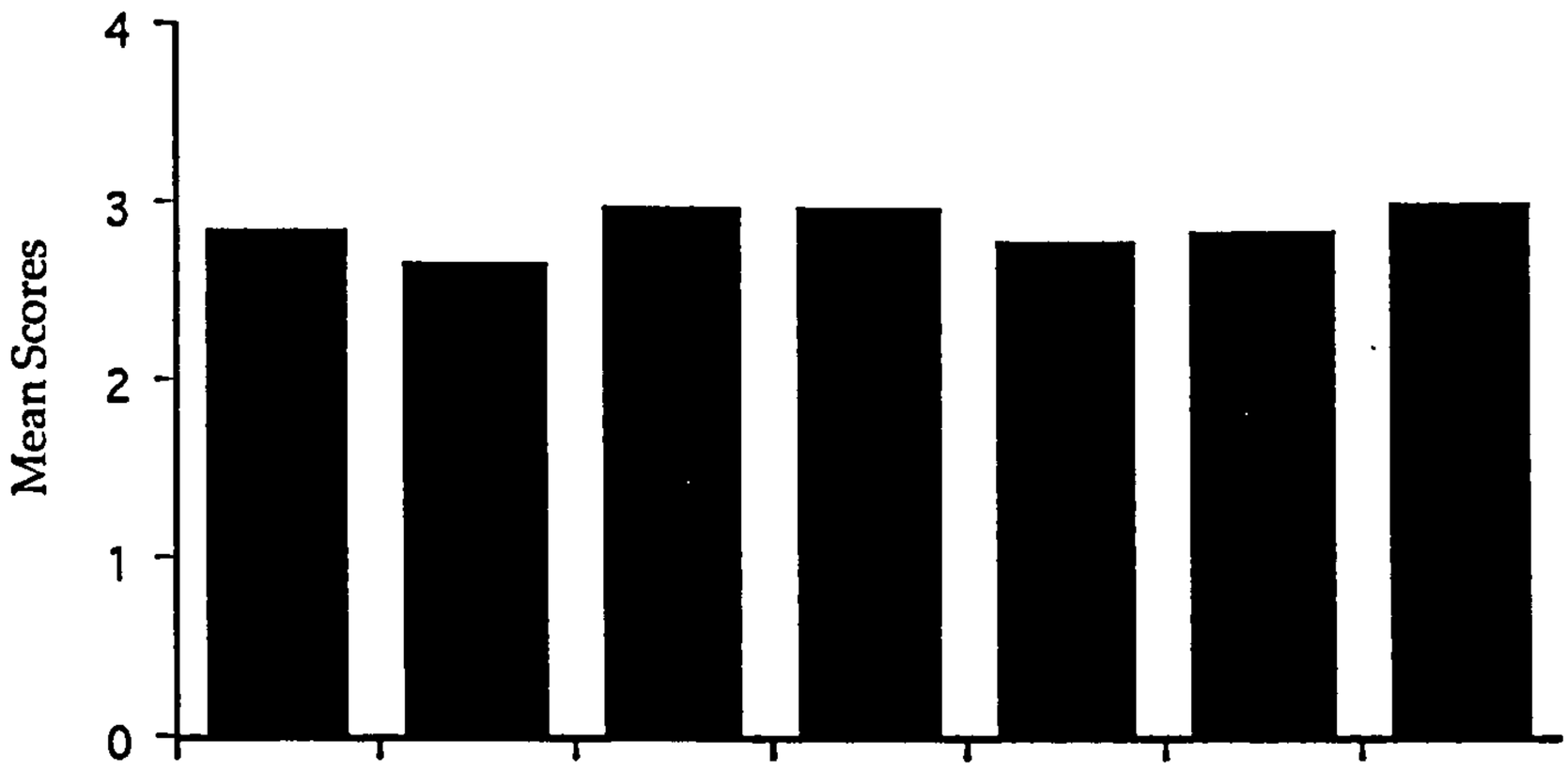


Figure 3.1 Simple pie chart showing variance accounted for by principle component factors.

3.7.1 Factor 1: "Listenability"

This accounted for 44.7% of the variance. Examination of the factor loadings indicated that seven of the scales were strong contributors to this factor.



Scales associated with listenability (listed in order of highest to lowest correlation with factor):

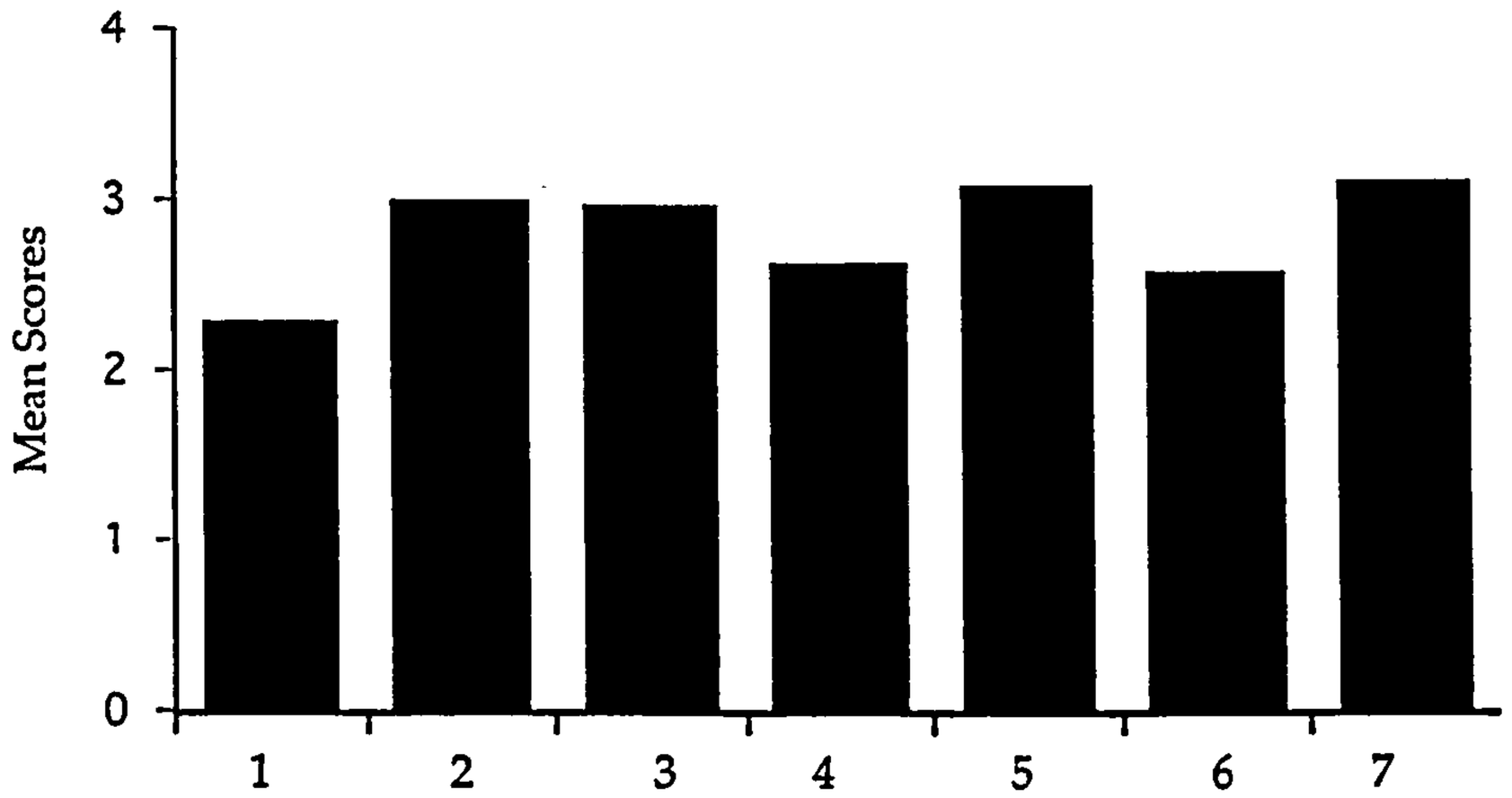
Scales	Correlation	Mean	S/D
1. Dissatisfied -- Satisfied	(.79)	2.851	0.43
2. Irritating -- Not Irritating	(.78)	2.659	0.54
3. Harsh -- Gentle	(.69)	2.978	0.39
4. Hostile -- Friendly	(.67)	2.976	0.38
5. Unpleasant -- Pleasant	(.65)	2.778	0.44
6. Disturbing -- Restful	(.65)	2.840	0.41
7. Crude -- Refined	(.50)	2.999	0.46

Regression analysis is required when we have a set of predictor variables (in this case the variations in the synthetic voice parameters), which correlate with a criterion variable (in this case listenability factor scores). This is done using a regression line, a line of best fit calculated for the data matrix.

For the purposes of this analysis, the extent that voice parameter predictor variables predict values of listenability was calculated using SPSS. The analysis revealed that the Richness parameter correlated with factor one to .57 ($p<0.005$), accounting for 33% of the variance. Smoothness correlated .53 ($p<0.005$), accounting for 27%. Together, these parameters accounted for 41% of the variance in factor one.

3.7.2 Factor 2: Assurance

This factor accounted for 12.6% of variance. Factor loading examination revealed 7 scales correlated with this factor:



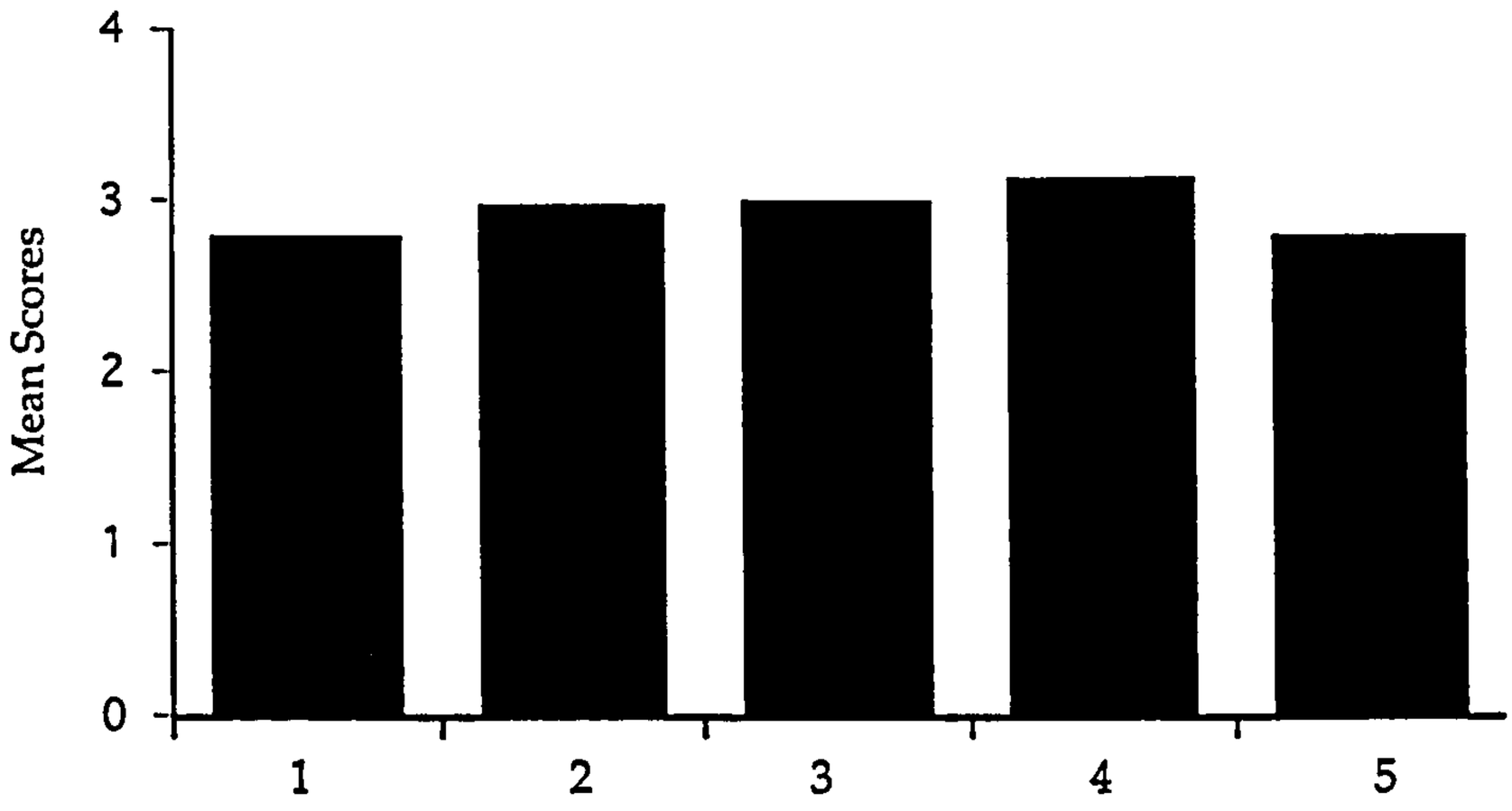
Scales associated with assurance (listed in order of highest to lowest correlation with factor):

Scales	Correlation	Mean	S/D
1. Calm -- Anxious	(.77)	2.920	0.42
2. Relaxed -- Tense	(.74)	2.996	0.43
3. Authoritarian -- Meek	(.69)	2.976	0.33
4. Clear -- Confusing	(.67)	2.627	0.45
5. British -- Foreign	(.62)	3.075	0.64
6. Composed -- Confused	(.55)	2.590	0.41
7. Knowledgeable -- Uneducated	(.52)	3.107	0.41

Regression analysis revealed the Average Pitch parameter correlated with the factor to .42 ($p<0.005$), accounting for 18% of the variance in factor two. Head size correlated .42 ($p<0.005$), accounting for 18%. The analysis showed that these parameters correlated so highly that they cannot really be said to make independent contributions to the variance.

3.7.3 Factor 3: 'Amiability'

This factor accounted for 5.2% of the variance. Factor loading examination revealed 5 scales correlated with this factor:



Scales associated with amiability (listed in order of highest to lowest correlation with factor):

Scales	Correlation	Mean	S/D
1. Fluent -- Halting	(.73)	2.777	0.55
2. Friendly -- Unfriendly	(.55)	2.966	0.36
3. Smooth -- Rough	(.55)	2.996	0.38
4. Comfortable - Frustrating	(.55)	3.129	0.41
5. Pleasant -- Unpleasant	(.52)	2.778	0.44

The Richness parameter correlated with factor 3 to .42 (p<0.005). Smoothness also correlated to .42 (p<0.005). Each of these parameters accounted for 18% of the variance in factor 3. The analysis showed that they correlated so highly that they cannot really be said to make independent contributions to the variance.

3.8 Discussion

When the scales associated with factor one are considered, they can be seen to apply to a reasonably distinct perceptual attribute. All but one seem to refer to the aesthetic listenability of the voices. Only one scale is potentially ambiguous, and that is Dissatisfied/Satisfied. Judging by the other scales associated with the factor it is likely that subjects were possibly interpreting this scale as a measurement of *their* satisfaction with the voices rather than how satisfied or dissatisfied the voices sounded. Subjective interpretation of this scale could be investigated as a side issue, but for the purpose of this analysis, it is enough to note that the scale is a very strong predictor of the criterion variable and there is certainly no ambiguity apparent within the data to suggest caution is required.

The richness and smoothness parameters made independent contributions towards explaining the variance in factor one. As found by Rosson and Cecela (1986), the strongest overall contribution to the variance for factor one came from the richness variable. The characteristics of the 'listenability' factor and associated scales were, however, qualitatively different from the reported "fullness" factor. Furthermore, head size, which contributed to 'fullness' made no significant contribution to 'listenability'. Reasons for the differences between the isolated factors and other results may reflect differences between the scale set, the voice set and the range of stimuli. Firstly, the richness parameter certainly had a strong effect on factor loadings in both studies, but the differences between extracted factor characteristics may, to some extent, be a consequence of the perceptual scale sets offered to subjects. Secondly, the voice set used in Rosson and Cecela's study consisted of only one default voice, varied with subtle and graduating changes of the modification parameters. This may have led to many of the stimuli being perceptually indistinct. The current study used six markedly distinct

default voices which were modified fairly dramatically. Consequently, the much wider and more varied selection of voices is likely to have resulted in more comprehensive ratings from subjects, the data reflecting a more general and conclusive representation of speech quality perception. Thirdly, the stimulus sentences in the current study consisted of the complete range of the DECtalk's capabilities and, indeed, balanced instances of the entire phoneme range in English speech that can be generated by the technology. Ratings obtained from such stimuli would be expected to elicit a greater range of perceptual responses than those obtained by confining the stimulus to a small subset of phonemes contained in a single sentence. Finally, the data obtained from forty subjects is likely to result in more robust conclusions than data obtained from only eight.

Rosson and Cecela stated in their conclusion: "There may have been important voice quality manipulations, and perceptual or situational variables that we failed to include" and consider their study to be only a "partial mapping" of the relationships among these kinds of variables. The current study builds on the findings of such research and, whilst in no way offering a complete analysis of the issues, expands and enhances the existing knowledge concerning perception of the aesthetic listenability of synthetic speech. If speech can be made more listenable, then listeners are likely to have a greater tolerance for speech output applications. This should in turn improve the chances of successful implementation of the technology.

When considering factor two, "Assurance", research has demonstrated that voice pitch can indicate the urgency of a message with higher pitch signalling greater urgency (Simpson et at 1984). Therefore it is appropriate that scales such as calm/anxious and relaxed/tense are correlated with changes in average pitch and head size. The scales suggest the factor to be associated with an assured,

calm, assertive style of voice. The inclusion of the scale "British - Foreign" suggests that subjects were rating voices of their own native language as being more assured than foreign-sounding voices. All subjects had been told to interpret this scale as "English is first language - English is not first language" in order to avoid confusion about how American-sounding voices should be rated on this scale. This was important as the DECtalk voices do have an American inflection, although this was minimised as much as possible during programming.

The correlation of this factor with the average pitch parameter and head size (which is associated with changes in overall pitch) would suggest then that assurance is determined by the perception of confidence and calmness within voices. Higher pitched voices tend to sound more anxious, tense etc. whereas lower pitched voices have a more assured and confident feel. This finding has an intuitive appeal. A consideration of the typical use of smooth, low-pitched voices in commercial advertising for persuasive purposes tends to reflect this.

Factor 3 "amiability" shows effects of richness and smoothness as does factor one. Furthermore, the scales associated with this factor are similar to scales associated with "listenability". Factor three is characterised as an agreeable, fluent voice. Interestingly, in factor one, the parameters richness and smoothness make independent contributions to the variance, whereas in factor three, the same parameters are so closely correlated that they cannot be said to make independent contributions. Unfortunately the variance accounted for by this factor is minimal and the correlation with the scales (in all but one case) rather weak. Consequently, detailed analysis here is inexpedient.

3.9 Factor analysis conclusions: Relevance to voice-output systems

Clearly the actual value of the ratings obtained are heavily dependent on the synthesiser under test, although it would not be unreasonable to expect the findings to apply to other formant synthesisers of comparable quality. Furthermore, DECtalk type synthesis forms the basis of a significant majority of synthesis-by-rule systems and whilst methods for manipulating the speech may vary between synthesizers, the rating methodology would remain valid.

This experiment utilised synthetic speech because the device used enables accurate and quantifiable modification of specific characteristics of speech. Exact replication is therefore possible, as is the option to adapt voices with precision and to quantify such adaptations for subsequent experimental investigation. With 'natural' recorded or digitised speech, such modifications would be technically problematic. Even highly trained professional vocalists would have difficulty producing the precise modifications in speech quality that are possible with synthetic speech. Digitised speech files can be processed and manipulated using various sound editing packages but this is a laborious and time-consuming process in comparison to the relative simplicity of synthetic speech modification.

A number of studies have highlighted differences between synthesised speech and natural speech but these have been more commonly concerned with assessing issues of the cognitive processing of the speech (see section 2.10), or selecting a particular voice for a task (Cox and Cooper, 1981) rather than investigating perceptions of speech quality or style. It should be noted that the technology of speech synthesis may eventually reach a point where any differences between the quality and naturalness of synthesised

speech and that of human speech will be neglible. At which point it should be possible to confidently apply findings from synthetic speech experiments to human speech output, assuming the objective features of human speech can be analysed and modified with comparative precision.

This stage of the study has illustrated that it is possible to manipulate the parameters which have been identified in order to help to design a synthesised voice which is significantly improved in terms of aesthetic listenability, pleasantness and an assurance quality. It is important, however, to be wary of anticipating the development of a single 'ideal' voice as being a unique solution to all intolerance and hostility to synthetic speech applications.

The tendency to see any particular breakthrough in technology as somehow being a universal solution to all implementation setbacks is unfortunately common within speech system development, as well as in other technological fields. In section 3.2 this type of reaction and the problems it can cause were discussed in relation to breakthroughs in synthetic speech intelligibility. Overestimation of the impact of 'major' breakthroughs in other areas of technology have had similar effects. For example, Laurel (1993) perceptively discusses the effects of technological innovations in a parallel scenario, describing how the relatively immature science of immersive virtual reality system development was hyped out of all proportion. When initial breakthroughs failed to fulfil the totally unrealistic expectations that they had generated, many investors withdrew funding and developers were faced with a whole new set of challenges. Like virtual reality, speech technology development is arguably still in its infancy and we must be wary of making over-ambitious claims for any one particular innovation.

It should be remembered that it is unlikely that any particular synthetic voice will be amenable to speech output in all situations, there is and cannot be an 'ideal' voice. Rather, the combination of

task and user characteristics associated with a particular system will dictate the applicability and distinguish any desirable, or undesirable features of speech displays if they are to be implemented within a specific installation.

As desirable voice characteristics may well be application dependent, it is important to consider how far we can generalise from experimental laboratory interactions with speech displays to real-world applications. As stated by Rosson and Cecela (1986), ratings tasks are abstract in nature and we do not know how the results reported here will correspond to an environment where 'real' users are interacting with 'real' voice-output installations. Findings from task-appropriateness ratings may indicate the potential reactions of users but do not preclude field studies of real users interacting with real speech output applications. It may be that an 'ideal' voice, in terms of user aesthetics, could be inappropriate in many settings due to problems with intelligibility and task efficiency. For example, it may well be necessary to find a balance between an aesthetically pleasing voice and one that meets the intelligibility demands required in a noisy office, or for use in a telephone-based installation. In another case, an 'ideal' voice that an elderly blind man may prefer for reading an electronic newspaper is likely to differ dramatically from the type of voice preferred by a young speech-impaired female child who is using the synthesiser to talk with. Further research is indicated to clarify such issues (these points are dealt with in greater depth in Chapter 6, sections 2, 3 and 4).

3.10 Factor analysis critique and justification for research progression

Factor analysis is a controversial technique, the application of which is not necessarily straightforward. Different variations of factor analysis and data rotation exist and their relative merits have been hotly debated. One criticism is that a particular style of factor analysis can yield a diverse range of possible solutions, based on the unique focus of the experiment and the mechanics of the mathematics used. In the case of experiment one, it is entirely possible that a range of voices modified using a different set of parameters may have yielded a completely different set of perceptual variables (DECtalk voices can be modified in many different ways). Furthermore, the data is shaped by the perceptual scales offered to the subjects. A different set of scales may elicit a completely different picture altogether.

In defence of the experiment, it should be pointed out that it would be impossible to ever come up with a 'perfect' set of ratings scales and a 'perfect' set of subjects who all interpret and use them in exactly the same way. As it is obviously impossible to investigate every single modification possible and have these rated on every conceivable adjective, it was necessary to start somewhere, and both the modifications used and the choice of the scales were guided by previous work in the field which has demonstrated that such adjustments may be particularly relevant in an investigation of speech perception (Including Cox and Cooper 1981, Rosson and Cecela, 1986, described in section 3.4).

When considering the mechanics of the technique it should be remembered that the factors from the analysis do not come ready labelled and the labels that have to be attached to the factors are only 'best guesses' about the interpretation of the factors. Merely to label

the factors does not necessarily provide a conclusive understanding of the issues but does give a clear starting point for ongoing enquiry.

Furthermore, there is more than one way of factor analysing a set of data and there is no 'best' way. The same data from the same sample can produce a number of different patterns of factors depending on which type of analysis is used (Gross, 1992). There has been much controversy in the intelligence literature about which type of factor analysis is most effective, so much so in fact that some researchers have concluded that the technique should not be used at all (Heim, 1970). It is beyond the scope of this thesis to explore the mathematical debate concerning the variations possible within factor analysis and their relative merits. Given that there is controversy concerning the technique, it is necessary for the progression of the research to emphasise that this first experiment has not identified the objective existence of 'listenability' 'assurance' and 'amiability'. Reification of factors, although tempting, should be avoided as in essence they are only mathematical abstractions (Gould, 1981). What the experiment *has* achieved, is the identification of a potentially fruitful focus for ongoing research and, with the ratings procedure, a practical and effective method for exploring speech perception in subsequent experiments. Experiments which will hopefully illuminate some of the psychological issues concerned with the perception of synthetic speech.

Having identified the limitations of the statistical techniques used, it is now necessary to establish whether or not the relationships that have been indicated are robust and stand up to sustained empirical scrutiny.

Chapter Four

Experiment Two

Empirical evaluation of the relationship between synthetic voice parameters and factor ratings

4.1 Summary

Having identified perceptual factors which were related to the modification of DECtalk voice quality parameters, it was necessary to determine the direction in which modifications of these parameters affect ratings on the extracted factors. It was also necessary to ensure that the factors uncovered were consistent and reliable through replication of the procedure with a new set of subjects. At this stage it was possible to streamline the ratings procedure by removing the superfluous scales and voices, using only those scales and voice parameters that were clearly associated with the extracted factors. This way, the individual factor/parameter relationships could be evaluated independently, rather than from within an amalgam of multiple unrelated variables.

In experiment 1 (chapter 3), factorial analysis had revealed that the voice parameters smoothness and richness made independent contributions to the variance within the listenability factor. For the assurance factor, perceptual ratings of variations in average pitch and head size correlated so highly that these parameters were not clearly shown to have made independent contributions to the variance. It is likely that subjects were perceiving modifications in

average pitch as virtually indistinguishable from modification in head size. This experiment concentrated exclusively on listenability and assurance and, using a concise and streamlined empirical ratings procedure, uncovered the way in which the manipulations in the synthetic speech signal can influence the two perceptual factors under investigation.

As well as the superfluous scales from the factor analytic study being removed the voice set was reduced to a single 'standard male' DECtalk voice. This was in order to eliminate the variability of a wide range of voices and allow an objective assessment of the influence of the voice parameters on factor ratings. The standard male voice was modified to produce a set of eight voices which contained examples of maximum and minimum settings of smoothness, richness, average pitch and head size.

For the factor listenability, a within subjects t-test revealed that when richness was set low, the voice was rated as significantly more listenable than when richness was set high ($p<0.001$). A second t-test also revealed that when smoothness is set high, the voice was rated significantly more listenable than when set low ($p<0.001$) (section 4.5 contains means and standard deviations).

Subjects clearly were perceiving both richness and smoothness modifications as having a direct effect on listenability, but in different ways. These parameters are not simply the reverse of each other, they clearly refer to separate parts of the synthetic speech signal and adjustments of them influence separate aspects of the auditory spectrum.

For the factor assurance, a t-test revealed that the voice with a low pitch was rated significantly more assured than the higher pitched voice ($p<0.001$).

Finally, the analysis revealed that voices with a large head size were rated significantly more assured than voices with a low head size ($p<0.05$) (section 4.4 contains the means and standard deviations).

This experiment clearly illustrates in which direction to modify the voices to produce either a more listenable, or a more assuring voice. In the case of assurance, it was shown that, although increasing head size does relate to an increase in ratings of assurance, a reduction of fundamental frequency is by far the most important manipulation for producing perceived assurance. Indeed the lower pitched voice was perceived as by far the most assured in the whole set.

These results corroborate the findings of the factor analytic study by providing further evidence of a consistent relationship between the specific modifications of the speech signal and the perceptual factors. Finally, the experiment inspires confidence in the validity of the streamlined ratings tool and the methodology developed in chapter 3.

4.2 Introduction

Experiment one uncovered the existence of relationships between the modification of synthetic speech parameters, certain perceptual scales, and the extracted factors. The next logical step was to determine empirically exactly how the modification of the speech parameters influence perceptions of listenability and assurance. This would have two functions, firstly as a test to ensure that the ratings methodology was producing consistent and reliable results, and secondly it would enable the development of a simple technique for enhancing listenability and assurance based on consistent empirical support.

The third factor, amiability, was not considered worthy of any further assessment at this stage for a number of reasons. Firstly, the scales that contributed to the perception of amiability were indistinct from the scales which contributed to listenability. That is, apart from a small suggestion of a 'fluency' characteristic (suggested by the scales: fluent/halting and comfortable/frustrating), the rest of the scales were qualitatively indistinguishable from the listenability scales. In two cases the scales associated with this factor used the same adjectives and the rest of the scale set appeared to be measuring a perceptual quality which could not be clearly differentiated from the perceptual factor characterised as listenability.

Secondly, there was ambiguity between the independence of the contributions the parameters smoothness and richness made amongst the factors. For the listenability factor, smoothness and richness made independent contributions. Whereas for factor three, amiability, the contributions of the two parameters were closely correlated. Listenability is characterised by separate influences of the richness and smoothness variables, whereas amiability was determined from the *combined* influence of the same two parameters.

Whilst it would perhaps be interesting at a later date to examine any subtle effects that combinations of smoothness and richness modification may make to perceptual evaluations, the data is really too weak to support sustained inquiry at this stage. Amiability was responsible for a minute proportion of the overall variance. Consequently, it is sufficient at this stage to briefly note that the factor illustrates an extremely marginal effect of the combined influence of the two listenability parameters, perhaps indicating that highly listenable voices are also, predictably, going to be considered friendly or likeable. At this level of analysis though with such weak data to draw conclusions from, the amiability factor was considered to be an anomaly of minimal importance to the study as a

whole `and further investigation of this factor at this point was considered inexpedient.

A more potentially productive line of enquiry was therefore taken, concentrating on firstly investigating the specific effects that parameter modification has on the speech signal, and secondly examining the relationship between the four parameters and factors one and two in greater depth in order to attempt to provide further support for the factor analysis conclusions.

4.3. Experimental design modification

The experimental procedure reported in chapter three was extremely tiring and repetitive for the subjects who participated. The overall procedure took approximately two hours and involved listening to over one hundred and ten minutes of more or less continuous samples of synthetic speech, the same sentences and range of voices repeated over and over again. This was deliberate, in as much as the intention was to duplicate - as close as is possible within a controlled laboratory setting - the kind of routine long term exposure that a user might experience if synthetic speech was a prominent feature in their daily working interface/environment.

Previous research examined closely in chapters one and two has clearly demonstrated that protracted exposure to even the most sophisticated synthetic speech can lead to less than favourable impressions of the technology and in some cases, outright hostility. It is possible that subjects may have become increasingly irritated with the experiment. The tediousness of the task may then have influenced the subjects ratings, especially on scales which relate to the aesthetic and pleasing characteristics of the voices. In effect, the monotony of the *experience* endured by the subjects may well have been reflected in their ratings of the voices. It is also possible that

some subjects may not have been able to concentrate on the task for such an extended period and some of their data may not have been an accurate and consistent reflection of their true perceptions of variations within the speech set.

In order to attempt to minimise any potential effects from this, and to spare the next participants from a protracted and tiresome ordeal, the experimental procedure was considerably streamlined. The modified ratings procedure lasted approximately thirty minutes in total. This was considered a long enough exposure to allow any effects to emerge, yet short enough to avoid fatigue influencing the subjects ratings. Another important consideration was that if, with a much shorter exposure to the speech, the main factors were still shown to be clearly linked to the same voice modification parameters, then this would provide further support for the reliability and validity of the ratings technique and confidence that perceived listenability and assurance could indeed be manipulated.

4.3.1 Voices

The voice set was reduced to eight variations on the standard male voice. The settings were chosen to give a voice set that demonstrated the maximum possible variation of the parameters under investigation. The parameters were therefore tuned to lower and upper extremes without exceeding DECtalk's capabilities or, in the case of average pitch and head size, without the voice sounding inhumanly high or low or overloading the synthesisers circuits.

Listenability set		Assurance set	
1. Smoothness	3%	5. Average Pitch	80Hz
2. Smoothness	100%	6. Average Pitch	160Hz
3. Richness	0%	7. Head Size	86%
4. Richness	100%	8. Head Size	115%

4.3.2 Scales

As the results of the factor analysis revealed the perceptual scales which made no significant contribution to the perception and measurement of the factors, it was now possible to remove these redundant scales and concentrate on determining the specific effects that the four voice parameters play in the measurement of validated listenability and assurance factor scales. This should enable a clarification of the relationship between the factors and the modifications of the parameters. Furthermore, if the results of the factor analysis are replicated, this will enhance the validity of the concepts of listenability and assurance and demonstrate the reliability of these particular ratings scales.

The scales which contributed significantly to the variance in the listenability and assurance factors were retained, totalling seven scales for each factor. Although the scale dissatisfied/satisfied seemed somewhat out of place in the listenability set (see chapter three, section 3.8), it was decided not to modify it in any way. It did, after all, have the highest relationship of all the scales with the listenability factor.

All other scales were those in experiment one which were clearly related to the two factors (see chapter three, sections 3.7.1 and 3.7.2)

The scale British/Foreign, although certainly related to the assurance factor, seems slightly ambiguous. A possible explanation for its inclusion in the assurance set may be that a voice which subjects perceive as characteristic of a native speaker of their first language is considered to be more assuring than one that is perceived as being foreign or alien in some way.

Listenability Scales

Dissatisfied	------	Satisfied
Irritating	------	Not Irritating
Harsh	------	Gentle
Hostile	------	Friendly
Unpleasant	------	Pleasant
Disturbing	------	Restful
Crude	------	Refined

Assurance Scales

Calm	------	Anxious
Relaxed	------	Tense
Authoritarian	------	Meek
Clear	------	Confusing
British	------	Foreign
Composed	------	Confused
Knowledgeable	------	Uneducated

4.3.3 Stimuli

None of the sentences used in Experiment one had elicited any kind of emotive reaction from the subjects (e.g. none provoked any reactions of amusement, a frequent common observation in

experiments when exposing subjects to some examples of synthetic speech). Furthermore, no subjects had reported any confusion generated by the combination of a particular stimulus sentence with a specific scale. Consequently, the suitability of the sentences was considered appropriate for further studies and they were retained in their original form.

4.3.4 Subjects

Twenty subjects (ten male and ten female) whose first language was English were used in the experiment, ages ranged between 18 and 34 years. None had taken part in any previous experiments with synthetic speech or had any prior knowledge of the aims of this particular experiment. They were not familiar with DECtalk or any other speech synthesis systems.

4.3.5 Method

A modified version of the control program from Experiment one was used to present the voices and collect and score the data. The practical procedure was therefore virtually identical to experiment one except in terms of the duration of the procedure and the range of voice modifications. Since none of the subjects had paused at any of the rest opportunities offered during Experiment one, these were considered redundant for the shorter procedure and were removed from the program. As in Experiment one, the subjects were shown a brief example of the simple procedure and then left alone in a soundproof laboratory to complete the task. Subjects worked at an even pace, completing the ratings procedure in approximately thirty minutes. Each subject made 56 ratings on the listenability scales for

voices 1 - 4 and 56 ratings on the assurance scales for voices 5 - 8. Presentation of scales, voices and sentences were, as before, randomised for each individual subject to avoid order and practice effects. Over the course of the experiment, each voice was rated twice on each individual scale for both factors.

4.4 Results:

Listenability

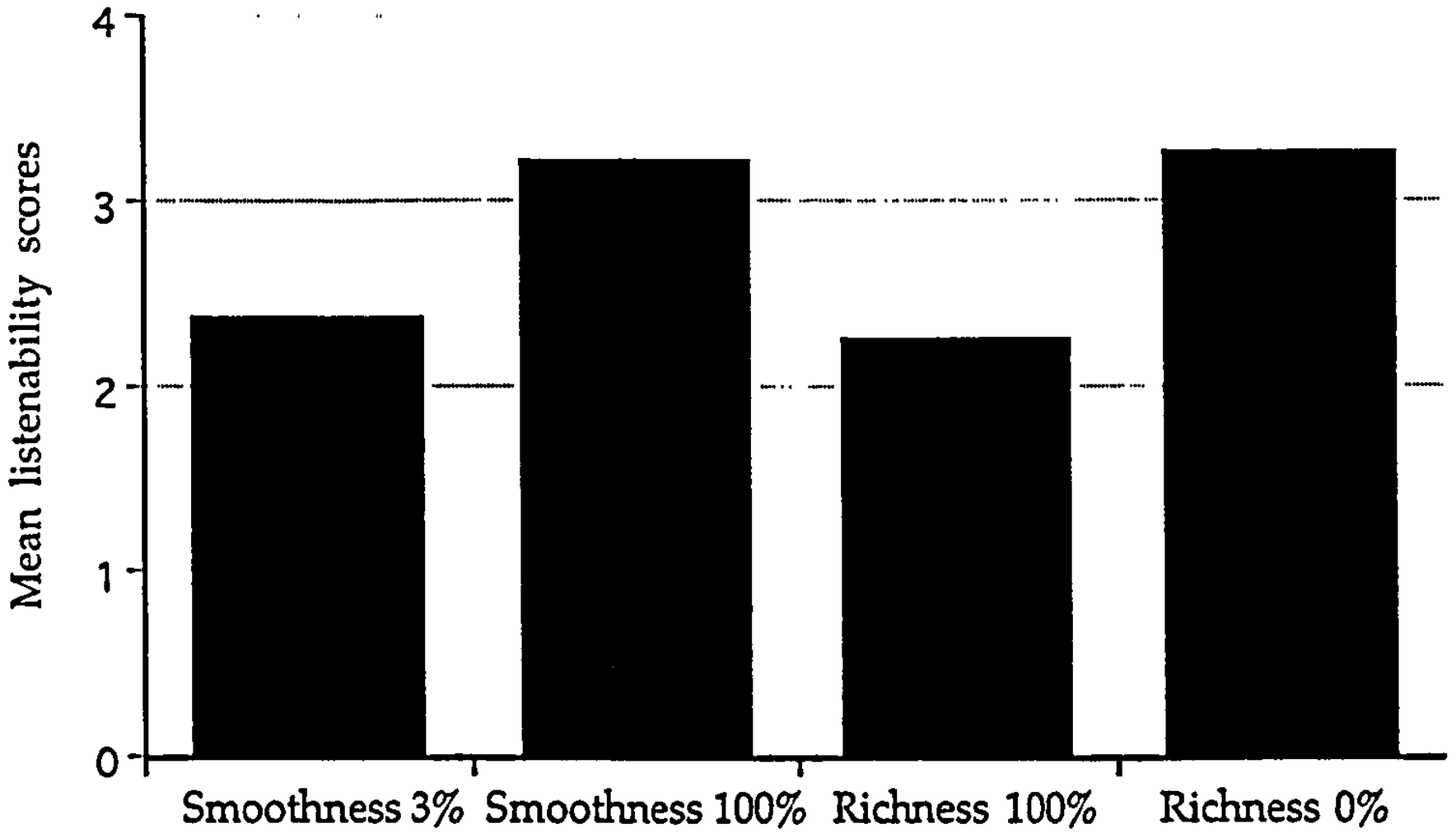


Figure 4.1 : Mean listenability scores for variations on the standard male voice.

Parameter	Mean	Standard Deviation
Smoothness 3%	2.353	0.51
Smoothness 100%	3.196	0.42
Richness 100%	2.250	0.49
Richness 0%	3.239	0.49

A within subjects t-test on the listenability data (see Appendix 4) revealed that when richness was set low, the synthetic voices were rated significantly more listenable that when it was set to maximum ($t(19) = 6.44$, $p<0.001$). Voices with smoothness set high were rated significantly more listenable that those with smoothness set low ($t(19) = -6.93$ $p<0.001$).

Assurance:

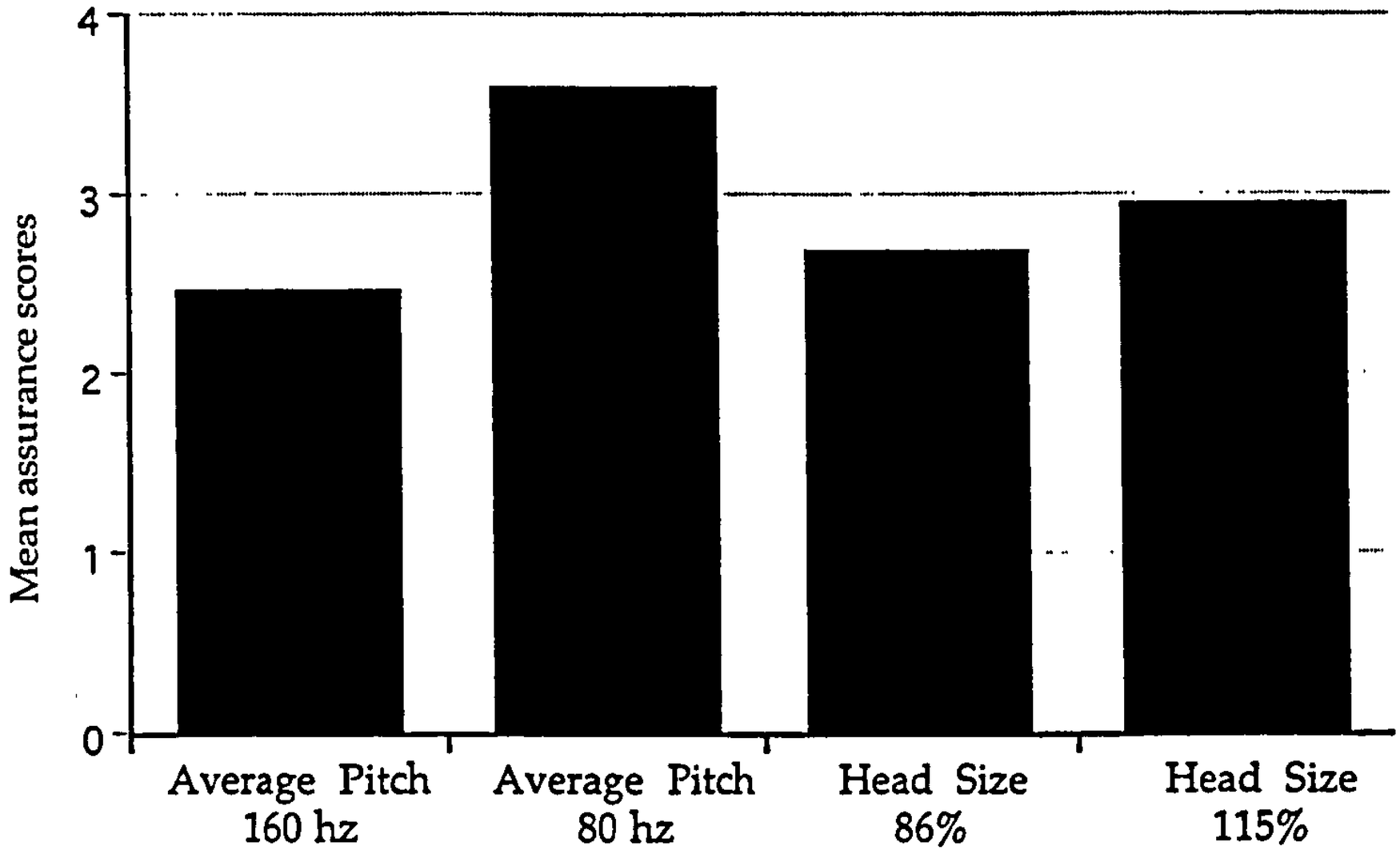


Figure 4.2: Mean assurance scores for variations on the standard male voice.

Parameter	Mean	Standard Deviation
Average Pitch 160hz	2.439	0.38
Average Pitch 80hz	3.560	0.43
Head Size 86%	2.664	0.32
Head Size 115%	2.924	0.38

t-tests on the assurance data (see Appendix 5) showed that voices with a low average pitch were rated significantly more assured that voices with a high average Pitch, or voices of either head size setting ($t(19) = 9.13$, $p<0.001$). Finally, voices with a large head size setting were rated significantly more assured that voices set with a small head size ($t(19) = -2.22$, $p<0.05$).

4.5 Discussion

The experiment has clearly illustrated the relationships between these speech signal manipulations and perceptions of listenability and assurance and reveals exactly how the DECtalk parameters need to be manipulated in order to achieve specific perceptual effects. In order to design a listenable voice, which is more likely to suit the majority of potential users and task settings, the smoothness parameter should be maximised. A comparable effect can be obtained by minimising the richness parameter. Furthermore, if listenability is of a high priority, as is likely to be almost always the case, the designer may do well to combine the two modifications to produce a voice which is likely to be especially pleasant and listenable.

For an assured or assuring voice an increase in head size is indicated but by far the greatest influence on this factor comes from a reduction of fundamental frequency. Lowering the pitch has been shown in this case to be the most important manipulation recommended for the design of voices high in assurance.

The results from this experiment suggest it may be possible to develop a technique for systematically enhancing the quality of Formant synthesis and indicate the progression of experimental work required in order to address this. However, before continuing the experimental program, an attempt was made to gain a simple understanding of what may be happening to the speech when the key parameters richness and smoothness are manipulated.

4.6 Spectrographic analysis

Although the DECtalk parameters (such as smoothness and richness) are cited in the research literature, there is a lack of technical information concerning the specific effects they have on the speech. For example, smoothness parameter modifications have been mentioned in association with changes in voicing energy at high frequencies (Bruckert, 1984) but the precise effects on the speech signal caused by parameter modification are not described in much detail in the general literature. Despite simple descriptions the parameters remain features specific to the DECtalk system and descriptions of their effects are vague (possibly due to commercial pressure).

In order to gain a basic understanding of the effects of the key parameters, recordings of various modified DECtalk speech stimuli with minimum and maximum varying richness and smoothness modifications were made. A spectographic analysis was carried out and the results revealed that smoothness and richness modifications affect different characteristics of the speech signal.

Samples of DECtalk's output were recorded on DAT tape and, using the Macintosh package "Signalyze", a number of spectrograms were produced to enable examination of visual representations of manipulations of the richness and smoothness parameters.

Prior to discussion of the spectrograms, it is important to note that a number of different electronic techniques may produce similar spectrographic representations. Therefore at this level of analysis we cannot be certain precisely what is being done to the speech signal to produce the changes shown. In effect, although the spectrograms show us *what* happens to the acoustic signal when the parameters are modified, they do not indicate *how* this has been

achieved (although obviously in this case they are the result of DECtalk's filtering technology).

Sound spectrographs were developed in the 1930s (Potter, 1946) and are frequently used in the analysis and processing of speech. The spectrographic speech-display contains the main, subjectively important features of the analysed speech (Pinter, 1996). Sound spectrographs are important tools in acoustic research, their output is basically a visual representation of the speech signal coming from a microphone or tape recorder connected to the machine. Spectrograms, or more recently *voice prints* are (basically) visible speech, that reflect the articulatory features of acoustic components (Fromkin and Rodman, 1993).

Two pairs of spectrograms are shown. The first pair (Figs 4.3 and 4.4). show the DECtalk saying the word "smoothness". Fig 4.3 has smoothness set at 3%, Fig 4.4 has Smoothness set at 100%. 3% is the lowest setting possible for DECtalk to say this word without producing a 'squawk' caused by an overload in the synthesisers circuits. The spectrograms are calculated as 2-dimensional frequency/amplitude plots, sampled at mid phoneme position for a single sample of the 'EH' sound.

A comparison of the amplitude spectrums shows a small effect on the lower frequencies (the peak to the left of the X axis which is sharpened but retains similar amplitude) but a pronounced attenuation of the higher frequencies (the peaks to the right).

The second pair of spectrograms (4.5 and 4.6) show the effect of modifications in the richness parameter. They show a strong effect on the lowest frequency peak (on the left which is greatly increased as richness is reduced).

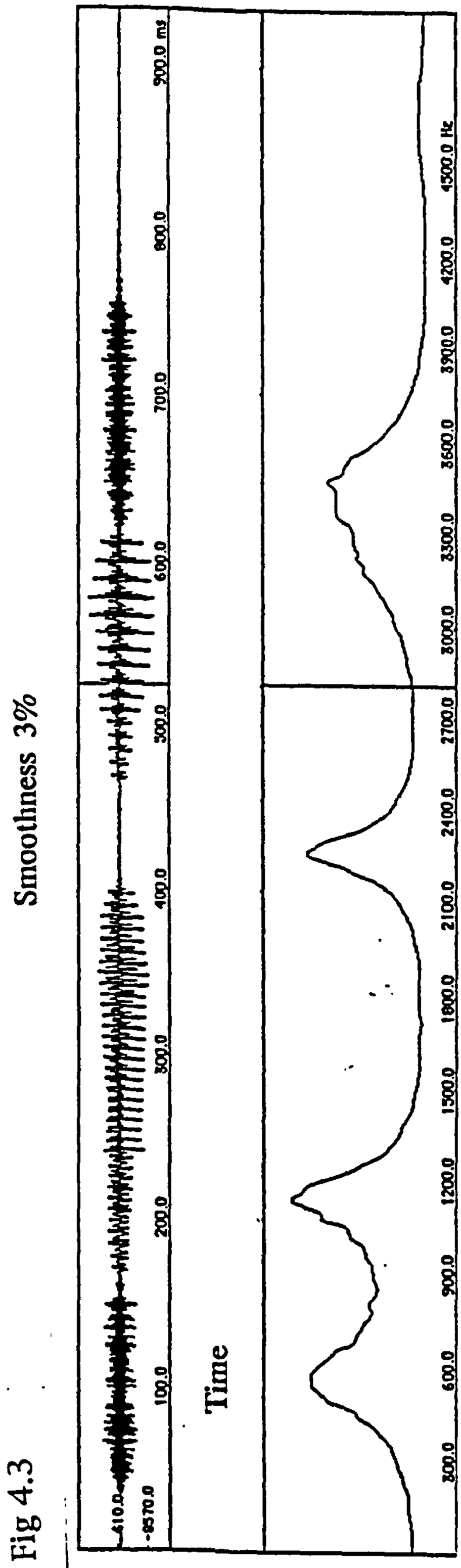

Smoothness 3%
900.0 ms
1500.0 Hz
300.0
400.0
600.0
700.0
800.0
2700.0
3000.0
3300.0
3600.0
3900.0
4200.0
200.0
1200.0
1500.0
1800.0
2100.0
2400.0
-410.0
-8570.0
100.0
800.0
900.0
Time
Frequency
Amplitude

Fig 4.3

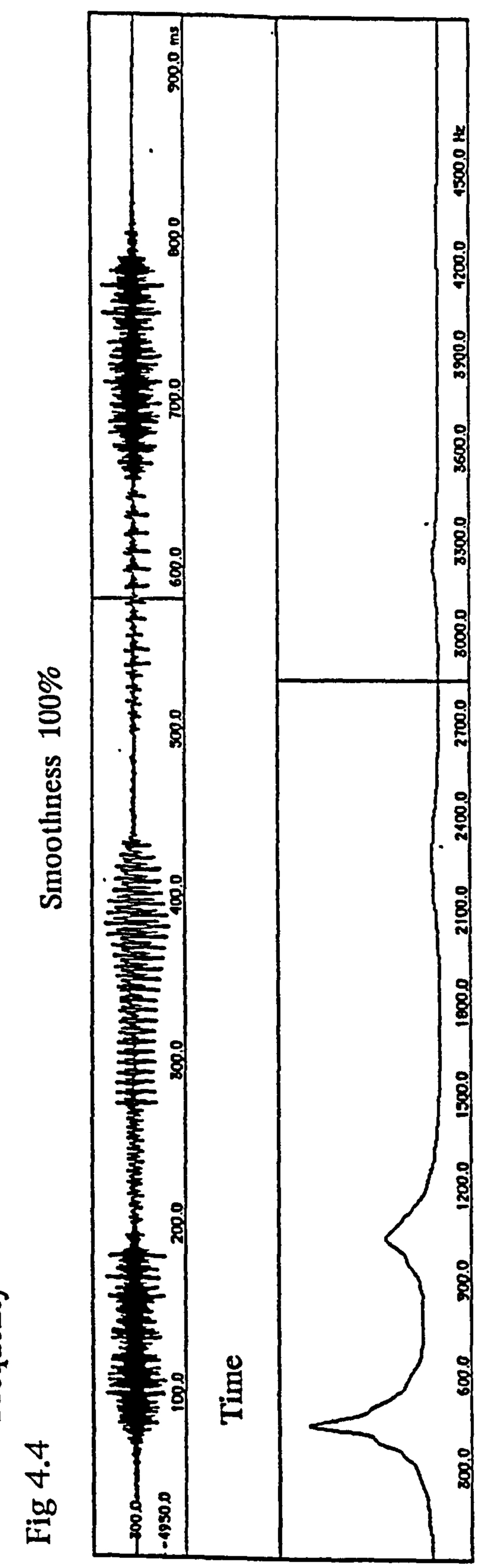

Smoothness 100%
900.0 ms
1500.0 Hz
300.0
400.0
600.0
700.0
800.0
2700.0
3000.0
3300.0
3600.0
3900.0
4200.0
200.0
1200.0
1500.0
1800.0
2100.0
2400.0
-300.0
-4950.0
100.0
800.0
900.0
Time
Frequency
Amplitude

Fig 4.4

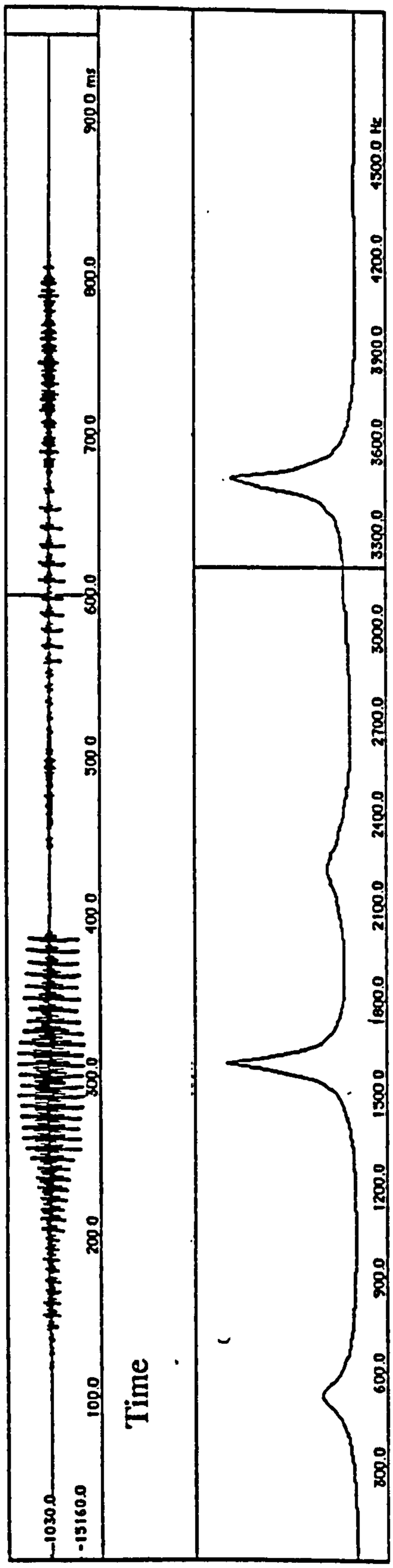

Richness 100%
900.0 ms
1500.0 Hz
600.0
700.0
800.0
300.0
400.0
200.0
3300.0
3600.0
3900.0
4200.0
2700.0
3000.0
2100.0
2400.0
1800.0
1500.0
1200.0
900.0
600.0
300.0
-1030.0
-15160.0
100.0
Time
Frequency
Amplitude

Fig 4.5

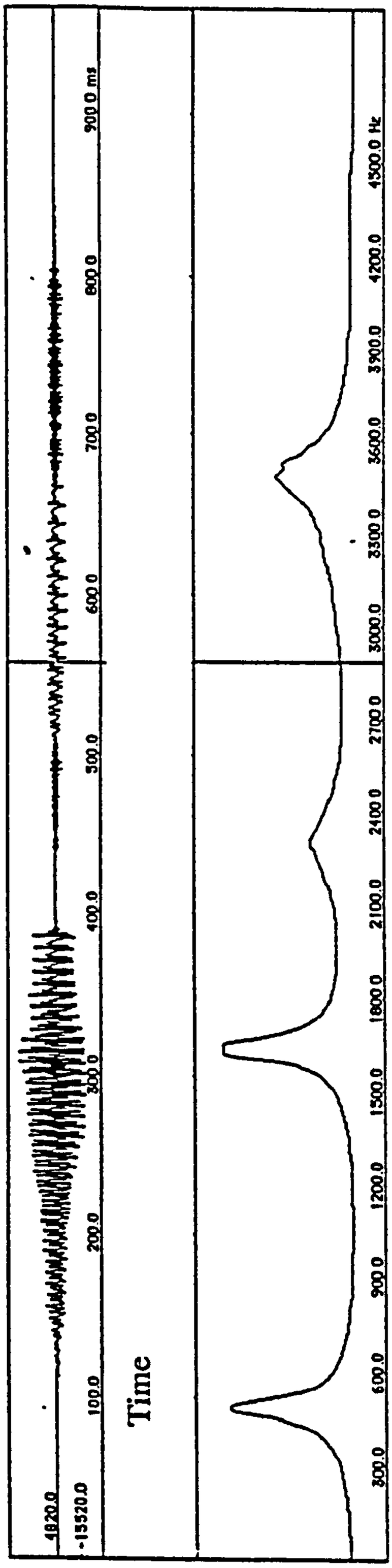

Richness 0%
900.0 ms
1500.0 Hz
600.0
700.0
800.0
300.0
400.0
200.0
3300.0
3600.0
3900.0
4200.0
2700.0
3000.0
2100.0
2400.0
1800.0
1500.0
1200.0
900.0
600.0
300.0
-1020.0
-15520.0
100.0
Time
Frequency
Amplitude

Fig 4.6

To conclude, the spectrograms give an early indication that an increase in the smoothness parameter appears to correspond with attenuation of the higher frequencies and has minimal effect on the lower frequencies. Whereas the richness spectrograms show that as richness is reduced, the lower frequencies are accentuated with little effect on the higher frequencies. However they are not mirror-images of each other, the vowel components of these signal occupy different regions of frequency space. While a certain similarity can be perceived when modifications are monitored acoustically, the two parameters (smoothness and richness) are actually manipulating very different parts of the speech signal.

As stated earlier, the spectrograms illustrate the effects of two levels of smoothness and richness on the synthesised speech signal. It may be possible to achieve comparable effects as those generated by richness and smoothness modifications by band pass filtering. If comparable results could be obtained using a filtering technique that was independent of a specific commercial product (in this case DECtalk), then a general tool for enhancing speech could be developed. The aim here would be to generalise the findings to all formant speech synthesisers, an interesting and potentially useful research avenue for speech science engineers with the appropriate technological resources (not currently available to the author).

As far as the current thesis is concerned however, it should be stressed that no major claims are being made from this simple analysis, its inclusion in the thesis serves to indicate how one might initially proceed if an engineering focus was employed in speech synthesis research. A systematic and extensive research program would be indicated if such an approach was to provide the backbone of a speech *science* thesis. This would involve a comprehensive sampling of all the relevant parameters in a full range of settings in order to develop a full understanding of formant speech signal engineering, especially if comparison with other techniques for

manipulating speech (both human and synthetic) were under investigation. Such an undertaking is outside the scope of this thesis, however, the findings are a useful starting point for other speech analysis researchers.

In this chapter, an effective technique for modifying the standard male DECtalk synthetic voice has been identified. This enables the production of specific perceptual effects. In the following chapter, we wish to establish whether these findings would generalise beyond the DECtalk male voice. This generalisation, if it occurs, would extend the value of the research. This would provide a useful test of the systematic reliability of the technique.

Chapter Five

Experiment three

Female synthetic voices: Listenability and Assurance.

5.1 Summary

The aim of the next stage of the research was to establish whether or not the voice modification procedure would continue to produce clear listenability and assurance effects and that the related ratings methodology would maintain general validity when applied to variations on a markedly different synthetic voice. The importance of replication within an experimental program is an established principle and is discussed in the introduction. It was also necessary to try and determine if there were any potential boundaries within the auditory ranges of the voice parameters, beyond which the perceptions of variations in the two factors become insignificant or disappear altogether.

In order to address these issues, variations of the DECtalk standard female voice were assessed using modified ratings software and an identical experimental methodology to that used in experiment two (chapter four). A set of eight voices were rated. These were modified from the standard female default settings and consisted of examples of the maximum and minimum possible settings of smoothness, richness, average pitch and head size that could be obtained within the constraints of the technology. The voice modification process was identical to that used in experiment two but some parameters needed

slight alteration (precise parameter settings are discussed in section 5.6). As before, twenty new subjects participated in the experiment, each making a total of fifty six ratings for each factor.

For the factor listenability, a within subjects t-test on the results for the female voice set revealed that when richness was set low at 0% the voice was rated as significantly more listenable than when richness was set the 100% maximum (p<0.001). The analysis also revealed that when smoothness was set high at 100%, the voice was rated significantly more listenable than when it was set low at 20% (p<0.001). These results are closely comparable to those obtained for the male voice and demonstrate that listenability can be significantly enhanced with the female voice with manipulations of richness and smoothness.

On examining results from the assurance scales, analysis showed that when average pitch was set low (160Hz) the voice was rated significantly more assured than when the voice was set at 240Hz (p<0.001). No significant assurance difference was shown between 95% and 115% head size manipulations. Once again, fundamental frequency has been shown to be a powerful indicator of perceived assurance.

This demonstrates, as in the previous experiment, that reduction of fundamental frequency is an effective method of producing a more assured and confident voice and that such an effect can be obtained consistently in both male and female voice ranges.

In experiment two, head size increases significantly enhanced listeners' perceptions of assurance in the male voice (although the effect was weaker than those observed from the other parameters), here in the female speech ranges, the effect is no longer apparent. This suggests that the use of head size increases to enhance

perceived assurance is limited in scope and is only likely to be effective within a standard male voice range.

Overall, the experiment clarifies the research issues and provides further support for the reliability of the principle speech modification techniques. The replication of the main effects, here shown on a set of female voices which are qualitatively different from the male voice used previously, suggests that the techniques are likely to have general validity across a range of different synthetic voices and can be employed with a considerable degree of confidence.

5.2 Introduction

The importance of replication

Within empirical psychological inquiry, the replication of experiments, to ensure the validity and reliability of conclusions, is generally considered to be an essential part of the methodological process. It is only through consistent replication of results that we can be certain that any effects observed are not anomalous artifacts of unique experimental conditions. In recent years academic and industrial researchers in the field of human-computer interaction have started to discover the power of controlled psychologically-orientated experiments in the field of interfaces (Barnard, 1991). As this approach is adopted, it is essential that rigorous methodological standards are maintained and, as such, the necessity for replication remains central to effective research.

Shneiderman (1998) notes that the transformation of HCI research from an introspective/intuitive model to an experimental one (which took two thousand years in physics) has occurred in just two decades. Because of this it could be argued that researchers have

started using an experimental approach without perhaps a proper understanding the basic principles and procedures which go to make up a fully rounded experimental program (drawing conclusions from an alarmingly small sample or an isolated experiment are typical weaknesses which crop up frequently in the literature). However, some prominent researchers have grasped the basic principles that make up effective experimental design, and one of these is the vital importance of replication. Shneiderman goes on to state that *multiple replications of experiments with similar tasks, subjects and conditions, will enhance reliability and validity,* and stresses interdisciplinary collaboration should be encouraged in order to effectively transfer such fundamentals of experimental psychology to the field of HCI.

5.3 Synthesis of the female voice

"The female voice has never been reproduced very convincingly in synthetic speech" (Monsen and Engbretson, 1977).

"In recent years it has often been said of work in speech synthesis that the field is ineluctably biased towards the adult male voice, and that attempts to synthesise the female voice have met with little real success." (Barry,1990)

Throughout the history of synthetic speech, research and development has been mainly concerned with the production of an adult male synthetic voice. In comparison, little attention has been paid to the development of sophisticated female synthetic voices, or children's voices. This is despite the fact that for most real world application domains, women and children are just as likely as men to encounter speech output systems. Furthermore, speech-impaired women and children are also just as likely as men to require the use of synthesisers as prosthetic devices in order to communicate. In

such cases, with the voice actually *personifying* the user, appropriate gender characteristics are likely to be not just a matter of aesthetic preference, but rather of essential importance to the user.

Typically, the literature shows research emphasis aimed at the production of a quality male voice via synthesis of a generic adult male vocal tract. Once this has been achieved to the designers satisfaction, female and/or children's voices have been developed by a process of adaptation of the male voice by implementing a series of parameter modifications associated with human female voice quality (Klatt, 1981). This process, in the case of female vocal tract simulation, typically involves a raising of the average pitch combined with a widening of the boundaries of the pitch range in order to produce a higher fundamental frequency and a more expressive quality. Furthermore an increase in breath sounds within the speech is included to model the relatively higher breathiness quality of most human female speech. Various other adjustments to the speech parameters are made but the results are not especially impressive.

The pioneering work of Fant and others in developing the acoustic theory of speech production dealt with a vocal tract of typical male proportions, and with a glottal source also characteristic of the male. Suggestions have come from many quarters as to how to break free of what Fant (1983) termed "the male dominance in speech synthesis".

It is likely that transformations *could* be devised for effectively converting rule systems between male and female speech although many attempts to generate acceptable voice quality of female speech from male rules have had only limited success (Sato, 1974, Holmes, 1988). It may well be that certain perceptually salient parameters of voice quality have been overlooked in synthesis work, and that

attention to these might significantly improve the quality of a female synthetic voice.

5.4 Differences between male and female synthetic voices

The characteristics which determine the perceived sex of a speaker, include pitch, head size, breathiness and sociolinguistic and dialect patterns. The female talker is perceived to have a smoother, higher, somewhat "breathy" voice than that of the male (Klatt, 1975).

A number of parameters whose origin is in the different character of the female glottal source have been shown to be of substantial importance in contributing to female voice characteristics in addition to the well documented importance of formant frequencies.

Although the phonetic descriptions of male and female speech for the same accent are very similar, their acoustic realisations are quite different. The fundamental frequency is also higher, usually by about 20%. The different dimensions of the vocal folds in a female larynx also cause the voiced excitation spectrum to be different in female speech, with far less power at the frequencies of the higher formants.

The female vocal chords vibrate in such a way that comparatively more breath noise is generated from an average female vocal tract than from a male tract of average proportions. In order to compensate for this variation DECtalk uses an increase in the parameter 'breathiness' to produce more realistic speech for the female voices within the default voice set (described in the DECtalk owner's manual, 1983).

It has been suggested that at least a part of the speech differences between the sexes is socially conditioned, in that the two sexes

actually learn different *styles* of speech, that there are clear differences in the socio-cultural articulatory characteristics between male and female voices (Guenzburger, 1984). If this effect is significant it could account for some of the difficulty in devising rule transformations between the sexes, especially if the attempted development of a female voice is based on an adaption of an existing male voice with an existing masculine style.

Due to the complexities of the issues here, the research necessary to produce a quality synthetic voice based entirely and directly on a female vocal tract would need to be considerably extensive. One recent project is worthy of mention here. Hanson (1997) has attempted the formulation of a set of acoustic parameters of the voicing source that reflect individual differences in the voice qualities of female speakers. The study attempts to describe and quantify normal variations of voicing characteristics across speakers and illustrates a continuing effort to improve the analysis and synthesis of female speech based on hypothesised glottal configurations.

Whilst certainly challenging, such a line of enquiry remains outside the scope of this thesis. The reader is therefore referred to Barry (1990) for an in-depth analysis of the acoustic properties of female voices, and to Hanson (1997) for insight into contemporary study.

5.5. Experimental aims

Having established a seemingly reliable technique for eliciting perceptual effects using the DECtalk standard male voice, the next logical stage was to broaden the perspective and assess the effectiveness of the technique and the associated ratings procedure on a markedly different synthetic voice. Clearly, if the technique were to have general validity, it would have to be applicable to a

variety of synthetic voices. The DECtalk standard female voice was chosen for a number of reasons. In practical terms, although the voice parameters are set at very different default levels than those used for the standard male, it is still possible to manipulate the characteristics previously shown to be associated with listenability and assurance with the same degree of precision used in the earlier experiments. Furthermore, the voice is markedly different from the standard male, certainly more so than any of the other DECtalk voices in the default voice set.

Aside of a test of general validity of the procedure concerning its ability to produce significant improvements to the speech and associated variations in perceptual attributes, this experiment was designed in order to address a number of related research issues.

Firstly, when considering listenability, the main purpose of this experiment was to determine whether or not a significantly more listenable female voice - than the standard default voice - could be created by modifying those parameters which resulted in a more listenable male voice in the previous experiment. A side issue related to this is the possibility that the neglect that designers have paid to the development of a convincing female synthetic voice of comparable quality to the male may have resulted in an overall listenability decrement. In order to investigate this, listenability scores obtained in the previous experiment could be compared with a similar set of scores for the standard female.

Secondly, in the case of the assurance factor, it was necessary to ascertain whether or not there might be a defined area, an active 'window' within the frequency spectrum where pitch modification has a clear effect on assurance ratings. Intuitively, it seemed unlikely that assurance perceptions could be consistently modified across the entire frequency spectrum and that there was likely to be some upper (and possibly lower) cut-off point, beyond which

modifications would no longer produce a noticeable effect. Experiment two clearly demonstrated a decrease in perceived assurance as pitch increased, it therefore seems possible that this effect becomes ambiguous or may be extinguished as the frequency escalates out of the limited scope of the male pitch ranges. Therefore the question arises as to whether there might be an upper *limit* to the effective manipulation of perceived assurance? There may be a cut-off point after which manipulations in the speech parameters produce no significant effect on perceptions of assurance. Or, will perceived assurance be just as relevant to voices in a female range, with the reduction of pitch within the higher range causing a similar increase in assurance perceptions? The relatively higher frequency ranges of the DECtalk standard female voice allow an assessment of the strength of the assurance effect as the frequency climbs beyond the ranges that subjects would normally perceive as associated with day-to-day speech and into extremes that would be impossible to achieve with a human vocal tract.

(N.B. this avenue of enquiry cannot be taken with the listenabilty factors because of technological considerations, the parameters being at their maximum and minimum settings for each voice in both experiments).

5.6. Voice parameter modification

There are a number of important differences in the glottal source wave between male and female speakers. The most obvious of these is the fundamental frequency, which, for a female voice, is almost always significantly higher than that of most men. Although raised fundamental frequency is arguably the most important variable which distinguishes female voices from male, there are a number of other variations which are also important. If the standard male voice

is simply modified to speak with a pitch level typical to female speakers, the voice is more reminiscent of a child or someone who is deliberately adopting a falsetto style, than that of an adult female.

Standard male voices with fundamental frequency set at typical female levels sound very unnatural and consequently, it seemed unlikely that the ratings procedure would elicit any useful information. If the voice was to sound like a child's, this could possibly have effects on ratings on both the factors based on variability between the subjects' experience and preferences of children's voices.

Other variables might affect the assurance ratings, with scales such as Knowledgeable/Uneducated and Authoritarian/Meek likely to be influenced by voices which are perceived as belonging to a child. Hence, it seemed necessary to include the other modifications to the speech signal which are important in creating a realistically *adult* female voice. In order to test for the factors in voices with fundamental frequencies higher than approximately 180Hz, the voices would need to be female voices in order to be credible. Any variability in results caused by the manipulation in the other parameters to achieve a set of adult voices is likely to be minimal in comparison to potential variation caused by perception of the voices as being those of children.

In the linguistics literature, average habitual vocal pitches for both sexes and all ages are suggested: 190Hz for a 17 year old female and 227Hz for an adult female (Wilson, 1979). Taking 200hz as an average baseline, the two average pitch settings were plus and minus 40Hz (as in experiment two). This therefore covers a distinction of fundamental frequency with the lowest setting equal to the highest setting for the male voice in experiment two. The other three parameters were modified to the same extent as in experiment two. i.e. as far apart as possible without the synthesiser 'squawking' or

the level of voicing becoming inaudible.

There are a number of other differences between the default parameter settings for the DECtalk standard male and standard female voices:

	Standard Male	Standard Female
Breathiness	0%	46%
Gain in Frication	73db	66db
Assertiveness	100%	65%

Both Breathiness and increased frication are characteristic of human female voices. Modification of these parameters affect the *style* of the voice arguably more than the quality (the distinction between style and quality was discussed in chapter two). Informal evaluation of modifications of these two parameters indicated that they appear to produce a minor and subtle improvement in the output, making the voice sound more convincingly female. Therefore, the different setting for the male and female voices should have no effect on the voice quality factors. As well as these there are also modifications to the distribution of the formant frequencies in order to reproduce speech which is a more realistic representation of speech produced from a vocal tract of different dimensions from that of the typical adult male. These variations are concerned with the creation of a voice with female characteristics rather than any enhancement or degradation of quality.

Assertiveness, on the other hand, has a pronounced effect on voice quality with an obvious effect on the timbre of the speech, although not one which is perceived by the listener as being especially representative of the femaleness of a voice. One can only assume that the designers may have attempted to somehow weaken, soften or reduce the stridency of the voice in order to enhance its female

character but were not noticeably successful with this particular adjustment.

Consequently, and in the absence of any documentation concerning the effects of assertiveness modifications on the speech signal, it was necessary to remove any potential variability in ratings caused by the different default settings for the standard male and female voices. Assertiveness was set to the same level as in the previous experiment (i.e. 100%), eliminating any potential bias and enabling a more accurate comparison of the findings from experiments two and three.

These modifications resulted in the standard female voice speaking with virtually identical style to the standard male voice used in the last experiment, the differences being only in terms of the characteristics of the voice (in terms of the expression of maleness and femaleness). This makes the comparison of voices with a higher fundamental frequency range possible.

The range of settings used to create the voice set in this experiment were:

Listenability set	Assurance set
1.Smoothness 20%	5. Average pitch 160Hz
2. Smoothness 100%	6. Average pitch 240Hz
3. Richness 0%	7. Head size 95%
4. Richness 100%	8. Head size 115%

(All voices using default standard female set at 100% assertiveness)

5.7 Method

The control program from experiment two was briefly modified to accommodate the parameter requirements. Scales and sentences remained identical to those used in experiment 2.

Twenty new subjects were used, aged between 17 and 26 (ten male and ten female). None had taken part in any of the previous experiments or were familiar with DECtalk or any other speech synthesis systems. The procedure followed was identical to experiment two. The same laboratory and equipment were used and there were no changes made to the methodology for presenting the speech and obtaining perceptual ratings.

5.8 Results

Listenability

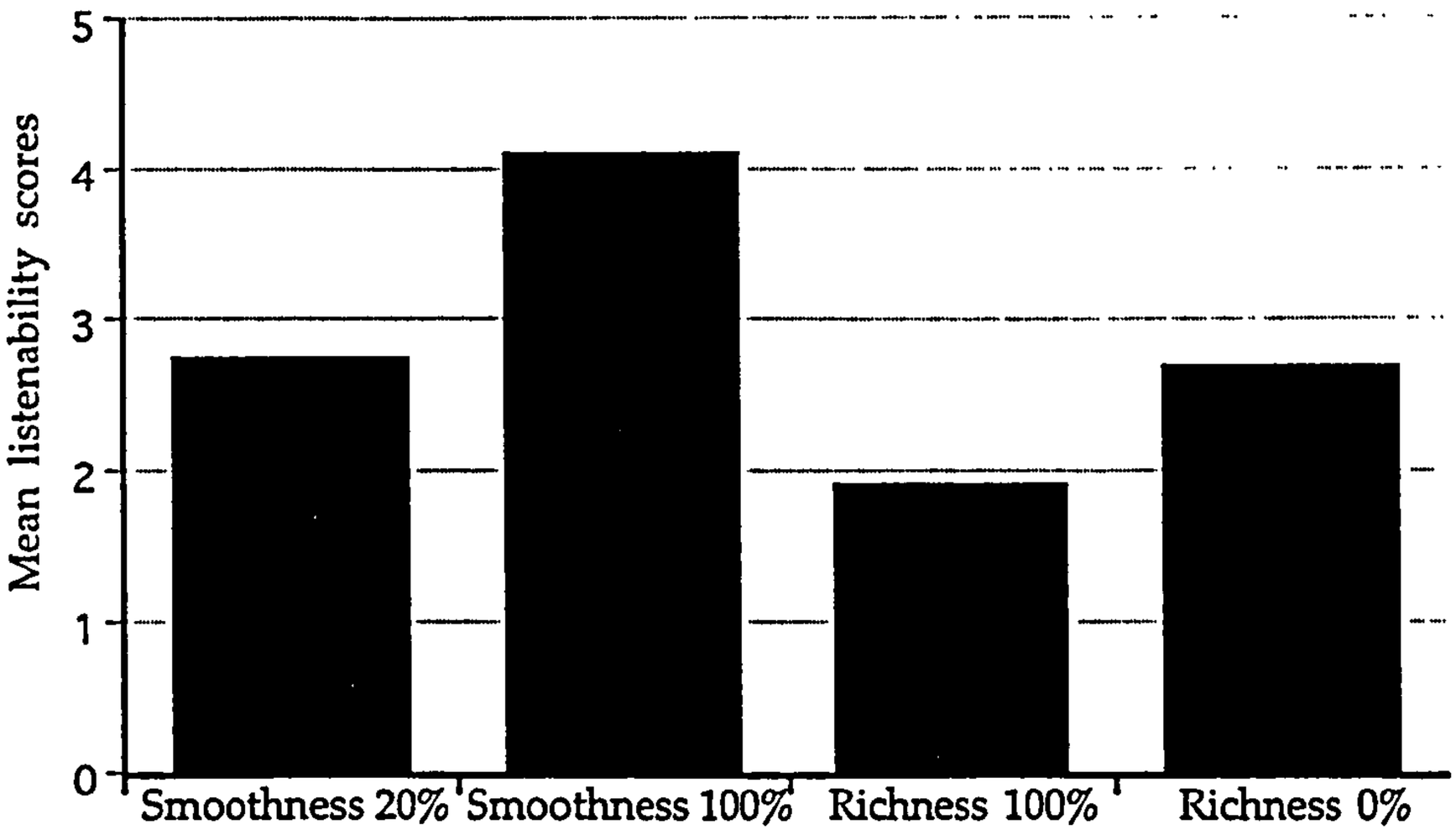


Figure 5.1 : Mean listenability scores for variations on the standard female voice.

Parameter	Mean	Standard Deviation
Smoothness 20%	2.725	0.29
Smoothness 100%	4.089	0.34
Richness 100%	1.878	0.44
Richness 0%	2.674	0.36

A within subjects t-test revealed that when richness was set low (0%) subjects listenability ratings were significantly higher than for those set at 100% ($t(19) = 7.13$ $p<0.001$). Furthermore, the voice high in smoothness (100%) was rated as significantly higher on the listenability scales ($t(19) = -16.21$, $p<0.001$. (see Appendix 6 for full analysis).

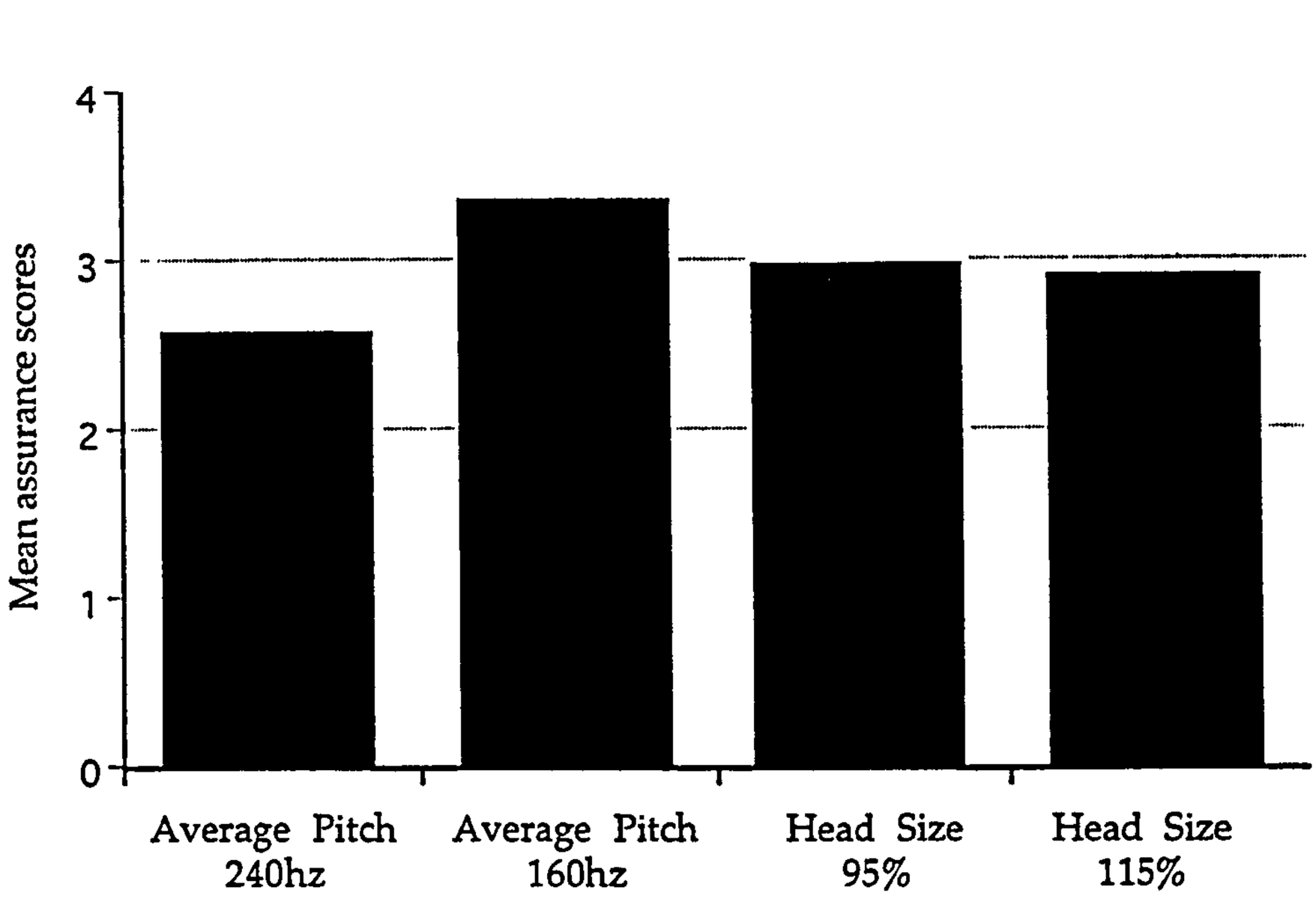


Figure 5.2: Mean assurance scores for variations on the standard female voice.

Parameter	Mean	Standard Deviation
Average Pitch 240Hz	2.557	0.35
Average Pitch 160Hz	3.332	0.34
Head Size 95%	2.953	0.26
Head Size 115%	2.892	0.291

A within subjects t-test revealed that when average pitch was set low (160Hz), the voices were rated significantly higher on the assurance factor than the voices with high average pitch (240Hz) ($t(19) = 6.28$ $p<0.001$). No significant differences were found between the 95% and 115% Head Size manipulations ($t(19) = 0.94$ $p<0.357$) (See Appendix 7 for the full analysis).

5.9 Discussion

Speaker gender and perceptual ratings

As with the standard male voice, the effect of modifying the smoothness and richness parameters produced closely comparable ratings on the listenability scales. As the modifications to the two parameters for the female voice were identical to those of the male voice, it is reasonable to conclude that listenability perceptions can be clearly modified with equal precision with the female voice as with the male voice. This result suggests that the listenability modifications are likely to enhance the aesthetic characteristics of a substantial range of varying synthetic voices and can be used with confidence in synthetic speech system design.

One interesting variation between the listenability results for male and female voice was revealed in the analysis. This is the comparative degree of listenability enhancement when considering the contributions of smoothness and richness. Whilst, for both voices, reduction of richness and increase in smoothness results in significantly enhanced listenability, for the female voice, smoothness modification has a comparatively greater impact. The high smoothness voices being rated as being significantly more listenable than not just the low smoothness voice but also both of the richness voices. In short, when developing a listenable male voice, smoothness and richness modifications appear to be both important, when developing a listenable female voice, smoothness modifications should take priority. They produce major effects which overshadow the relatively minor effects of richness modification.

Manipulations of average pitch did indeed affect perceived assurance in a similar direction as in the last experiment. The female voice with average pitch set low at 160Hz was rated significantly higher on the assurance scales than than the 240Hz voice. Once again, lowering

the fundamental frequency of the synthetic speech is shown to be the most effective method under consideration in eliciting higher scores on the assurance scales.

The lack of significant difference between assurance ratings of the head size variations is likely to be due to there being limited variability between the two voices under test. As was mentioned in the procedure section, the DECtalk was incapable of producing speech at anything lower than a 95% modification to the standard female voice head size without resulting in an overload of the synthesisers circuits. Lowering the head size to 86%, as was done for the male voice in experiment two, was therefore impossible. The DECtalk resonator gain parameters enable attenuation of the signal at critical points and then amplification of the signal back to normal later in the synthesis process. This can result in a 'cleaner' text-to-speech conversion which is less prone to overload. Despite a number of attempts to manipulate these parameters, in almost all examples of female synthetic output, the lowering of the head size beyond 95% led to an overload in DECtalk's filters and caused a piercing 'squawk'. This obviously was not suitable output for the subtle perceptual rating required for the experiment. Furthermore, raising of the upper setting beyond 115% to increase variability produced a bizarre voice that was completely inhuman-sounding. Here again a useful evaluation and rating could not be reasonably anticipated.

So, it is reasonable to conclude that the two voices used to test head size manipulations were not sufficiently distinct to elicit a significant variation in ratings scores. Unfortunately, until the synthesiser technology has advanced to compensate for this limitation, it is not possible to significantly manipulate perceived assurance for female synthesised voices using the head size parameter. If the technology can be improved to enable such an evaluation of head size modifications, it is possible that the results obtained would be comparable to the previous experiment, with voices with a significantly larger head size eliciting significantly

higher scores on the assurance scales.

It is also worth considering the possibility that the relative lack of sophistication of the female voices in comparison to the male ones disguises subtle changes made to the voices in the assurance modification. Smoothness/richness modifications are still distinguishable as changes in aesthetic quality, even with the cruder female voice, but the subtle changes in pitch/head-size may not be profound enough to be clearly perceived by subjects even if the technology allowed it.

The male and female listenability data came from different experiments using different participants. These were designed as two separate experiments to test the effects of parameter modification on the two factors. It would therefore be inappropriate scientific practice to attempt to establish statistical verification of synthetic speech gender preferences with data collected in experiments designed for quite different purposes (i.e. one to determine how subjects perceived listenability of male voices as compared to *other* male voices, and the second experiment examining changes within an exclusively female set). Furthermore, some of the parameters modified in the experiment were those that are considered to be directly related to perceived gender differences, i.e. head size and average pitch. With both male and female voice sets used varying in precisely those characteristics which determine perceived gender, a comparison should be undertaken with some caution.

In order to examine gender preferences for synthetic speech it would be necessary to conduct further experiments using a wide range of voice manipulations of the characteristics of the speech which have been suggested as being associated with perceptions of gender. It would also be necessary to establish specific techniques to determine gender preference. There is likely to be more to determining preference than simply scaling for listenability. This

may be an interesting and fruitful line of enquiry but remains outside the scope of this research.

However, there is no reason why mean listenability ratings cannot be portrayed descriptively in order to illustrate the perceived listenability variations between DECtalk standard male and standard female voices, bearing in mind that overall preferences for a particular gender cannot be established from this data and would require a separate line of enquiry. Figure 5.3 shows mean ratings scores for the four parameter settings used in the male and female listenability experiments. As can be seen, for the male voice, both smoothness and richness modifications make clear contributions to variation in perceived listenability, whereas with the female voice, the effect of high smoothness appears more pronounced.

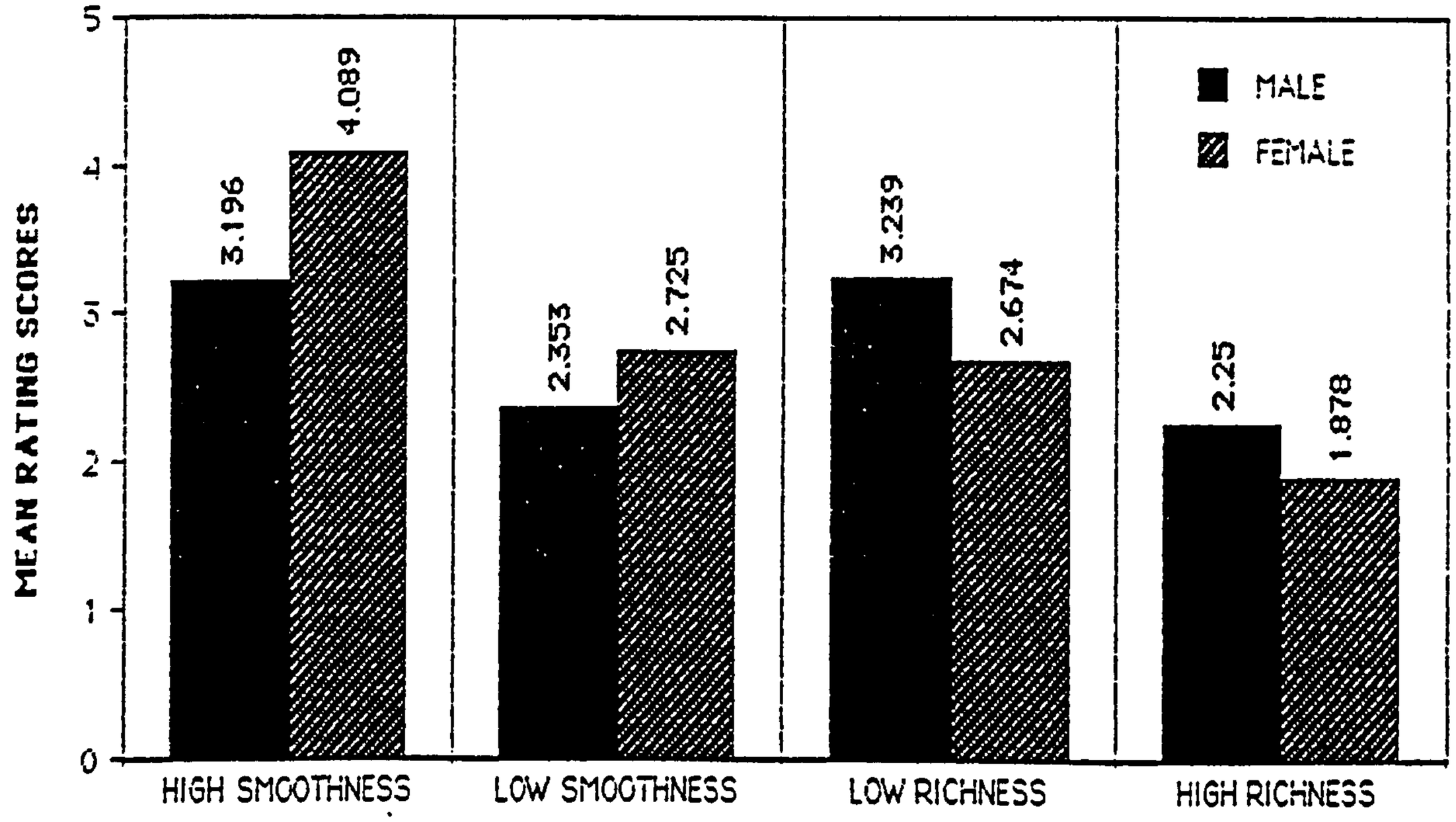


Figure 5.3: Listenability means for male and female voices, a comparison of data from experiment two and three.

This type of illustration is not suitable for the assurance data for a number of reasons. With listenability, the smoothness and richness voice manipulations were exactly the same for both the male and the female voice sets: 3% and 100% smoothness, 0% and 100% richness, for both male and female sets, allowing a direct comparison. This was not the case with assurance as pitch is modified by DECtalk to produce the default female voice and the manipulation used in experiment two could not be precisely replicated.

The only useful comparison was in the difference between the ratings for high and low pitch, how *far* a decrease in pitch in the female voices enhances assurance ratings compared to a similar (though at a different scale) modification in the male voice set. In the case of the female data, as has been discussed, the relationship between assurance modification and pitch levels was relatively minor when compared with the male data.

Finally it was not useful to compare the head size data between the male and female voices as the variations in the parameter settings were not equivalent due to the constraints of the technology. The relatively minor variation in the voices possible when head size is manipulated within the female ranges cannot therefore be usefully compared with the more extensive variations in the male set, which elicited a significant difference in rating scores. Until such time as the synthesis technology can overcome this limitation on head size manipulations in the female ranges, potential effects cannot be tested empirically.

Chapter Six

Experiment Four

Empirical evaluation of the relationship between parameter modification and perceived listenability

6.1 Introduction

At this point it was necessary to determine if the relationship between smoothness and richness modifications and listenability ratings was linear. Listenability was now adopted as the direct focus of the research (for reasons discussed later) and it was now necessary to refine and concentrate the line of inquiry and to clarify and validate the empirical findings to date. In order to achieve this, the relative merits of the two factors and their measurement were considered and the next experiment dealt exclusively with listenability. The main aim of the experiment then was to test and hopefully establish a direct method for enhancing or degrading the aesthetically pleasing characteristics of the speech using an incremental range of specific parameter manipulations.

Although the prolonged study of assuring voices may yield some interesting results, the study and development of aesthetically pleasing synthetic speech is likely to be of comparatively greater benefit to potential users of such technology. Certainly the need for voice quality enhancement and a reliable tool for measuring it is a common complaint in the literature across many application domains (discussed in chapter 3). Additionally, the listenability factor uncovered in experiment one was by far the most prominent

component by a considerable degree. The results clearly showed this and the following experiments demonstrated that this could be manipulated fairly dramatically. As the development and study of a tool for measuring synthetic speech quality and a robust technique for improving it were the principle aims at this point in the research program, the relatively minor contribution of the assurance factor was left for another program of research.

Another important justification for focussing the next experiment exclusively on the study of listenability relates to the precision and validity of the scaling procedure. As has already been discussed, previous research into speech quality has not always managed to successfully separate the rating of quality from the rating of speaker personality.

Indeed, the difficulty in generating scaling adjectives that allow subjects to rate voice quality independently of personality and/or style was encountered during experiment one (discussed in chapter 3). As we have seen, the factor analysis eliminated many of the major culprits from the original list of scales, and the decision to exclude 'amiability' from subsequent experiments was an additional refinement. However an examination of the listenability and assurance scales reveals that some ambiguity may still remain but, arguably, *only* to any significant degree with the assurance scales. With the exception of clear/confusing, all of the assurance scales can only refer to the perception of the speaker rather than the speech. On the other hand, with the exception of hostile/friendly (and providing subjects are using satisfied/dissatisfied as a rating of their own opinion of the speech), *all* of the listenability scales can only refer to the listener's perception of the aesthetic quality of the speech. This is a powerful reason for concentrating exclusively on listenability, it is a necessary step which is required in order to reduce ambiguity concerning the ratings methodology and add clarity into the progression of the research.

Experiments two and three have clearly shown that a maximum setting of smoothness and a minimum setting of richness both result in a significantly more listenable voice, but, whilst it may be tempting to assume that in each case the effect occurs incrementally in a linear fashion as the parameters are increased/decreased, the results do not necessarily support this. It may potentially be the case that whilst, for example, it is clear that 100% smoothness is significantly more listenable that 0%, there may perhaps be a mid range setting that is even more listenable than the top setting (or less listenable than the lowest setting). Additionally, richness may well produce an optimal effect at a higher setting than 0%. In each case, comparing only extreme values with a two point sampling technique does not give a definitive and comprehensive enough view of the overall picture.

A further necessity is a consideration of how the parameter settings might interact with each other. For example, does the optimal richness or smoothness parameter remain constant as the other parameter is modified? These questions need to be addressed with the next experiment, which, using a slightly different ratings scenario assessed the listenability of voices modified across the full spectrum of both of the parameters in question.

6.2 Method

This time, the experiment was conducted at a different institution, Bournemouth University, however attempts were made to ensure that conditions for the subjects remained closely comparable to those experienced in previous experiments.

6.3 Voices

Twenty five variations on the DECtalk default male voice were used. Five settings of smoothness, each compared with five settings of richness:

Richness:	0%	25%	50%	75%	100%
Smoothness:	0%	25%	50%	75%	100%

The voices were modified across the full range at equal increments in a fully crossed design. For example, for richness 0%, there would be five variations across the smoothness range. This meant that every possible combination of the settings was included and the results should reveal any interactions that are occuring. The voice set was then varied enough to examine both the relationships of the factors to listenabilty and any potential interactions, without being so large as to make the task too tiresome or tedious for the subjects.

6.4 Scales

Listenability has been consistently and effectively measured with the seven, five-point ratings scales derived from experiment one and used in the previous two experiments. All seven of these were used again here for each of the 25 voices. Although the total 175 scales were presented in a random order, it was ensured that during the experiment, each of the 25 voices would be rated on all seven of the scales. Although the scales appear below with all the negative adjectives on the left hand side, for the experiment, the poles were randomised in order to avoid any response bias that may possibly occur.

Listenability Scales

Dissatisfied	------	Satisfied
Irritating	------	Not Irritating
Harsh	------	Gentle
Hostile	------	Friendly
Unpleasant	------	Pleasant
Disturbing	------	Restful
Crude	------	Refined

6.5 Stimuli

None of the combinations of sentence and parameter settings overloaded the synthesizers circuits. Therefore, the suitability of the sentences previously used was considered appropriate for further studies and they were retained in their original form.

6.6 Apparatus

The range of synthetic speech samples were generated using the DECtalk synthesizer controlled by a Macintosh computer. The speech samples were recorded onto a high-quality audio cassette using a JVC KD-D30 cassette deck. This machine was also used to present the speech during the experiment, in this case using a pair of Sony headphones.

6.7 Subjects

Twenty subjects (ten male and ten female) whose first language was English were used in the experiment, ages ranged between 18 and 50 years. None of them had any known hearing impairment. None had taken part in any previous experiments with synthetic speech or had any prior knowledge of the aims of this particular experiment. Although some of the subjects had heard examples of synthetic voices before, none of them were especially familiar with DECtalk or any other speech synthesis system.

6.8 Method

The practical procedure was virtually identical to previous experiments except for two minor modifications required due to technical constraints. In this case the speech was presented using a high-quality cassette recording rather than 'live', direct from the DECtalk, and ratings were recorded on printed score sheets rather than on a computer. After careful consideration it was decided that this minor change should not make any difference to the subjects. In both cases, the subjects are sitting in a room, listening to speech and scoring scales. Whether this is done by pressing a key or by ticking a box was considered to be irrelevant and it is impossible to distinguish 'live' DECtalk speech from a high quality recording.

In each case the subjects were seated in a laboratory and given instructions concerning how to proceed with the experiment. They were played a short section of synthetic speech at the default setting in order to eliminate any surprise elements from initial exposure. Subjects were informed that the semantic content of the sentences was irrelevant and that they should concentrate on rating the

quality of each speech sample, rather than assessing the characteristics of the speaker. As with the previous experiments, subjects were instructed to imagine routine, daily exposure to the individual voices and to rate their judgements accordingly.

Subjects were each given seven score sheets. Each one had 25 scales printed on it. 175 speech samples were recorded on the tape, with a 5 second gap of silence after each sentence. In order to ensure that subjects remained synchronized with the tape, after every 25 presentations, a synthetic voice stated "please turn to the next page" and there was a short pause to allow the subjects to do this before starting on the next page. The DECtalk standard female voice was used for this instruction as it is clearly distinguishable from the voices being rated. When the procedure was clearly understood, the tape was turned on and the subjects were left alone to complete the task.

Although headphones had not been used in the previous experiments their use was considered appropriate in this case because of the relatively subtle modifications of the voices. It was vital that subjects played extremely close attention to voices which, in cases where parameter settings are close, are not always easy to distinguish between.

One concern with the use of a tape to present the speech was that the subjects were forced to make their ratings at a specific pace, they could not complete the experiment at their own pace as they had with the automated procedure used previously. In order to address this, the first few subjects were observed through a one way mirror to determine if the pacing of the speech samples on the tape was comfortable or if they might lose their place on the scoring sheets. However, it was soon observed that they very quickly got into a steady rhythm and, if anything, subjects would have been able to proceed with the task efficiently at an even faster presentation rate. When asked, after the experiment, subjects stated that the pace of the

tape gave them sufficient time to make each rating, they did not feel pressurized to respond too quickly.

Left alone to complete the task, subjects worked at an even pace, completing the ratings procedure in approximately twenty five minutes. Finally, subjects were thanked and debriefed and the results were unscrambled, scored, tabulated and analysed.

6.9 Results and discussion

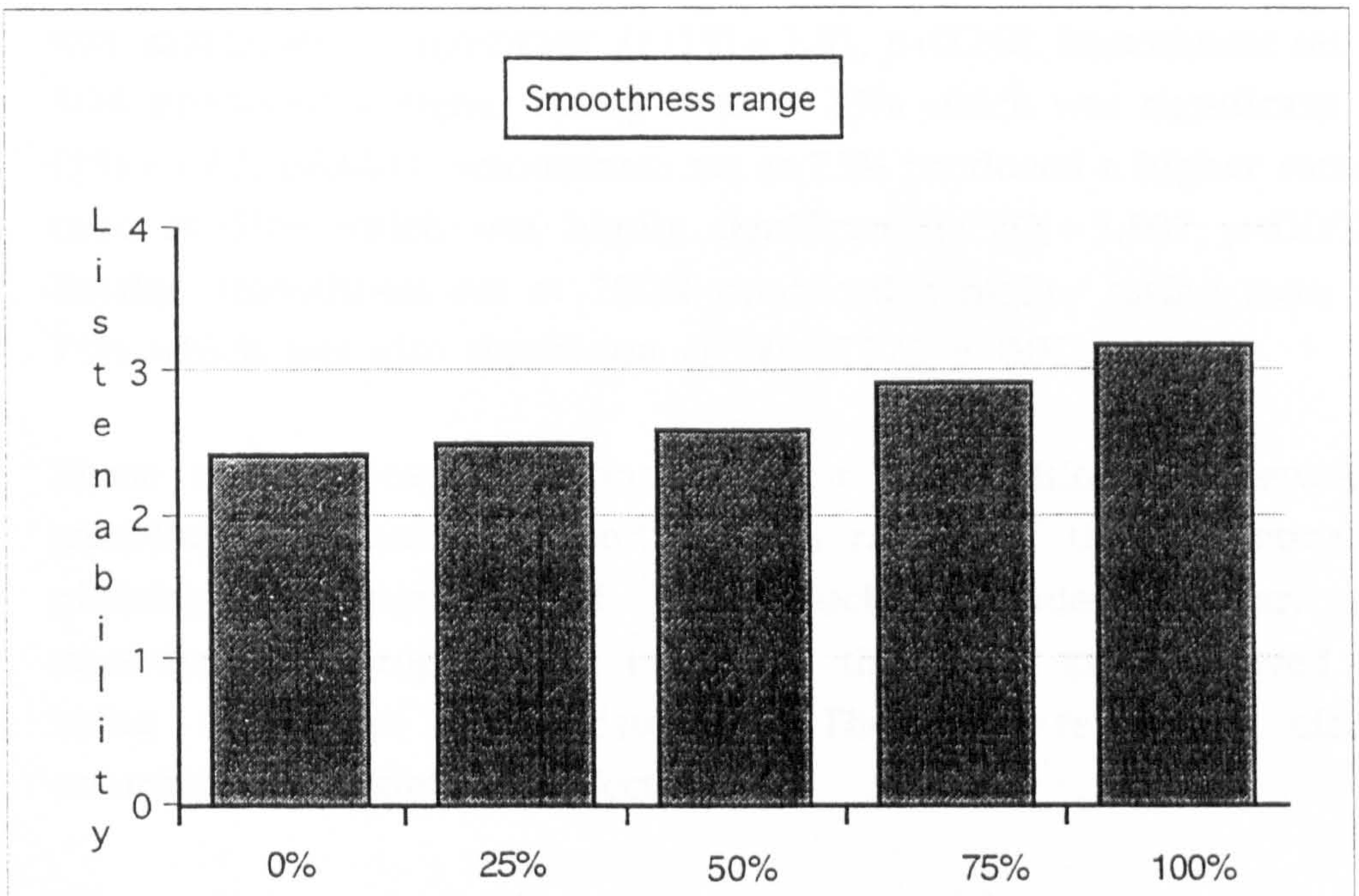


Fig 6.1: Mean listenability scores for variations on the smoothness range for 20 subjects (for each setting, the richness modifications are collapsed across the full range).

Parameter	Mean	Standard Deviation
Smoothness 0%	2.43	0.36
Smoothness 25%	2.49	0.35
Smoothness 50%	2.63	0.48
Smoothness 75%	2.94	0.46
Smoothness 100%	3.22	0.37

In order to test for within-subjects effects, a two way analysis of variance was applied to the data using SPSS. This revealed a highly significant effect for smoothness modification:

$F(4,76) = 28.9 \quad p<0.001$

Planned comparisons were conducted between the different levels of smoothness collapsed across richness. The two-tailed analysis revealed that a 25% level produced a higher rating than 0% which was marginally significant (t (19) = 1.81, p<0.10). Smoothness set at 50% produced a higher rating than at 25% which was significant (t (19) = 1.87, p<0.01). Smoothness set at 75% produced a higher rating than at 50% which was highly significant (t(19) = 3.807, p<0.001). Finally, smoothness set at 100% produced a higher rating than at 75% which was also significant (t(19) = 2.325, p<0.01).

These results clearly demonstrate that the relationship between smoothness enhancement and subjects ratings of the aesthetically pleasing characteristics of the speech is indeed linear. As smoothness is progressively increased, the voices are perceived as being more and more listenable. The data reveals a clear, unambiguous pattern of effects.

(The graph and results for richness modifications are shown grouped together below, in Fig 6.2 for ease of viewing).

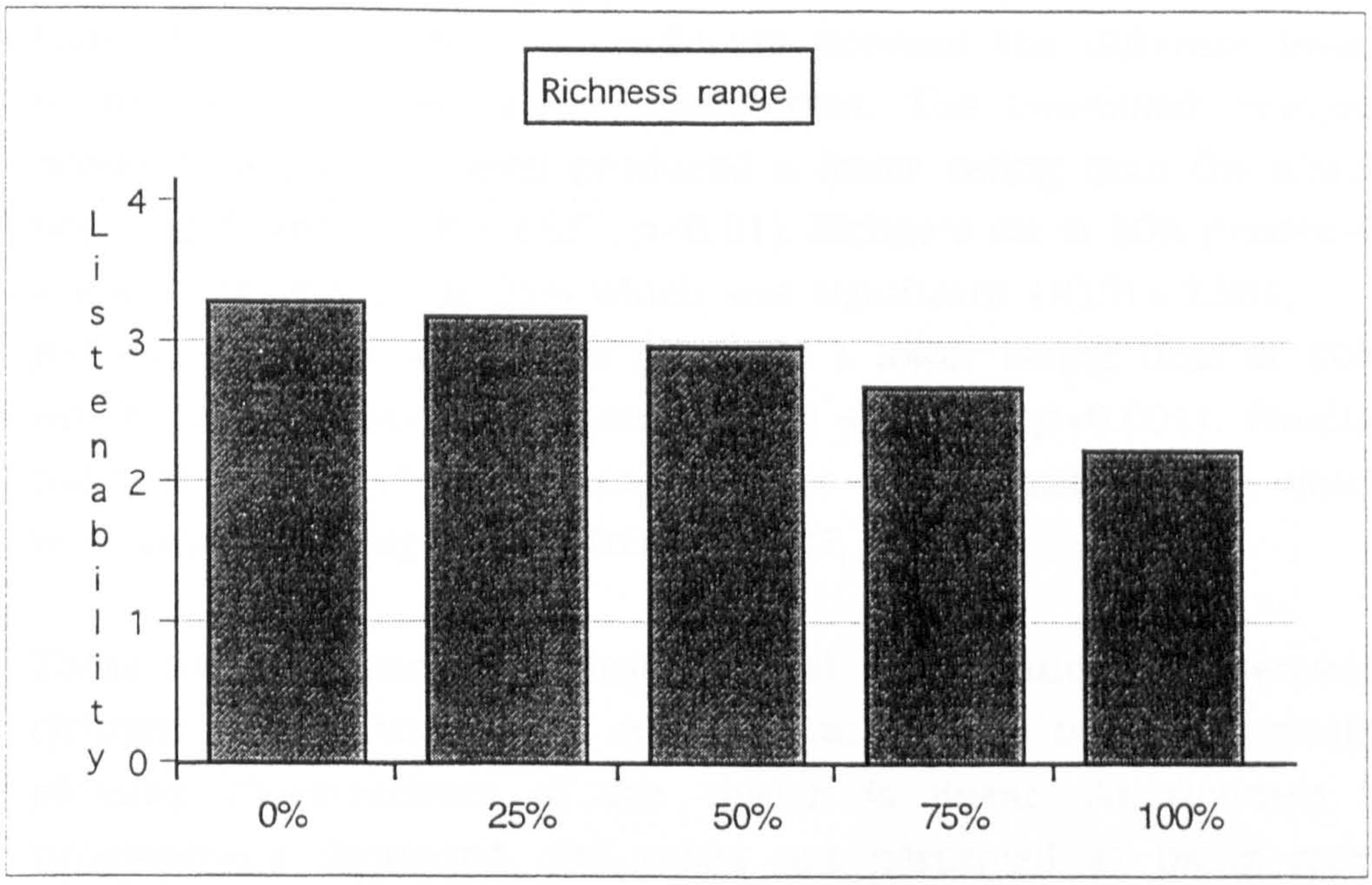


Fig 6.2: Mean listenability for variations on the richness range for 20 subjects (for each setting, the richness modifications are collapsed across the full range).

Parameter	Mean	Standard Deviation
Richness 0%	3:15	0.50
Richness 25%	3.04	0.51
Richness 50%	2.87	0.42
Richness 75%	2.57	0.33
Richness 100%	2.16	0.40

In order to test for within-subjects effects, a two way analysis of variance was applied to the data using SPSS. This revealed a highly significant effect for richness:

$F = (4,76) = 23.287 \quad p < 0.001$

Planned comparisons were conducted between the different levels of richness collapsed across smoothness. The two-tailed analysis revealed that a 25% level produced a lower rating than 0% which was significant ($t(19) = 2.557, p<0.01$). Richness set at 50% produced a lower rating than at 25% which was significant ($t(19) = 2.881, p<0.01$). Richness set at 75% produced a lower rating than at 50% which was highly significant ($t(19) = 4.489, p<0.001$). Finally, richness set at 100% produced a lower rating than at 75% which was also highly significant ($t(19) = 3.847, p<0.001$).

These results clearly demonstrate that the relationship between richness modification and subjects ratings of the aesthetically pleasing characteristics of the speech is linear. As richness is progressively decreased, the voices are perceived as being more and more listenable. The data again reveals a clear, unambiguous pattern of effects.

Finally, the analysis revealed that there was a two way interaction between smoothness and richness ($f(16,304) = 2.77, p<0.001$).

6.10 Parameter interactions

In order to develop a more comprehensive picture of the overall results consideration of parameter interactions is indicated. Up to this point, analysis has focussed on the individual effects of the two parameters, in each case with the other parameter modifications collapsed over all settings. Whilst this has revealed a clear picture of the way smoothness and richness can be used individually to produce specific improvements in a set of voices, it is also worth considering how the various settings may interact with each other. Optimal settings have been established for the parameters individually but, as the parameters affect the speech signal in

different ways, we cannot assume that by setting both parameters to their most listenable setting, we will necessarily produce the most postively rated voice. This needs to be established by additional examination of the data.

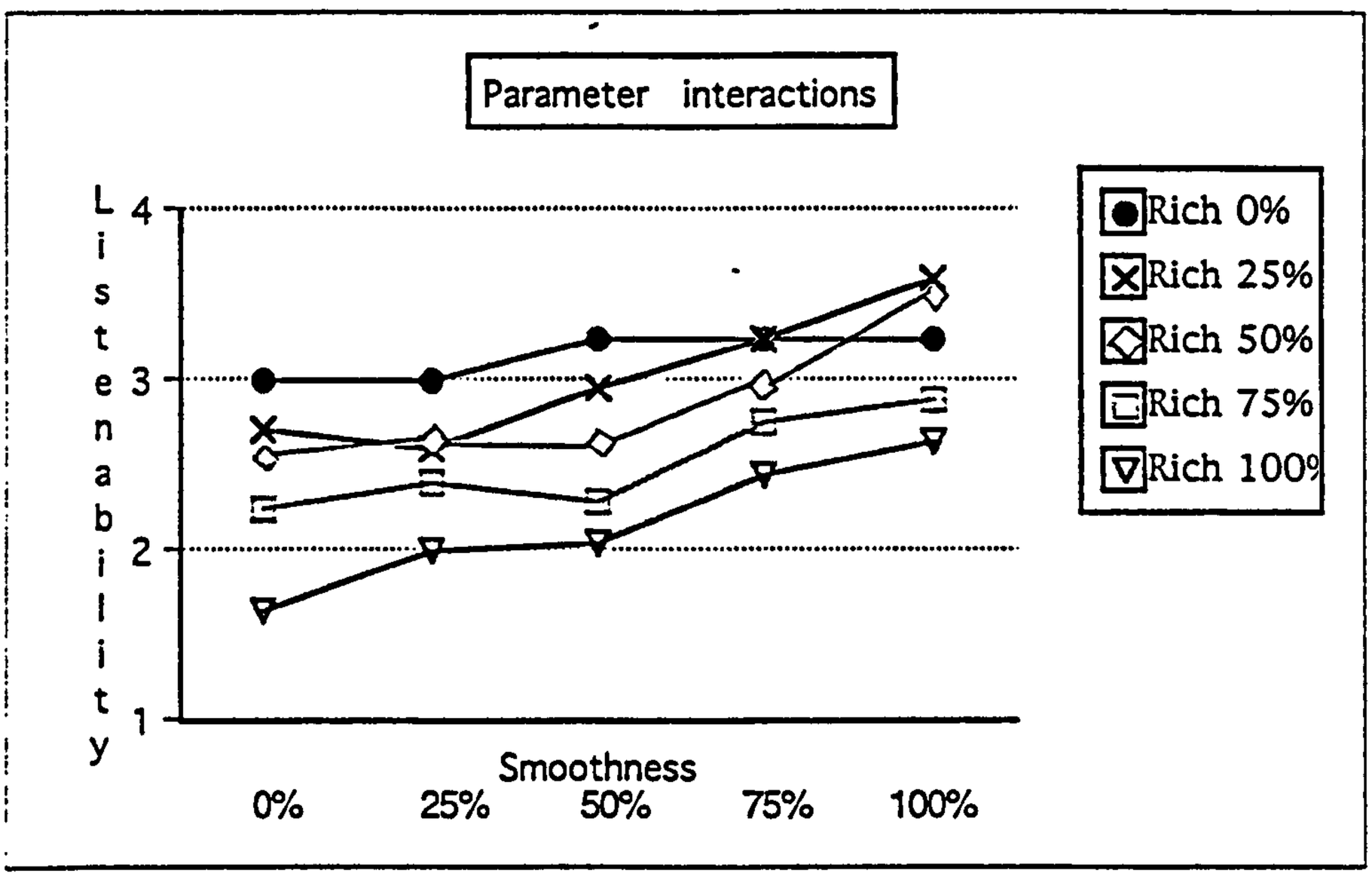


Fig 6.3: Mean listenability scores for smoothness variations with richness modified over 5 settings.

Examination of the graph reveals that for most combinations, a fairly standard pattern has been revealed. For the richness settings 100%, 75%, 50% and 25% the pattern is similar with the combined effects of increasing smoothness with decreasing richness resulting in clear listenability enhancements. There is however, one interesting feature shown, when richness is set at 0%, the pattern of enhancements virtually dissapears. At this setting, smoothness manipulations appear to make little difference. Although all the voices are rated consistently more pleasing, additional enhancements caused by smoothness are not apparent.

Overall, the graph illustrates that the richness parameter appears to be a slightly stronger determinant of listenability, when it is set at its highest setting, the ratings produced are the least acceptable set of all (richness 100% with smoothness at 0% was rated the worst voice in the entire set). Smoothness changes make a marginal effect when richness is set to 0% (smoothness 100% paired with richness 25% produce the most positively rated voice of all).

To summarise, while the graphs show some interesting patterns in the data and the relative strengths and weaknesses of the parameters at different levels, no dramatic or surprising interactions are revealed that would indicate hidden complexities in the overall pattern of effects. The graphs illustrate that richness and smoothness can both be manipulated to enhance listenability individually, and that when combined, those same manipulations result in the highest quality voice of all.

The experiment has revealed a clear picture of the linear nature of the relationship of the listenablility factor for both smoothness and richness parameters. The data is unambiguous, providing highly significant results. Clearly, a gradual increase in smoothness corresponds with increased perceptions of voice quality. The same effect occurs as richness is decreased. Consequently, the experiment has demonstrated an effective technique for enhancing or degrading the aesthetically pleasing qualities of DECtalk speech.

The simplicity of the experiment and the clarity and consistency of the data suggests that the technique used was sufficiently rigorous and robust. The experiment could be repeated using finer graduations of the modified speech, for example sampling voice modifications at 10% intervals rather than 25% but it seems extremely unlikely that a different pattern would emerge. Indeed, an

experiment with many more samples may produce distorted results because the procedure would be much longer and subjects judgements could be effected as they become bored and irritable.

When considering the ratings scales, the fact that the subjects were specifically rating voice quality rather than speaker personality (or any other opinion about the speech) can be accepted with reasonable confidence for all but one of the scales. Interestingly, one subject reported that he had experienced some difficulty rating voices on the hostile/friendly scale without making a conscious judgement about the speaker. No other subjects reported any particular difficulty with this scale and examination of the raw data reveals that scores on this scale fall clearly within the overall pattern of quality rating. However, in the light of the scaling issues discussed in chapter 3, it should be noted that the interpretation of this particular scale may be slightly ambiguous for certain individuals.

The data that emerged was encouragingly free of noise, although there was one subject who was more or less completely neutral about all of the voices and rated them virtually all the same. This subject appeared to either be unable to distinguish the voices or to be completely indecisive. The inclusion of his data, and the resultant dilution of effects, merely serves to illustrate the strength of the overall pattern. The fact that the effect observed was clear and consistent in virtually all cases and the results found to be highly significant testifies both that the experiment was conducted rigorously and that this is a highly robust phenomenon. It clearly demonstrates that when synthetic speech is modified in this way, there is a linear relationship between the modification of the parameters and enhancement/degradation of perceived voice quality. There is certainly no ambiguity about results such as these.

To conclude, it has now been clearly shown that it is possible to significantly affect synthetic speech quality and, in the case of

DECtalk, a technique has been established to achieve this. However, there is one issue that may have a profound influence on the ecological validity of the overall findings. In each case, speech quality perception has been investigated in a *passive* listening scenario where subjects were instructed to pay no attention to the semantic content of the messages. Indeed they were constructed to be deliberately bland and easy to ignore. However, in almost every conceivable 'real-world' implementation of computer speech output, users would always be *actively* processing the speech for meaning. Rarely, if ever, would they be ignoring the message. This is rather a different task scenario and the question of how this might influence perceptions of quality requires investigation. This should strengthen the ecological validity and overall contribution of the research. This consideration will form the next stage in the investigation.

Chapter Seven

Synthetic speech comprehension: The effects of cognitive load on listenability.

"...it is possible for some very well-established phenomena that have robust effects in abstracted laboratory tasks to have very different levels of influence when embedded in more complex tasks." (Landauer,1988)

7.1 Introduction

The meaning of any knowledge gained from experimental procedures and the confidence we can have in any conclusions are contingent on the methods used (McGrath, 1994). The experiments in this study have so far used a passive scaling methodology. They involved subjects reporting their perceptual impressions of variations across a range of synthetic speech characteristics. The empirical work was conducted as rigourously and as scientifically as possible and it is unlikely that this type of data could be obtained in any other way. Nevertheless we should be aware that the chosen methodology is vulnerable to a number of possible influences which could result in a distorted picture of the issues under question.

One of the most challenging problems faced by researchers who wish to apply findings from empirical research to real world settings is achieving a satisfactory representativeness in the experimental scenario (Landauer, 1988). When considering the present study, all of the ratings experiments took place with the subjects isolated in featureless laboratories. There was nothing to distract the subjects, no background noise, no variety in the visual field, just a table, chair

and computer (with a simple, repetitive display) or tape recorder and ratings sheets (for experiment 4). In all cases, apart from experiment 1 being a rather tedious experience due to its extreme length, the task was also non-demanding, passive and repetitive, requiring nothing more of the subjects other than for them to listen to a wide variety of speech samples (with irrelevant semantic content) and then to make relatively simple judgments about them. There were no specific skills to develop or learning involved, no real challenge or objectives to achieve. This is clearly qualitatively different from virtually *any* 'real-world' application of the technology where processing the speech for meaning would almost always be required, whatever the implementation.

Consequently it is necessary to consider *how* different is a passive rating scenario in a sterile environment from the reality of day to day interaction with speech synthesis devices across a wide variety of specific goal-orientated tasks in real world settings? Here, a number of points need to be examined and solutions to potential design weaknesses identified.

Firstly, the task itself. The ratings procedure itself was repetitive and tedious, especially in the first experiment where subjects spent a particularly long time in the laboratory. It could be argued that this type of experience provides a reasonably close approximation of the regular and repetitive exposure that an individual would have in certain interaction situations, for example at a synthetic speech installation for a supermarket check-out, or on an industrial production line. Perhaps, in these examples, processing the information for meaning - on top of the long term exposure - would influence their perceptions of speech quality in ways that may be difficult to predict.

For various task scenarios the length of time the user is exposed to a particular voice coupled with the requirements of task-specific goals

may result in substantial variations in levels of tolerance to the speech. In some situations, users may be exposed to the synthetic voice for a very brief period, possibly on a single occasion only, for example when using a tourist information service, or they may have occasional exposure to a voice, such as in the case of a customer enquiry telephone facility to obtain information concerning their bank account. In such cases, tediousness may not be a factor and issues of clarity and intelligibility are likely to be more pertinent. A user trying to copy down a telephone number or account details over the telephone is likely to have considerably less concern for the aesthetic characteristics of the voice than that of a visually impaired user who is using a synthesiser to read his or her daily newspaper or electronic novel. In the latter case, a pleasant, listenable, human-style voice may be preferable. In other circumstances, some users may have little use for a convincingly human voice, for example, pilots who may require synthetic status messages that are clearly distinguishable from cockpit radio communications.

If speech is used for uncommon alarms, such as in hazardous industrial settings where workers might rely on synthetic speech to deliver urgent warnings requiring immediate action, the purposes of the implementation and the motivations of the listeners could be different again. Here, a *deliberately* irritating, low-listenability voice may actually enhance the effectiveness of the system.

The tasks that have been described vary in terms of the function of the information elicited from the synthesizer, however, in spite of the diversity between the various tasks, one thing they all have in common is that the speech has to be actively processed for meaning. What users do with the information may vary, but in all cases they must be able to understand the meaning of the message in order to respond appropriately.

To summarise, the important distinction then between the laboratory experience of speech synthesis in passive rating experiments and

the 'real-world' use of such speech is that in the latter situations, in all cases where non-redundant speech is present within a task context, the user is actually utilising the information in the spoken message to achieve a specific goal. Whilst it could be argued that making a ratings decision about a particular voice *is* a specific goal, the semantic information in the message is not actually necessary for the completion of the task. This stage of the study aims to rectify any potential weakness of conclusions drawn from passive ratings data by introducing a dual-level cognitive load task into the experimental procedure.

7.2 Working memory involvement in speech processing

Baddeley and Hitch's (1974) model of working memory has been widely accepted by the academic community and has stimulated intense research activity. Now, 25 years later, the model (with some refinements) is still considered a valid conception of short-term memory processing by many researchers. Working memory consists of a number of components. Put simply, the basic structure of the model includes, firstly, the central executive which is an attentional system that performs a number of functions including the allocation of processing resources to the other components. The visuo-spatial scratchpad is concerned with visual short term memory. Finally the articulatory or phonological loop is concerned with short-term memory for spoken information. The phonological loop was investigated by Baddeley and his associates using word-span experiments.

They discovered that subjects' ability to reproduce a sequence of words was better for short words than with long words. This suggested that the capacity of working memory is limited and that as the quantity and complexity of spoken information increases, so does the load on the listeners' information processing resources. As well as longer words in a sentence, longer sentences have a similar

effect. Baddeley and Gathercole (1993) state that for long sentences, it is likely that the listener has to maintain in phonological memory the full word sequence in order to interpret the full sentence form, so the processing of longer sentences directly taxes the limited capacity of working memory.

A variety of experimental procedures can be adopted in order to study the role of the phonological loop in speech processing. These include comprehension load, articulatory suppression and irrelevant speech (these techniques will be discussed later). Detailed discussion of the extensive research into working memory and the phonological loop is outside the focus of this study. So, for the purposes of the current research, two basic points will be taken as having been clearly established in the literature. Firstly, that human beings have limited processing resources for spoken information, and secondly, that the load on those resources increases as both semantic and syntactic complexity of spoken information increases.

7.3 Synthetic speech processing demands

As discussed in chapter 2, the processing of synthetic speech is particularly demanding, especially when compared with the processing of natural human speech. The study cited to illustrate this was conducted by Luce et al in 1983. More recently in 1995 Paris, Gilson, Thomas and Silver conducted a comparative investigation of the comprehension of synthetic and natural speech. Results showed that whilst the comprehension of highly intelligible synthetic speech (DECtalk) was equal to that of a natural human speaker, when subjects had to shadow passages of prose in both conditions, accuracy was significantly better for natural speech than for synthetic. The comparative difficulty subjects had in actively processing the synthetic speech again illustrates that there is something particularly taxing or demanding about the cognitive processing of machine generated speech, and, as stated in chapter 2, this extra

effort is only apparent when the person is using the speech to undertake a task which draws on their cognitive resources.

Processing synthetic speech then can be highly taxing and places considerable demands on the listener. Such demands are likely to increase as messages get longer and more complex. It is therefore essential to determine if the pressure and demands on the listener who is processing the speech for meaning are influencing their perceptions of the aesthetics of the speech. If preferences for synthetic voices change as the demands of the task change, listenability enhancement may not be as straight forward as the previous experiments have demonstrated. It may indeed be task specific, which would mean that an extensive analysis of the cognitive demands of particular tasks would be required in order to determine an ideal voice for a given implementation.

To summarise, the main aim of this experiment is to determine whether the introduction of cognitive load (of two distinct levels) influences listenability perception and changes the pattern of related voice effects required to optimise speech.

7.4 Design

In order to construct an appropriate set of stimuli sentences it was necessary to conduct a pilot study. As the aim of the next experiment was to determine the effects of cognitive load on perceptions of speech quality it was necessary to generate statements, the comprehension of which requires high and low amounts of cognitive resources to process accurately.

Due to the logistics of the planned experiment (the mathematical interactions of variables), an even number of stimuli statements were needed. Consequently, and in order to slightly relieve the

demands on the subjects (from what would be a far more taxing experiment to participate in than previously), 150 stimuli statements were needed. These would take the form of true or false statements, 75 being simple and easy to answer, the other 75 to be clearly much more difficult to process, understand and answer correctly. Each of the 25 voice manipulations would be rated on the six strongest predictors of listenability. Scale seven, the weakest predictor according to the initial analysis, was eliminated to balance the relationships between the number of voice manipulations, high/low processing demands and number of scales, to make the experiment. possible.

Of the six presentations of each modified voice, three would be low demand and three would be high demand. This would reveal the relationships between voice manipulations, ratings of quality and the cognitive effort required to process the information accurately. Furthermore, any variations between quality ratings between accurately and inaccurately processed stimuli (as measured by error rates) would be revealed. Overall the experiment is intended to provide a highly detailed picture of the perception of variations of synthetic speech quality whilst undertaking a task with two clear levels of cognitive load.

7.5 Pilot study for Experiment five

In order to establish two distinct difficulty levels for the stimuli statements, a pilot study was conducted. Firstly, 160 true/false statements were generated (the extra ten were included to allow room for ambiguous ones to be excluded). Of these, 80 were highly simple, for example "circles are round" or "three is less than four". The other eighty were longer and not so straightforward, for example "the age of people has no direct association with time", or

"five is more than two but not actually a greater number than three" (the full list is presented in appendix eleven).

Ten subjects were given lists of the 160 statements. The eighty pairs were divided equally so that subjects would evaluate an equal number of pairs where both statements were true, both false, or one true and the other false. Each pair consisted of statements that were similar in context (for example, mathematical) or using the same or similar nouns and verbs in either a simple or longer and more complex formation.

Subjects were instructed to read the pairs of statements and then asked to identify which of the statements in the pair was the easiest to understand. After the subjects had evaluated the lists, their ratings were compiled and the 10 most ambiguous pairs were eliminated, leaving the 150 statements required for the experiment. The final list consisted of pairs of statements where at least 8 out of 10 subjects were in agreement over the easy/difficult distinction. This resulted in two distinct difficulty levels which had been independently verified by multiple assessors.

7.6 Stimuli

Twenty five variations on the DECtalk default male voice were used in order to investigate a wide range of potential interactions. Five settings of smoothness, each compared with five settings of richness in a fully crossed design:

Richness:	0%	25%	50%	75%	100%
Smoothness:	0%	25%	50%	75%	100%

This meant that every possible combination of the settings was included and the results should reveal any interactions that are occuring. The voice set was then varied enough to examine both the relationships of the factors to listenabilty and any potential interactions associated with high and low processing load.

With two levels of processing load included in the design as an additional variable an equal number of stimuli presentations was required. In order to achieve this, the six most powerful listenability predictor scales were used. The seventh scale, which had the weakest correlation with listenability, was not used this time.

The statements were randomised within the constraints of the design. Each voice modification was included 6 times, with a different statement to be evaluated coupled with each of the 6 listenability predictors. Of the 6 sentences presented in each modified voice, 3 were taken from the high processing load or "difficult" set and 3 from the low or "easy" set.

7.7 Apparatus

DECtalk voices were recorded onto high-quality audio cassette using a JVC KD-D30 cassette deck. The machine was also used to present the speech during the experiment over Sony headphones.

7.8 Subjects

Twenty subjects (ten male and ten female) whose first language was English were used in the experiment, ages ranged between 17 and 35 years. None of them had any known hearing impairment. None had taken part in any previous experiments or were unusually

familiar with synthetic speech. They had no prior knowledge of the aims of this particular experiment. Although some of the subjects had heard examples of synthetic voices before, none of them were especially familiar with DECtalk or any other speech synthesis system.

7.9 Method

The practical procedure for gathering the ratings data followed the same format as experiment 4 to ensure consistency within the experimental program (see section 6.8). However, the major difference this time though concerned the introduction of the two levels of cognitive loading incorporated into the task. Subjects were given the ratings score sheets which had now been adapted to allow a true/false decision to be recorded for each statement alongside each of the listenability scales.

Subjects were instructed to listen carefully to each sentence on the tape and firstly to indicate on the score sheet whether or not the sentence would generally be considered to be true or false. They were told not to scrutinise the meaning of each statement for any potential exceptions to the common rule and to deal with the material in a straightforward way. For example, the statement "all houses have doors" is commonly true and subjects were instructed they should rate such a statement as being true, even though they may be able to come up with a rare exception (eg. a tree-house). Having made this rating they were then instructed to rate the voice for listenability on the appropriate scale. Once subjects were confident with the procedure they were played a short introductory sample of synthetic speech and then left alone to complete the experiment.

Despite the increased demands on the subjects from having to rate

the two levels of comprehension load, all of the subjects managed the task at an even pace. Afterwards subjects were thanked and debriefed. When asked, some subjects stated that they had found the task required a considerable degree of concentration, but all had managed to follow the prescribed procedure. Results were then unscrambled, scored, tabulated and analysed.

7.10 Results and discussion

Before a thorough analysis of the results could be undertaken, it was necessary to investigate the possibility that the ratings may have been influenced by an additional variable. The subjects were making ratings concerning the aesthetically pleasing characteristics of the various speech samples and, providing that they could interpret the sentences correctly, the fact that they were actively processing for meaning would be unlikely to have had a confounding influence on their ratings. However, there were (relatively few) cases where the subjects made errors. Such errors are likely to have been caused by a small number of particularly ambiguous statements rather than through any particular combination of parameter settings, indeed, patterns amongst the errors in the raw data support this assumption.

Irrespective of the cause of the errors, the fact that subjects were making some mistakes during the task introduces the possibility that their failure to deal with the information accurately may have had a negative influence on their ratings of the voice. Their assessment of the voice could have been coloured by their irritation at being unable to respond decisively and accurately. If ratings for errors were significantly lower than for accurate responses, then this could introduce ambiguity into conclusions drawn about the results and analysis. It would be impossible to determine whether a negative

rating was the result of a combination of unfavourable parameters or due to irritation caused by mistakes being made (or even some interaction between the two).

In order to eliminate this potential confound, the mean listenability ratings for errors and non-errors were calculated for each subject and a paired-samples within subjects t-test was conducted. The data shows that far from causing more negative ratings, the overall mean score when errors were made was actually slightly higher than for accurate responses (for clarity, the results table is on the next page).

Subject	Mean (error)	Mean (accurate)
1	2.7	2.7
2	2.9	2.7
3	2.7	1.9
4	2.6	2.8
5	3	3
6	3.8	3.5
7	2.7	2.6
8	2.3	2.7
9	3.3	3.2
10	1.8	2.4
11	2.9	3.4
12	3	2.9
13	2.9	2.1
14	3	2.9
15	2.8	2.7
16	2.3	2.5
17	2.4	2.1
18	2.7	2.4
19	2.9	2.6
20	3.5	3
Total Mean	2.81	2.7

A within subjects t-test revealed no significant differences between listenability ratings when errors were made and ratings where responses were accurate ($t(19) = 1.267, p<0.22$).

The lack of impact of errors on listenability ratings means that the analysis can be continued with confidence. It is possible that, in a number of cases, the subjects were not even aware of the fact they were making errors. This would be in cases where the *majority* of subjects made an error on a specific sentence, this implies that they were not making mistakes due to perceptual ambiguity but rather to semantic ambiguity.

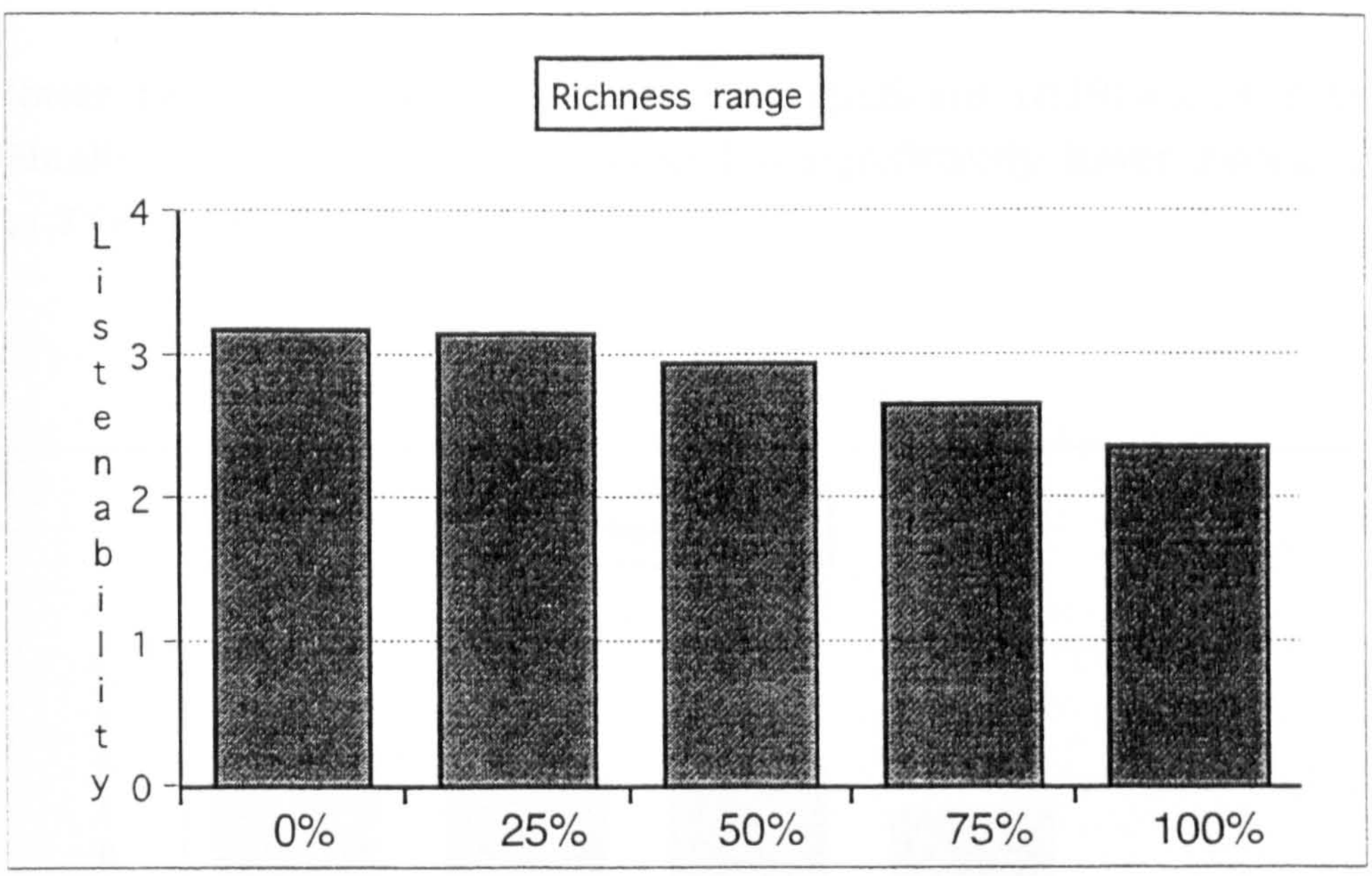


Fig 7.1: Overall richness scores with both cognitive loadings combined (on each setting, the smoothness modifications are collapsed across the full range)

Parameter	Mean	Standard Deviation
Richness 0%	3.19	0.45
Richness 25%	3.17	0.38
Richness 50%	2.95	0.65
Richness 75%	2.66	0.26
Richness 100%	2.38	0.38

An analysis of variance was applied to the data using SPSS. This revealed a highly significant effect for richness modification:

$F(4,76) = 35.36,\ p<0.001$

Planned comparisons were conducted between the different levels of richness (with ratings in both high and low cognitive load conditions combined) collapsed across smoothness. The two-tailed analysis revealed no significant difference between 0% and 25%. The 50% level produced a lower rating than at 25% which was marginally significant ($t(19) = 1.92, p<0.069$). Richness set at 75% produced a

lower rating than at 50% which was significant ($t(19) = 2.14$, $p<0.04$). Finally, richness at 100% produced a significantly lower rating than at 75% ($t(19) = 4.13$, $p<0.01$).

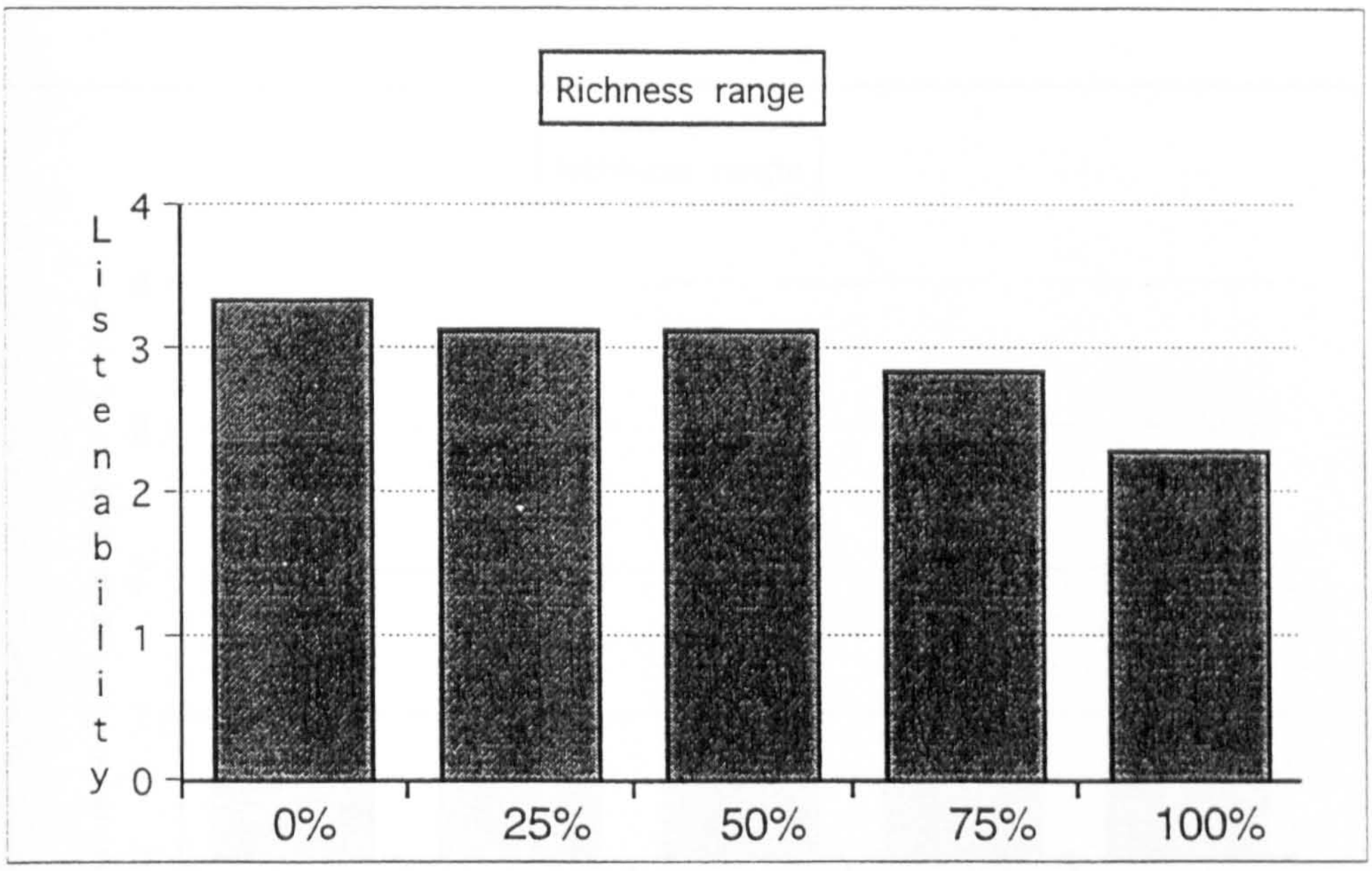


Fig 7.2: Richness scores for low cognitive load ratings (on each setting, the smoothness modifications are collapsed across the full range)

Parameter	Mean	Standard Deviation
Richness 0%	3.33	0.43
Richness 25%	3.13	0.33
Richness 50%	3.14	1.05
Richness 75%	2.85	0.33
Richness 100%	2.3	0.38

Planned comparisons were conducted between the different levels of richness in the low cognitive load condition collapsed across

smoothness. The two-tailed analysis revealed richness set at 25% was rated significant lower than set at 0% ($t(19) = 3.04$, $p<0.007$). There was no significant difference between 50% and 25% or between 75% and 50%. However richness set at 100% was rated significantly lower that at 75 ($t(19) = 6.92$, $p<0.001$).

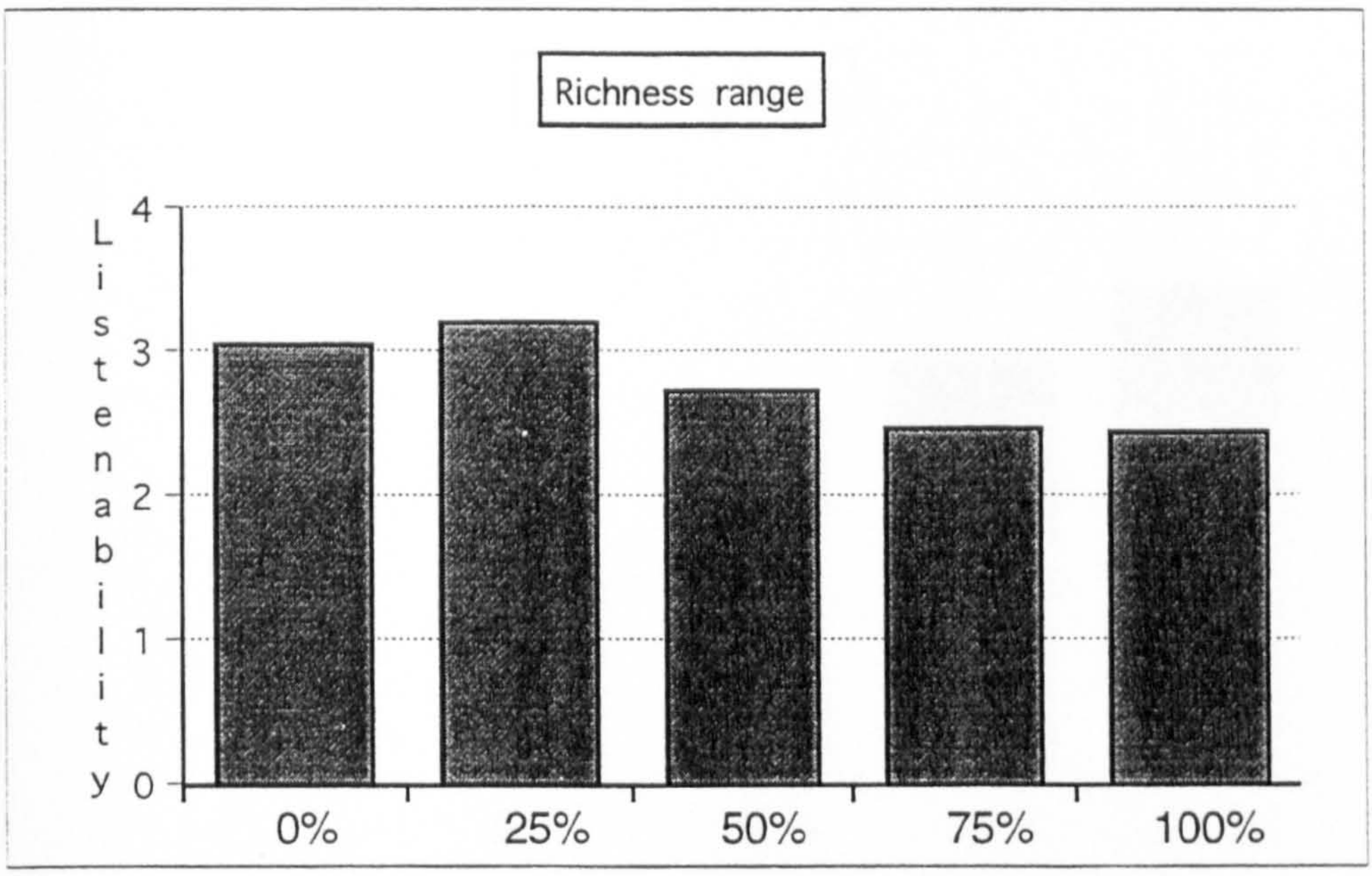


Fig 7.3: Richness scores for high cognitive load ratings (on each setting, the smoothness modifications are collapsed across the full range)

Parameter		Mean	Standard Deviation
Richness	0%	3.04	0.55
Richness	25%	3.21	0.51
Richness	50%	2.75	0.48
Richness	75%	2.48	0.33
Richness	100%	2.45	0.45

Planned comparisons were conducted between the different levels of richness in the high cognitive load condition collapsed across

smoothness. The two-tailed analysis revealed richness set at 25% was not rated significant lower than set at 0%. Richness set at 50% was rated significantly lower than 25% ($t(19) = 6.36$, $p<0.001$). There was no significant difference between 75% and 50%, nor was there a significant difference between 100% and 75%.

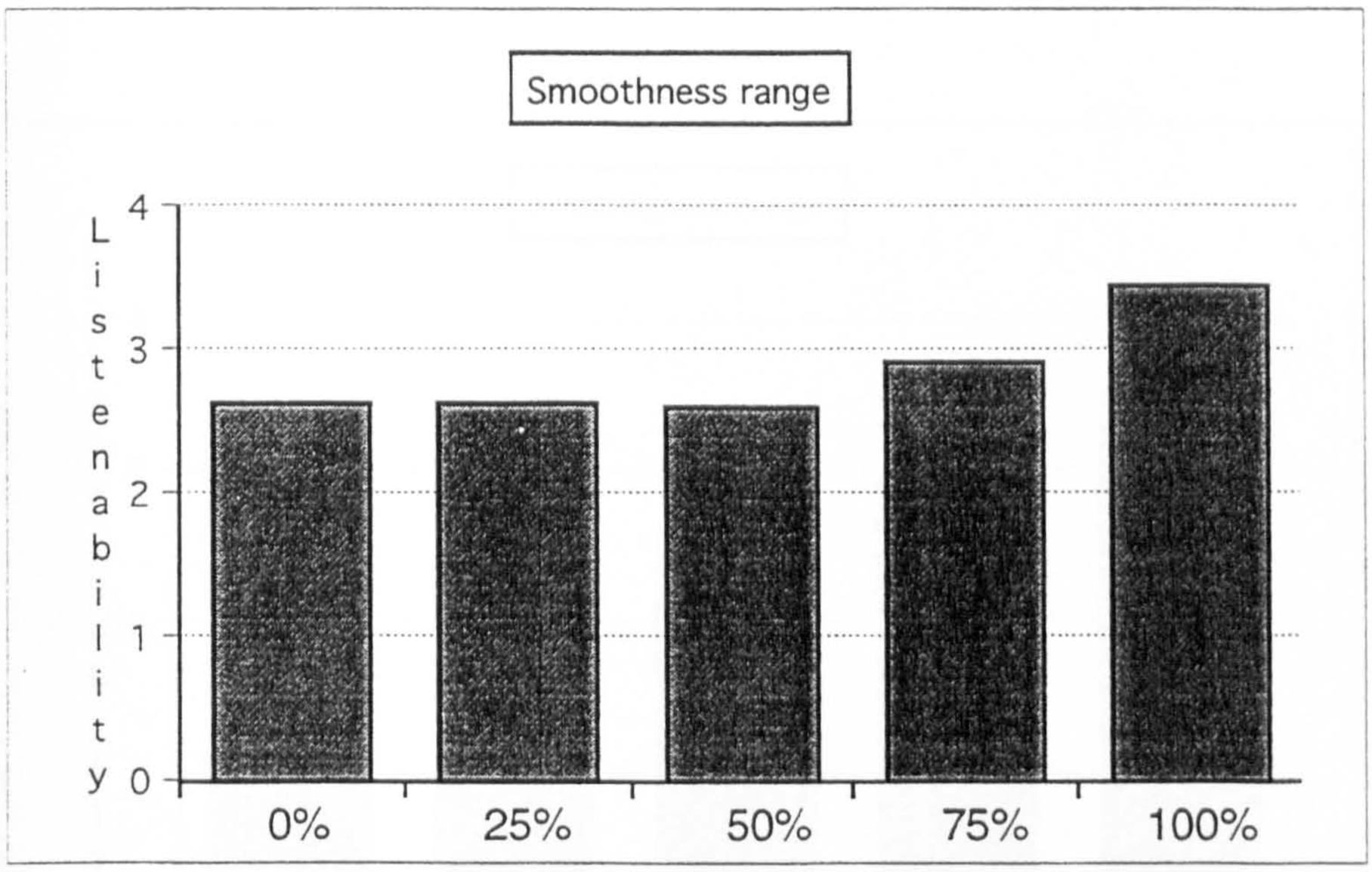


Fig 7.4: Overall smoothness scores with both cognitive loadings combined (on each setting, the richness modifications are collapsed across the full range)

Parameter	Mean	Standard Deviation
Smoothness 0%	2.63	0.33
Smoothness 25%	2.63	0.38
Smoothness 50%	2.60	0.49
Smoothness 75%	2.91	0.31
Smoothness 100%	3.45	0.44

An analysis of variance was applied to the data using SPSS.
This revealed a highly significant effect for smoothness modification: $F(4,76) = 28.78$, $p<0.001$

Planned comparisons were conducted between the different levels of smoothness (with ratings in both high and low cognitive load conditions combined) collapsed across richness. The two-tailed analysis revealed no significant difference between 0% and 25% or between 50% and 25%. Smoothness set at 75% was rated significantly higher than at 50% ($t(19) = 4.08$, $p<0.001$). Smoothness set at 100% was rated significantly higher than at 75% ($t(19) = 5.46$, $p<0.001$).

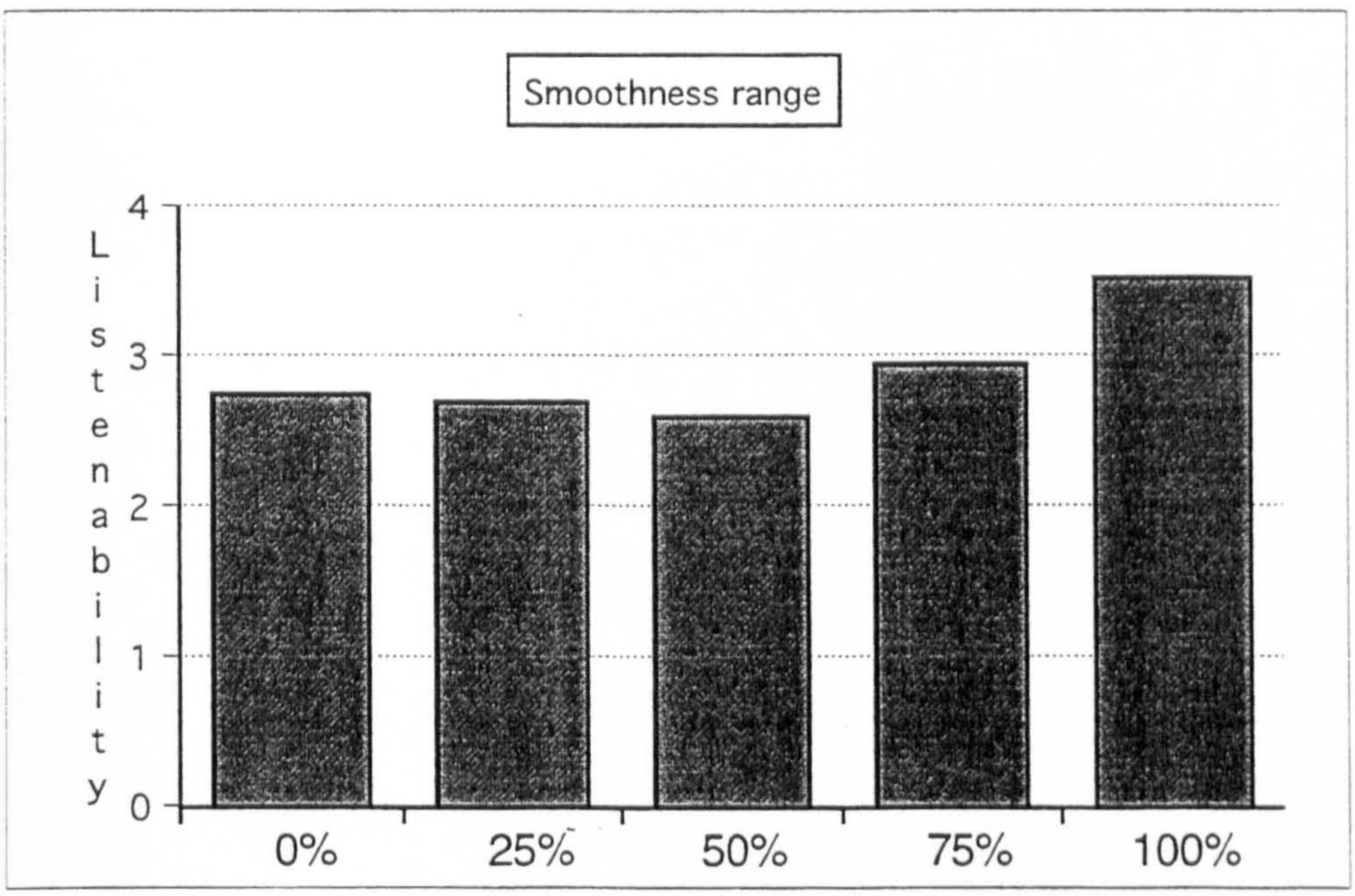


Fig 7.5: Smoothness scores for low cognitive load ratings (on each setting, the richness modifications are collapsed across the full range)

Parameter	Mean	Standard Deviation
Smoothness 0%	2.74	0.34
Smoothness 25%	2.71	0.44
Smoothness 50%	2.61	0.55
Smoothness 75%	2.96	0.35
Smoothness 100%	3.53	0.45

Planned comparisons were conducted between the different levels of smoothness in the low cognitive load condition collapsed across richness. The two-tailed analysis revealed no significant difference between 0% and 25% or between 50% and 25%. Smoothness set at 75% was rated significantly higher than at 50% ($t(19) = 3.1$, $p<0.006$). Smoothness set at 100% was rated significantly higher than at 75% ($t(19) = 6.74$, $p<0.001$).

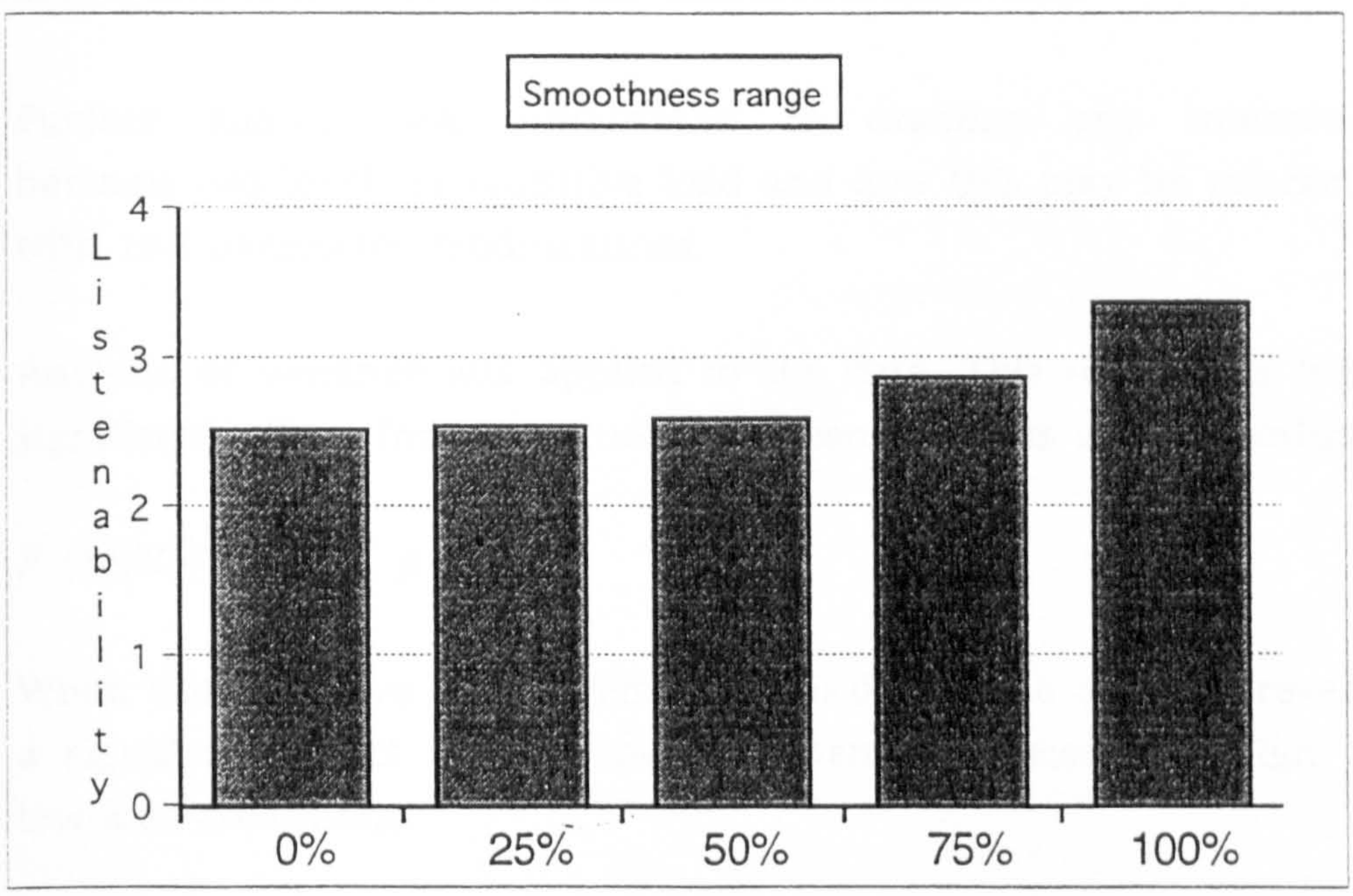


Fig 7.6: Smoothness scores for high cognitive load ratings (on each setting, the richness modifications are collapsed across the full range)

Parameter	Mean	Standard Deviation
Smoothness 0%	2.53	0.44
Smoothness 25%	2.56	0.42
Smoothness 50%	2.60	0.52
Smoothness 75%	2.87	0.39
Smoothness 100%	3.38	0.51

Planned comparisons were conducted between the different levels of smoothness in the high cognitive load condition collapsed across richness. The two-tailed analysis revealed no significant difference between 0% and 25% or between 50% and 25%. Smoothness set at 75% was rated significantly higher than at 50% ($t(19) = 3.62$, $p<0.002$). Smoothness set at 100% was rated significantly higher than at 75% ($t(19) = 4.03$, $p<0.001$).

7.11 Parameter and cognitive load interactions

Further analysis was undertaken to examine any interactions between two levels of cognitive load and how this may be interacting with the parameter modifications.

Analysis of variance was applied to the data. This revealed a highly significant effect for interactions between richness and smoothness:

$F\ (16,304) = 4.571,\ p<0.001$

When the cognitive load levels are considered, the analysis revealed a significant effect between overall listenability scores in high and low load conditions:

$F\ (1,19) = 5.648,\ p<0.05$

When interactions between difficulty level and the richness parameter were considered, the analysis revealed a highly significant effect:

$F\ (4,76) = 8.43,\ p<0.001$

However, when interactions between difficult level and smoothness are considered, no significant effects were observed.

$F\ (4,76) = 0.834,\ p<0.508$

Finally, when interactions between difficulty level, richness and smoothness were analysed, highly significant effects were observed:

$F(16,304) = 7.55,\ p<0.001$

7.12 Graphical depictions of interactions

With five levels of smoothness and five of richness in the voice set and two levels of cognitive load, achieving a clear graphical depiction of all of the interactions on a single chart is not possible. Putting all the data points on to one graph resulted in a cluttered and confusing picture that was difficult to interpret. In order to address this, a number of different tables and charts are included which illustrate some of the complex patterns of interactions that have emerged from the data.

N.B. The graph plotting resources available place the data points vertically above the number between the ticks on the X axis, rather than above the ticks themselves. A ruler placed vertically on the paper can facilitate viewing and comparing the graphs.

	Loading	Smooth 0%	Smooth 25%	Smooth 50%	Smooth 75%	Smooth 100%	Mean	Combined load
Rich 0%	Low	2.99	3.06	3.56	3.41	3.64	3.33	Means :
	High	2.92	2.55	2.86	3.11	3.75	3.03	Rich 0% 3.18
Rich 25%	Low	3.1	2.63	2.7	3.15	4.05	3.12	
	High	2.59	2.86	3	3.87	3.73	3.21	Rich 25% 3.16
Rich 50%	Low	2.62	2.92	2.37	3.18	3.58	2.93	
	High	2.69	2.79	2.64	2.3	3.32	2.74	Rich 50% 2.83
Rich 75%	Low	2.97	2.71	2.21	2.82	3.53	2.84	
	High	2.32	2.27	2.4	2.39	3	2.47	Rich 75% 2.65
Rich 100%	Low	2.01	2.2	2.19	2.23	2.83	2.31	
	High	2.1	2.33	2.07	2.66	3.08	2.44	Rich 100% 2.37
Smoo low	Mean	2.73	2.7	2.6	2.96	3.53		
Smoo high	Mean	2.53	2.56	2.59	2.86	3.37		
Smoo comb	Mean	2.63	2.63	2.6	2.91	3.45		

Figure 7.7: Overall data table showing average scores for all fifty parameter combinations as well as mean scores for the five levels of richness and smoothness at low, high, and combined (total) cognitive load.

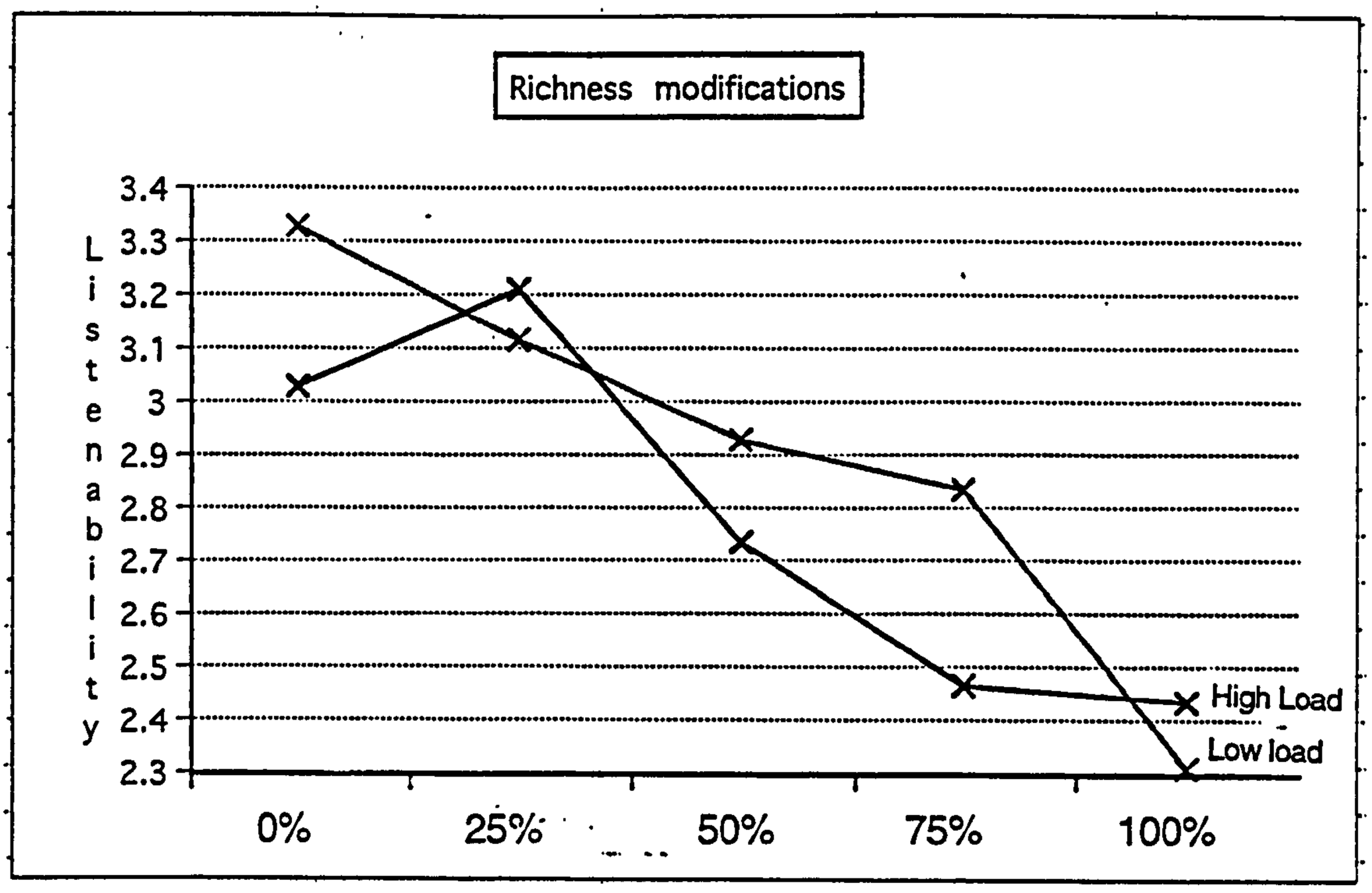


Figure 7.8: Listenability scores in both high and low cognitive load conditions for richness modifications (for each setting, smoothness is collapsed across the full range).

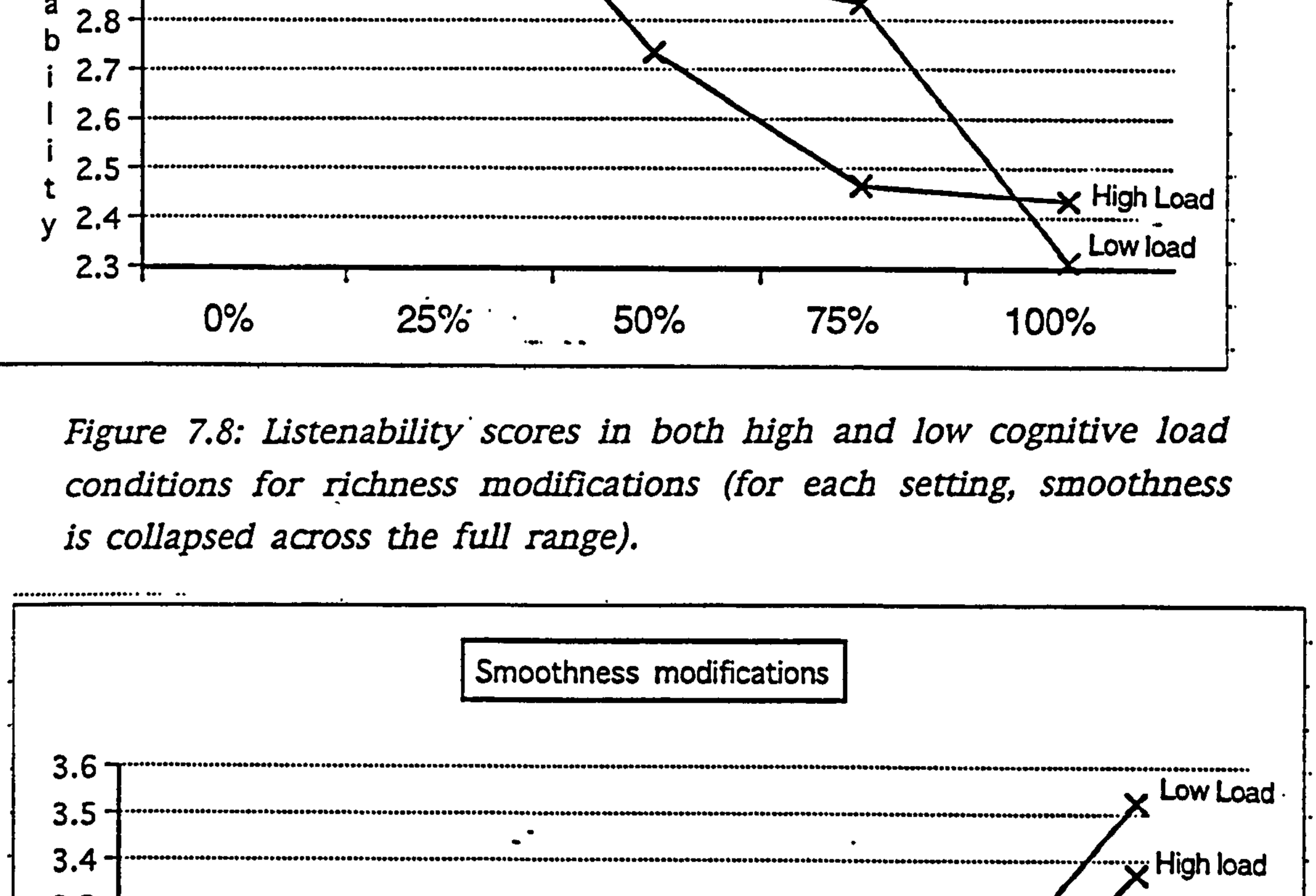


Figure 7.9: Listenability scores in both high and low cognitive load for smoothness modifications (for each setting, Richness is collapsed across the full range).

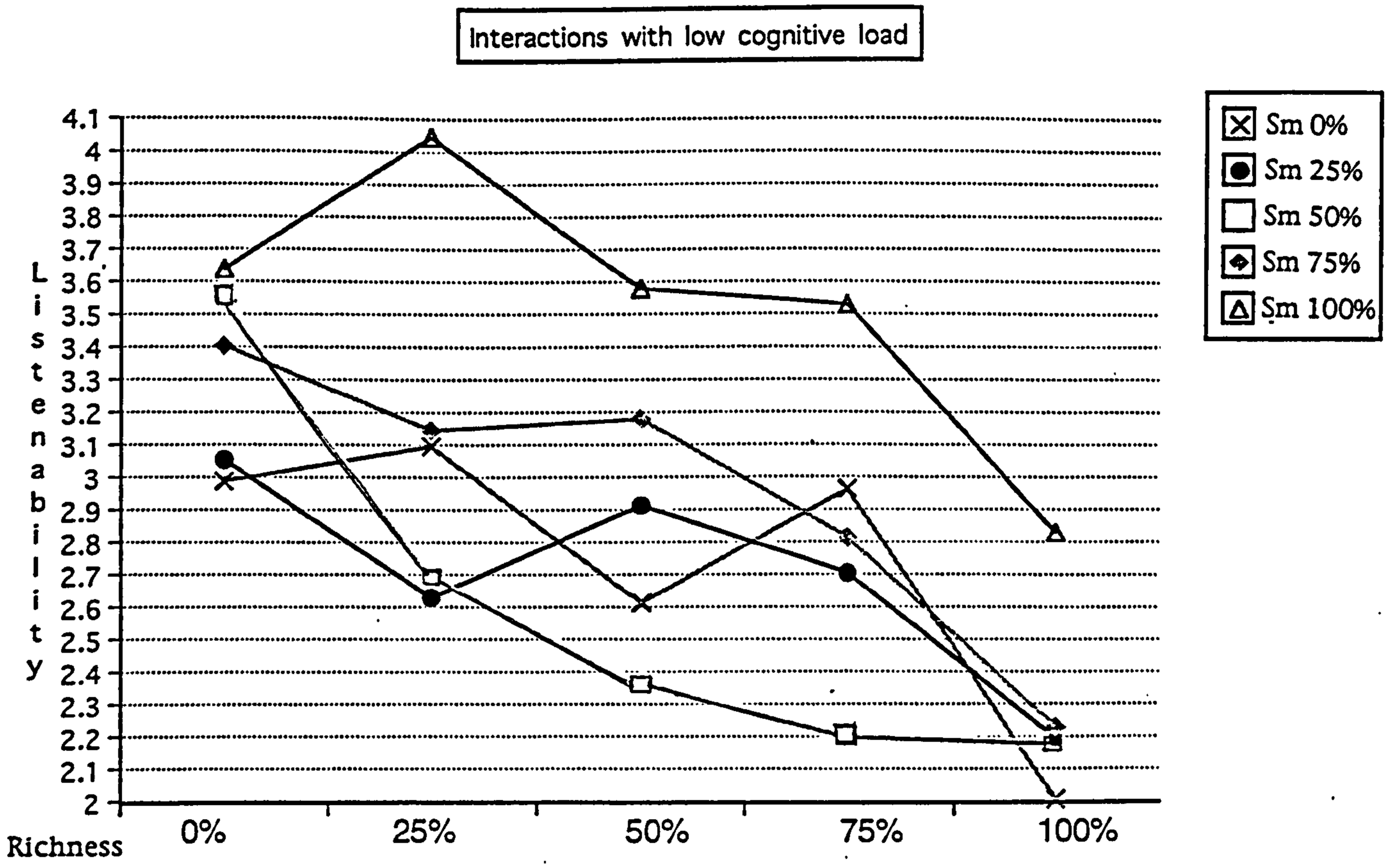


Figure 7.10: Listenabilty scores for parameter interactions between smoothness and richness in the low cognitive load condition.

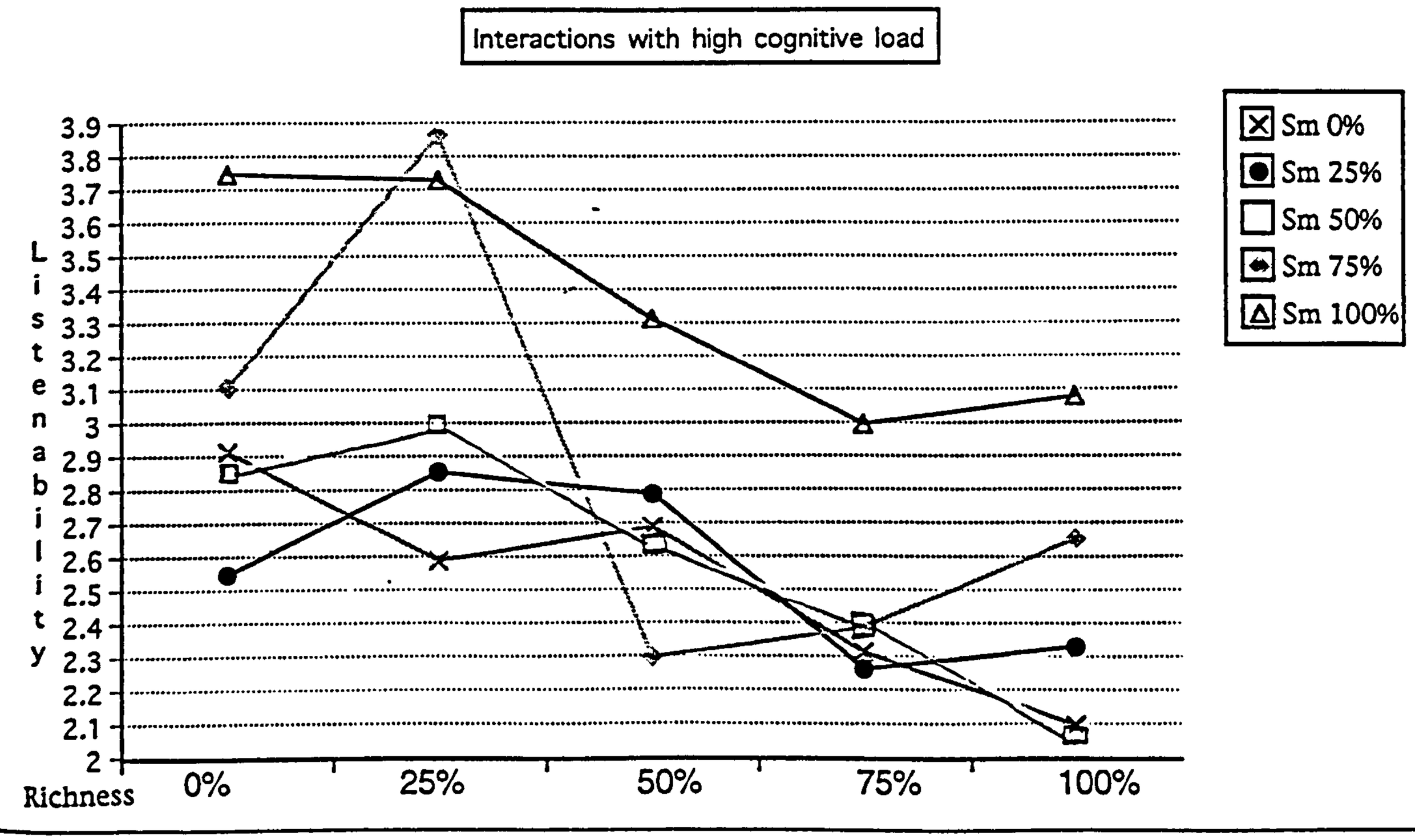


Figure 7.11: Listenability scores for parameter interactions between smoothness and richness in the high cognitive load condition.

7.13 Discussion of results and contributions to knowledge

This experiment generated a considerable amount of data and the results from the analysis have revealed a number of patterns and interactions which can be intepreted in various ways. In terms of a contribution to theoretical knowledge, one of the most interesting findings concerns the relationship between cognitive load and listenability. Here, the results show that subjects perceived the synthetic voices in the high cognitive load condition to be significantly less listenable than those in the low cognitive load condition. As cognitive load increases during synthetic speech processing, perceived listenability decreases significantly. Furthermore the analysis reveals that this is independent of the number of errors made. This result may be of general value because conclusions drawn from this finding are independent of any particular parameter modifications that may be exclusively available to DECtalk users.

As far as DECtalk is concerned, when interactions between difficulty level and the richness parameter were considered, the analysis revealed a highly significant effect:

$F(4,76) = 8.43,\ p<0.001$

Richness has been shown to be the major contributor to listenability effects and the results here concerning the optimal settings demonstrate that the DECtalk synthesiser would need to be set up in different ways to produce the best results for tasks of varying cognitive complexity. While the method of producing the optimal settings and the settings themselves may perhaps be different if another synthesiser was being used, *these results clearly indicate that the optimal settings for listenable speech are likely to vary between conditions of high and low cognitive load.* This has important implications for the development of optimal speech

systems in different application domains. It demonstrates that it is not enough just to develop a voice that has been modified to produce optimal perceived listenabilty ratings, this has to be done within the context of the specific task that the speech will be used for. While the processing of synthetic speech has been shown to require greater cognitive resources than for natural speech (see chapter 2), the cognitive load associated with the specific task that the speech is to be used for (and how modifications interact with them) has not yet been systematically researched. Here, the current research makes a significant contribution.

Moving onto the results that are more specifically concerned with DECtalk parameter modifications, for low level load, once again the results show once again a similar pattern that has emerged before with listenability ratings of richness and smoothness parameter modifications. The optimal settings remain 25% richess paired with 100% smoothness. This was shown in experiment four and, in addition to providing a futher illustration of a reliable effect, it also makes theoretical sense. Passively rating voices on the listenability scales can be considered to be a task requiring relatively low cognitive resources, therefore the pattern emerging again in the low cognitive load condition here might have been anticipated.

In the high cognitive load condition, a different picture emerges. The effect of parameter modifications and interactions on listenability is different. Richness still appears to have the greatest effect, with an overall significant reduction of listenability as richness is increased. Smoothness however does not provide significant enhancements. It may be that the comparatively subtle effects of smoothness modifications are lost because, whilst concentrating in completing the taxing high-load task, the detection of the minor changes caused by smoothness modifications is beyond the capabilties of the subjects perceptual abilities. Richness

modifications, having a much more pronounced influence on listenability, remain detectable, even under the strain of the high-load task. Possibilities for future research in these areas are discussed in chapter 8.

The interactions between richness and smoothness are rather more difficult to interpret. Study of figures 7.10 and 7.11 reveals that for the lower levels of smoothness, although variations in the pattern occur, there are no especially striking differences between ratings in the high and low load conditions. This is most likely because smoothness does not really make much of an impact on listenabilty until it is increased beyond 50%. Figure 7.9 which shows smoothness effects (with richness collapsed) clearly illustrates this.

Once smoothness settings reach higher levels a rather distinctive effect occurs. The combination of smoothness 75% and richness at 25% gives the highest rating, but only in the high cognitive load condition. This is followed by a substantial drop in ratings for 50% richness paired with 75% smoothness

The explanation of this is not clear cut. One possibility might be that in the high cognitive load condition, while richness remains low and the voice is reasonably pleasant to listen to, the additional enhancements of the higher levels of smoothness remain in place. However when richness reaches the point where the voices start to become unpleasant, towards 50% and beyond, then the combination of the irritating high richness voice with the taxing high cognitive load overides and eliminates listenability enhancements from high smoothness settings. This would provide a credible explanation for the sudden drop in listenability effects shown at the smoothness 75% setting as richness reaches 50%. Unfortunately this explanation is somewhat undermined when smoothness at 100% is considered. At this level of smoothness, there is no sudden drop, the pattern in low and high cognitive load conditions remains reasonably consistent.

Another possibility may be related to intelligibility. Intelligibility can be considered to be a component in an overall conception of

speech quality. Indeed it should be noted that for some researchers, this has always been the case. Dilts (1984) describes intelligibility as one of the "main factors of overall speech quality".

High levels of smoothness tend to reduce and/or eliminate some major features in the speech signal, attenuating parts of the signal which may well contribute to overall intelligibility. If this is the case then it could be argued that, in the high load condition, where intelligibility may be more important for task success, when smoothness reaches 75%, the speech becomes less intelligible, making the task harder and annoying the subjects (who rate the speech less favourably as a result). This would explain the radical dip in rating at 75% smoothness in the high load condition. However, if this was the case, smoothness at 100% should degrade intelligibility even more and, as we have already seen, the 100% level shows little difference between low and high load. Consequently, this explanation also seems unlikely.

An alternative explanation could come from an examination of the cognitive complexity of the particular stimuli sentences offered where richness 50% is combined with smoothness 75%. Although the stimuli sentences were objectively classified into two levels of cognitive complexity, within those levels there may be variations in the amount of effort required to process the information and determine the authenticity of the statements. If, during the random allocation of sentences to particular parameters some (or all) of the sentences associated with this combination were especially excruciating or taxing to process, then this might account for the sudden drop in listenability ratings. However, an informal examination of the specific sentences did not bear this out. Statements for this particular combination of levels did not stand out in any way as being uniquely difficult to comprehend. Therefore this seems an unlikely explanation for the observed effect.

To summarise, while the interpretation of the main results is clear and straightforward, the detail of the interaction results cannot be

easily interpreted. Each potential explanation for it has clear weaknesses. It may be that subjecting the data to such a thorough and fine level of scrutiny has amplified or exaggerated what is a relatively insignificant blip in the pattern of the data.

As to why cognitive load affects listenabilty, it may be that the general irritation people experience when listening to synthetic speech is compounded by their struggle to comprehend and process the higher load sentences. Comprehension of synthetic speech has been shown to place a significant load on processing resources, it seems likely that this would make it especially irritating if listeners are engaged in a cognitively taxing task. In general, perhaps people do not like tasks that require high degrees of concentration, possibly the sheer effort of processing synthetic speech messages combined with the effort of processing material with a high comprehension load results in an especially negative evaluation of the quality. This takes us back to Galer (1974) who stated that the annoyance of sound is very much dependent on the information that the sound brings with it. If that information requires considerable processing effort, this may negatively influence the perception of the quality of the voice.

To conclude, the results demonstrate that the listenability of synthetic speech varies significantly under different levels of cognitive load and that this should be taken into account if optimal settings are required for tasks of varying cognitive complexity. This is likely to have generality beyond the DECtalk synthesiser. As far as DECtalk is concerned, although smoothness still has an enhancing effect (especially in low load conditions) the parameter richness appears to be the strongest contributor to listenability. The results suggest that for DECtalk-based systems optimal richness levels vary for tasks which require different levels of cognitive processing. For other synthesizers, different modification techniques are likely to produce optimal settings, but the fact that the experiment clearly indicates that such settings may vary under different levels of cognitive load offers a considerable contribution to research.

Chapter Eight

Major findings and final conclusions

8.1 Major findings and general methodological comments

All laboratory-based studies which ask subjects directly about the perceived quality of speech samples need to address certain methodological considerations. For example, it can be argued that every listener will have different values and preferences and that their opinions may be influenced by many factors. However, provided a sufficient number of listeners are used, then results can be obtained which have generality. Many of the key findings in the present study have been shown clearly to be consistent, reliable and replicable.

The use of alternative objective techniques (such as observation or physiological measures) would be mostly ineffective for gathering data concerning perceptual impressions of synthetic speech quality, as well as being subject to their own limitations and weaknesses. For these reasons, opinion tests form the basis of all modern speech quality assessment methods (Johnston, 1996).

Roe (1995) identifies the diversity of practical problems encountered in the real world which can lead to a perceived gap between laboratory and real-world performance. He suggests that the degree of task difficulty depends on performance of the speech system and the expectations of the user of the system. A task of moderate difficulty in a controlled and focussed laboratory environment may prove to be much more demanding when real-world objectives with real-world consequences have to be achieved and obstacles

overcome. This suggests we should use caution when applying conclusions about task performance derived from laboratory settings to real-world scenarios. We may make significant discoveries using empirical techniques in the laboratory but in order to build a more complete picture we must ultimately involve the end-user while at the same time attempting to maintain high scientific standards.

All of the experiments in this thesis were tightly-controlled laboratory studies which used experimental subjects rather than real-world users of speech systems. However the fact that the effects found were consistent and unambiguous throughout the progression of research demonstrates that these are very real and replicable effects. Furthermore the inclusion of comprehension load in experiment five considerably enhanced the real-world nature of the experimental procedure.

The first major finding came from Experiment one which generated a significant amount of data concerning the perceptual rating of synthetic speech. From a large set of scales, factor analysis revealed a number of factors, the most important were characterised as listenability and assurance. A specific set of scales was associated with each and certain speech parameter manipulations correlated with them. Factor analysis is a controversial technique, output from such a procedure is very much dependent on input. However a number of the voice parameters used in the voice set were those that have been shown in previous literature to be associated with perceptual factors (Rosson and Cecela, 1986). Furthermore, the wide variety of voices (and scales) offered gave listeners an opportunity to experience very many potential modifications/customisations of synthetic speech which they could rate in very many different ways. This was done in order to offer the subjects an unconstrained exposure to synthetic speech so that any perceptual effects observed would not be the product of a restricted voice set.

The experiment was an attempt to quantify perceptual impressions of synthetic speech and develop a synthetic speech assessment tool. Within the constraints of an exploratory factor analytic approach, this was achieved.

Having discussed the potential drawbacks of factor analysis (see section 3.10), Experiment two was designed to consolidate the implications and conclusions that had arisen. Here, the ratings procedure was streamlined and listeners were offered only the voice modifications and scales associated with listenability and assurance. The standard DECtalk male voice was used at maximum and minimum settings of the four parameters associated with the two factors. The results showed significant improvements in listenability when smoothness is increased and richness decreased, and significant improvements in assurance ratings when average pitch was lowered and headsize increased.

Experiment three took the research further by investigating whether or not the modification techniques applied to a markedly different synthetic voice, in this case a female voice, would result in significant enhancements. The results demonstrated that not only did the modifications prove effective with a female voice (although with some variations partly due to DECtalk's technical constraints), but also that the ratings procedure was a reliable, replicable method of assessing the perceptual attributes of synthetic speech.

Experiment four concentrated exclusively on listenability. Speech quality has been shown to be the major, if not main, requirement for the acceptability of synthetic speech systems (see section 3.2). Experiment four was designed to determine the precise effects of richness and smoothness as they relate to listenability. Although enhanced smoothness and decreased richness had been shown to enhance listenability in experiments two and three, it was necessary to determine whether these effects were linear. To address this, five levels of each of the parameters were tested using the listenability

scaling procedure. In each case a clear linear pattern emerged. Analysis revealed significant listenability enhancements and planned comparisons show these to be linear.

Experiment four then demonstrated exactly how the aesthetically pleasing characteristics of synthetic speech can be either enhanced or degraded by specific adjustments of the speech signal, an important finding for synthetic speech research.

Experiment five progressed the research into new territory. While the first four experiments consisted of a progressive analysis of synthetic speech optimization, the ratings tasks used in the experiments were passive procedures, qualitatively different to the 'real-world' use of synthesis where comprehension load would be part of almost any task. In order to assess this, a pilot study identified sentences which had two defined levels of comprehension load. In the main experiment subjects not only had to rate the voice set (modified over five levels for each of the parameters) for listenability, but also had to process the material for meaning over two distinct levels of cognitive load.

The results showed that listenability enhancements caused by parameter modification remained in the data at both levels of comprehension load. However variations in the pattern emerged and a significant difference was revealed between optimal parameter modifications for conditions of high and low cognitive load. This finding is of importance to people working with synthetic speech in general, not just those using DECtalk. Researchers using different systems may have to modify them in different ways in order to optimise the speech, but the fact that they will have to make additional modifications for tasks of varying cognitive load is an important and useful finding.

8.2 Suggestions for future research

Experiment five demonstrated that task complexity influences the design of an optimal set of speech characteristics. Such variation suggests that a valid and challenging future research avenue might involve a systematic quantification of user/task/environmental requirements as they relate to synthetic speech perception. Ideally, voices could be assessed in a wide variety of scenarios whilst listeners undertake a number of different interactive tasks of varying degrees of complexity. These might be both goal-orientated productive tasks and experiential tasks where the activity is undertaken purely for the experience offered (as in speech implementation within systems designed for games and recreation applications). Such complexities suggest on-going research in these areas would present a considerable challenge.

Cognitive load is likely to be a key determinant of system acceptability. There is considerable potential for research here. One potentially fruitful approach would be to examine the working memory span of subjects who then undertake tasks of varying levels of cognitive complexity. It is likely that differences in working memory span may interact with the relationship of task load to optimal parameter modification. Fine tuning the voice to both the cognitive load of the task and the cognitive capacity of the user may well prove to be the ideal procedure for developing highly-useable, customised synthetic speech systems. While the implementation of fully adaptive systems that can achieve this may be currently beyond the reach of commercial applications, this at least offers the potential for some interesting research developments.

Clearly it would be impossible to study examples of all possible tasks that exist or might exist, so a practical approach would be to study several systems where the tasks are very clearly qualitatively different (Landauer, 1988). Criteria other than ratings procedures

might also be used to enhance our overall understanding of synthetic speech processing. These could include task speed and efficiency measures under differing levels of performance demands. Investigation in all of these areas is likely to provide us with a more expansive and comprehensive understanding of the research domain.

A number of factors can influence cognitive load and various techniques can be used to measure and manipulate processing abilities. Individual differences can be determined by measurement of working memory span and cognitive load can be induced by the introduction of irrelevant speech and/or articulatory suppression. One potential experimental procedure might be to increase the cognitive load even further, up to and beyond the subject's cognitive capacities. Experiment five demonstrated that the subtle effects of smoothness enhancements were less pronounced at a high level of cognitive load. It would be interesting to see if the powerful richness effects would also change or eventually disappear as task-related cognitive effort becomes highly intense, or whether they remain up to and beyond the subjects ability to complete the task/s effectively.

As far as introducing greater levels of cognitive load using articulatory suppression is concerned, this is not possible with the ratings task undertaken in the experiments reported in this thesis. While considering how best might cognitive load be introduced into the task for experiment five, five subjects were asked to attempt to undertake the rating of voices while suppressing their articulatory processing resources by repeating the word "the" over and over. All subjects who tried simply did not have the ability to do this, interference was too great and they could not complete the task. As a consequence, the introduction of cognitive load using comprehension load was selected for experiment five, and articulatory suppression is probably not the ideal technique to use.

When considering the diversity of potential task environments, intelligibility is a significant factor and can be affected by background speech. A soundproof laboratory offers optimum listening conditions with no background noise, minimal acoustic reverberation/distortion or any distracting visual stimuli. A real-world interaction situation is certain to be substantially different. Irrelevant background speech has been clearly shown to have a disruptive effect on serial short term memory (Jones et al. 1993), therefore a listener's attempts to process synthetic speech messages in real-world settings are likely to vary in success due to interference from environmental factors which are not normally present in a standard laboratory setting.

If subjects hear an irrelevant stream of speech during presentation of verbal information, recall accuracy is disrupted as there is interference with phonological processing. This has been shown to affect cognitive load but perhaps not to such an extent as to make the ratings task impossible. Irrelevant speech is present in much of our working life and day to day experience. In an open-plan office installation and other public settings the many different types of background noise (including speech) will almost certainly interfere with the perception and understanding of synthetic messages in complex ways, ways which are likely to defy accurate and reliable prediction of intelligibility and ratings of perceptual preferences. The potential for research in this domain is considerable and may lead to important findings for synthetic speech designers.

To summarise, evaluation of speech samples in a controlled setting using a precise, validated and reliable ratings procedure within an exploratory research paradigm has been shown to provide valuable insights into the synthetic speech characteristics which will help us select an appropriate voice for an interface. The introduction of varying levels of cognitive load is an essential stage of the developmental and implementation cycle for synthetic speech systems and provides promising opportunities for future research.

On consideration of future empirical work and the relevance of the results to voice output system development, other avenues for further inquiry/progress are revealed. These will be discussed in relation to speech listenability because this factor was the main focus of the study. Listenability has been clearly shown to be the result of specific manipulations (modifications of low and high frequency ranges in the speech signal). Similar modifications could easily be made to different synthetic voices, and, with a certain amount of ingenuity, to digitised human speech samples. The study clearly demonstrates that it is possible to adapt synthetic voices in order to significantly enhance (or degrade) the strength of their aesthetically pleasing characteristics or ability to convey assurance and confidence. As the modifications needed to do this (at least in the case of listenability) have been clearly identified, it seems reasonable to hypothesise that similar modifications could be applied to human/digitised voices, perhaps by band-pass filtering, and that these might result in comparable perceptual ratings of the speech. It it tempting to make the intuitive leap and suggest that it may be possible to modify *any* voice (not just synthetic, but recorded/digitised human voices as well) and produce similar perceptual effects (providing the appropriate technology was available).

An interesting line of research might be an investigation into whether or not the listenability enhancements identified in the experimental program would result in synthetic speech which is perceived as being as natural and listenable as human speech. Might such optimised speech overcome the general rejection of speech synthesis reported in the literature? If the listenability enhancements are suitably powerful, perhaps modified synthetic speech may be as acceptable as natural human speech?

The major stumbling block to this line of enquiry are the huge problems associated with any kind of objective description and quantification of what, exactly, *is* a listenable human voice, or even a *standard* or *normal* human voice.

Synthetic voices can be precisely quantified and modified. Experimental calibrations can be precisely replicated. Furthermore most modern formant synthesis systems have a consistent style, it is hard to tell them apart. No such generic human voice has been established. It was for precisely these reasons that a synthesiser was used in this study and human speech was avoided (except in terms of the implementation criteria discussed in chapter two).

Consequently, in order to deal constructively with any comparisons and contrasts between human and synthetic voices, a major investigation would need to be undertaken in order to try and construct a picture of exactly what is a 'standard' or 'average' human voice. This would involve the sampling and standardisation of a wide range of different voices and entail complex analysis of the variations and differing characteristics of speech quality. This would not be a simple confirmatory experiment in order to test whether modified synthetic voices are perceived as being as (or more) listenable as human voices, but rather a mammoth undertaking involving extensive quantification of multiple variables.

It can be argued that many of the studies that have been discussed in this thesis and which have used human voices in comparative analysis with synthetic speech are potentially undermined by their failure to attempt to describe/quantify the human voices used to provide the stimuli in the experiments. In many cases, the human voices used in experiments are chosen by opportunity (or if this is not the case, no mention is made of any particular selection criteria being used).

For example, Pisoni (1982) repeatedly talked about the "natural speech" used in a number of experiments but failed to specify which voice was used and why. The same applies in Pisoni et al (1985). Pratt (1987) used the word "human" in a results table and mentioned "six male speakers" as providing speech samples, no further mention was made of how these voices were selected. There is a wide variation in

human speech quality and characteristics and this should at least be addressed if human speech is to be used in such experiments.

Nusbaum and his associates seemed to be aware of this problem as early as 1985, stating "In natural speech there are many acoustic cues that change as a function of context, speaking rate and talker". Later, the researchers conclude "...the differences in perception of natural and synthetic speech are largely the result of differences in the acoustic-phonetic structure of the signals."

A more recent study, (Nusbaum, Francis and Henly, 1995), tackled the problem directly. Here, an attempt was made to precisely control and quantify the human speech samples used in the experiment and to propose a new methodology for accurately comparing the naturalness of human and synthetic speech. Rather than use words or sentences for stimuli, the researchers used short, isolated glottal source samples from a number of synthetic and natural vowels to produce a number of "snapshots of the naturalness of source characteristics". Subjects were required to identify these samples as being either produced by a synthesiser or a human source.

The researchers successfully used this technique to distinguish amongst different levels of naturalness found for different text-to-speech systems and human speech. They demonstrate that the perception of naturalness is affected by information contained within the smallest part of speech and by information contained within the prosodic structure of a syllable. Such an approach illustrates an effective methodology for distinguishing speech quality from speech style and might possibly lead to substantial improvements for synthesis.

One interesting finding from this work can be related to the current study. That is, Nusbaum and his associates (1995) found that although DECtalk was shown to be able to accurately model some of the low frequency source characteristics of human speech, there

"may be a problem in the shape of the glottal pulse as revealed by sensitivity to the higher frequency components". The researchers do not elaborate on this but do stress the need for further research. Since in the current study, variations in high frequency signals were responsible for enhancing perceptions of listenability (through attenuation of the high frequencies via increased smoothness), this suggests that, at least as far as DECtalk is concerned, the high frequency components clearly appear to have a potentially negative influence on perceptual variables.

It may be possible therefore, in a further series of experiments using both voice source comparisons and a perceptual rating methodology to unpack Nusbaum et al's "naturalness" in terms of its contributing factors and arrive at a detailed and precise psychological profile of synthetic and natural speech perception. This would be an extensive inquiry and a considerable challenge.

To summarise, whilst it is possible to compare various synthesisers, providing they each have detailed and clearly specified voice parameter modification facilities, any comparison of these with human speech would be difficult to control. With words and sentences used as stimuli, there is far too much variation between different human speakers to allow us adequate experimental control. One solution to this is Nusbaum et al's (1995) methodology using glottal source samples and this is certainly worthy of further inquiry. Finally, an investigation into modification of human voices using the criteria developed in this study could prove enlightening, providing the modifications could be clearly quantified and the technology allowed precise adaption strategies for digitised speech samples that are analogous to those developed here for synthetic speech.

Having established a clarification and improvement in the human-factors field of implementation criteria in speech system development, and demonstrated how the customisation of specific

speech parameters can result in significant improvements in synthetic voice quality for both male and female voices, the way is now open for research to elaborate on this and to investigate the many potential avenues for future human-factors investigation. As stated earlier, the subjective preferences of actual and potential users are of tantamount importance to the success of future implementation of speech synthesis systems. Consequently, potentially fruitful research initiatives should focus on the requirements, expectations, motivations and preferences of users within the context of task and environment.

The importance of synthetic speech quality and performance is beginning to be recognised, and new technology is being developed with this in mind. Increasingly, sophisticated models of linguistic structure are being developed and computer speech synthesis is indeed reaching very high levels of performance (Liberman, 1995).

8.3 Technological limitations: DECtalk

DECtalk was used to generate the various speech modifications in this study. DECtalk is certainly the most well-known and widely used system. The DECtalk speech style and general quality personifies synthetic speech and remains, for now, arguably state of the art. The machine does however have certain limitations. Certain modifications of the speech parameters often interact and overload DECtalk's circuits, producing an unpleasant 'squawking' sound. On occasion this complicated the development of balanced sets of modified voices, especially in the early experiments. Furthermore the particular combinations of words or settings that cause overloading are impossible to predict. For the same sentence and settings DECtalk will produce good output on one occasion and a dreadful squawking on another. This problem made the development of recordings for the experiments occasionally frustrating and always time-consuming.

A further, and more serious, disadvantage of DECtalk is that the specific effects of the parameters on the speech signal are not clearly described in the DECtalk literature. Names such as richness and smoothness tell little about what these parameters do to the speech signal and where in the synthesis process this takes place. One assumes this is because the information is commercially sensitive, but this does not make research using DECtalk especially easy. A systematic spectrographic study, as discussed in Chapter four (section 4.6), would be required to address this.

Recent research has (tentatively) suggested that DECtalk's standing of many years as being the best quality synthesiser available may finally be in jeopardy. Klaus, Fellbaum and Sotscheck (1997) announced that they have evaluated a synthesis system which "produces better speech quality than others, above all compared with formant synthesis techniques" i.e. DECtalk (interested readers are referred to Klaus et al, 1997 where "Time Domain Pitch Synchronous Overlap and Add" techniques are compared with other methods of synthesis).

Discussion of the technicalities of the synthesis methodologies they are investigating is beyond the scope of this study but suffice to say, such advances suggest that the research field is still evolving and developing. The misconception that speech synthesis research was somehow finished or complete when their intelligibility reached levels comparable with human speech (discussed in section 3.2) appears to have been overcome. Indeed, Cole et al (1995) identify speech synthesis/generation as being one of the key research challenges for on-going investigation into the development of spoken language systems. They state that the research field is indeed a rapidly expanding area and highlight the need for multidisciplinary research, for the development of related resources and for rapid communication among researchers.

The computer speech research field is an amalgam of many

disciplines. Westall, Johnston and Lewis (1996) propose a multidisciplinary model (Fig 8.1) and argue that a broad-based, holistic approach (with the various disciplines interacting and complimenting one another) is increasingly required in order to realise systems which will be acceptable to their users. Such an approach is likely to boost the progress of speech technology considerably.

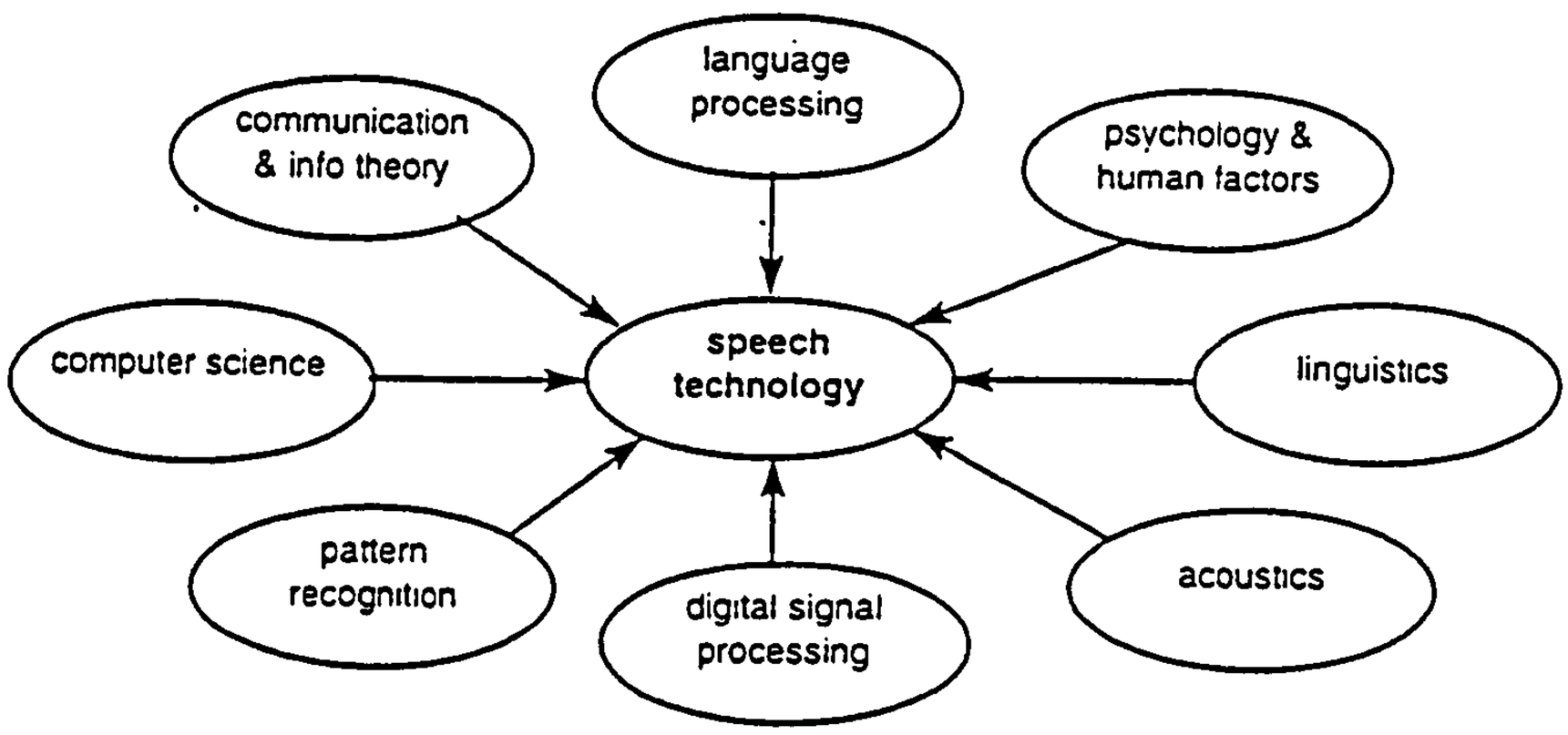


Fig 8.1 Contemporary speech technology, "an amalgam of many difference disciplines." Adapted from Westall, Johnston and Lewis, (1996)

The importance of future research opportunities for speech synthesis is also stressed by Schafer (1995), whilst other researchers suggest that speech technology will be the catalyst that introduces a new era of human computer interaction and that todays commercial applications for speech interface systems represent:

".....a small preview of a rich future for speech technology that will eventually replace keyboards with microphones and loudspeakers to give easy accessibility to increasingly intelligent machines." (Oberteuffer ,1995)

While such glowing optimism should always be viewed with a healthy degree of scepticism, there is no doubt that the research community has not given up on speech output as a valid area of research and may well be poised to exploit the considerable potential for speech systems in human computer interaction across a wide range of application domains. Research into the improvements of the technological aspects of synthesis, coupled with the reliable and precise quantification of the perceptual characteristics of synthetic speech, should considerably enhance the overall quality of applications and ultimately lead to much wider implementation and acceptance of speech systems.

8.4 Ergonomic conclusions

The first stage of this study was concerned with a critical evaluation of design strategies and implementation criteria needed for effective multi-modal system development, identifying and refining ergonomic guidelines and drawing conclusions that might help the technology of synthesis and digitisation to achieve some of its unique and mostly untapped potential. The consistent use of flexible ergonomic guidelines during an iterative and cyclic design process is certain to enhance the likelihood of potential speech output systems being implemented in suitable environments and for appropriate tasks.

Human beings are multi-modal creatures and the powerful and familiar audio channel is one of the primary modalities of human communications. It is therefore likely that systems designers *can* capitalise on speech in human computer interaction, and that the use of flexible and dynamic human-factors based guidelines is likely to be a significant benefit in the achievement of this goal. Adherence to some of the basic recommendations concerning the functionality of speech messaging for varying users, tasks and environments is certainly likely to prevent the designer from making some

fundamental usability errors.

It is of further importance that *evaluative* ergonomic and psychological studies are conducted and the subsequent results published. So often in the literature, new applications are announced enthusiastically yet their eventual success or usefulness is not consistently researched or reported. This observation was made by Pisoni et al back in 1985 and the situation has not yet been clearly improved. Liberman (1995) identifies evaluative strategies as having had a crucial role in the rapid progress made in speech recognition research over the last decade. He states that similar approaches in synthesis research are an especially promising trend. An extension of an ergonomically-focused research strategy would therefore involve the development of evaluative criteria to ensure that guidelines, recommendations and ergonomic implementation strategies can be shown to be of consistent benefit when applied to a wide range of speech implementations.

8.5 Empirical conclusions

The empirical work in the second stage of this thesis was designed to tackle the issue of user irritation and dissatisfaction with the quality of synthetic voices, which seems to arise despite high levels of intelligibility and sophisticated text-to-speech conversion techniques. This was clearly shown by examination of the relevant literature which highlighted that such irritation was usually the major factor in the rejection of synthetic speech devices by potential users.

Examination of the results of the factor analytic study and the later complimentary Experiments 2 and 3 which refined and validated the technique clearly demonstrates that the quality of synthetic speech *can* be substantially improved in terms of both its aesthetically pleasing characteristics and the assurance or confidence of the voice as perceived by listeners.

"If an individual using a particular text-to-speech system cannot tolerate the sound of the speech or does not trust the information provided by the voice output device, the usefulness of this technology will be reduced"
Pisoni, Nusbaum and Greene (1985).

Although the issues of tolerance and trust were not the exclusive targets of inquiry at the onset of the empirical stages, it can be clearly seen that the points made in the above quotation have been directly addressed by the results. Enhancement of listenability will obviously increase a listener's tolerance of a speech system and a voice with enhanced assurance will inspire trust. Consequently, the findings concerning the manipulation of perceived listenability and assurance through specific parameter modification have demonstrated a precise methodology for increasing the usefulness of the technology.

The findings from the experimental results elaborate on previous research in terms of the identification of factors and their relationships with particular voice parameters. Consistency throughout the empirical stages of this study gives us confidence that the ratings methodology is an effective procedure for accurately quantifying these particular factors and investigating speech perception in the laboratory. Modified versions of this procedure could, and should be used to evaluate user preferences in a variety of settings and for a wide range of tasks. Furthermore, the results from the study suggest it may be appropriate to use the ratings software to test *any* voice, synthetic or human/digitised for the strength of its aesthetically pleasing characteristics or ability to convey assurance and confidence.

Experiment four concentrated specifically on an examination of the aesthetically pleasing characteristics of the speech signal, as measured by listenability scaling. This required a refinement in the research methodology, with direct focus now on the production of ·

voices which are pleasing to the listener. The vital importance of this for a wide variety of application domains having clearly been established in the literature. The unambiguous results showed the synthetic speech could be consistently improved by modification of richness and smoothness parameters, and that such modifications improved speech in a linear fashion.

Experiment five expands the focus of the research considerably by addressing the issue of cognitive load and synthetic speech processing and how this might interact with listenability modifications. Cognitive load has already been considered in the first part of this thesis (chapter two where the processing of natural and synthetic speech is compared). The findings from Experiment five extend our knowledge of the relationship between cognitive load and synthetic speech perception. The results clearly show that while the enhancements to the synthetic speech remain when the listener is processing the speech for meaning, the optimal settings for the speech vary between different levels of cognitive load, an important finding for anyone working with synthesis systems, not just those researchers using DECtalk.

8.6 Summary of contributions to the field

The current study has achieved the stated aims and objectives (identified in forewords to sections one and two, pages 9 and 56 respectively) and has made the following contributions to the research field:

1. Section one of this study is a successful clarification and integration of the evolution and development of speech output research over the last three decades. This has highlighted the necessity for ergonomic and human factors research to be included within the research paradigm (which had been traditionally limited by technological orientation).

Chapters one and two both conclude with comprehensive ergonomic guidelines. These are based on the identification and analysis of previous research into the effects of the implementation of speech systems on users, tasks and environments. The guidelines illustrate the functional role of speech output in human-computer interaction in relation to the visual presentation of information as well as the quantification and assessment of implementation strategies for speech output (as determined by the different available techniques and their appropriateness for various functions).

These guidelines, when adhered to during implementation of multi-modal systems, will almost certainly result in more successful and acceptable interactions. The guidelines, although seemingly quite straightforward or obvious at first glance, have been cited occasionally in the literature. For example, Baeker et al (1995), and seven years after their original publication in Stanton and Baber (1997).

2. Section two of this study was an extensive exploratory research program which successfully identified a number of specific acoustic variables that can characterise synthetic voices and which are consistently related to defined perceptual factors.

Section two involved:

a) The development of an efficient and reliable laboratory-based speech ratings environment and an associated assessment methodology for the investigation and quantification of speech perception. This involved the construction of an elegant and usable interface which enabled the modification and presentation of synthetic speech samples and the collection and recording of perceptual ratings. Ease of use and the consistency of results over a series of experiments illustrate the reliability and validity of this ratings system for data collection for this type of study.

b) The identification of salient perceptual factors which are associated with exposure to synthetic voices. The first experiment revealed that firstly, the factors 'listenability' and 'assurance' were indicative of specific perceptual characteristics which were attributed to variations on synthetic voices, and secondly that the ratings methodology could effectively quantify these factors using a defined range of perceptual scales.

(c) The identification of a linear relationship between modifications of specific DECtalk parameters and repeated enhancements of listenability.

(d) A detailed analysis of how such modifications affect listenability ratings at two quantified levels of cognitive load, thus demonstrating how synthetic speech can be optimised for tasks of varying cognitive complexity.

In effect, the main contribution of section two of the study was the development of a reliable and consistent methodology for modifying synthetic speech signals which consistently results in the production of voices which are perceived as more (or less) listenable and aesthetically pleasing for specific tasks of varying cognitive load. The findings from both section one and and two can be used by synthetic voice designers in order to significantly improve the overall quality of both their speech output implementation strategies and interface design, and also to significantly enhance the quality of the actual voices used in systems.

As quality improves, products should become more widely available. If the speech they produce is appropriate for the task and actually pleasant and reassuring to listen to, then speech synthesis may finally make the kind of impact within human-computer interaction that has been anticipated for so long.

C.Cowley
1999

Appendix One

Factor Analysis Ratings Program

Overview

The program was written and compiled in the high-level Pascal programming language and utilised the Digital Equipment Corporation's prototype Dialogue-Bus environment (c). This allows streams of ASCII code to be sent to the synthesiser to be processed. The code consists of phonetically-balanced stimulus sentences and the embedded commands needed to modify the voice parameters and specify which member of the voice set is required for each individual trial. The program was designed to elicit ratings of DECtalk synthetic speech efficiently and elegantly

As was discussed in chapter two, the encoding of synthetic speech can place considerable demands on information processing capacities, leaving fewer resources available for any additional tasks that listeners are asked to perform simultaneously. It was therefore essential that the operation of the program be extremely simple, allowing the subjects to concentrate fully on the task of listening to the voices. This objective was achieved as all subjects understood the procedure within one or two practice trials and none of the subjects experienced any difficulty whatsoever in operating the program at any stage of the experiment.

The program also includes a number of safety checks to eliminate any potential problems with the procedure. These include a software 'masking' of all of the keys that have no part in the experiment and the use of a simple self-correction facility in the unlikely occurrence of a subject mistakenly entering an unintentional

ratings score. Furthermore, it was necessary to develop a strategy to eliminate the unfortunate possibility that a subject, bored by the repetitive nature of the experiment (especially in the two hour factor analysis study) may perhaps be tempted to race through the procedure, entering scores at random, in order to complete the experiment in the shortest possible time. In early versions of the program it was possible to enter a rating for each presentation before the DECtalk had produced the output (rendering the data meaningless). This would cut down the length of the experiment to the shortest possible time. Whilst it was hoped that none of the chosen subjects would be unscrupulous enough to do this, it was necessary to ensure that this situation did not arise. To achieve this, the software was developed so that it was impossible to enter a rating until the DECtalk had finished speaking the stimulus sentence. If a subject attempted to enter a rating too early, the computer would reject the score and produce an error 'bleep'. As a result, all the subjects completed the procedure correctly and there were no incidences of the program hanging up or losing any data.

Appendix 2A

Factor Analysis Ratings Program

Summary of program architecture:

Procedures and functions

Procedure: open_results_file:

A file is opened to accept data at the conclusion of the experiment.

Procedures: assign_scales/voices/sentences.

Scales, voices and sentences are assigned to numerical arrays.

Procedure: setup_terminal.

The terminal is then set up to accept input from the ratings keys automatically, and without the need for the subject to press the <return> key after every rating.

Procedures: random_seed_from_time and get_integer.

The program interrogates the VAX "Time" utility, getting the exact time of the onset of the experiment to hundredth of a second. This is transformed into a unique six figure integer.

Procedures: scramble_scales/voices/sentences and random/initial.

A random number generator uses this figure as a seed, the algorithm scrambling the presentation order of the scales, sentences, and

voices. This way, each subject is presented with a unique random order of ratings trials (the possibility of the experiment starting at the exact time, to one hundredth of a second, on consecutive days was assumed unworthy of serious consideration).

Next, to prevent the subject entering a rating before the scale and stimuli have been presented, the Dialogue_Bus is rewired to send input to the null channel. Whilst in this state, any key-presses from the subject are automatically rejected.

Procedures: present_scale/speech (containing the nested procedures: draw_scale, present_speech, wait).

The appropriate scale is presented and, after a brief pause allowing the subject to read and understand which scale he/she will be rating the voice on, a sentence of synthetic speech is spoken. As soon as the sentence has been spoken, the Dialogue_Bus is rewired to accept input.

Procedure: do_rating (containing the function: record_rating, and nested procedures: correct_rating, confirm_rating and reject_rating).

The subject is then prompted to enter a rating. On pressing an appropriate key, the scale is highlighted at the chosen point in reverse video. The subject then has the option either to press a correction key which removes the highlighting and allows the subject to enter a different rating, or press the <return> key to move on to the next trial.

Procedure: rest_period.

At six points in the experiment, the program pauses allowing the subject to take a brief rest before carrying on with the next block of trials. When the final trial has finished, the results file is opened and

the data is written to the file in an unscrambled array. Finally this file is closed and the program terminates.

The programs for the subsequent experiments had exactly the same architecture and used the worked in the same manner. The only difference being that the voice sets were modified and the inappropriate ratings scales were removed in order to test the specific voice modifications and scale sets under investigation.

Appendix 2B: Stimuli

Phonetically balanced sentences

A number of the stimuli are spelled phonetically and use DECtalk's phoneme codes to enhance clarity and minimise American inflections.

[1] the ["m][ae]["p] is in the [b][ae][g]
[2] [ay] took the briefcase home
[3] his voice is rather loud
[4] they [ch]["aa][jh] a [l][ow][ax] fee
[5] a fish swam through the net
[6] the boy rang up [h]["ih][s] friend
[7] some-thing good will happen
[8] she"s using her ["leh]["zh][ax] time
[9] the ply-wood was very thin
[10] the blood-hound has the scent
[11] the cheese is wrapped in tin foil
[12] the universe ["ih][s] enormous
[13] the thief left a ransom note
[14] [sh]["iy] searched for a vacancy
[15] you should change trains at ban["g][ax]
[16] he stitched up ["th]["ax] loose ["b]utton
[17] collect a few more shells
[18] we are leading the way two the shelter
[19] I begged them to change the [ah]["p]["oy][n][t][m]["eh][n][t]
[20] I think ["w][h][iy]["l] reach wales by morning
[21] these coins are danish cur-en-see
[22] the new ones were only [hx][aa][f] cooked
[23] the handsome [d][aa][g] was chained to the kennel
[24] such books give people food for thought
[25] when making bread, use good fresh yeast
[26] the seizure of the [k][aas][el] brought victory
[27] the [pleh][zh][ax] boat goes up [axn] down the river
[28] the changes here will spoil our future
[29] the judge found the witnesses annoying
[30] trooth-full leaders get more votes.

Appendix 2C: Perceptual Scales

30 bi-polar 5 point rating scales.

[1] Happy 1 — 2 — 3 — 4 — 5 Unhappy
[2] Sluggish 1 — 2 — 3 — 4 — 5 Energetic
[3] Relaxed 1 — 2 — 3 — 4 — 5 Tense
[4] Dull 1 — 2 — 3 — 4 — 5 Alert
[5] Active 1 — 2 — 3 — 4 — 5 Passive
[6] Depressed 1 — 2 — 3 — 4 — 5 Cheerful
[7] Composed 1 — 2 — 3 — 4 — 5 Confused
[8] Disturbing 1 — 2 — 3 — 4 — 5 Restful
[9] Calm 1 — 2 — 3 — 4 — 5 Anxious
[10] Unenterprising 1 — 2 — 3 — 4 — 5 Enterprising
[11] Interesting 1 — 2 — 3 — 4 — 5 Boring
[12] Dissatified 1 — 2 — 3 — 4 — 5 Satisfied
[13] Comfortable 1 — 2 — 3 — 4 — 5 Frustrating
[14] Harsh 1 — 2 — 3 — 4 — 5 Gentle
[15] Clear 1 — 2 — 3 — 4 — 5 Confusing
[16] Rough 1 — 2 — 3 — 4 — 5 Smooth
[17] British 1 — 2 — 3 — 4 — 5 Foreign
[18] Uneducated 1 — 2 — 3 — 4 — 5 Knowledgeable
[19] Stimulating 1 — 2 — 3 — 4 — 5 Tiring
[20] Distracting 1 — 2 — 3 — 4 — 5 Aids Concentration
[21] Fluent 1 — 2 — 3 — 4 — 5 Halting
[22] Hostile 1 — 2 — 3 — 4 — 5 Friendly
[23] Melodious 1 — 2 — 3 — 4 — 5 Grating
[24] Timid 1 — 2 — 3 — 4 — 5 Assertive
[25] Authoritarian 1 — 2 — 3 — 4 — 5 Meek
[26] Discrete 1 — 2 — 3 — 4 — 5 Intrusive
[27] Irritating 1 — 2 — 3 — 4 — 5 Not Irritating
[28] Refined 1 — 2 — 3 — 4 — 5 Crude
[29] Unpleasant 1 — 2 — 3 — 4 — 5 Pleasant
[30] Friendly 1 — 2 — 3 — 4 — 5 Unfriendly

Appendix 2D: Stimuli

Showing modifications to head size, average pitch, richness and smoothness for each of the twenty four voices.

```
voice[1] := '[:np :dv hs 115]'; {paul}
voice[2] := '[:np :dv ap 170]';
voice[3] := '[:np :dv ri 30]';
voice[4] := '[:np :dv sm 84]';

voice[5] := '[:nb :dv hs 115]'; {betty}
voice[6] := '[:nb :dv ap 230]';
voice[7] := '[:nb :dv ri 50]';
voice[8] := '[:nb :dv sm 74]';

voice[9] := '[:nh :dv hs 135]'; {harry}
voice[10] := '[:nh :dv ap 128]';
voice[11] := '[:nh :dv ri 36]';
voice[12] := '[:nh :dv sm 64]';

voice[13] := '[:nk :dv hs 105]'; {kit}
voice[14] := '[:nk :dv ap 256]';
voice[15] := '[:nk :dv ri 90]';
voice[16] := '[:nk :dv sm 74]';

voice[17] := '[:nf :dv hs 105]'; {frank}
voice[18] := '[:nf :dv ap 203]';
voice[19] := '[:nf :dv ri 30]';
voice[20] := '[:nf :dv sm 66]';

voice[21] := '[:nr :dv hs 110]'; {rita}
voice[22] := '[:nr :dv ap 156]';
voice[23] := '[:nr :dv ri 50]';
voice[24] := '[:nr :dv sm 94]';
```

Appendix 2E:

Factor Analysis Ratings Progam

Source code

```
 [inherit   ('sys$library:starlet')]
 program   speech_rating(input,output);

const
    esc       = chr(27);
    control_u  = chr(21);
    control_g  = chr(7);
    bell       = esc+control_g;
    pmi        = esc+control_u;
    maxscales  = 30;
    maxvoices  = 24;
    maxsentences = 30;
    terminal   = 'TT:'; {getkey facility}

type
    string        =  varying [512] of char;
    timestring = packed array [1..11] of char;
    scales        = array [1..maxscales] of string;
    voices        = array [1..maxvoices] of string;
    sentences       = array [1..maxsentences] of string;
    short         = [WORD] 0..65535;       {      }
    io_block      = record                 {      }
    io_stat, count : short;  {getkey  }
    dev_info       : integer {facility}
end;                         {      }

  presentations  = array [1..maxvoices,1..maxscales]
   of integer;
```

```
voice_index    = array [1..maxvoices]
    of  integer;
            sentence_index = array [1..maxsentences]
                of  integer;
                        txt_file = file of char;
var
    screen_file  :  string;
    results_file : txt_file;
    time_seed : timestring;
    sx, x, next, final_number, first, second,
    third : integer; fourth, fifth, sixth : integer;
    seedstring, bit1, bit2, bit3, bit4, section
    : string;
    seed_real : real;
    v      : voice_index;
    s      : presentations;
    sin    : sentence_index;
    response: array [1..maxvoices,1..maxscales]
    of char;
    scale   : scales;
    voice   : voices;
    sentence : sentences;
    request : string;
    rating, resume   : char;
    seed    : real;
    random_choice, number, num : integer;
    finish  : boolean :=false;
    {getkey  variables}
       term_chan : short;
            sys_stat  : integer;
                        iostat_block  : io_block;

procedure LIB$STOP(%immed cond_value : integer);
    extern;
```

```
{**********************************************}
procedure  get_integer;
begin
    case section[next] of
                '1' : sx:= 1; '2' : sx:= 2; '3' : sx:= 3;
                '4' : sx:= 4; '5' : sx:= 5; '6' : sx:= 6;
                '7' : sx:= 7; '8' : sx:= 8; '9' : sx:= 9;
                '0' : sx:= 0;
end;
        next := next+1;
end;
{**********************************************}
procedure  random_seed_from_time;
begin
   time(time_seed);
    seedstring := time_seed;
      bit1 := substr(seedstring, 4, 2);
      bit2 := substr(seedstring, 7, 2);
      bit3 := substr(seedstring, 10, 2);
  next := 1; section := bit1;
get_integer; first := sx; get_integer; second := sx;
  next := 1; section := bit2;
get_integer; third := sx; get_integer; fourth := sx;
       next :=1; section := bit3;
get_integer; fifth := sx; get_integer; sixth := sx;
final_number := first * 100000 + second * 10000
+ third * 1000 + fourth * 100 + fifth * 10 + sixth;
        seed_real := final_number;
          seed_real := seed_real/100000;
end;
{**********************************************
*}
procedure  setup_terminal;
     {assign channel to terminal}
```

```
begin
 sys_stat := $ASSIGN(terminal, term_chan, , );
  if not odd(sys_stat) then begin
       writeln('cannot assign channel --- exiting');
      LIB$STOP(sys_stat);
                  end
 end;
{**********************************************************************
}
function getkey : char;
 var
  c : char;
 begin
    sys_stat := $QIOW(chan := term_chan,
          func := int(UOR(IO$_READVBLK,
IO$M_NOECHO)),
 iosb := iostat_block, P1 := c, P2 := 1, P4 := 0);
   if not odd(sys_stat) then begin
      writeln('QIO failed --- exiting');
    ` LIB$STOP(sys_stat);
end;
 getkey := c
end;
{**********************************************************************}
procedure  assign_scales;
begin
scale[1]  := 'Happy                Unhappy';
scale[2]  := 'Sluggish              Energetic';
scale[3]  := 'Relaxed                Tense';
scale[4]  := 'Dull                    Alert';
scale[5]  := 'Active                Passive';
scale[6]  := 'Depressed            Cheerful';
scale[7]  := 'Composed            Confused';
scale[8]  := 'Disturbing               Restful';
```

```
scale[9]  := 'Calm                    Anxious';
scale[10] := 'Unenterprising      Enterprising';
scale[11] := 'Interesting             Boring';
scale[12] := 'Dissatified           Satisfied';
scale[13] := 'Comfortable        Frustrating';
scale[14] := 'Harsh                   Gentle';
scale[15] := 'Clear                 Confusing';
scale[16] := 'Rough                  Smooth';
scale[17] := 'British                 Foreign';
scale[18] := 'Uneducated     Knowledgeable';
scale[19] := 'Stimulating             Tiring';
scale[20] := 'Distracting  Aids Concentration';
scale[21] := 'Fluent                  Halting';
scale[22] := 'Hostile                 Friendly';
scale[23] := 'Melodious               Grating';
scale[24] := 'Timid                  Assertive';
scale[25] := 'Authoritarian              Meek';
scale[26] := 'Discrete                Intrusive';
scale[27] := 'Irritating         Not Irritating';
scale[28] := 'Refined                   Crude';
scale[29] := 'Unpleasant              Pleasant';
scale[30] := 'Friendly              Unfriendly';

end;
{*************************************************}
procedure  assign_voices;
begin
voice[1] := '[:np :dv hs 115]'; {paul}
voice[2] := '[:np :dv ap 170]';
voice[3] := '[:np :dv ri 30]';
voice[4] := '[:np :dv sm 84]';
voice[5] := '[:nb :dv hs 115]'; {betty}
voice[6] := '[:nb :dv ap 230]';
voice[7] := '[:nb :dv ri 50]';
voice[8] := '[:nb :dv sm 74]';
```

```
voice[9] := '[:nh :dv hs 135]'; {harry}
voice[10] := '[:nh :dv ap 128]';
voice[11] := '[:nh :dv ri 36]';
voice[12] := '[:nh :dv sm 64]';

voice[13] := '[:nk :dv hs 105]'; {kit}
voice[14] := '[:nk :dv ap 256]';
voice[15] := '[:nk :dv ri 90]';
voice[16] := '[:nk :dv sm 74]';

voice[17] := '[:nf :dv hs 105]'; {frank}
voice[18] := '[:nf :dv ap 203]';
voice[19] := '[:nf :dv ri 30]';
voice[20] := '[:nf :dv sm 66]';

voice[21] := '[:nr :dv hs 110]'; {rita}
voice[22] := '[:nr :dv ap 156]';
voice[23] := '[:nr :dv ri 50]';
voice[24] := '[:nr :dv sm 94]';

end;
{**********************************************************************
}
procedure   assign_sentences;
begin
sentence[1] :='[:ra 160] the ["m][ae]["p] is in the [b][ae][g].';
sentence[2] :='[:ra 160] [ay] took the briefcase home.';
sentence[3] :='[:ra 160] his voice is rather loud.';
sentence[4] :='[:ra 100] they [ch]["aa][jh] a [l][ow][ax] fee.';
sentence[5] :='[:ra 140] a fish swam through the net.';
sentence[6] :='[:ra 160] the boy rang up [h]["ih][s] friend.';
sentence[7] :='[:ra 160] some-thing good will happen.';
sentence[8] :='[:ra 160] she"s using her ["leh]["zh][ax] time.';
sentence[9] :='[:ra 160] the ply-wood was very thin.';
sentence[10] :='[:ra 150] the blood-hound has the scent.';
```

```
sentence[11] :='[:ra 160] the cheese is wrapped in tin foil.';
sentence[12] :='[:ra 160] the universe ["ih][s] enormous.';
sentence[13] :='[:ra 160] the thief left a ransom note.';
sentence[14] :='[:ra 160] [sh]["iy] searched for a vacancy.';
sentence[15] :='[:ra 160] you should change trains at ban["g][ax].';
sentence[16] :='[:ra 150] he stitched up ["th] ["ax] loose ["b]utton.';
sentence[17] :='[:ra 160] collect a few more shells.';
sentence[18] :='[:ra 160] we are leading the way two the shelter.';
sentence[19] :='[:ra 160] I begged them to change the
[ah]["p]["oy][n][t][m]["eh][n][t].';
sentence[20] :='[:ra 160] I think ["w][h][iy]["l] reach wales by
morning.';
sentence[21] :='[:ra 160] these coins are danish cur-en-see.';
sentence[22] :='[:ra 130] the new ones were only [hx][aa][f] cooked.';
sentence[23] :='[:ra 140] the handsome [d][aa][g] was chained to the
kennel.';
sentence[24] :='[:ra 160] such books give people food for thought.';
sentence[25] :='[:ra 160] when making bread, use good fresh yeast.';
sentence[26] :='[:ra 140] the seizure of the [k][aas][el] brought
victory.';
sentence[27] :='[:ra 130] the [pleh][zh][ax] boat goes up [axn] down the
river.';
sentence[28] :='[:ra 160] the changes here will spoil our future.';
sentence[29] :='[:ra 160] the judge found the witnesses annoying.';
sentence[30] :='[:ra 160] trooth-full leaders get more votes.';
end;
{**********************************************************}
}
procedure random (LargestRandomInteger   : integer;
                  var RandomInteger : integer;
                  var seed          : real);
begin
        seed := sqr(seed + 3.1415927);
        seed := seed - trunc(seed);
        RandomInteger  :=  trunc(LargestRandomInteger
```

```
        * seed) +1
end;
{**********************************************************************}
}
procedure  initial;
var
I, J : integer;
begin
for I := 1 to maxsentences do
  sin[I] := i;
for I := 1 to maxvoices do
   begin
     v[I] := I;
     for J := 1 to maxscales do
        begin
         s[I,J] := J;
    end;
 end;
end;
{**********************************************************************}
procedure  scramble_scales;
  var i, j, t, temp : integer;
begin
  for i := 1 to maxvoices do
     for j := 1 to maxscales do begin
       temp := s[i, j];
        random(maxscales, t, seed);
       s[i, j] := s[i, t];
       s[i, t] := temp
     end;
end;
{**********************************************************************}
procedure  scramble_voices;
  var i, t, temp : integer;
begin
```

```
for i := 1 to maxvoices do begin
    temp := v[i];
     random(maxvoices, t, seed);
    v[i] := v[t];
     v[t] := temp;
  end;
end;
{**********************************************************}
procedure   scramble_sentences;
  var i, t, temp : integer;
begin
   for i := 1 to maxsentences do begin
      temp := sin[i];
          random(maxsentences, t, seed);
           sin[i] := sin[t];
           sin[t] := temp;
 end;
end;
{**********************************************************}
procedure  wait;
 var delay : integer;
begin
        delay := 0;
              while delay<500000 do
        begin
              delay := delay + 1;
        end;
end;
{**********************************************************}
procedure  draw_scale;
begin
  write(esc,'[14;0H');
   write(esc,'#6');
   writeln('[1]------[2]------[3]------[4]------[5]');
end;
```

```
{**********************************************************}
procedure present_scale(x : integer);
begin
   writeln(esc,'[1;1H',esc,'[J');
   write(esc,'#6');
   writeln('       Speech Perception Scale');
   write(esc,'#6');
   writeln('              ------------------');
   write(esc,'[11;0H');
   write(esc,'#6');
   writeln('',scale[x],'');
   draw_scale;
   write(esc,'[23;0H');
   writeln('  ');
end;
{**********************************************************}
procedure present_speech(x, y : integer);
begin
     writeln(pmi,'{dectalk}',voice[x],'');
      writeln(pmi,'{dectalk}',sentence[y],'');
end;
{**********************************************************}
function record_rating : boolean;
begin
        write(esc,'[22;0H');
        write('                              ');
        write(esc,'[22;0H',esc,'#6');
        writeln('Please enter a number (1 to 5)..');
        rating := getkey;
        if rating in ['1','2','3','4','5'] then
        record_rating := true
    else
         record_rating := false
end;
{**********************************************************}
```

```
procedure   correct_rating;
begin
        write(esc,'[21;0H');
        writeln('                                    ');
        write(esc,'[22;0H',esc,'#6');
        writeln('Press RETURN to continue                ');
        write(esc,'#6');
        writeln('Press "C" to change rating              ');
        request := getkey;
end;
{*****************************************************}
procedure   confirm_rating;
begin
  draw_scale;
  case rating of
        '1' : write(esc,'[14;2H',esc,'[7m1',esc,'[0m');
        '2' : write(esc,'[14;11H',esc,'[7m2',esc,'[0m');
        '3' : write(esc,'[14;20H',esc,'[7m3',esc,'[0m');
        '4' : write(esc,'[14;29H',esc,'[7m4',esc,'[0m');
        '5' : write(esc,'[14;38H',esc,'[7m5',esc,'[0m');
         end
end;
{*****************************************************}
procedure   reject_rating;
begin
        write(bell);
        write(esc,'[21;0H',esc,'#6');
        writeln('Please try again..                  ');
end;
{*****************************************************}
procedure do_rating(i, j : integer);
var
  done : boolean;
begin
   while not record_rating do
```

```
reject_rating;
  confirm_rating;
done := false;
 while not done do begin
    correct_rating;
    if request = chr(13) then      {return}
     done := true
else
  if (request = 'c') or (request= 'C') then begin
   write(esc,'[23;0H');
    writeln('                              ');
       draw_scale;
        while not record_rating do
          reject_rating;
        confirm_rating
    end
     else
        write(bell)
  end;
   response[i, j] := rating;
end;
{*****************************************************************}
procedure  open_results_file;
begin
writeln('Enter filename for results : ');
      read(screen_file);
         open(results_file, screen_file, NEW);
            rewrite(results_file);
end;
{*****************************************************************}
procedure rest_period;
var keypress : string;
 begin
    writeln(esc,'[1;1H',esc,'[J');
     write(esc,'#6');
```

```
writeln('Press return to continue');
      resume := getkey;
end;
{**********************************************************}
{main  program}
var i, j, k, np, rest : integer;
continue : boolean := false;
begin
  rest := 0;
  random_seed_from_time;
  seed := seed_real;
     write(esc,'[?25l'); {cursor off}
   initial;
         open_results_file;
         assign_scales;
         assign_voices;
         assign_sentences;
         scramble_scales;
         setup_terminal;
         scramble_sentences;
 np := 0;
 rest := 0;
rest_period;
 for j := 1 to maxscales do begin {maxscales begin}
  scramble_voices;
  for i := 1 to maxvoices do begin {maxvoices begin}
     {to prevent too-early rating, rewire bus}
 writeln(pmi,'{dialogue_bus}save');
present_scale(s[v[i], j]);
  writeln(pmi,'{dialogue_bus}connect
 db$input db$null');
   wait;
    present_speech(v[i], sin[np+1]);
 np := np + 1;
  if np = maxsentences then begin
```

```
scramble_sentences;
      np := 0;
   end;
      wait;
    writeln(pmi,'{dialogue_bus}restore');
   {write(esc,'[22;0H');}
      do_rating(v[i], s[v[i], j]);
   rest := rest + 1;
         if rest = 100 then begin
           rest_period;
             rest := 0;
        end;
    end;
end;
for i := 1 to maxvoices do begin
     for j := 1 to maxscales do
         write (results_file,' ',response[i, j]);
       write(results_file, chr(13), chr(10));
    end;
       close(results_file);
         write(esc,'[?25h'); {cursor back on}
         finish := true;
end.
```

Appendix 3

Full range of extracted factors

With the correlation cut-off point set to +/- 0.5, six factors were extracted after the variamax rotation, these accounted for an accumulative percentage of 62.5% of the overall variance. The scales are listed in order of highest to lowest correlation with the factor. Only factors 1, 2 and 3 were retained for further analysis as their Eigenvalue was greater than one. Factors 4, 5 and 6 were not retained but are included here in order to give a fully comprehensive picture of the results.

It should also be noted that the classification of the factors as specific named entities is by no means straightforward and relies heavily on the number of, and the perceived similarity of, the associated scale adjectives. As the number of associated scales diminishes, classification and naming becomes less precise and is more vunerable to subjective interpretation. Examination of the following factors and associated scales shows that it is only really factors 1 and 2 which appear relatively invunerable to ambiguity in classification. Factors 3, 4, 5 and 6 have a decreasing number of associated scales and are much more difficult to characterise with any strong degree of confidence.

Factor 1: "listenability"

Accounted for 44.7% of the variance

Dissatisfied	-----	Satisfied	(.79)
Irritating	-----	Not irritating	(.78)
Harsh	-----	Gentle	(.69)

Hostile	-----	Friendly	(.67)
Unpleasant	-----	Pleasant	(.65)
Disturbing	-----	Restful	(.63)
Crude	-----	Refined	(.50)

Factor 2: "Assurance"

Accounted for 12.6% of variance.

Calm	-----	Anxious	(.77)
Relaxed	-----	Tense	(.74)
Authoritarian	-----	Meek	(.69)
Clear	-----	Confusing	(.621)
British	-----	Foreign	(.620)
Composed	-----	Confused	(.55)
Knowledgable	-----	Uneducated	(.52)

Factor 3: "Aimiablity"

Accounted for 5.2% of variance.

Fluent	-----	Halting	(.73)
Friendly	-----	Unfriendly	(.558)
Smooth	-----	Rough	(.553)
Comfortable	-----	Frustrating	(.550)
Pleasant	-----	Unpleasant	(.548)

Factor 4: "Usability"

Accounted for 4.8% of the variance.

Melodious	-----	Grating	(.79)
Stimulating	-----	Tiring	(.71)
Interesting	-----	Boring	(.57)
Aids concentration	-----	Distracting	(.52)
Cheerful	-----	Depressed	(.50)

Factor 5: "Positivity/motivational"

Accounted for 4.3% of variance.

Enterprising	-----	Unenterprising	(.81)
Active	-----	Passive	(.80)
Happy	-----	Unhappy	(.67)
Alert	-----	Dull	(.59)
Cheerful	-----	Depressed	(.56)

Factor 6: "Animation"

Accounted for 3.5% of the variance.

Assertive	-----	Timid	(.87)
Energetic	-----	Sluggish	(.71)

Appendix 4

Experiment two

Listenability analysis for the male voice set

Experiment Title: Listenability analysis: Male voices, t-test on smoothness manipulations

```
DATA LIST FREE/ smlow smhigh.
BEGIN DATA
23 53
25 44
33 50
39 45
18 34
31 49
40 45
38 41
50 55
33 49
33 39
34 41
40 52
35 44
40 41
31 38
26 44
29 42
32 37
29 52
END DATA.
T-TEST/PAIRS  smlow  smhigh
```

Variable	Number of cases	Mean	Standard Deviation	Standard Error
smlow	20	32.9500	7.185	1.607
smhigh	20	44.7500	5.839	1.306

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
-11.8000	7.620	1.704	.330	.156	-6.93	19	.000

t-test on richness manipulations

```
DATA/LIST FREE/ smlow smhigh.
BEGIN DATA
45 21
49 29
50 35
47 27
47 26
45 35
54 40
39 35
54 36
46 42
39 28
43 38
46 25
42 24
36 38
31 27
46 37
39 27
```

48 40
61 20
END DATA.
T-TEST/PAIRS rilow rihigh

Variable	Number of cases	Mean	Standard Deviation	Standard Error
rilow	20	45.3500	6.784	1.517
rihigh	20	31.5000	6.794	1.519

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
13.8500	9.621	2.151	-.004	.987	6.44	19	.000

Appendix 5

Experiment two

Assurance analysis for the male voice set

Experiment Title: Assurance analysis: Male voices
t-test on Average Pitch manipulations

```
DATA LIST FREE/ lopitch hipitch.
BEGIN DATA
57 31
47 32
56 32
58 37
45 31
50 40
50 46
62 35
61 32
44 43
45 30
52 45
47 30
53 28
43 29
40 34
51 32
55 34
51 31
60 31
END DATA.
T-TEST/PAIRS lopitch hipitch.
```

Variable	Number of cases	Mean	Standard Deviation	Standard Error
lopitch	20	51.3500	6.368	1.424
hipitch	20	34.1500	5.314	1.188

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
17.2000	8.421	1.883	-.031	.896	9.13	19	.000

t-test on Head Size manipulations

```
DATA LIST FREE/ smhead hihead.
BEGIN DATA
27 47
33 37
39 44
37 41
45 30
42 46
44 43
37 45
36 43
40 44
34 35
43 36
40 54
37 37
33 44
37 37
39 41
36 35
```

```
35 39
32 41
END DATA.
T-TEST/PAIRS smhead hihead.
```

Variable	Number of cases	Mean	Standard Deviation	Standard Error
smhead	20	37.3000	4.414	.987
hihead	20	40.9500	5.385	1.204

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
-3.6500	7.365	1.647	-.121	.611	-2.22	19	.039

Appendix 6

Experiment three

Listenability analysis for the female voice set

Experiment Title: Listenability analysis: Female voices
t-test on Smoothness manipulations

```
DATA LIST FREE/ smlow smhigh.
BEGIN DATA
40 57
38 59
43 58
38 53
40 55
41 62
39 62
36 50
35 64
33 55
34 52
36 57
33 61
37 61
36 61
36 61
32 44
45 58
47 57
42 60
38 59
END DATA.
T-TEST/PAIRS  smlow  smhigh.
```

Variable	Number of cases	Mean	Standard Deviation	Standard Error
smlow	20	38.1500	4.056	.907
smhigh	20	57.2500	4.767	1.066

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
-19.1000	5.271	1.179	.295	.207	-16.21	19	.000

t-test on Richness manipulations

```
DATA LIST FREE/ rilow rihigh.
BEGIN DATA
39 32
37 28
42 21
32 20
40 28
41 27
44 34
30 18
38 22
35 20
37 18
44 23
27 31
32 35
39 37
36 32
41 28
46 31
```

```
33 20
36 21
END DATA.
T-TEST/PAIRS  rilow  rihigh.
```

Variable	Number of cases	Mean	Standard Deviation	Standard Error
rilow	20	37.4500	4.979	1.113
rihigh	20	26.3000	6.114	1.367

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
11.1500	6.991	1.563	.218	.355	7.13	19	.000

Appendix 7

Experiment three

Assurance analysis for the female voice set

Experiment Title: Assurance analysis: Female voices
t-test on Average Pitch manipulations

```
DATA LIST FREE/ lopitch hipitch.
BEGIN DATA
44 40
43 38
46 32
42 35
43 34
48 32
49 46
45 36
55 30
51 28
41 39
41 29
50 35
48 44
55 31
39 38
43 37
46 40
50 40
54 32
END DATA.
T-TEST/PAIRS lopitch hipitch.
```

Variable	Number of cases	Mean	Standard Deviation	Standard Error
lopitch	20	46.6500	4.804	1.074
hipitch	20	35.8000	4.905	1.097

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
10.8500	7.727	1.728	-.267	.256	6.28	19	.000

t-test on Head Size manipulations

```
DATA LIST FREE/ smhead hihead.
BEGIN DATA
39 42
45 43
40 37
43 39
43 43
42 33
44 46
47 48
41 39
29 35
41 39
40 41
39 41
42 40
40 38
44 42
41 41
46 41
```

40 48
41 34
END DATA.
T-TEST/PAIRS smhead hihead.

Variable	Number of cases	Mean	Standard Deviation	Standard Error
smhead	20	41.3500	3.689	.825
hihead	20	40.5000	4.072	.910

(Difference) Mean	Standard Deviation	Standard Error	2-tail Corr.	Prob.	t Value	DF	2-tail Prob.
.8500	4.030	.901	.464	.039	.94	19	.357

Appendix Eight

Experiment 1 Covariance Matrix

Data from all subjects, thirty mean scale scores (shown horizontally) from each of forty subjects (shown vertically)

Subject 1:
3.417 2.333 3.000 2.333 3.583 2.583 2.542 2.250 3.250 2.833 3.625 2.583 3.333
2.500 2.125 3.292 2.167 3.708 3.917 2.208 2.333 2.875 3.667 2.750 2.750 3.042
2.167 2.417 2.417 3.167

Subject 2 (etc)
3.333 2.917 3.125 2.792 2.958 2.542 3.083 2.792 2.958 2.792 3.333 2.958 3.208
3.083 2.833 2.917 3.250 2.958 3.125 2.625 2.792 2.958 3.208 2.958 3.125 3.250
2.625 2.875 2.708 2.917

2.958 2.833 2.958 2.750 3.333 2.583 2.750 3.208 2.542 2.375 3.417 3.375 2.875
3.167 2.292 3.250 3.958 3.667 3.708 2.667 2.583 3.583 2.792 3.417 2.917 3.125
3.125 2.333 2.875 3.208

3.417 2.167 2.500 2.208 3.542 2.500 2.875 2.875 2.500 2.125 3.708 2.500 3.750
3.583 2.708 3.208 2.292 2.750 3.958 2.292 3.125 3.000 2.833 2.792 3.500 3.417
2.125 2.875 2.500 3.208

3.167 2.667 3.042 2.792 3.000 2.750 2.667 2.917 2.625 2.875 2.833 2.667 3.250
3.417 2.500 3.125 3.417 3.375 3.542 2.917 2.875 3.167 2.625 2.958 2.583 2.875
2.375 2.542 2.875 2.875

3.292 2.458 3.333 1.792 3.708 1.875 2.500 3.083 2.250 2.583 3.625 2.000 3.500
3.042 3.000 2.792 2.292 3.292 3.917 2.792 3.083 2.667 2.917 2.667 3.250 2.625
2.833 3.125 2.625 3.083

2.917 2.625 2.917 2.958 2.833 2.958 2.667 2.708 2.958 3.292 3.000 3.042 3.208
2.792 2.625 2.958 2.958 3.542 3.375 2.792 2.500 2.875 2.875 2.750 2.875 3.042
2.417 2.833 2.667 2.667

3.417 2.792 3.042 3.500 2.542 2.292 2.292 2.792 2.875 3.125 3.042 3.417 3.125
2.875 2.292 2.917 3.958 3.167 3.542 2.958 2.542 3.333 3.500 3.333 2.792 3.000
2.500 3.583 2.833 2.833

3.625 2.625 3.000 2.500 3.875 2.458 2.708 3.083 3.083 2.708 3.542 3.667 3.042
3.125 2.542 3.208 3.375 3.167 3.500 2.833 3.375 3.208 3.250 3.125 3.333 2.833
4.083 2.375 3.333 2.917

3.208 3.042 2.750 3.667 2.875 2.875 2.750 2.792 3.000 3.083 3.292 3.125 2.958
3.375 2.958 3.458 2.750 3.208 3.375 2.750 2.583 3.167 3.000 3.083 3.125 2.625
2.250 2.833 2.333 2.625

3.417 2.625 3.500 2.458 3.000 2.542 2.208 2.375 2.833 1.917 3.458 2.875 2.958
2.542 2.833 3.333 2.625 3.000 3.083 2.125 2.792 2.583 3.333 3.208 2.917 3.583
3.083 3.167 2.667 3.125

3.208 2.583 3.250 2.750 3.333 2.708 2.792 2.667 3.208 3.042 3.458 2.708 2.792
3.208 2.500 3.083 2.667 2.833 4.042 2.750 2.708 3.042 3.083 2.917 2.917 3.583
3.417 3.625 3.083 3.125

3.500 2.500 2.583 2.208 3.250 2.417 2.208 3.250 2.417 2.833 3.667 3.208 3.042
3.292 2.208 3.208 2.292 3.625 3.750 2.708 3.000 2.917 3.167 2.792 3.500 2.625
2.833 2.792 3.125 3.375

3.208 2.375 3.250 2.625 2.792 2.583 2.708 2.542 3.458 3.000 3.292 2.958 3.750
2.708 3.292 2.792 3.250 3.000 3.417 2.458 3.000 2.667 3.750 3.333 2.500 3.333
2.292 3.167 2.292 3.375

3.542 2.542 2.833 2.750 3.375 2.833 2.417 3.125 2.458 2.875 3.333 3.042 3.000
3.167 2.583 3.458 2.708 3.542 3.667 2.625 2.792 3.167 3.333 3.042 3.375 3.083
2.875 2.833 2.750 2.958

3.417 1.625 2.750 2.625 3.083 2.125 2.917 2.792 2.708 2.917 4.000 2.833 3.250
2.875 3.000 2.833 2.792 3.083 3.917 2.708 3.417 2.875 2.917 2.375 3.292 2.958
2.667 3.458 2.875 3.083

3.292 2.208 2.792 2.083 3.042 2.292 2.458 3.000 2.625 2.792 4.250 2.875 3.417
3.042 2.667 2.875 2.667 3.000 3.708 2.375 2.875 2.875 3.792 2.750 3.250 3.333
2.375 3.083 2.333 2.958

3.167 2.417 2.833 2.833 3.375 2.792 2.625 2.958 3.083 2.833 3.000 3.000 2.875
2.958 2.500 2.833 3.000 3.292 2.958 3.042 2.625 3.125 2.792 3.250 2.542 2.792
3.042 2.500 3.042 2.833

3.708 2.917 2.500 2.500 3.542 2.625 2.542 3.000 2.542 2.917 3.833 2.500 2.458
3.125 3.000 3.250 3.208 2.958 3.625 2.583 2.792 3.333 3.167 3.167 3.042 3.333
2.708 3.208 3.042 2.708

3.542 2.833 2.333 2.417 2.833 2.625 2.167 3.583 2.292 3.417 2.625 3.417 2.708
3.458 2.125 3.542 2.625 3.708 3.458 2.958 2.042 3.250 2.917 3.042 2.917 2.417
3.000 2.583 3.625 2.375

3.417 2.583 3.083 2.750 3.125 2.542 2.375 3.000 2.958 2.750 3.333 3.417 2.750
3.417 2.917 3.208 4.292 3.292 3.250 3.000 2.250 3.292 2.958 2.917 2.750 2.833
2.833 2.750 3.042 2.667

3.500 2.750 3.333 2.333 2.875 2.083 2.500 2.542 3.750 2.542 3.458 2.042 3.208
2.875 2.750 3.375 3.333 3.083 3.708 2.417 2.542 3.125 3.042 2.667 3.250 3.000
2.208 2.833 2.667 2.792

3.333 2.792 2.542 3.083 3.083 2.750 2.042 2.792 2.708 3.125 3.500 3.250 3.167
2.958 2.292 2.792 3.333 3.333 3.542 2.375 2.750 2.917 3.417 3.375 2.917 3.292
2.667 2.542 2.667 3.333

3.042 2.750 2.500 2.250 2.958 2.875 2.708 2.917 2.667 3.042 3.542 2.875 2.875

3.292 2.250 2.792 2.917 2.917 3.125 2.833 2.958 3.083 2.750 2.833 3.042 2.792
2.750 2.792 3.250 2.750

3.542 2.125 2.958 2.250 3.708 2.250 2.250 2.708 2.917 2.375 3.958 3.083 2.792
3.500 2.542 3.000 3.083 2.875 3.667 2.250 2.833 3.417 3.375 2.417 3.375 3.208
3.083 2.917 3.250 2.458

3.208 3.042 2.958 2.542 3.292 2.208 2.417 2.500 3.125 2.750 3.750 2.917 3.208
2.875 2.167 3.208 3.000 3.500 3.500 2.292 2.500 2.750 2.875 2.833 2.875 3.167
2.833 3.000 3.250 2.750

3.625 2.583 4.000 2.583 2.875 2.375 2.917 2.625 3.833 2.542 3.458 2.417 3.625
2.292 3.167 3.042 3.167 2.542 3.958 2.458 3.417 2.417 3.500 3.042 2.875 3.125
2.292 3.917 2.500 3.625

3.208 2.625 2.833 3.292 2.792 3.042 2.125 2.958 3.042 2.875 2.833 2.833 2.500
2.875 2.875 3.125 2.708 3.292 3.292 2.708 2.792 3.083 2.958 3.042 2.708 3.083
2.917 2.583 2.875 2.792

3.667 2.583 2.958 2.125 3.167 2.083 2.667 2.708 3.333 2.083 3.583 2.667 3.375
3.042 2.333 2.917 3.208 3.083 3.750 2.583 3.500 2.667 3.292 3.208 3.042 3.250
2.750 3.167 2.583 3.250

3.667 2.792 2.125 3.000 3.375 2.542 1.917 3.708 2.250 2.750 3.542 3.333 2.917
3.458 2.000 3.667 1.417 3.458 3.292 3.792 1.250 3.500 3.292 3.208 3.375 2.750
4.042 2.292 3.292 2.542

3.458 2.792 3.250 2.583 3.208 2.292 2.458 2.625 3.083 2.667 3.333 2.542 2.875
2.542 2.500 2.708 2.875 2.667 3.875 2.208 1.750 3.000 3.000 3.167 2.792 2.875
2.333 3.208 2.958 2.875

3.583 2.375 3.625 2.667 3.375 2.833 3.292 3.083 3.083 2.750 3.292 3.292 3.292
3.042 3.125 2.458 3.250 3.042 3.125 2.708 3.250 2.750 3.208 2.458 2.917 3.292
2.458 2.667 2.667 2.917

2.958 2.500 3.000 3.083 2.917 3.125 2.958 3.125 3.208 3.208 2.792 2.875 3.000
3.458 2.667 3.208 3.167 3.000 3.292 2.750 2.667 3.417 2.625 3.083 2.708 3.042
3.250 3.208 3.417 2.833

2.875 2.792 3.292 3.333 1.917 3.250 1.792 3.458 2.917 3.792 3.667 3.083 3.042
3.042 2.292 2.708 3.833 3.083 3.542 2.708 3.167 3.167 3.542 3.500 2.417 3.542
2.333 3.417 2.750 2.792

4.000 2.542 4.375 1.333 3.250 1.458 4.208 1.167 4.167 2.625 4.708 1.542 4.875
1.625 4.583 1.542 4.958 1.417 4.708 1.083 4.542 1.458 4.625 2.667 1.917 4.750
1.042 4.583 1.042 4.458

3.167 2.250 2.833 2.417 3.000 2.458 2.750 2.417 2.792 2.583 4.250 2.708 3.125
2.500 2.208 2.750 3.458 2.833 3.292 2.208 2.792 2.375 3.167 2.917 3.500 3.125
1.917 3.542 2.208 3.125

3.458 2.792 3.250 2.583 3.208 2.292 2.458 2.625 3.083 2.667 3.333 2.542 2.875
2.542 2.500 2.708 2.875 2.667 3.875 2.208 1.750 3.000 3.000 3.167 2.792 2.875
2.333 3.208 2.958 2.875

3.375 3.083 2.750 2.667 3.167 2.958 3.000 2.833 2.917 2.542 3.375 2.333 3.458
2.500 2.667 2.250 3.375 2.667 2.625 3.167 2.917 2.667 2.667 3.208 2.875 2.917
2.125 3.292 2.417 3.125

3.167 2.750 3.333 2.542 3.167 2.625 2.583 2.917 2.917 3.000 3.625 3.000 3.042
2.958 2.167 2.917 4.042 3.083 3.458 2.750 3.125 2.958 3.333 2.583 3.333 2.792
2.458 3.167 2.417 2.833

3.250 3.000 2.542 2.792 2.792 2.750 2.292 3.083 2.375 2.792 2.917 2.542 2.667
3.000 2.500 3.125 2.458 3.583 3.042 3.292 2.500 3.333 2.667 2.958 3.083 2.917
2.958 2.667 2.917 2.458

Appendix Nine
Mean scores for Experiment Two

Listenability means for male voice set (Individual subject means and total mean for each voice manipulation)

Smoothness 3%	Smoothness 100%	Richness 100%	Richness 0%
1.642	3.785	1.5	3.214
1.785	3.142	2.071	3.5
2.357	3.571	2.5	3.571
2.785	3.214	1.928	3.357
1.285	2.428	1.857	3.357
2.214	3.5	2.5	3.214
2.857	3.214	2.857	3.857
2.714	2.928	2.5	2.785
3.571	3.928	2.571	3.857
2.357	3.5	3	3.285
2.357	2.785	2	2.785
2.428	2.928	2.714	3.071
2.857	3.714	1.785	3.285
2.5	3.142	1.714	3
2.857	2.928	2.714	2.571
2.214	2.714	1.928	2.214
1.857	3.142	2.642	3.285
2.071	3	1.928	2.785
2.285	2.642	2.857	3.428
2.071	3.714	1.428	4.357
Total	Total	Total	Total
2.3532	3.19595	2.2497	3.2389

Assurance means for male voice set
(Individual subject means and total mean for each voice manipulation)

Average Pitch 160hz	Average Pitch 80hz	Head Size 86%	Head Size 115%
2.214	4.071	1.928	3.357
2.285	3.357	2.357	2.642
2.285	3	2.785	3.142
2.642	4.142	2.642	2.928
2.214	3.214	3.214	2.142
2.857	3.571	3	3.285
3.285	3.571	3.142	3.071
2.5	3.428	2.642	3.214
2.285	4.357	2.571	3.071
3.071	3.142	2.857	3.142
2.142	3.214	2.428	2.5
3.214	3.714	3.071	2.571
2.142	3.357	2.857	3.857
2	3.785	2.642	2.642
2.071	3.071	2.357	3.142
2.428	2.857	2.642	2.642
2.285	3.642	2.785	2.928
2.428	3.928	2.571	2.5
2.214	3.642	2.5	2.785
2.214	4.285	2.285	2.928
Total	Total	Total	Total
2.4388	3.5674	2.6638	2.92445

Appendix Ten
Mean scores for Experiment Three

Listenability means for female voice set (Individual subject means and total mean for each voice manipulation)

Smoothness 20%	Smoothness 100%	Richness 100%	Richness 0%
2.857	4.071	2.285	2.785
2.714	4.214	2	2.642
3.071	4.142	1.5	3
2.714	3.785	1.428	2.285
2.857	3.928	2	2.857
2.928	4.428	1.928	2.928
2.785	4.428	2.428	3.142
2.571	3.571	1.285	2.142
2.5	4.571	1.571	2.714
2.357	3.928	1.428	2.5
2.428	3.714	1.285	2.642
2.571	4.071-	1.642	3.142
2.357	4.357	2.214	1.928
2.642	4.357	2.5	2.285
2.571	4.357	2.642	2.785
2.285	3.142	2.285	2.571
3.214	4.142	2	2.928
3.357	4.071	2.214	3.285
3	4.285	1.428	2.357
2.714	4.214	1.5	2.571
Total	Total	Total	Total
2.72465	4.0888	1.87815	2.67445

Assurance means for female voice set
(Individual subject means and total mean for each voice manipulation)

Average Pitch 240hz	Average Pitch 160hz	Head Size 95%	Head Size 115%
2.857	3.142	2.785	3
2.714	3.071	3.214	3.071
2.285	3.285	2.857	2.642
2.5	3	3.071	2.785
2.428	3.071	3.071	3.071
2.285	3.428	3	2.357
3.285	3.5	3.142	3.285
2.571	3.214	3.357	3.428
2.142	3.928	2.928	2.785
2	3.642	2.071	2.5
2.785	2.928	2.928	2.785
2.071	2.928	2.857	2.928
2.5	3.571	2.785	2.928
3.142	3.428	3	2.857
2.214	3.928	2.857	2.714
2.714	2.785	3.142	3
2.642	3.071	2.928	2.928
2.857	3.285	3.285	2.928
2.857	3.571	2.857	3.428
2.285	3.857	2.928	2.428
Total	Total	Total	Total
2.5567	3.33165	2.95315	2.8924

Appendix 11

Experiment five stimuli sentences

Instructions given to subjects:

Please read the following pairs of statements and decide whether they are commonly true or false. Sometimes both statements will be true, or both false, or one true and the other false. For some, you may be able to think of odd exceptions to the rule, in such cases just put the obvious answer that you think most people would agree on. Having indicated your response with a T or F at the end of each sentence, decide which of the two sentences is *the easiest to understand* and mark this by circling the number. Thankyou.

For example:

21. Fish live in the sea. T

(22). The sea is consistently an unnatural habitat for many fish F

1. Most dogs have four legs
2. Most dogs have between 3 and 5 legs each

3. Ships sail on the sea
4. The sea is commonly a typical and suitable medium for shipping

5. The production of cream does not ever require newspapers
6. Cream is not made from newspapers

7. Beer is often sold in pubs
8. Pubs are establishments for the frequent vending of beer

9. Trees have lots of branches
10. Branches are common attributes of most trees

11. Sand is never nice to eat
12. The consumption of sand is consistently unpleasant

13. Eggs production requires the presence of chickens
14. Eggs are laid by chickens

15. Asia is not the real geographical location of England
16. England is not in Asia

17. Vision is unavailable for all blind individual people
18. Blind people cannot see

19. Most birds have wings
20. Wings are a standard component feature of most birds

21. Manufacture of trumpets commonly requires brass
22. Trumpets are made of brass

23. Most tables have legs
24. Legs are used as vertical supports on most tables

25. Most windows are made of glass
26. Glass is a manufacturing component used for windows

27. People sit in chairs
28. Chairs provide an opportunity for people to sit down

29. Pens are used for writing
30. Writing can always be achieved with the use of pens

31. Two plus two equals four
32. Four is the combination of two in addition to two

33. Heat is a common and consistent attribute of fire
34. Fire produces heat

35. Roundness is a property of the wheels of vehicles
36. Cars have round wheels

37. Mortality is inevitable for all individual people
38. Everybody will die

39. Fish do not ride bicycles
40. The operation of bicycles is possible for any fish

41. Heat is a consistent product of the sun
42. The sun is cold

43. Houses all have doors
44. Doors are always uneccesary features of houses

45. There are seven days in a week
46. Each week does not ever feature less than eight separate days

47. A solitary heart is almost always possessed by every individual
48. People have two hearts

49. A round configuration is the standard design for footballs
50. Footballs are square

51. Grass cutting is accomplished using lawnmowers
52. Lawnmowers are not used to cut grass

53. Three is less than four
54. Five is more than 2 but not actually a greater number than three

55. Black is not white
56. White is exactly equivalent to black

57. Bananas are yellow
58. The colour blue is typically common to bananas

59. To stay alive, you have to breathe
60. Breathing is not a neccessity for maintaining existence

61. Everyday we get older
62. The age of people has no direct association with time

63. Time is a type of fruit
64. The classification of time is not in the category of fruit

65. Circles are round
66. Squareness is a common attribute of every circle

67. Three sides are possessed by every triangle
68. Triangles have 4 sides

69. Day is a direct contrast and opposite to night
70. Night is the same as day

71. The number four is a combination of three added to one
72. One plus three equals four

73. Each calender year consists of 12 seperate months
74. There are thirteen months in a year

75. All houses are level in all cases at all times
76. Houses are always flat

77. Five even vertices or sides are possessed by all squares
78. A square has five sides

79. Two seperate heads are possessed by every person
80. All people have two heads

81. Books are made of cheese
82. Cheese is essential during process of book manufacture

83. Lobsters roost in trees
84. Trees provide a natural habitat for crustaceans

85. Elephants are yellow and square
86. Elephants all possess some square and yellow attributes

87. Branches are not usually a common attribute of trees
88. Trees do not have branches

89. The sea is the standard habitat frequented by cats
90. Cats live in the sea

91. Money is not kept in banks
92. Banks never provide a depository for any money

93. Bricks are very light
94. There is a relatively minimal weight possessed by bricks

95. Orange is the typical colouration of most grass
96. Grass is usually orange

97. Wood is a consistent property of all music
98. Music is made of wood

99. All people sleep in the day
100. Sleep is always undertaken in daytime by every person

101. Things fall sideways
102. The direction that objects fall is always horizontal

103. Honey comes from spiders
104. The manufacture of honey is undertaken by spiders

105. Cricket is a smell
106. A common aroma is known as cricket

107. People fly in televisions
108. Television sets provide flying transportation for humans

109. Telephones swim in the river
110. Rivers provide swimming environments for telephones

111. Seven consists of a combination of one added to five
112. Five plus one equals seven

113. Cinemas never show films
114. Motion pictures are never projected at cinemas

115. Trains do not run on rails
116. Rails are usually necessary for trains to travel on

117. All people wear fruit
118. An excess of fruit is not worn by every individual

119. The letter b comes before the letter a
120. The letter a typically precedes the following letter b

121. Cameras do not take photos
122. Photographic images are generated with cameras

123. The moon floats in the sea
124. The sea does not ever contain the floating moon

125. Violins are used to play music
126. Musical compositions are not performed using violins

127. You need fish to play chess
128. Fish are not essential for undertaking the game of chess

129. Most cats have whiskers
130. The majority of felines do not possess whiskers

131. Most teacups have handles
132. Teacups almost always never possess handles

133. Eleven is the combined sum of six and two
134. Six plus two is not eleven

135. The nose is a receptor for auditory information
136. People do not hear with their noses

137. Elephants are smaller than ants
138. The physical size of ants is less than that of elephants

139. People don't drive on the roads
140. Roads provide a surface for the driving of cars

141. A day has thirty hours
142. Thirty hours are not usually contained within each full day

143. Germany is type of hat
144. People don't wear germany as a garment on their heads

145. Books do not have pages
146. Multiple sheets of paper are always contained in books

147. Food is not cooked in ovens
148. Ovens are frequently involved in the process of cooking food

149. Rain is dry
150. Moisture is a component feature of rain

151. Eskimos never live in igloos
152. Igloos provide many Eskimos with traditional accommodation

153. Bicycles have no handlebars
154. Handlebars are essential component features of most bicycles

155. Frames are not used to contain and display pictures
156. Pictures often have frames

157. Typewriters do not have keys
158. Keys are always used as typewriter components

159. Fridges make food hot
160. The temperature of food can be decreased in fridges

Thank you.

N.B. Of the total 160 the following pairs were eliminated because of ambiguity (i.e. subjects did not agree on which was the easiest to understand):

27/28, 41/42, 51/52, 67/68, 73/74, 75/76, 133/134, 139/140, 147/148 and 151/152

Appendix 12: Mean scores from Experiment four

Smoothness				
subject scores				
Smoothness 0%	Smoothness 25%	Smoothness 50%	Smoothness 75%	Smoothness 100%
2.6	2.4	2.8	3	3.4
2.2	2	2.4	3	3.8
2.2	2.2	2	2	3.4
3	3	2.8	3.6	3.8
2	2.2	2	2.2	3
2.2	2.8	2.8	3	2.8
2.4	2.6	2.6	3	3
2.2	2.6	2.8	3.2	3.4
1.8	2.2	2	2.8	3.4
2	2.2	2.2	2	2.6
2.4	2.2	2.4	3	3.4
2.4	2	2	2.6	2.8
2.8	3	3.4	3.6	3.2
3	3	3.6	3.2	3.6
2.6	2.6	3	3.2	2.8
2.8	2.6	3	3.2	3.2
3	3	3	3.2	3
2.4	2.4	2.6	3.4	4
2.6	2.6	2.8	2.6	2.8
2	2.2	2.4	3	3
2.43	2.49	2.63	2.94	3.22

Richness				
subject scores				
Richness 0%	Richness 25%	Richness 50%	Richness 75%	Richness 100%
3.4	3.6	3.6	2.4	1.4
3.2	3.2	2.8	3.4	2
2.4	2.4	2.4	2.2	2.6
3.6	4	3.4	3	2
2.4	2.4	2.2	2.2	2.2
3.2	3	2.8	2.6	2
3.2	3	2.8	2.6	2.2
3.6	3.4	3	2.2	1.8
2.8	2.6	2.6	2.4	1.8
2.4	2.2	2.2	2	2.2
3.2	2.8	2.8	2.2	2.4
2.6	2.4	2.4	2.4	2
3.8	3.6	3.6	2.8	2.4
4	3.4	3.2	3	2.8
3	2.8	3	2.8	2.6
3.4	3.6	3	2.8	1.8
3	3	3	3.2	3
4	3.6	3.2	2.6	1.6
3	3	3	2.2	2.2
2.8	2.8	2.4	2.4	2.2
3.15	3.04	2.87	2.57	2.16

Interaction means from Experiment four

Variable	Mean	N
R100S0	1.65	20
R100S25	2.00	20
R100S50	2.05	20
R75S0	2.25	20
R75S50	2.30	20
R75S25	2.40	20
R100S75	2.45	20
R50S0	2.55	20
R25S25	2.60	20
R100S100	2.65	20
R50S25	2.65	20
R50S50	2.65	20
R25S0	2.70	20
R75S75	2.75	20
R75S100	2.90	20
R25S50	2.95	20
R0S25	3.00	20
R50S75	3.00	20
R0S0	3.00	20
R0S50	3.25	20
R0S100	3.25	20
R0S75	3.25	20
R25S75	3.25	20
R50S100	3.50	20
R25S100	3.60	20

References

Allen, J. (1980) Speech synthesis from text. In: Simon, J.C (editor) Spoken language generation and understanding. Dordrecht, Holland: Reidel.

Aucella, A (1986) Voice: technology searching for communication needs. In Carroll, J. Human factors in computing systems Vol.4: Graphic interface. Elseview Science Publishers B.V.

Baber, C. (1993) Speech output. In: Baber, C. and Noyes, J.M. (editors) Interactive speech technology. Taylor and Francis

Baddeley, A. D. and Hitch, G.J. (1974) Working memory. In G.Bower (Ed.), The psychology of learning and motivation. Vol 8, pp. 47-90. New York: Academic Press

Baddeley, A.D. and Gathercole, S.E. (1993) Working memory and language LEA ltd

Baecker, R.M., Grudin, J., Buxton, W.A.S. and Greenberg, S. (1995) Human-Computer Interaction: Toward the year 2000 2nd edition page 530 Morgan Kaufmann

Barnard, P (1991) The contributions of applied cognitive psychology to the study of human-computer interaction. In Shackel, B. and Richardson, S (Eds) Human factors for informatics usability, Cambridge University Press, Cambridge, U.K.

Barry, M. (1990) Synthesis of female voice quality. Unpublished postgraduate dissertation presented in partial fulfillment for the degree of Mphil, Cambridge University.

Bastien, J.M.C. and Scapin, D.L. (1993) Preliminary findings on the effectiveness of ergonomic criteria for the evaluation of human-computer interfaces. Interchi'93 Adjunct proceedings. pp.187-188

Bernsen, N. O. Dybkjaer, H. and Dybkjaer, L (1998) Designing interactive speech systems Springer

Boden, M. (1984) Artificial intelligence and natural man. Hassocks, Sussex: Harvester press

Bristow, G. (editor) (1986) Electronic speech recognition. London: Collins

Bruckert, E (1984) A new text-speech product produces dynamic human-quality voice. Speech Technology Jan/Feb pp. 114- 119

Burns, W. (1979) Physiological effects of noise. In: Harris, C.M. (editor) Handbook of noise control. New York: McGraw-Hill

Cole, R., Hirschman, L, Atlas, L. et al (1995) The challenge of spoken language systems - Research directions for the nineties. IEEE Transactions on speech and audio processing. Vol.3, No. 1, pp.1-21

Cooper, A. M., Whalen, D.H., and Fowley, C.A. (1986) P-centers are unaffected by phonetic categorization. Perception and Psychophysics. Vol. 39. pp.187-96.

Cox, A.C. and Cooper, M. B. (1981) Selecting a voice for a specified task: The example of telephone announcements. Language and speech Vol. 24, part 3 pp. 233-243

Deatherage, B. H. (1972) Auditory and other sensory forms of information presentation. In: Gott, H.P. and Kinsdale, R.G. (editors) Human engineering guide to equipment design. Revised edition. Washington: U.S. government printing office

DECtalk Owner's Manual (1983), Copyright (c) Digital Equipment Corporation. Order number: EK-DTC01-OM

Delogu, C., Paoloni, A., Ridolfi, P. and Vagges, K. (1995) Intelligibility of speech produced by text-to-speech systems in good and telephonic conditions. Acta Acustica Vol 3, No.1, pp. 89-96

Dilts, M (1984) Text to speech In Bristow, G. Ed. Electronic speech synthesis Granada publishing

Dix, A., Finlay, J., Abowd, G., and Beale, R. (1993) Human-computer interaction. Prentice Hall

Dybkjaer and Bernsen (1998) A methodology for diagnostic evaluation of spoken human-machine dialog. International Journal of Human-computer studies. Vol 48. No 5.

Edman, T. R. and Metz, S. V. (1983) A methodology for the evaluation of real-time speech digitization. Proceedings of the human factors society 27th meeting. pp. 104-107

Edwards, A. D.N. (1991) Speech synthesis. technology for disabled people Paul Chapman publishing, London

El-Shinnawy, M. and Markus, M.L. (1997) The poverty of media richness theory: explaining people's choice of electronic mail vs. voice mail. International journal of Human-Computer Studies. Vol 46. No 4.

Fant, C. G. (1983) Preliminaries to analysis of the human voice source. Speech transmission laboratory. Quartery progress report Vol. 4, pp. 1-27

Frankish, C., and Noyes, J. (1990) Sources of human error in data entry tasks using speech input. Human Factors. Vol.32, pp. 697-716

Frankish, C., and Noyes, J. (1993) Feedback in automatic speech recognition. In: Baber, C. and Noyes, J.M. (editors) Interactive speech technology. Taylor and Francis

Fritzel, B., Hammarberg, J., Gauffin, I. Karlsson, and Sundberg, J. (1986) Breathiness and insufficient vocal fold closure. Journal of phonetics Vol 14, pp. 549-553

Fromkin, V. and Rodman, R. (1993) An introduction to language. 5th edition, Harcourt Brace and Co

Frude, N. (1983) The intimate machine. Century Publishing: London.

Galer, I. (1974) Applied ergonomics handbook. Butterworth Heinemann

Guenzburger, D. (1984) Perception of some male-female voice characteristics. Progress Report, Institute of Phonetics. University of Utrecht Vol 9/2. pp. 15-26

Giles, H. and Powesland, P.F. (1975) Speech style and social evaluation. London.

Gobl, C. and Chassaide, A, N. (1992) Acoustic characteristics of voice quality. Speech communication Vol.11, pp 481-490

Gould, J. D. (1978) How experts dictate. Journal of Experimental Psychology: Human Perception and Performance Vol. 4.4 pp. 648-661.

Gould, J. D. (1982) Writing and speaking letters and messages. International journal of man-machine Studies Vol 16, pp. 147-171.

Gould, J. D. and Boies, S. J. (1984) Speech filing. An office system for principles. IBM systems Journal Vol. 23, No 1, pp. 65-81.

Gould, S.J. (1981) The mismeasure of man. Harmondsworth, Middlesex: Penguin.

Gray, T (1984) Talking computers in the classroom Ln Bristow, G. Ed. Electronic speech synthesis Granada publishing

Greene, B.G., Logan, J.S., and Pisoni, D.B., (1986) Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. Behaviour Research Methods, Instruments and Computers. Vol. 18, (2), pp. 100-107

Gross, R.D. (1992) Psychology: the science of mind and behaviour Hodder and Stoughton

Gutcho, L. (1985) DECtalk - A year later. Speech Technology Aug/Sep. pp. 98-102

Hanson, H.M. (1997) Glottal characteristics of female speakers: Acoustic correlates. Journal of the Acoustical Society of America. Vol.101, No.1, pp.466-481

Hayes, P.J. and Reddy, D.R. (1983) Steps towards graceful interaction in spoken and written man-machine communication. International journal of man-machine studies. Vol.19, pp. 231-284

Heim, A. (1970) Intelligence and personality - their assessment and relationship. Harmondsworth, Middlesex: Penguin

Helander, M. (1993) Foreward to: Baber, C. and Noyes, J.N. (editors) Interactive speech technology. Taylor and Francis

Holmes, J, N. (1988) Speech synthesis and recognition. Wokingham: Van Nostrand Reinhold

Hone, K. and Baber, C. (1999) Modelling the effects of constraint upon speech-based human-computer interaction. International journal of Human-Computer studies Vol 50, No 1.

Hunnicutt, S. (1980) Grapheme-to-phoneme rules, a review. Speech transmission laboratory, Royal Institute of Technology, Stockholm, Sweden. QPSR 2-3, pp.38-60

Hunnicutt, S. (1995) The development of text-to-speech technology for use in communications aids. In: Syrdal, A, Bennett, R. and Greenspan, S. (editors) Applied Speech Technology CRC Press. pp.547-563

Johnston, R. D. (1996) Beyond intelligbility - the performance of text-to-speech synthesizers. BT Technology Journal. Vol.14, No.1, pp.100-111

Jones, D. M. (1989) The Sonic Interface. In: Smith, M. and Salvendy, G. (editors), Proceedings of HCI International '89, Work with Computers: Organizational, Management, Stress and Health Aspects Amsterdam Elsevier

Jones, D.M., Macken, W.J., and Murray, A.C. (1993) Disruption of short-term memory by changing-state auditory stimuli: The role of segmentation. Memory and Cognition Vol. 21 (3), pp 318-328

Jusczyk, P. (1986) Speech perception. Handbook of Perception and Human Performance New York: Wiley Vol 2

Kidd, A. L. (1982) Problems of man-machine dialogue design. Proceedings of the 6th international conference on computer communication. London, pp 531-536

Kim, J. and Mueller, C.W. (1978) Factor Analysis: Statistical Methods and Practical Issues. Beverly Hills, CA: Sage Publications

Klatt. D. H. (1981) Synthesis of a female voice. Paper presented at the 101st meeting of the Acoustical Society of America Ottawa

Klatt, D.H. (1982) The Klatttalk text-to-speech system. Proceedings of the international conference of Acoustic Speech Signal Processing. ICASSP-82, pp.1589-1592

Klatt, D.H. and Klatt, L.C. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America Vol. 87, pp. 820-857

Klaus, H., Fellbaum, K., and Sotscheck, J. (1997) Subjective evaluation and comparison of the speech quality of text-to-speech systems for the German language. Acustica. Vol. 83, No.1, pp.124-136

Landauer, T.K. (1988) Relations between cognitive psychology and computer system design. In: Preece, J. and Keller, L. (editors) Human-computer interaction. Prentice Hall

Laurel, B. (1993) Computers as theatre. Addison Wesley, pp. 199-202

Lauretta, D. and Deffner, G. (1996) Dialogue style for hybrid telephone-based interfaces. Behaviour and information technology, Vol 15, No 1.

Liberman, M. (1995) Computer speech synthesis - its status and prospects. Proceedings of the National Academy of Sciences of The United States of America Vol.92, No.22, pp.9928-9931

Linggard, R. (1985) Electronic synthesis of speech. Cambridge, Cambridge University Press

Logan, J.S., Greene, B.G. and Pisoni, D.B. (1989) Segemental intelligibility of synthetic speech produced by rule. Journal of the Acoustic Society of America. Vol.86, No.2, pp.566-581

Luce, P. A, Feustel, T. C. and Pisoni, D. B. (1983) Capacity demands in short term memory for synthetic and natural speech. Human Factors. Vol 25, No 1, pp. 17-32

Marcus, S. M. (1981) Acoustic determinants of perceptual center (P-center) location. Perception and Psychophysics. Vol.30. pp. 247-56.

Michaelis, P. R, and Wiggins, R. H. (1982) A human factors engineer's introduction to speech synthesizers. In: Badre, A. and Schneiderman, B. (editors) Directions in Human-computer Interaction. Norwood NJ: Ablex Publishing

McGrath, J.E. (1994) Methodological matters: Doing research in the behavioural and social sciences. In: Baecker, R.M., Grudin, J., Buxton, W.A.S. and Greenbergy (editors) Human Computer Interaction. Morgan Kaufmann publishers inc.

McPeters, D.L., and Tharp, A.L., (1984) The interference of rule-generated stress on computer-synthesized speech. International Journal of Man Machine Studies. Vol. 20, pp. 215-226

Monsen, R. B. and Engbretson, A. M. (1977) Study of variations in the male and female glottal wave. Journal of the acoustic society of America. Vol. 62, pp. 981-993

Morton, J., Marcus, S., and Frankish, C. (1976) Perceptual centers (P-centers). Psychological Review. Vol. 83, pp. 405-8.

Murray, I., Arnott, J. L., Alm, N., and Newell, A.F. (1991) A communication system for the disabled with emotional synthetic speech produced by rule. Proceedings of EUROSPEECH Genova, Italy. Vol 1, pp. 311-314

Negroponte, N. (1995) Being Digital. New York, NY: Random House.

Nooteboom, S.G. (1983) The temporal organisation of speech and the process of spoken-word recognition. IPO (Eindhoven) progress report. Vol.18, pp. 32-36

Nusbaum H.C., Schwab, E.C., and Pisoni D.B. (1984) Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability. Research on Speech Perception Progress Report Vol.10 Bloomington, IN, Speech Research Laboratory, Indiana University.,

Nusbaum, H.C., Francis, A.L. and Henly, A.S., (1995) Measuring the naturalness of synthetic speech. International Journal of Speech Technology. Vol.1, pp.7-19

O'Malley, M.H. (1990) Text-to-speech conversion technology. Proceedings of IEEE Computer. Vol. 23(8), pp. 17-23

Oberteuffer, J.A. (1995) Commercial applications of speech interface technology - An industry at the threshold. Proceedings of the National Academy of Sciences of the United States of America. Vol. 92, No. 22, pp.10007-10010

Paris, C.R., Gilson, R.D., Thomas, M.H and Silver, N.C. (1995) Voice intelligibility on speech comprehension. Human Factors. Vol. 37 No 2, pp. 335-340

Peacock, G. E. (1984) Humanising the man/machine interface. Speech Technology. Vol 2, pp. 106-108

Pinter, I. (1996) Perceptual wavelet-representation of speech signals and its application to speech enhancement. Computer speech and language. Vol.10, No.1 pp. 1-22

Pisoni, D.B. (1981) Speeded classification of natural and synthetic speech in a lexical decision task. Journal of the acoustical society of America. Vol. 70, S98

Pisoni, D.B, Nusbaum, H.C. and Greene, B.G. (1985) Perception of synthetic speech generated by rule. Proceedings of the IEEE. Vol.73, NO. 11, pp. 1665-1675

Pisoni, D.B., Nusbaum, H. C., Luce, P. A. and Schwab, E. C. (1983) Perceptual evaluation of synthetic speech: Some considerations of the user/system interface. Proceedings of ICASSP '83 pp.535-538

Potter, R.K. (1946) Technical aspects of visible speech. Journal of the Acoustical Society of America. Vol. 17

Pratt, R.L. (1987) Quantifying the Performance of Text-to-Speech Synthesizers. Speech Technology. March 1987.

Robert, J. Fiset, J. and Bergeron, G. (1989) Impact of five task related factors on the choice of a vocal or a manual input modality. Proceedings of the European Conference on Speech Communication Technology. Edinburgh: CEP

Radford (1980) cited in: Gross, R.D. (1992) Psychology: the science of mind and behaviour Hodder and Stoughton

Roe, D.B. (1995) Deployment of Human-Machine dialogue systems. Proceedings of the National Academy of Sciences of the United States of America. Vol. 92, No. 22, PP.10017-10022

Roe, P (1984) Speech synthesis in telecommunications In Bristow, G. Ed. Electronic speech synthesis Granada publishing

Rosson, M. B. and Mellon, N. M. (1985) Behavioural issues in speech-based remote information retrieval. In: Lerman, L. (editor) Proceedings of the voice I/O systems applications conference. San Francisco: AVOIS

Rosson, M.B., and Cecela, A.J. (1986) Designing a quality voice: An analysis of listeners' reactions to synthetic voices. Proceedings of CHI'86 pp 192-197

Sato, H. (1974) Acoustic cues of female voice quality. Transactions A of the Institute of Electronics and Communications Engineers of Japan Vol. 57, pp. 23-30

Schwab, E. C., Nusbaum, H. C. and Pisoni, D. B. (1985) Effects of training on the perception of synthetic speech. In: Pisoni, D.B., Nusbaum, H.C., and Greene, B.G. Perception of speech generated by rule. Proceedings of the IEEE. Vol.73, No. 11, pp 1665-1675

Schafer, R.W. (1995) Scientific bases of human-machine communication by voice. Proceedings of the National Academy of Sciences of the United States of America. Vol. 92, No. 22, pp.9914-9920

Shneiderman, B (1998) Designing the user interface Addison Wesley

Silverman, K. (1985) What can be done to improve the intonation of synthetic speech? Proceedings of the 11th symposium on human factors in telecomunications. Cesson-Sevigne, France: CCETT

Simpson, A. (1986) Attitudes and responses to spoken interactions with computers. Unpublished postgraduate dissertation presented in partial fulfillment for the degree of MSc, University of Wales.

Simpson, C. A. (1981) Evaluation of synthesized voice approach callouts (Syncall). In: Morall, J., and Kraiss, K.F. (editors) Manned system design: methods, equipment and applications. Plenum, pp. 375-393

Simpson, C. A, Marchionda-Frost, K, and Navarro, T. N. (1984) Comparison of voice types for helicopter voice warning systems. Proceedings of the third Aerospace Behavioural Engineering Technical Conference. SAE Technical Paper Series 841611 SAE Aerospace Congress and Exposition, Warrendale, PA

Simpson, C. A, McCauley, M. E, Roland, E. F, Ruth, J. C, and Williges, B. H. (1987) Speech controls and displays' In: Salvendy, G. (editors) Handbook of Human Factors. John Wiley: Chichester

Stanton, N.A. and Baber, C (1997) Comparing speech versus text displays for alarm handling Ergonomics Vol 40 No 11 pp. 1240-1254

Stern, K. R. (1984) An evaluation of written, graphics, and voice messages in proceduralized instructions. Proceedings of the Human Factors Society. 28th annual meeting. Santa Monica

Sutcliffe, A. (1988) Human-Computer interface design Macmillan Education Ltd pp. 166-169

Tatham, M. (1993) Voice output for man-machine interaction In: Baber, C. and Noyes, J.N. (editors) Interactive speech technology. Taylor and Francis

Tucker, P. (1989) Human-computer vocal interaction. Unpublished undergraduate dissertation presented in partial fulfilment for the degree of BSc. University of Wales.

Van Ness, F. L. (1986) Human factors engineering of interfaces for speech and text in the office. IPO Annual Progress Report No 21

Van Ness, F. L. (1988) Multimedia workstations for the office. IPO (Eindhoven) Progress Report No 23 pp 104-111

Vine, D. S. G. (1998) Time-Domain Concatenative Text-to-Speech Synthesis. PhD Thesis. Bournemouth University.

Voiers, W.D. (1964) Perceptual bases of speaker identity. Journal of the Acoustical Society of America. Vol. 36(6), 1065-1073.

Vries, G. de and Johnson, G.I. (1997) Spoken help for a car stereo, an exploratory study. Behaviour and information technology Vol 16, No 2

Waterworth, J. A. and Holmes, W. J. (1986) Understanding machine speech. Current psychological research and reviews. Vol. 5, pp. 228-245

Waterworth, J. A. and Talbot, M. (1987) Speech and language-based interaction with machines. Ellis Horwood Ltd

Westall, F.A., Johnston, R.D. and Lewis, A.V. (1996) Speech technology of telecommunications. BT Technology Journal. Vol.14, No.1, pp.9-27

White, G.M. (1990) Natural language understanding and speech recognition. Communications of the ACM. Vol 33, No.8

Williges, B H and Williges, R C. (1982) Structuring human/computer dialogue using speech technology. Proceedings of the workshop on standardization for speech i/o technology. Gaithersburg MD: National Bureau of Standards. pp. 143-151

Wilson, G. (1996) Voices, Attitudes and emotions in speech synthesis. From:
http://www2.shef.ac.uk/uni/projects/vaess/documents/index.html

Wolf, C.G, Kassler, M. Zadronzny, W. and Ophyrchal, L. (1997) Talking to the conversational machine: an empirical study. In Howard, S. Hammond, J. and Lindgaard, G. (Eds) Human Computer Interaction. Interact 97 pp. 461-468