IMPACT OF DIFFERENT SPEECH INTERFACES OF PERSONAL DEVICES ON USERS' PERCEPTION

Mazen Wadea

A thesis submitted to
Auckland University of Technology
in partial fulfilment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2011

School of Computing and Mathematical Sciences

Primary Supervisor: Dr Judith Symonds

TABLE OF CONTENTS

TABLE OF CONTENTS i
LIST OF FIGURES iii
LIST OF TABLES iv
LIST OF ABBREVIATIONS v
ATTESTATION OF AUTHORSHIP vi
ACKNOWLEDGEMENTS vii
ABSTRACT viii
CHAPTER I INTRODUCTION 1
1.1 Background to the Study 1
1.2 Research Goal and Aims 3
1.3 Research Objectives 4
1.4 Research Question 5
1.5 Research Contributions 5
1.6 Definitions of Speech Interfaces Technology 6
1.7 Design of Research Equipment 8
1.8 Research Methodology 10
1.9 Important Literature 11
1.10 Thesis Outline 15
1.11 Chapter Summary 15
CHAPTER II LITERATURE REVIEW 17
2.1 Introduction 17
2.2 Interface Design for Speech Synthesis Systems 18
2.3 Applications of Speech Synthesis 26
2.4 Impact of Voice Characteristics on the Acceptability of Speech Interfaces 32
2.5 Perception of Different Types of Speech Interfaces 35
2.6 Conclusions 39
CHAPTER III RESEARCH METHODOLOGY 41
3.1 Introduction 41
3.2 Research Type and Paradigms 42
3.3 Research Method 43
3.4 Design, Development, and Testing of the Equipment 46
3.5 Research Design and Preparation of Data Collection 57
3.6 Listening Task Conditions and Procedure 60
3.7 Conducting the Data Collection Phase 63
3.8 Data Analysis 64
3.9 Generalizability, Reliability and Validity of the Findings 65

3.10 Conclusion 67

CHAPTER IV RESEARCH FINDINGS AND DISCUSSIONS 68

4.1 Introduction 68

4.2 Background of the Participants 69

4.3 Research Results and Findings 72

4.4 Analysis of the Findings 77

4.5 Discussion 90

4.6 Conclusions 93

CHAPTER V CONCLUSIONS 94

5.1 Introduction 94

5.2 Research Design 94

5.3 Research Findings 96

5.4 Contribution to Knowledge 97

5.5 Limitations of the Study 98

5.6 Recommendation for Future Research 99

5.7 Implications for Research 101

REFERENCES 103

GLOSSARY 106

APPENDICES 108

Appendix A – Demographic Questionnaire 108

Appendix B – Participant Invitation Letter 110

Appendix C - Participants Consent 113

Appendix D – Listening Task 114

Appendix E – Evaluative Questionnaire 115

Appendix F – Interview Questions 116

Appendix G – Stimulus Passage for Readers 117

Appendix H – The Evaluative Data 118

Appendix I – Full Interview Analysis of Couple Y 123

Appendix J – Full Interview Analysis of Couple Z 126

Appendix L – The VoiceTester Code Introduction 129

LIST OF FIGURES

Figure 1.1 - Overview of speech interfaces (source: author) 6

Figure 2.1 - Functional diagram of a general TTS synthesizer (Dutoit & Stylianou, 2003, p. 326) 19

Figure 2.2 - Example of natural language voice interaction with machines (Schroeter et al., 2002, p. 212) 27

Figure 2.3 - Framework of the research question (source: author) 35

Figure 2.4 - Dimensions which are relevant to development of speech interfaces (source: author) 40

Figure 3.1 - Research Methods Flowchart (source: author) 45

Figure 3.2 – VoiceTester prototype showing the frontend and the backend (source: author) ... 48

Figure 3.3 - Voice Memos Application 52

Figure 3.4 - Shows the actions taken on the signal output by Camtasia 53

Figure 3.5 - Speech Samples Integration Process (source: author) 54

Figure 3.6 - Screenshot showing the VoiceTester icon (source: author) 55

Figure 3.7 - Play and Pause interactive screens (source: author) 56

Figure 3.8 - The application alerts whenever a step is completed (source: author) 57

Figure 4.1 - Diagrammatical emergence of the category ‘FV perception’ for Couple X 83

Figure 4.2 - Diagrammatical emergence of the category ‘NV perception’ for Couple X 83

Figure 4.3 - Diagrammatical emergence of the category ‘CV perception’ for Couple X 83

Figure 4.4 - Emergent concepts in ‘FV perception’ derived from key point analysis of interview data in 3 couples (source: author) 87

Figure 4.5 - Emergent concepts of ‘NV perception’ derived from key point analysis of interview data in 3 couples (source: author) 88

Figure 4.6 - Emergent concepts of ‘CV perception’ derived from key point analysis of interview data in 3 couples (source: author) 89

LIST OF TABLES

Table 1.1 - Recording and categorizing the most important resources 11

Table 1.2 - Best practice of mobile TTS interface from related books....................................... 12

Table 2.1 - Strengths and weaknesses of mixed-voice and TTS-only interfaces...................... 22

Table 3.1 - Research question and interview questions relationship 58

Table 4.1 - Participants background 69

Table 4.2 - Subjects description.. 71

Table 4.3 - Key points analysis and codes from the data in Couple X 79

Table 4.4 - Emergence of concepts from the codes in Couple X data 82

Table 4.5 - Emergence of categories from the concepts in the data from 3 Couple 85

Table 5.1 - Link between the research objectives and research question................................. 95

LIST OF ABBREVIATIONS

Some of these are included because they are in common usage, others because they are big words and had been shortened to save space.

API	Application Programming Interface
AT	Assistive Technology
CASLT	Computer-Aided Speech and Language Therapy
CV	Computer-generated voice
CSS	Computer-Synthesized Speech
FV	Familiar voice
GMT	Goal Management Training
IDE	Integrated Development Environment
iOS	iPhone Operating System
MOS	Mean Opinion Score
NV	Natural voice
NZ	New Zealand
PDA	Personal Digital Assistant
S2ST	Speech-to-Speech Translation
SDK	Software Development Kit
SGDs	Speech Generating Devices
TBI	Traumatic Brain Injury
TTS	Text-To-Speech
UI	User Interface
VQM	Voice Quality Measurement
VTTS	Virtual Text-To-Speech

ATTESTATION OF AUTHORSHIP

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Mazen Wadea

ACKNOWLEDGEMENTS

First of all, I would like to thank the Health Research Council for funding this project. This research is my best academic achievements so far. Being a part of the GMT (Goal Management Training) project, especially developing the iPhone application, was the most exciting thing.

I wish to express my deepest gratitude to my supervisor, Dr. Judith Symonds, for her supervision, her encouragement and ongoing support, and especially her precious advices and knowledge were of great practical importance and guided me substantially throughout the writing of this thesis.

Most importantly, my gratefulness goes to the participants who participated in the survey, took the acceptance testing and the interview, also to the talented speakers who provided the stimuli, without whom this research would not have been possible.

I would also like to thank Dr. Andrew Ensor for setting up the apple developer programme license for the iPhone, where the application has been developed by.

Special thanks to the proof reader, (Ms Marg. Fortescue) for the advice and proof reading of my work.

My colleagues and friends at Auckland University of Technology were of great importance in their assistance, kindness, and encouragement throughout my studies.

My thankfulness goes to my family, who were of great importance in their support, prayers, and encouragement throughout this project.

This research study has been approved by the Auckland University of Technology Ethics Committee on 6 August 2010 AUTEC Reference number 10/118.

ABSTRACT

Because of Text-to-Speech (TTS) lacks both clarity and prosody of normal human speech, TTS sounds unnatural and is unpleasant to listen to. It is generally accepted using natural speech for a static prompts, whereas synthetic speech for dynamic content. However, most commercial applications on the market adopt mixing human speech and TTS within the same sentence and/or between sentences. But, this mixing approach led to inconsistent interface (Gong & Lai, 2001). So that, an immediate issue in the design of such speech interface is what type of speech should be used.

The goal of this project is to explore users' perception towards different types of speech in order to investigate the acceptability of personal speech interfaces. This study is aimed for the public users of mobile applications. This project explored redevelopment of the speech interface of the Goal Management Training (GMT) system based on results from testing different speech samples by the delivered VoiceTester mobile application. The VoiceTester application has been developed on the iPhone in this study, to facilitate the listening task therefore adding validity to the responses from participants by simulating environment of speech interfaces on personal devices. The contribution of this study is to provide some knowledge to the developers and health researchers about exploring the impact of different types of speech interfaces on users' perception. The findings are ultimately helpful to the Traumatic Brain Injury (TBI) patients. As the recommended software will assist them undertake activities with support to help prevent them from making errors (McPherson, Kayes, & Weatherall, 2009).

Six participants from different age groups have been chosen in the form of 3 couples, each couple construct of both genders. The examined types of speech are computer-generated voice (CV), natural voice (NV), and familiar voice (FV). The synthetic voices were generated by computer software, the natural speech samples were provided by two native speakers of New Zealand English, and the familiar voices for each couple were simply the recording of each other voices. Participants completed three times a post paper-and-pencil self-perception of task performance scales after each listening test, and then followed by an interview. The evaluative data were used to inform the participants and the researcher about the study and to guide the interview process. The main methods were largely qualitative through the use of semi-structured interviews to explore the users' perception about manner of speaking and the speaker of the three examined speech samples, as well as, to investigate the importance of the used voice characteristics. The interviews are analysed to discover themes and patterns related to an analysis framework structured from the literature review.

The findings revealed differences between three couples in their perceptions of different types of speech. The effect of gender was slightly present, as the subjects revealed a more positive attitude to their opposite gender. Both human voices, NV and FV, were acceptable to the majority of participants with many reporting improved mood and goal attainment. Participants found working with CV both challenging and rewarding. NV seemed particularly helpful in engaging people in the task process, while FV appeared particularly helpful in providing a structured framework for error prevention in attempting goal performance.

CHAPTER I INTRODUCTION

1.1 Background to the Study

Information technology is making our lives easier, in particular improving and bringing new mobility to people who are visually impaired and suffering speaking impairments to do their daily activities or needs. Recently, Neff, Kehoe, and Pitt (2010) stated that web designers face accessibility issues for visually impaired users, however, the same issues are applicable to sighted users accessing the web with small screen mobile devices. Both Moser and Melliar-Smith (2008) and Neff et al. (2010) stated the importance of auditory rich content for the web accessibility as it can facilitate the use of Web services from mobile devices with small screens and keyboards, and can also make the user's interaction with a mobile device more user friendly. Moreover, it made accessing the Web anytime and anywhere, whether at work, at home, or on the move easier. Text-to-Speech (TTS) has improved over several years. TTS synthesis has made personal applications and devices easier to access and use by different users. According to Dutoit and Stylianou (2003), there are several potential applications of TTS systems used for telecommunications services, language education, aid to handicapped people, measurement or control systems, communication between man and machine or between people speaking different languages, and fundamental and applied research on speech. These are discussed in the next chapter (Subsection 2.3).

Speech synthesis has long been a vital assistive technology tool and its application is significant and widespread. Assistive technology (AT) refers to any type of device that is designed to assist people with disabilities (Bunnell & Pennington, 2010; Liu et al., 2008; Symonds, Parry, & Briggs, 2007). AT enables greater independence to people with difficulties because these people will be able to perform tasks that they could not have performed previously. These people could have had difficulty in completing certain tasks. Others were unable to perform such tasks. TTS technology provides enhancements assisting them to interact with technology that is required to complete such tasks. Similarly, disability advocates state that technology is created without regard to disabled people. This creates pointless barriers to several people with disabilities. The developers of assistive technology will still complain that general designs are preferable to the need for AT and that they should continuously develop universal designs and concepts. Mullennix and Stern (2010b), reviewed the important issues for researchers and practitioners using computer-synthesized speech (CSS) as an assistive aid, where CSS refer to text that is synthesized and amalgamated on the computer and is also known as TTS. Individuals with speech disorders who cannot communicate normally with others are using CSS devices to ease communication. There are different speech disorders in the world and therefore researchers are always on the lookout on ways to improve the devices they have in order to accommodate everyone. Speech generating devices (SGDs) and Voice output Communication Aids are clinically referred to as user's devices. This is because the software contained in them is personally controlled. The user's devices are normally user friendly and this makes it easier for the person to interact with the device in a natural and intuitive way. All the above devices are speaking aids that experts have been using to conduct their research.

From the last decade, Gong and Lai (2001) proposed that speech perception is an important aspect of users' acceptance of interface through showing the impact of each synthetic speech, human speech, and mixing both types on users' performance, perception and attitude. Moreover, in terms of speech perception, there are numerous experimental methods to study the perception of acoustic analysis and synthesis of speech, clear speech, intonation and the role of intonation as suggested by (Sawusch, 2005; Uchanski, 2005; Vaissière, 2005). Where speech perception is the process of hearing, interpreting and understanding sounds of language ("Definition of perception," 2011). Therefore, studying the phonetics in linguistics and perception in psychology has to be done, in order to study perception of speech. Hence, the focus of this research will be exploring best practice for development of mobile phone application of TTS interface (see Table 1.1 and Table 1.2). An important design aspect of this project is that the researcher developed an application on the iPhone that basically runs computer-generated voice (CV), natural voice (NV), and familiar voice (FV) in an easy and simple interface to minimize the distraction of the interface and let the participants focus on the voice quality. The answer to this question will be covered in Chapter 4.

The issue of speech perception has been an area of discussion and research in several branches of speech interfaces, such as speech synthesis, and speech recognition. Consequently, each sub-field has been defined in a few words and focus of research interest. Since this research draws ideas from almost all existing perspectives on accent and gender of different speech types, the researcher considers it is important to elaborate on the relevant technologies in the context of each of the speech interface branches of speech synthesis and recognition, and identify the aspects that were put together to serve as the theoretical framework of reference for this study. Therefore, the following section discusses how speech interface technologies have been defined in light of the aforementioned areas of research. Following this, delimits the main research focus by progressively outlining the main area of the examined types of speech (CV, NV, and FV) which the researcher intends to inspect.

Attitudes toward CSS may differ depending on purpose such as Machine used to help people to communicate when they can't communicate for themselves vs. interface giving direction. People view CSS less favourably than natural human speech. However, the listener's reactions change dramatically when the disability status is indicated to the listeners, although this forgiveness can disappear as it depends on listener's patient and on purpose that CSS is used for as suggested by Mullennix and Stern (2010a). Therefore, this is considered an important issue for researchers and practitioners using computer synthesized speech as an assistive aid. Additionally, the incorporation of speech interfaces in the assisted living and augmented devices will be explained in the next chapter.

1.2 Research Goal and Aims

The goal of this project is to investigate the impact of different speech types on the utility, functionality, and acceptability of personal TTS applications. The outcome of this study will guide and inform a parent project that is developing a system to assist Traumatic Brain Injury (TBI) patients undertake goal management training (McPherson, Kayes, & Weatherall, 2009). The aim of this project is to inform redevelopment of the TTS interface for the Goal

Management Training (GMT) system based on the results from testing different speech samples by the developed VoiceTester application which is developed by the research and deployed on the iPhone for the purpose of this study.

1.3 Research Objectives

The information from this study will explore how types of speech should be used in speech interfaces. In general, this study is targeted for the public users of mobile applications as this information will help the researchers to understand the benefits of putting extra effort into the overall quality of TTS voices of personal applications. This prompts the establishment of several research objectives. The first main research objective is exploring the impact of different types of speech on the acceptability of personal TTS applications. To empirically test how these types of speech should be used in speech interfaces, a listening test experiment has to be conducted and then interviewing the listeners, as implied previously by Gong and Lai (2001). The interfaces have to be evaluated according to users' task performance, self-perception of task performance, and attitude towards the TTS interface. Therefore, the main research objective of this project is dealing with the users' perspectives and experiences:

> *To explore the users' perception towards different types of speech and measure its impact on the acceptability of personal speech interfaces.*
>
> *To identify tools and techniques used that potentially support exploring the perception of TTS interfaces, as well as determine their efficacy.*

However, the objectives of exploring the impact of different speech samples seem to be broad and need to be specified. Therefore, exploring the impact of users' perception of each type of speech samples (CV, NV, and FV). Accordingly, the research objectives are expanded to:

> *To investigate participants' perception and attitude of importance regarding the types of speech, as well as review their self-perception of task performance that allocated to related activities.*

To identify the voice characteristics used to explore its impact towards the different types of speech that needs to be shared.

To identify barriers that would hinder users from accepting the TTS interface, as well as their methods for overcoming these obstacles.

In addition, to identify the gap between what is being reported in literature and what is currently in practice, and to investigate the reasons for any differences.

To compare what is being reported in literature and actual practice, in order to identify gaps or provide supporting empirical evidence.

1.4 Research Question

Research question has been established in line with the research objectives described earlier. Thus, this research will address the following research question:

What are the users' perceptions of different types of speech of speech interfaces? This research question is divided to three issues:

- What is the users' perception towards the Computer voice (CV)?
- What is the users' perception towards the Natural voice (NV)?
- What is the users' perception towards the Familiar voice (FV)?

This research identifies characteristics of importance and investigates how human listeners perceive speech sounds and use this information to recommend an accessible TTS interface. The different speech types that are going to be considered in this research are computerized, natural, and familiar voices.

1.5 Research Contributions

This research is contributing to explore, compare and contrast, the impact of different types of speech (CV, NV, FV) on the user's perception of personal mobile applications based TTS interface. Moreover, it is making a contribution by developing the VoiceTester application of the iPhone that can be used in any future speech perception and performance studies, as it simulates an appropriate environment of personal devices. This study not only provides important guidelines for the design of VoiceTester-like speech interfaces but also sheds light on understanding how users respond to and interact with speech interfaces in general.

This research may explore users' perception of the voice characteristics of gender and accent towards the speech types occupied. And also, may reveal some of the causes of the research-practice gap in the literature.

1.6 Definitions of Speech Interfaces Technology

This section illustrates hierarchically the classification of speech interfaces technologies and some general applications, and briefly defines each of the related speech interface techniques. Figure 1.1, shows the two types of speech interfaces.

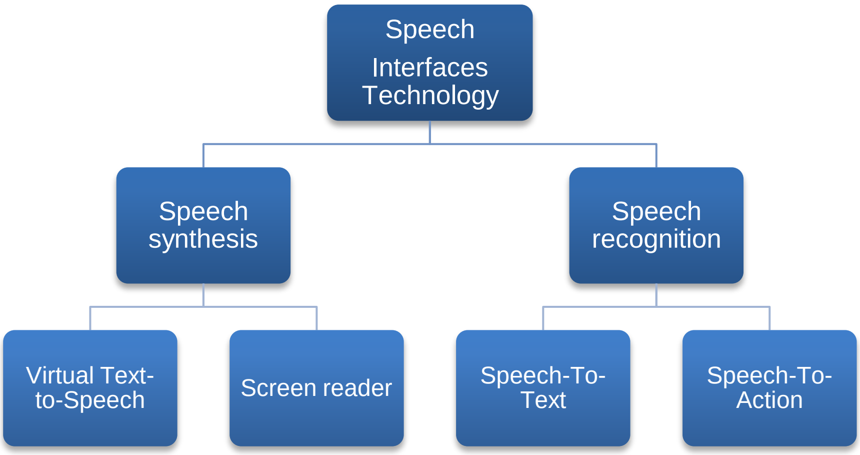


Figure 1.1 - Overview of speech interfaces (source: author)

There are two techniques of speech interfaces, which they are speech synthesis and speech recognition. Speech synthesis is the artificial production of human speech, which is also known as Text-to-Speech (TTS). Speech synthesis uses a system known as speech synthesizer, and this system can be implemented in software or hardware. A speech synthesizer performs several operations once the raw text is fed to it. Basically, it has two parts; namely, the front end, which normalizes raw text, and the back end, which converts the normalized text to speech; this functionality is described in the next chapter under Subsection 2.2.1.

Visual Text-to-Speech (VTTS) is the combination of facial animation with synthesised speech. It involves lip reading and facial animations into producing a speech. This method uses software that can convert facial expressions such as surprise, anger, and joy into speech through analyzing each expression. It also involves lip reading service where the software also determines speech through analysing expressions produced by the lips. Whereas, screen reader is the ability of a TTS interface to speak items on the screen for people with visual impairment by using the TTS technique such as Read Out Loud facility of the Adobe Reader and TTS add-on of web browsers that read aloud particular content elements is where the synthesizer converts these elements into speech.

These are examples of speech synthesis systems that use TTS technique to produce speech; it has several potential applications discussed in the next chapter under Subsection 2.3. On the other hand, speech recognition is the capability of a device to recognize the voice, which is exclusive such as a fingerprint of an individual. Here the speech recognition system is able to distinguish words only and not a person's voice features. The most popular use of speech recognition is in Speech-to-Text technique (text messaging and email services). Speech-to-Text is the ability of a machine to recognize words in a given voice and convert it to a text. According to Lamel and Gauvain (2003), it involves converting speech waveform into sequence of words. Speech-to-Action is the ability of a machine to convert a speech within the voice into action. For instance commanding a machine to perform an action such as a delete key in the computer involves pressing the key with the word delete and it deletes a word.

Many applications can benefit from voice input and output on a mobile device, including applications that provide travel directions, weather information, restaurant and hotel reservations, appointments and reminders, voice mail, and e-mail. Moreover, the use of a speech interface, along with textual, graphical, video, tactile, and audio interfaces, can improve the experience of the user of a mobile device as suggested by Moser and Melliar-Smith (2008). Generally, it is important to provide several modes of interaction, so that the user can use the most appropriate mode, depending on the application and the situation. Some modern speech interfaces have used both techniques within its functionality. For example, Speech-to-Speech translation (S2ST) system involves both speech interface techniques of recognition of a speech by a machine and converts it to a synthesised speech in another language. According to Dutoit and Stylianou (2003) this is normally used in areas where the machine is required to translate a certain speech from language to another. With regards to a recent study by Wester et al. (2010) on a system called Effective Multilingual Interaction in Mobile Environments which performs personalised S2ST, such that a person's spoken input in a language is used to generate spoken output in different language, while continuing to sound like the person who spoke the voice. A personalized system means that a system where the person who spoke the input speech sounds like the output synthesized speech. However, judging speaker similarity to the synthesized speech concerns a big issue for such system, as there is no system which can deliver a good cross-lingual speaker similarity yet.

1.7 Design of Research Equipment

In regard generating the TTS speech samples, there are several speak text editor programs such as SwiftTalker that comes within Cepstral package (Cepstral Corp., 2011). SwiftTalker has a number of options that can be used to fine-tune the intonation and rhythm as well as the volume, pitch and rate of the speech can be configured. Also, the speech synthesizer can be controlled by using a special phoneme alphabet.

Pre-recorded speech samples will be integrated within the developed iPhone application (VoiceTester) in order to explore the users' perception toward each type of speech. SwiftTalker is a speech synthesis software used to generate a short set of questions as audio files for both genders (David and Callie), and Voice Memos is an iPhone application that is used to record two native speakers from both gender. These are then generated and recorded as files which have been integrated within the VoiceTester application as CV and NV respectively. The FV will then be recorded by the Voice Memos application and will follow the same process to be integrated with the VoiceTester application, for more information about the process (see Figure 3.5). The gender of the first two speech samples, CV and NV, is controlled to reduce bias of gender (Lee, Nass, & Brave, 2000), so that the male voice for the male participant and vice versa regardless the gender of the FV which preferably will be from an opposite gender to form a couple. However, an Apple developer program license will be needed to deploy the developed application on the iPhone. Through this application the participants will be able to interact and listen to the types of speech samples (CV, NV, and FV). It has the option for both male and female voices for each of the three speech interfaces.

The volunteers completed a demographic questionnaire (see Appendix A – Demographic Questionnaire). In the demographic questionnaire, volunteers could indicate their willingness to participate further in the study. Each participant will be briefed according to the general test protocols ensuring that they have all the necessary information about the study (Appendix B – Participant Invitation Letter) and have given a consent form (Appendix C - Participants Consent). Also, the participants would be able to gain a report of the results, as they have the ability to indicate their interest on the consent form. Each participant will be summarized about the developed application and the specific listening test that they are about to complete. A listening test, according to Uchanski (2005), is an approach of testing specific speech sample or recording through human listening experiments. The listening task will be undertaken by the researcher, who will be the facilitator during the listening test. Throughout the listening test, the researcher will encourage the participant to 'think aloud' and a general note will be taken. Each of the participants will be informed that he/she will be interviewed regarding their responses towards each of the listened speech samples. The interviews will be recorded and later transcribed. The listening task will be based on similar protocols of usability testing, suggesting by Carter (2007), in particular, to ask participants specific questions relating to a specified aspect in order to guide the listening task. Some of the expressions commonly used with positive feedback are:

- *What are you thinking?* Has to be asked when the participant takes more than the normal time to do a specific task.
- *What you are experiencing?* Has to be asked when the participant does something and something come up as this would remove the fair of not being in line.
- *What's happening?* Has to be asked when the participant does something different to what I'm expecting or what the tasks are asking the donor to do. This statement would generate confidence to let the donor continue and then give feedback.

- *Something up, eh?* This has to be asked with an emotive tone to encourage the participant to focus more. Basically it needs to be asked before the donor tries to leave or give up letting the user to stay in line.

1.8 Research Methodology

This research is considered exploratory, since the aim of this study and the limited knowledge available regarding the user perception of the examined speech types. This study will use mainly a qualitative research approach, to collect and analyse the interview data in order to identify characteristics of importance, since it is the peoples' perceptions and experiences of the phenomenon suggesting therefore an interpretive approach has been chosen. A basic demographic questionnaire will be used as a selection tool for six participants in the form of three couples, as well as to get the background of the participants include participants' gender and their preferred FV. Evaluative scales will also be used to inform the researcher and the participants about the listening task and to facilitate the process of the interview process.

In regards the data collection and data analysis approaches planned for this project, three types of data will be collected from the six participants in the form of demographic data, evaluative data and interview data. Following the briefing session, each participant will then undertake the task experience of listening to three speech samples. After each listening test, a paper-and-pencil questionnaire rating scales is to be filled by the listener in order to evaluate the users' self-perception of task performance. The focus is on gaining insights to guide the interview questions prior to a more rigorous investigation. A semi-structured interview will be used because of its flexibility and ability to combine open questions with investigations to gain deeper understanding of the phenomenon. Once the participants have experienced all the three speech samples, they will be interviewed to collect their perception, experience, and attitude toward speech interfaces.

The recorded interview data from each of the participants will be transcribed and prepared for coding by identifying key ideas. The process of analysis starts with content analysis, then identifying, coding, categorizing, classifying, and labelling the primary patterns in the data Patton (2002). This is a qualitative approach to analyse the interview data collected from the participants, which will result in categories of feedback from the participants.

1.9 Important Literature

In terms of TTS interfaces, Gong and Lai (2001) described designing TTS interfaces of the virtual assistant and discuss how users respond to and interact with that interface. Besides, Lee et al. (2000) justified and modified the impact of gender of speech synthesis on the user's perception, which relates to its conformity.

During this chapter, a checklist was conducted for reading the literature to realize the purpose of a literature review by comparing the purpose of the study, the conducted research, findings, the limitations and the future work with other studies and my own research. A general, analytical approach was applied by recording the author and date, topic, and significance of the study in order to record and categorize the most important previous studies and resources from the past 10 years as shown in Table 1.1.

Table 1.1 - Recording and categorizing the most important resources

Source	Subject	Significance
• Voice characteristics		
Lee et al. (2000)	Gender stereotype of Computer-generated speech	Male genders prefer male computerized voice, and female genders prefer female computerized voice. However, there are some exceptions
Bayard and Green (2005)	Evaluating English Accents Worldwide	The Evaluating English Accents Worldwide project is a multinational research project collaborating how different English accents are perceived cross-culturally.
• Perception of different types of speech		
Gong and Lai (2001)	Impact each of Synthetic speech, human speech, and mixed type on users' performance, perception and attitude.	Mixing types lead to inconsistent interface. Interface with TTS-only may contribute in the consistency of the interface that led to better users' understanding and interaction, generally accepted, using human speech for a static text and

		synthetic speech for dynamic content.
• User interface for individuals with cognitive impairments		
Liu et al. (2006)	Developing interface of the indoor way finding system for individuals with cognitive impairments	Recommended using image, audio, and text messages with both directions and prompts for individuals with cognitive impairments by basing on results from the pilot tests
Sánchez and Aguayo (2007)	Development of mobile messenger for blind people.	Found that the end users prefer to listen to natural voice instead of synthetic voice, even though the quality of the synthetic voices is good and accepted
• Methodology		
Patton (2002)	Qualitative data analysis	The recorded interview data from each of the participants will be transcribed and prepared for coding by identifying key ideas
Carter (2007)	Usability testing protocols	The listening task will be based on similar protocols of usability testing by using recommended expressions encouraging the participants to think aloud
Collis and Hussey (2009)	Research paradigm and methodology	Qualitative research design for the data collection and data analysis approaches

Another analytical approach was applied by recording the source and title, and significance of the study of each relevant chapter under each book in order to record and categorize some previous books, from various themes and authors as presented below in Table 1.2. The frequency of such themes in the theory and practice of TTS interfaces has increased in recent years.

Table 1.2 - Best practice of mobile TTS interface from related books

Source and Title	Significance
• Improvements in speech synthesis / edited by Keller et al. (2002)	
Keller, 2002 'Towards greater naturalness: future directions of research in speech	The quality of speech synthesis systems has been improved dramatically through improving its intelligibility and naturalness

synthesis’	
Beskow, Granstrom, and House, 2002 ‘A multi-Modal speech synthesis tool applied to Audio-Visual prosody’	The perceived feeling of naturalness about visual speech synthetic systems is mainly based on adding facial expressions. Obviously, transmitting non-verbal information also, contributes to the liveliness of the face
Flach, 2002 ‘Interface design for speech synthesis systems’	The speech synthesis applications proved its acceptability and covered a wide range of areas, therefore requiring speech synthesis systems with various characteristics
• The Oxford handbook of computational linguistic / Edited by Mitkov (2003)	
Dutoit and Stylianou, 2003 ‘Text-to-Speech synthesis’	High-quality TTS systems have several potential applicaitons, such as in telecommunications services, language education, aid to individuals with disabilities, vocal monitoring, mulitmedia, man-machine communication, and fundamental and applied research
• The Handbook of Speech Perception / Edited by Pisoni and Remez (2005)	
Sawusch, 2005 ‘Acoustic Analysis and Synthesis of Speech’	Synthesizing natural voice by rule would be practical in studies of spoken language processing rather than natural speech since the generated stimuli is controllable
Uchanski, 2005 ‘Clear Speech’	What is clear speech? Is age or hearing status of the listener important? Is there a benefit in typical listening environments? Does language experience matter?
Vaissière, 2005 ‘Perception of Intonation’	Studying and understanding intonation is considered as a complex process. At the time of writing, there is no speech synthesis system that can produce perfect natural attitudinal and emotional nuances carried by intonation.
• Handbook of Research on User Interface Design and Evaluation for Mobile Technology / Edited by Lumsden (2008)	
Schmidt-Nielsen et al., 2008 ‘Speech-Based UI design for the automobile’	The Speech-In List-Out interfaces (SILO), can result in lower driving interference than the menu driven interface. Whereas SILO is speech-based user interface that returns a shortlist of responses from which the user can select

Moser and Melliar-Smith, 2008 'Voice-Enabled user interfaces for mobile devices'	Providing multiple ways in which the users can interact with the applications on mobile devices brings a new level of convenience to the users of those devices
• Computer Synthesized Speech Technologies: Tools for Aiding Impairment / Edited by Mullennix and Stern (2010)	
Mullennix and Stern, 2010b 'Important issues for researchers and practitioners using computer synthesized speech as an assistive aid'	People perceive CSS less satisfactorily than natural human speaker. However, this would be reversed when people indicate the disability status of the CSS users, but still this fact depends on certain conditions
Bunnell and Pennington, 2010 'Advances in CSS and implications for assistive technology'	Using CSS devices that assist people with speech disorders to communicate with others around them. These assisted technology devices make use of speech output. The quality of output given is very important, as this is what determines the efficiency with which a person communicates with another. For a CSS to be applicable, it has to be intelligible, natural and must have technical considerations
Neff, Kehoe, and Pitt, 2010 'Considering the perceptual implications of auditory rich content on the web'	The importance of auditory rich content for the web accessibility as the issues are applicable for visually impaired users and for sighted mobile users who are accessing the web with mobile devices
Koul and Dembowski, 2010 'Synthetic speech perception in individuals with intellectual and communicative disabilities'	By repeatedly exposing individuals with speech impairments to synthetic speech, the level of speed and accuracy is increased. However, every time the listening task became complicated, the level of accuracy of the synthetic speech reduced
Saz et al., 2010 'The use of synthetic speech in language learning tools'	Use of CSS in the development of Computer-Aided Speech and Language Therapy (CASLT) tools for the improvement of the communication skills in handicapped individuals
Mullennix and Stern, 2010a 'Attitudes toward computer sythesized speech'	Generally, people perceive CSS less favourably than natural human speech. However, the listener's reactions change dramatically when the disability status is indicated to the listeners, although this forgiveness can disappear as it depends on listener's patience and the purpose for which CSS is used

• iOS Human Interface Guidelines - User Experience (2011)	
Apple Inc., 2011 'MobileHIG'	Using the standard user interface (UI) elements that provided and follow the recommended usage would determine the behaviour and default appearance of UI views and controls

1.10 Thesis Outline

This thesis consists of five chapters:

- **Chapter 1 – Introduction:** Introduces the background of the topic and states the research foundation for conducting this research.
- **Chapter 2 – Literature review:** Provides a review of previous works relevant to this research in order to identify gaps this study will address.
- **Chapter 3 – Methodology:** Motivate and justify the research methodology and the research methods occupied in conducting this research, as by clarifying the methods used for collecting data and the techniques used to analyse the collected data.
- **Chapter 4 – Research analysis, discussion and findings:** Presents a summary of analysis of the findings, states the emergent model in addition to answering the research question.
- **Chapter 5 – Conclusions:** Presents the conclusion of the research and reflects on the design and execution of this study that provides some shortcomings, some possible extensions to the research and some lessons learnt.

1.11 Chapter Summary

This chapter provides the reader with the background and foundation of this research with its goal, aim, and objectives. Also, it structures the research question according to the research objectives. This chapter presents contributions to the developers and researchers of TTS interface of personal mobile devices. Moreover, this chapter covers a brief introduction on the research methodology of this research, and some related literature to the aim of this research. Lastly, it presents an outline of the thesis structure.

Another interesting dimension that this study offers is a definition of the types and some common applications of speech interfaces. That is, this research examines how perceptions towards different types of speech can be explained by the voice characteristics closeness (similarity of gender and familiarity of accent) between the participants and those of the speech samples providers. This specific area of research has a wide array of perspectives and can be accessed from a diversity of fields.

The next chapter reviews the major studies that have been done in the area of TTS interfaces, and more specifically focuses on those that have examined users' performance, perception, and attitude towards different speech types.

CHAPTER II
LITERATURE REVIEW

2.1 Introduction

This chapter reviews the relevant literature in order to demonstrate theory and practice related to the study and evaluates the existing body of knowledge on this topic to support the research question. Literature review is “a critical evaluation of the existing body of knowledge on a topic, which guides the research and demonstrates that the relevant literature has been located and analysed” (Collis & Hussey, 2009, p. 100). Throughout this chapter, the literature reviews the impact of different speech types on the acceptability of TTS interfaces according to user’s perception, as well as, shows the contribution of voice characteristics of gender and accent into the acceptability of speech interfaces in order to guide the implementation of these characteristics. In the conclusion of this chapter, different factors that may have impact on the development of TTS interfaces were identified and presented in a general framework leading to the research question. Consequently, the developed framework was used to guide the selection of the methods and techniques. These methods and techniques are going to be used for gathering and analysing data.

The next section, Section 2.2, discusses the interface design for speech synthesis systems in the context of presenting the functional diagram of a general TTS conversion system, an overview of the improvements on TTS technology, an introduction of voice quality measurement. Section 2.3, describes several potential applications of high-quality TTS synthesis systems. Section 2.4, shows the impact of voice characteristics of gender and accent on the acceptability of speech interfaces. Section 2.5, sets the framework of the literature review by presenting the aspects of the research question. Finally, Section 2.6, concludes the chapter with a summary of the main studies to identify gaps or supporting empirical evidence in the literature and ending with introduction to the research methodology chapter.

2.2 Interface Design for Speech Synthesis Systems

Speech synthesis is the production of speech by machines and is known as TTS. A TTS system converts typed text into spoken voice output. The ultimate goal of a TTS system is to read any introduced text out loud (Atkinson, 2008; Flach, 2002). The introduced text is any text including numbers, abbreviations, acronyms, and idiomatic, in any format (Dutoit & Stylianou, 2003). Speech synthesis is not a sequence of pre-recorded words. Dutoit and Stylianou claim that recording and storing all the words of the related language is impracticable and ineffective. While this is a brilliant idea, in practice, it is not possible to record and store all the words of the focus language. In the context of TTS synthesis, producing a natural-sounding speech is to generate continuous and coarticulated speech. Coarticulation is the process of generating sequences of articulators' movements by regularly altering them to reduce the necessary efforts within a given context.

The interface design for speech synthesis systems is discussed in the following subsections. The next subsection, Subsection 2.2.1, presents the functional diagram of a general TTS conversion system and each part has been briefly described. Subsection 2.2.2, discusses the development and improvement in the generated signal of TTS speech synthesis systems in regard of naturalness and intelligibility reported in the field. Subsection 2.2.3, introduces voice quality in terms of perception, and the implementation of voice quality measurement. Subsection 2.3, the numerous potential implementations of high-quality TTS speech synthesis systems are discussed.

2.2.1 Overview of TTS Processing

According to Dutoit and Stylianou (2003) a general TTS system composes of two parts (see the Figure below): a natural language processing module (NLP) consists of four sections (preprocessor, morphosyntactic analyzer, phonetizer, and prosody generator); NLP generates a phonetic transcription with the required prosody from the introduced text. And the other part being a digital signal processing module (DSP) consists of one main section which is the speech signal synthesizer; DSP transforms the received symbolic information into speech.

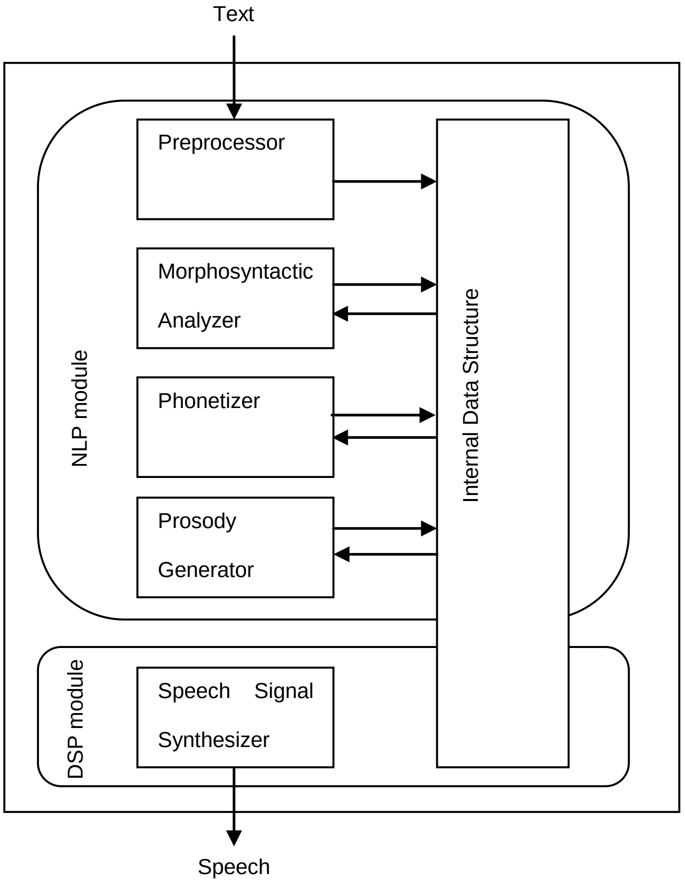


Figure 2.1 - Functional diagram of a general TTS synthesizer (Dutoit & Stylianou, 2003, p. 326)

Dutoit and Stylianou (2003) briefly defined each of the above commponets:

- **Preprocessor:** It is identified as the front-end of the NLP module. In other words it is called text normalization, as it is responsible for finding the end of sentences in the introduced text. It sorts the input text into lists of words and then saves them in the internal data structure.
- **Morphosyntactic analyzer:** Provides speech tagging and arranges the input text into related categories of words.
- **Phonetizer and prosody generator:** Phonetizer is pronouncing the series of phonemes. And, prosody generator is identified as the audible modifications to the speech characteristics of pitch, loudness, and syllable length. However, Lee, Nass, & Brave (2000) stated its advantageous for the flexibility in modifying different voice parameters of speed

rate, intonation, and rhythm that made TTS gender stereotype closer to the characteristics of natural, human speech.

- **Speech signals synthesizer:** After figuring out the phonemes and prosody by the NLP module, the speech signal synthesizer of the DSP module will take care of finishing and creating the speech samples. A digital-to-analogue converter could play these samples. There are two primary technologies for generating synthetic speech. First is the synthesis by rule as formant-based and articulation-based rule systems. This method means that it is based on a set of rules to generate high-quality speech naturalness; however, the improvements to the naturalness are still not as excellent as human speech. Second is the concatenative synthesis by data based unit concatenation system; this method concatenates auditory units of natural speech to generate signal. Each technology has strengths and weaknesses, and which approach to use depends on the intended uses of a synthesis system (both technologies have been discussed below). However, Formant synthesis and concatenative synthesis are the two dominant approaches to CSS devices as suggested by Bunnell and Pennington (2010).
  - Formant-Based speech synthesis has the ability to produce a high-speaking rate without reducing the utterance intelligibility. Therefore, the synthesized speech is used as reading assistance for persons who are visually impaired. Furthermore, it generates stimulus which could be used for studying the perception of speech. For both reasons, it is commonly used in a broad range of applications as in speech practising, hearing support, speech recognition, and additional refining to the synthesis speech.
  - Concatenative synthesis is based on the concatenation of segments of recorded speech. Generally, concatenative synthesis produces the most natural-sounding synthesized speech. Unit selection synthesis uses large database to synthesize and store the voice recording are more successful than others. This is because their level of accuracy is higher than those with smaller databases. However, for concatenative synthesis such as diphone synthesis that uses a little database of documentation speech, it is unlikely that all of the units needed to create natural-sounding prosody would be presented in the database. In that case, the structure may well use methods to change prosodic features in the initially recorded speech to compel it to match a

detailed prosodic formation. This means that, phonetic content is concatenated, whereas the prosodic configuration of the sound is synthesized by altering the naturally recorded speech.

2.2.2 Improvements in Speech Synthesis Technology

In the last decade, the quality of speech synthesis systems has been improved dramatically in terms of naturalness and intelligibility, but it does not match that of human speech (Duggan & Deegan, 2003; Dutoit & Stylianou, 2003; Keller, 2002; Lee et al., 2000). In addition, Lee et al. (2000) converse that TTS interfaces have become increasingly acceptable and popular to users through delivering a high-quality output with a high level of naturalness and intelligibility. Naturalness is the similarity of the generated voice to human speech, while intelligibility is the level of users understanding of synthetic utterances. There are, however, several developments over the past few years. When it was invented, the main target group was people with reading difficulties. However, the cost-performance ratio causes speech synthesizers to become cheaper and more accessible to the people, so that, they have increased from commercials to teaching in classrooms and educating the public in several ways.

Many commercial companies, currently, use pre-recorded human speech for fixed prompts only and TTS for dynamic content. However, TTS is considered the way of generating prompts for fixed and dynamic text, because recording human speech as prompts for a dynamic text is not practical, if not impossible. Thus, fixed prompts could be CSS or pre-recorded human speech, depending on the cost and speech quality. However, as pointed out by (Ralston, Pisoni, & Mullennix, 1995 as cited from Gong & Lai 2001), listening to synthetic speech is unpleasant as it is unnatural. The ideal speech synthesizer is both natural and intelligible. Although, Gong and Lai (2001) argue, a speech synthesis with a high level of naturalness and intelligibility is feasible through mixing natural human speech with TTS whenever possible. Moreover, stated there are two ways to generate a speech in dynamic text. Firstly, mixing human speech vs. TTS, and secondly, using TTS only. Each technique has strengths and weaknesses related to overall task performance, self-perception of task performance, clarity and liking of TTS interfaces, repetition of email messages and calendar listings, and ease of use the system (see Table 2.1).

Table 2.1 - Strengths and weaknesses of mixed-voice and TTS-only interfaces

	Mixed natural voice vs. TTS and TTS-only
Task performance	Users had poorer task performance when they interacted with the mixed-voices than they did with TTS-only, as well as less effort required for the mixed interface.
Self-perception of task performance	The mixing of TTS and human voice had better self-perception of task performance but with negative impact on task performance. As the participants though, they did better and had more positive attitudinal responses.
Clarity and liking	Was significantly and positively correlated with self-perception of task performance, participants though TTS in the TTS only interface clearer than in the Mixed interface and liked the TTS more in the TTS only than in the mixed interface.
Repetition	With TTS only, users had greater number of repeating the task. However, users in mixed-voices interface were reluctant to complete the task because of the inconsistent interface. It means that the system is costly to the users.
Ease of use	Using TTS only in both fixed and dynamics helps to maintain consistency in the interface therefore, facilitating smooth interaction of the user with the interface. Also, users stay focused because it is the only form of speech in the interface.

However, the chosen approach depends also on consistency. Gong and Lai have showed in their study that mixing human speech and synthetic speech is not the optimal solution to achieve a similar perception as for only natural voice, as the big difference in their quality would cause inconsistent speech interface. The problem, moreover, mixing these two types of voices would be hard to modify by the developers. Nevertheless, inconsistent interface may affect the user's interaction and perception with the interface. It seems that interface with TTS-only may contribute in the consistency of the interface leading to a better user's understanding and interaction.

However, Dutoit and Stylianou (2003) stated that the naturalness level is regarded as unimportant in most applications of telecommunication services, so that, speech synthesis gets accepted and becomes popular among customers. Alternatively Flach (2002) discusses, high-quality speech synthesis systems are required with various characteristics to cover a wide range of areas: Telecommunications services, Language education, Aid to persons with disabilities, Vocal monitoring, Multimedia, man-machine communication, and Fundamental and applied research. These areas are further discussed in this chapter (under Section 2.3). Moreover, Gong and Lai (2001) argue, synthesized speech even sounded robotic but the potential uses outweigh them.

2.2.3 Voice Quality Measurement

Lee et al. (2000) claimed that human voices are distinguishable from the synthesized voices by the bizarre influences that are generated through synthesizing a speech. This is because despite the many developments over the years, speech synthesizer has not met the quality of a natural human voice.

Voice quality, according to Ibrahim (2006), is the perceived result of the immediate judgment by a subject on a specific live or recorded spoken language about his/her experience, motivation, and expectation of voice quality. Measuring voice quality is based on two factors: similarity to human voice and ability to be understood. Voice quality measurement (VQM) is the customer's perception of a service or device. In regards of voice communication systems, it is the average of responses from the subject's experience of a voice telephony services. There are, however, subjective measures and objective measures to measure the voice quality. The subjective measures are tests that are performed to measure the voice quality by using a number of people's opinions and experiences for different network conditions. The voice quality is performed by measuring the average quality perceived by those people whom are selected for tests in the service/system. And the objective measures are the same as subjective measures but the use of an algorithm to compute the MOS value is done by using a small portion of the speech in question instead of selecting a number of people. Mean opinion score (MOS) is the average score of responses by subjects for specific rated condition (Ibrahim, 2006).

In regards to the needs for voice quality measurement, there has been much debate on the topic of what is the best way to measure quality of a voice. However, Ibrahim (2006) addressed a method in measuring voice quality, through understanding the customers' perception. Moreover, the needing of the state-of-the-art VQM algorithms is vital for the service providers to evaluate the speech quality automatically. However, price and quality are the most crucial aspects customers consider when opting for a purchase or to subscribe to such a service or product, therefore, service providers need VQM to reduce the price and respond faster to customers' needs in fixing or maintaining any issues.

Improvements of speech synthesis depend on improving the signal quality, and coherence and naturalness. The synthesized speech differs from human speech in prosody and clarity (Gong & Lai, 2001). There are, however, high-quality speech synthesis systems that almost sound real, but still there are many people who perceive it as a machine voice. Some authors conclude that listeners are unable to describe what exactly contributes into judgments of naturalness (Keller, 2002; Lee, et al., 2000). So that, it is not easy to list the characteristics that naturalness of speech synthesis system depends on. As pointed out by (Sluijter et al., 1998 as cited from Keller, 2002), an approach to measure naturalness, listeners were asked to answer 11 questions on 5-point rating scales. These questions were about general quality, ease of comprehension, comprehension problems for individual words, intelligibility, pronunciation and/or occurrence of deviating speech sounds, speaking rate, voice pleasantness, naturalness, liveliness, friendliness, and politeness. Also, it was impossible to have participants agree to the naturalness of speech without first getting acceptability. Therefore, both acceptability and naturalness were incorporated in their approach. However, it is possible to modify and edit these parameters. In contrast, Tatham and Morton (2005) believe that the concept behind naturalness of speech synthesis is about reflecting how speech was generated. Naturalness is known as a variable of speech synthesis characteristics, if we concentrate on perception when generating speech. Focusing on the listener's perceptions as well as the general environment that the speech is perceived would produce speech that is appropriately more natural than 'just natural' (Tatham & Morton, 2005).

Ibrahim (2006) claimed that VQM is the most accurate listening test. VQM is done through the subjects' opinion after perceiving speech signals processed through various distortion conditions.

The intelligibility and naturalness, according to Wester et al., (2010), are the key properties of evaluating acceptable synthetic speech. However, Bunnell and Pennington (2010) beleive a CSS to be applicable must have technical considerations as well, below discussed and compared with different views of different researchers.

- **Intelligibility:** is the level of users understanding of synthetic utterances. The SGDs intelligibility relies mostly on the quality of synthetic speech. Whereas, measuring the intelligibility of a particular speech sample or recording can be through using human listening experiments as suggested by Uchanski (2005). However, to evaluate intelligibility according to Wester et al., (2010), the subjects have to be asked to transcribe semantically unpredictable sentences.
- **Naturalness:** is the similarity of the generated voice to human speech. Over the years, research has proved that sound produced through the data base system is more natural than the sound produced using the rule based system. However, Sawusch (2005) stated that synthesising natural voice by rule would be practical in studies of spoken language processing than natural speech since the generated stimuli is controllable. Naturalness varies depending on the voice sample quality, accuracy and consistency. This is because all human beings are distinctive in their manner of talking. In line with this fact, Sawusch (2005), mentioned that listeners perceive naturalness of a voice upon similarity to an actual voice, as human exposed to many voices. Listeners would perceive the synthetic stimulus in a different manner of phonetic distinction than the natural speech. Moreover, this can be shown through comparing the perception with synthetic stimuli to those found with natural speech. However, to evaluate naturalness and similarity to the target speaker according to Wester et al., (2010), a mean opinion score (MOS) of 5-point rating scales has to be used.
- **Technical constraints:** The used technology plays a key role in the development of AT. This is because with the emergence of new and better technology, the AT improves. This means that better devices come up with time leading to devices that are more sophisticated.

Therefore, after much research was conducted, it was proven that the data based synthesis system was the best technology used to create CSS devices. This is because scientists have found ways of improving software that has been used to develop these devices.

2.3 Applications of Speech Synthesis

For many years now, there have been, however, numerous potential applications of high-quality TTS synthesis systems:

- **Telecommunication services**

Also known as the telephone-based speech applications by Gong and Lai (2001), these applications give users the capability to hear textual information over the telephone as well as those that are visual textual information, for example, web pages. Texts vary in length from normal messages to faxes. Moreover, Dutoit and Stylianou (2003) describe how this can be done through incorporating TTS systems and telephone services. This information can be accessed through either voice recognition or the keypad of the telephone.

There are some applications established by AT&T providing speech synthesis with a reasonable intelligibility and a moderate naturalness. A naturalness level is noted as unnecessary in most applications; therefore, speech synthesis is accepted and popular among customers (Dutoit & Stylianou, 2003). An example of such application is the integrated messaging service that gives users the ability to listen to email over the telephone. And also Gong and Lai (2001) said that a speech based application, functions as a virtual-assistant system to access business data, such as email messages, voicemail messages, and calendar entries over the telephone Also mentioned are possible uses, such as retrieving information about stocks, weather, news, nearest restaurant, or assisting with phone calls. Furthermore, according to Dutoit and Stylianou, the Telephone Relay Service that was introduced by AT&T uses the text-to-voice and voice-to-text to enable people with speech or hearing impairments being able to efficiently understand a telephone conversation.

The diagram below illustrates an example of a new technology that uses TTS is the Voice Enabled Telecom Services. A conversation like “Welcome > to> New Zealand Airways>

Flights" where the quotation marks are the slots that are filled from the inventory of allowed fillers, which have been recorded previously. These mixed systems play a crucial role in first rate yet easy natural language dialogue systems.

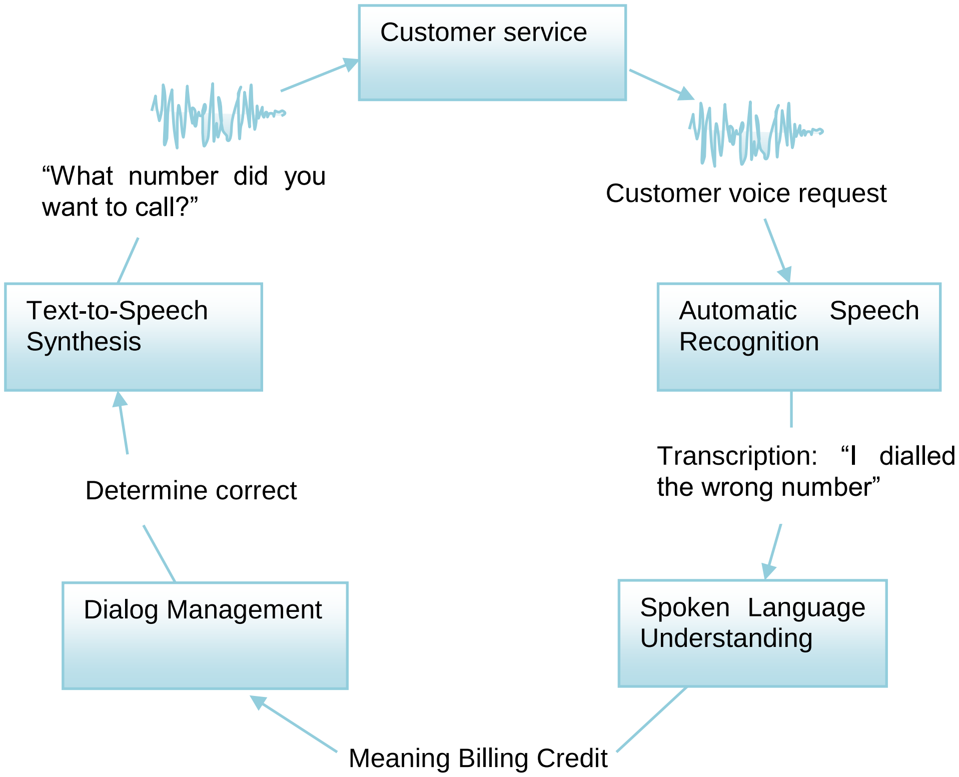

Customer service
Customer voice request
Automatic Speech Recognition
Transcription: "I dialled the wrong number"
Spoken Language Understanding
Meaning Billing Credit
Dialog Management
Determine correct
Text-to-Speech Synthesis
"What number did you want to call?"

Figure 2.2 - Example of natural language voice interaction with machines (Schroeter et al., 2002, p. 212)

- **Language education**

In recent times, TTS is being used as second languages learning support through aided learning applications with a high-quality TTS synthesis installed on a personal device such as Personal Digital Assistant (PDA), mobile or computer. It is generally acceptable to use compatible disks of pre-recorded conversations for learning new languages. However, TTS systems could perform as a native speaker who is accessible anytime and anywhere as described by Dutoit and Stylianou (2003). And also, Keller (2002) mentioned that speech synthesis systems used in second language learning skills could be performed by a native speaker who's accessible anytime and anywhere. TTS is used in literacy training. It is a fact that speech synthesis systems are not human, and learners perceive it as the unbiased system. In addition, TTS has been implemented in educative talking toys. According to Pandzic (2002), facial animations with the use of TTS are increasingly becoming a teaching aid. Teachers and other corporate entities use it as an enlightening tool in matters related to business or education. In addition, according to Saz et al. (2010), using the synthetic speech as assistive technology is described in the following section.

- **Aid to individuals with disabilities**

Recently, CSS is used in the development of Computer-Aided Speech and Language Therapy (CASLT) tools for the improvement of the communication skills in handicapped individuals. CSS is strongly required in these tools for two reasons: providing alternative communication to users with different impairments and reinforcing the correct pronunciation of words and sentences. Different possibilities have arisen for this goal, including pre-recorded audio, embedded TTS devices or talking faces (Saz et al., 2010). Dutoit and Stylianou (2003) describe TTS systems as extensively assisting handicapped individuals with speech, hearing, or visual impairments. In today's world, speech or hearing-impaired persons are able to have an effective telephone conversation as the mentioned telephone-relay service that provides information for visually impaired people. TTS can be helpful to individuals who have speaking difficulties by using a particular keyboard with excellent word prediction software. Although, Mullennix and Stern (2010) mentioned that patients who use CSS devices are constantly faced with social stigma that may at times be discouraging to them. This is because the society lacks the proper understanding of CSS. This problem needs to be addressed in order to ensure the CSS users are in an environment where the people around understand them and what they are going through. Specialists hope that one day, they will acquire a technology that will enable CSS devices for sounding natural and also be easier to use. This will enable CSS users to live almost normal lives in their community without any stigma (Mullennix & Stern, 2010).

In fact, web designers face accessibility issues for visually impaired users; however the same issues are applicable to sighted users accessing the web with small screen mobile devices (Neff, Kehoe, & Pitt, 2010). Moreover, TTS has been employed to help visually impaired people to access written information over the telephone. At the time of writing, TTS systems are also used to assist people with reading difficulties, such as pre-literate youngsters, elderly, and visually impaired. There are numerous studies where TTS technology is used to assist blind users to be connected with the world-wide-web through the internet. According to

Sánchez and Aguayo (2007), a virtual keypad consists of only nine buttons that enable TTS technology designed for blind users through applying usability tests on their modules to provide a messenger system especially customized to users with visual disabilities.

In regards assisted living technology, as claimed by Bunnell and Pennington (2010), AT devices enable these people to perform tasks that would ordinarily be difficult for them to perform. One such device is the CSS devices that assist people with speech disorders to communicate with others around them. These AT devices make use of speech output. The quality of output given is very important, as this is what determines the efficiency with which a person communicates with another. However, it is mentioned by Mullennix and Stern (2010b) that the greatest obstacle facing CSS over the years is the fact that specialists were unable to make CSS voices sound natural. This hindered listeners from understanding what was being said hence the usability of the CSS device became questionable. However, over the years, the intelligibility of CSS devices has improved with the emergence and improvement of new and better technology. This makes users more comfortable when accessing communication output from these devices. Moreover, scientists have devised ways of personalizing CSS. This means that they tailor make each device in accordance to the listeners' special needs and specifications. This has encouraged people with communication impairments to take up the new technology. In addition, intelligible speech synthesis systems have the ability of voice-over to speak out words, letting the persons who are visually impaired listen to written words on personal devices. As Sánchez and Aguayo (2007) developed the application for the visually impaired people, and they found that TTS technology has a critical impact on the users. For that reason, TTS is considered as a fundamental tool for system output of users with vision impairment. Moreover, they have focused on creating sound-based emotions.

Different people have different needs. Therefore, it should not be assumed that two people would automatically have the same speech disorder. In regard the issues that both the clinicians and the patients are facing with every new assistive technology, Mullennix and Stern (2010b) described this as every system should be made flexible and adaptable to the user needs and environment. As a result, Alternative and Augmentative Communications (AACs) and CSSs technologies are rapidly expanding to accommodate the changing trends in the world. Changing a system used by a patient is challenging for both the patient and the clinician. However, this in the long run will be more beneficial and efficient to the patient.

- **Vocal monitoring**

Oral information is more efficient in urgent problems than written messages to grab the user's attention (Dutoit & Stylianou, 2003), so that integrating speech synthesizers in measurement or control systems is necessary. For example, using speech synthesis in an airplane cockpit in order to minimize the pilot's distraction may prevent him focusing on other vital visual information. Another example is an in-car GPS navigation system that can provide location information automatically for location-aware services. GPS technology already exists on many mobile devices, and can be used to provide location-aware services (Moser & Melliar-Smith, 2008). And also an indoor way finding system is designed to be suitable for use on personal devices such as a PDA for people with cognitive impairments (Liu et al., 2008). Moreover, this application considered as a GPS for blind people in case of using prompts only, and it's useful in tour guide systems inside the mall or office buildings.

- **Multimedia**

In this section, the current implementations of TTS in visual animations are presented and the advantage of incorporating these animations is explained. Visual animations are facial animation, talking agent, and an image with a bubble text. With the increasing online services, in the coming years, the VTTS mobile applications would play a major role in the acceptability of these services. Human-like virtual characters could give a perception of a personality and a more human touch (Beskow, Granstrom, & House, 2002). In the same meaning, integrating a virtual talking person within mobile applications could be more pleasant and convincing than just text and graphics (Pandzic, 2002). VTTS applications require low bandwidth and high-interaction ability, hence, it can be used in the talking head scenarios, as an alternative to the video streaming. Pandzic (2002) categorized the possible applications of VTTS in sections, where these sections are: entertainment, personal communications, navigation assistant, news casting, commerce, education, and advanced multimodal human-computer interface. Recently, a new technology for movie making was introduced called Text-To-Movie. It generates 3D animated movies from entered text. It can be a text written by a person and then the voice that produces this message resembles that of the actor, although, the characters in the movie speak the typed conversation, and act in response to the selected icons, just like smiley's in IM/chat.

The combination of TTS and facial animation is known as Visual Text-To-Speech (VTTS) (Pandzic, 2002). The ideal VTTS system has to generate content of a cartoon character that speak and behave automatically. Moreover, the visual quality of the generated VTTS has to look attractive and animate naturally. However, as suggested by Pandzic (2002), no need for a marvellous quality, a straightforward cartoon character could do the job efficiently. Beskow et al. (2002) believe that lips, jaw, and tongue are the main three visual articulating movements in the face during speech. Nevertheless, there are other noticeable facial expressions, for example, eye movements, eyebrows, and head for expressing emotions, stress, and blinking. In fact, the feeling of naturalness about visual speech synthetic systems is mainly based on adding facial expressions with the purpose of transmitting nonverbal information, as well as contributing to the liveliness of the face. These movements are more complicated to form in a universal way than the articulators' movements, since they are optional and highly dependent on the speaker's personality, mood, purpose, of the utterance, etc. as noted by (Cave et al., 1996 as cited from Beskow et al., 2002). Meanwhile, audio-visual prosodic signals are used in applications of spoken dialogue systems and automatic language learning.

- **Man-machine communication**

A natural language has the ability to minimize the level of the required training, and for this reason, a TTS synthesis with high naturalness simplifies the human interaction with machine. In the near future, high-quality TTS synthesis and improvements in speech recognition are required, to facilitate the communication between people and computers or between people speaking different languages with the use of S2ST system (Dutoit & Stylianou, 2003).

- **Fundamental and applied research**

Controlling and measuring TTS systems are easy and accurate. TTS gives researchers the ability to run exactly the same experiment with spoken words. Such a level is not possible with a human voice as each time there will be small differences in emphasis, tone, etc. This reasoning made TTS an accurate laboratory tool (Dutoit & Stylianou, 2003; Sawusch, 2005; Vaissière, 2005). Hence, the speech synthesis has an advantage over natural speech production for evaluating the contribution of each individual parameter through controlling one parameter at a time, for example, fundamental frequency (F_0).

2.4 Impact of Voice Characteristics on the Acceptability of Speech Interfaces

2.4.1 Impact of Gender on the Acceptability of TTS Interfaces

There are two genders in speech synthesis systems, male and female. Lee et al. (2000) evaluated the masculinity and femininity of a speech synthesis to identify the gender stereotype. However, Lee et al. found that users accurately recognized the gender. Nevertheless, we must acknowledge that this result does have some positive effects on the conformity of the gender stereotype. They have noticed that individuals found voices with their same gender more attractive than those of the opposite gender. There was, however, an obvious crossover interaction between participants. Social identification has been realized according to attractiveness by which voice represents as pleasant, likable, and friendly.

The integration of male and female voices in commercials has been a research thath has proved to be useful. Lee et al. (2000) believe that a male voice perceives by participants more convincing than a female voice. Based on this, choosing a voice gender for application speech-based interface became easier in case of unknown gender of the targeted users. However, Lee et al.'s results show that the gender assignment varies, depending on the goal or the stereotype topic of the delivered interface. For example, research shows that a male voice is preferred in shopping sites, as here the goal is persuasion. However, the topic has an effect on the persuasiveness of a male voice and if the topic doesn't fit stereotypical gender roles, it may become ineffective. Therefore, if the topic of a website is stereotypically female, such as cosmetics, then definitely a female voice is preferred by users.

Despite the two features of voice that discussed in the previous paragraphs, it is also beneficial to understand that familiarity has some effect on user perception. Male voices are powerful and exhibit some form of confidence while female voices are loud and clear. However, Gong and Lai (2001) studied only the impact of a male gender in both conditions (mixed human voice and TTS vs. TTS-only), while they did not measure the impact of mixing gender. Nevertheless, their participants were males and females separated equally. In addition, they stated that there are a few current TTS systems mix the gender as well as the type of speech (see Figure 2.2). However, research on areas of mixing gender voices is still unexplored.

2.4.2 Impact of Accent on the Acceptability of TTS Interfaces

Accent is the outside reflection of an individual's way of speaking and is a noticeable characteristic of speech. Bayard and Green (2005) contributed in the Evaluation English Accents Worldwide project, and studied how different English accents are perceived cross-culturally. Four accents were evaluated in 19 countries, using English accents. These are Australia, NZ, North American, and UK. They found that the North American accent is growing universally, and the prestige of the UK English is rather reduced. Additionally, they categorized the results by region, influence of the media and intonation.

- **Regional results**

'Attitudes to the NZ accent as compared to recognised prestige are more likely to reflect new Zealanders' attitudes towards themselves' (Bayard & Green, 2005). However the study on the attitudes to regional varieties of New Zealand English showed that participants tended to give a higher rating for correctness and particularly pleasantness if the region was where they come from.

English-speaking countries, both the North American and UK accents were readily identified by all of the samples. In Europe the male UK English accent retained high scores in status, prestige, and power. In Asia the male NZE accent had low ranking for its status and power. In the Pacific (Fiji) the female NZE voice had high rating equivalent to the North American accents. In South America they ranked the North American accent in manner of fashion; with the UK English accents had low rating in regards of solidarity.

- **Media influence**

Generally, the consistent theme emerging from participants about the influence of the English-language television programs is overwhelming the North American accent. Interestingly, as pointed out by (Mugler, 2002 as cited from Bayard & Green, 2005), the NZ television programs in Fiji motivate part of the difference in perception by Fijians of the NZ accent relative to other countries.

- **Intonation influence**

In regards of speech perception, there are numerous experimental methods to study the perception of intonation and the role of intonation. However, Vaissière (2005) believes that studying and understanding intonation is considered as a complex process. Moreover, he states that there is still no speech synthesis system which can produce perfect natural attitudinal and emotional nuances carried by intonation. To study the intonation influence, according to Bayard and Green (2005), the fundamental frequency has to be removed from the speech samples to generate flat intonation, and then map the intonation patterns of the North American accent (expressive intonation) on to the other accents. The NZE male rated higher with flat intonation (Bayard, Weatherall, Gallois, & Pittam, 2001) where the NZE male is described as monotonic and South American described intonation as opposed to all other accents which was lower as has expressive intonation.

The following subsection discusses the other factors that might have an impact on the perception.

2.4.3 Perception of Listeners, Environments, and Language Experience

The style of speaking habitually adopted by talkers when speaking in difficult communication situation is called clear speech. As suggested by Uchanski (2005) “the details of utterances depend on the environment, the physical and emotional state of the talker, and the composition of the audience” (p. 207). In other words, it is true to say that the speech produced by an individual varies to a great extend during a day. For example, the vocal level depends on the noise and distance between talker and listener, therefore, in the case of a noisy environment and lengthened distance, the vocal level would increase accordingly. In addition, volume has impact on the pace of the speech as speech produced in a hall is likely to be produced more slowly than in a small enclosure. Moreover, the fundamental frequency of a speech would increase more than usual in the case of an individual under emotional stress or the influence of alcohol. Furthermore, In the case of a speech directed to a child, the simplicity of the sentence and the likeness of vocabulary are adjusted accordingly, while in the case of speech directed to an infant, exhibits large coal pitch variation, and also, in the case of a hearing-impaired listener, a talker will typically aim to speak clearly.

2.5 Perception of Different Types of Speech Interfaces

It's fair to say that these days it's difficult to live without some of TTS technologies. Researchers debate how types of speech affect speech interfaces acceptability. In GPS navigation systems there are different types of languages, accents for both genders which can be used; these factors reduce the bias and increase the acceptability. In order to get an acceptable TTS interface, according to Gong and Lai (2001), users' perceptions of task performance and attitudinal responses have to be measured.

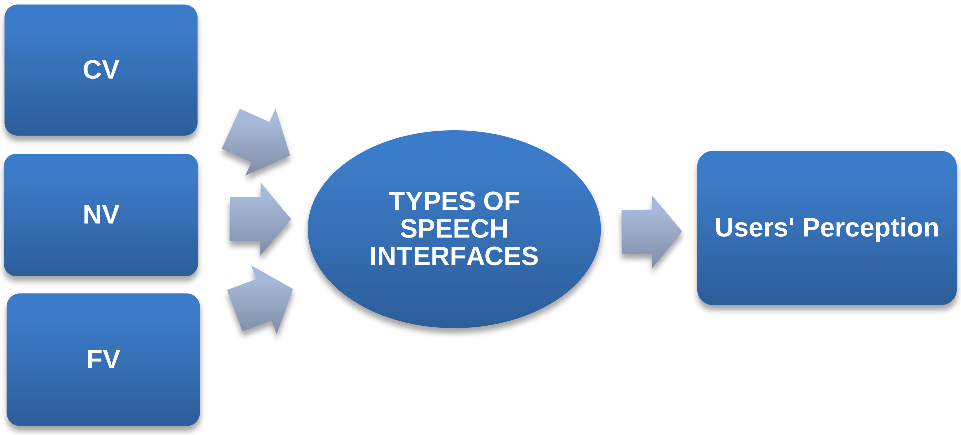


Figure 2.3 - Framework of the research question (source: author)

Figure 2.3 illustrates the framework of the research question that identifies the significant types of speech from literature. This section investigates their significance to the acceptability of applications with speech interfaces in more detail. These have been categorised into CV, NV, and FV of speech interfaces for the purposes of this discussion. The following subsections discuss these categories in more detail.

2.5.1 Perception of Synthetic Speech Interfaces

It is generally acceptable to use human speech for a fixed text and synthetic speech for a dynamic text, since it is impossible to record human voice for a dynamic text. However, Gong and Lai (2001) cliamed that mixing both voices together within an interface proved better perception of performance, easier to use, less repetition in understanding the email messages, and less effort (see Table 2.1). Currently, there are a few applications that mix TTS and human speech within the same sentence and/or between sentences. For example, most telecommunication services (e.g., Vodafone, Telecom) use this approach in accessing voicemails and checking credit balance over the telephone, whereas the fixed text would be recorded with natural voice and the credit balance would be synthesized by TTS system.

Regarding recent studies by Koul and Dembowski (2010), the development of SGDs has been one of the most significant communication devices developed to assist people with communication impairments. The percentage of people who have turned to SGDs for interpersonal communication has been increasing over the years. This is an indication that people have gained confidence in the technologies used and they are more comfortable while using them. The difference in cognitive domains as well as languages between different individuals creates an obstacle for researchers and developers in generalizing findings in synthetic speech perception. The recent statistics on discernment of short sentences presented in synthetic speech show that people with mild-to-moderate learning disabilities show signs of greater complexity in perception of even high-quality synthetic speech than matched typical individuals. However, individuals with learning difficulties who utilize synthetic speech can appreciate sentences formed by their SGDs. These findings were supported by Koul and Dembowski (2010) who agreed with the fact that people who had mild to moderate learning difficulties exhibited notable improvements in sentence accuracy as well as in single words scores. This was as a result of repeatedly conducting voice trials hence producing a synthetic speech that was eloquent and more practical to use. Through rigorous research, scientists eventually came up with the conclusion that by repeatedly exposing individuals with speaking impairments to synthetic speech, the level of speed and accuracy will be increased. However, every time the listening task becomes complicated, the level of accuracy of the synthetic speech will be reduced. Therefore, the response latency tends to increase in individuals including those who do not have any disorders. Research had proved that the discrepancy of the outcome across subjects might have been due to procedural restrictions that increased task complexity beyond a desirable level.

Mullennix and Stern's (2010a) study reviews people's reactions to speaking computers. Mullennix and Stern focus on attitudes of social psychological, including people's reactions to CSS, their attitudes toward the spoken message being conveyed, whether they believed the message was effective and whether they were persuaded by the message. However, people recommend the natural speaker on the CSS speaker, in cases of unawareness of the disability status of the CSS speaker. Mullennix and Stern's findings show that their users didn't like CSS speakers that much and even they expressed negative reactions. People trusted the natural speaker more, believing that the natural speaker is more knowledgeable and the message is more convincing, etc.

Therefore, it can be concluded that by repeatedly listening to synthetic speech outcomes in persons with intellectual disabilities and devoting minimal capital to processing the acoustic-phonetic qualities of synthetic speech, we need better resources to extract connotation from the synthetic speech signal. The likely shifting of cognitive resources away from processes by identifying syllables, phonemes and vocabulary to processes that are involved in extracting semantic content of the message, may conclude both accurate and faster recognition of synthetic speech stimuli. In conclusion, it is very important to probe into the discernment of synthetic speech in people with hearing and speaking impairment. This is because many people with acquired and developmental disabilities that may profit from a SGD will also exhibit hearing loss. Investigations addressing insight on synthetic speech in folks with hearing impairment shows that hearing loss may not have a damaging influence on the meeting out of synthetic speech.

2.5.2 Perception of Natural Speech Interfaces

There is at all times some debate on whether it is essential to spend a large sum of money on the acceptability of TTS interfaces or whether the greater amount should be directed toward the synthetic speech quality in general evidence. According to a recent study by Bunnell and Pennington (2010), the cost of developing a personal voice determines the type of technology an individual chooses to use. This is determined by the required testing, recording time and other expenditure if applicable. The key challenge for users using these devices is that sometimes the audio signal quality, below the user's standard. This could be as a result of cultural background of the speaker, the pitch of the speaker, the recording devices and cheap sound cards in the market. To reduce this problem users are advised to record a series of recordings ensuring that the developers select the best recording for their use. Moreover, the user is encouraged to speak fluently and accurately.

According to Sánchez and Aguayo (2007), the usability testing results of their study showed that the end users prefer to listen to natural voice instead of synthetic voice, even though the quality of the synthetic voice is good and acceptable. Moreover, according to a recent study on attitudes towards speaking computers by Mullennix and Stern (2010a) indicates that people view CSS less favourably than natural human speech. However, research has showed that people's attitudes towards CSS users have some exceptions as the previous statement perceived the opposite by people when the CSS users are with speaking impairments. However, this statement is not always true. However, according to Stern et al.'s (2007) study, negative attitudes toward CSS reappear if the CSS speaker is engaged in an activity people don't like, for example a telephone campaign. In addition, the purpose of the technology can be used to modify people's reactions.

TTS interfaces become more common. Therefore it is important to improve TTS interfaces usability, effectiveness, voice quality, and accuracy. Schroeter et al. (2002) stated that evaluating TTS systems correctly is still a mainly unsolved research problem. However, natural voice has proved all these benefits. In other words, using natural voice in speech interfaces leads to a better performance than using computerized voice. Moreover, there are some negatives associated with using TTS-only for fixed and dynamic text such as low quality, synthesized speech with less clarity than the natural voice, but consistency outweighs them (Gong & Lai, 2001). On the other hand, the use of human voice should let contemporary speech interfaces be intelligent. As Gong and Lai's result showed that natural voice improves both the effectiveness and the user's interaction, which it is a vital point in the field of human-computer interaction. However, the computerized voice still contributes in the flexibility and dynamicity of the interface itself. In addition, the ease of controlling the voice characteristics of gender and accent are considered as advantageous to TTS systems.

Regarding familiar perception, people tend to be more aligned to familiar voices in any way. Just like in accent, native speakers of a particular language tend to listen to voices that have their native accents. In a crowd, an individual will tend to identify more with a familiar voice than unfamiliar voices. An individual, on the other hand, can perceive speech in a voice that he/she is familiar with more than unfamiliar voices. For example, children perceive their mothers voice well because of familiarity.

Music can have a positive effect on enjoyment of and motivation for performing physical exercise (Wijnalda, Pauws, Vignoli, & Stuckenschmidt, 2005). Lots of research has also been done in order to develop further products and devices that would benefit users with regards to exercise. However, there are many devices that are now being used to help aid exercise such as Nike Plus. This product gives an audio feedback of the current progress. Either a male or female can be selected, as the voice details the current time, distance and pace. This spoken feedback is pre-recorded speech samples of celebrities. The celebrity voice was far more enjoyable to run whilst listening to music (Byrne, 2007). Also has a power song, which it is one song on the list that can be selected as an extra motivations aid. This is extremely beneficial at times where motivation is low, and hearing the favourite or most inspirational song helps to keep exercising. Also after completing a personal best an auditory message congratulating will be received.

2.6 Conclusions

A review of current related literature confirms that speech type has an impact on the user's perception of personal TTS applications. However, there is limited literature that focuses on the practice of developing speech interface based FV. Also, there is limited research on areas that improve strength of speech perception and performance. This is not just strength against noise but also strength against any condition that influences performance. For example, the research on acceptability of TTS voices is still under research. Besides, there is still unsolved research problem in evaluating TTS systems accurately. This, therefore, supports the contention of this study to investigate this area.

An analysis of the literature reveals that there are several factors influencing the acceptability and depth of development of speech interfaces with specific options depending on voice characteristics of gender and accent. There is, however, one key factor that cannot be overlooked in the discussion, which is, computerized voice is not a human voice. However, the use of speech synthesizers within personal devices made today goes a long way toward helping humankind to enjoy a better tomorrow.

These ideas have been illustrated in Figure 2.4, which represents a model of these factors. These dimensions provide a structure that presents some insights that contribute in the accessibility of TTS interfaces, in addition providing some insights to the comprehension of current practice in TTS implementations. This model will form the basis of design for a field study and a comparison of the findings with literature. Before presenting the findings and discussing them, the research methodology, and the research methods and techniques that used in this study are described and justified in the next chapter.

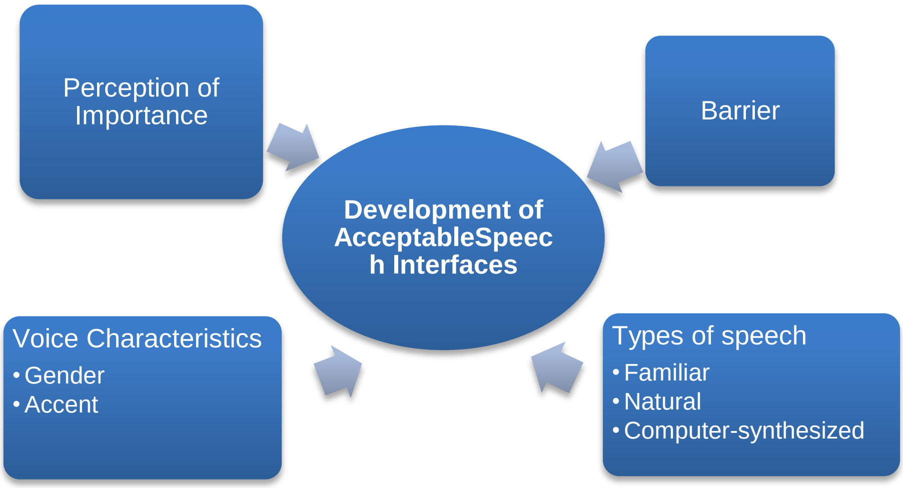


Figure 2.4 - Dimensions which are relevant to development of speech interfaces (source: author)

CHAPTER III
RESEARCH METHODOLOGY

3.1 Introduction

This chapter addresses the research question and guides the selection of the methods and techniques. The selection of the research methods depends on the selected research methodology, which has to be chosen according to the theoretical framework of the research question (Figure 2.3). Methodology, according to Collis and Hussy (2009), is an approach refers to the process of the research, encompassing a body of methods. A method is a technique for collecting and/or analysing data. These methods and techniques will be used for gathering and analysing data.

The next section, Section 3.2, describes the identification of paradigm and justifies the choice of the methodology in the context of qualitative approaches, numerous methods are evaluated, and research approaches are justified. Section 3.3, justifies the choice of the methods and techniques, also presents the main features of the research design with drawing on previous studies that have been used similar approaches. Section 3.4, illustrates the process for the design, development, and testing of the equipment and apparatus. Section 3.5, discusses the methods used to collect and analyse the research data, and argues its relation to the research design. It also describes the sampling method and the participants' selection criteria. Section 3.6, describes the listening task procedure. Section 3.7 presents the data collections methods. Section 3.8 presents the data analysis methods. Section 3.9, discusses the limitations in the research design with reference to generalizability, reliability and validity of the results and findings. Finally, Section 3.10 concludes the chapter with a summary of the main research decisions and ending with introduction to the results and findings chapter.

3.2 Research Type and Paradigms

Research paradigm is a philosophical framework that guides how research should be conducted. There are two main paradigms, positivism and interpretivism. A positivism paradigm tends to use large samples and produce results with high reliability but low validity, as well as, it can be generalized from the sample to the population. Whereas, an interpretivism paradigm tends to use small samples, and produce findings with low reliability but high validity, as well as, it can be generalized to other settings (Collis & Hussy, 2009). The interpretive research involves an inductive process with a view to providing interpretive understanding of social phenomena within a particular context.

Research approaches can be either qualitative or quantitative. A qualitative research approach attempts to gain an in-depth opinion from participants about users' perceptions and experiences. On the other hand, quantitative research methods will be used to generate statistical analyses through evaluative scales to inform the participants and the researcher about the interview process. A number of research approaches could be taken to achieve the aims and goals, and to answer the research question described in Chapter 1. Mainly qualitative approach is taken to collect interview data, which this data are going to be analysed by key point analysis. A demographic questionnaire will be used to form a background of the participants, a selection tool of the three couples (six participants) for the listening test, and also, for the nomination of a familiar voice for each participant is by giving three names of possible familiar people and rating on a semantic differential rating scale of familiarity to each person. A listening test will be taken, as an approach of testing specific speech samples or recording through human listening experiments (Uchanski, 2005). Thus, after each listening test, evaluative rating scales will be used in order to evaluate the users' self-perception of task performance. The focus is on gaining insights prior to a more rigorous investigation. This type of rating scale measures intensity of opinion to allow respondents to give a more discriminating response, indicating if they feel neutral (Collis & Hussey, 2009).

It is common to distinguish research as either deductive or inductive. Deductive research is a theoretical and conceptual structure well developed and then tested using empirical observation and in this case deductions of particular instances are made from general inferences. Whereas, inductive research involves development of models from observation of empirical reality and therefore induction of general inferences is made from particular instances (Collis & Hussy, 2009). In terms of this study, an inductive research is taken, and will then be underpinned by an exploratory research approach. As, this type of research, inductive research, is normally conducted when the research problem is not clearly defined. This will enable the researcher to manipulate the available variables of voice characteristics while controlling the others by basing on evidences from previous studies (Bayard & Green, 2005; Lee et al., 2000). In exploratory study, the researcher has deliberate manipulation towards the involved voice characteristics (gender and accent) of the speech samples to see how it contributes to the acceptability of each interface. It will make use of the primary data obtained. It will also help the researcher to determine suitable data collection and analysis methods. This approach is used to achieve the objectives of this study and hence help in answering the research question described in Chapter 1.

Within this research paradigm there are a number of methods and techniques for gathering and analysing the data. These are discussed in the next section and the use of a qualitative approach involving semi-structured interviews is justified.

3.3 Research Method

This study adopts an interpretive research approach, since this involves the investigation of peoples' interaction within their social and cultural context. The adoption of an interpretive epistemology is common and is well accepted within information systems research. In terms of this study the research approach to be taken is exploratory, because the researcher will be dealing with a new issue and will be conducting a research in a field that has not been previously researched. The approach will gain insight and understanding before more rigorous investigation is conducted in future. Other researchers will use the findings of the study as a base to conduct further research in this area.

It is predicted that this study has to be taken over a period of twelve months. Interviews have to be organised and documentation has to be gathered over this period to analyse the data and build a clear picture of how events unfolded. Therefore, inductive research has been chosen on deductive research because this study is designed to make a contribution to general knowledge and theoretical understanding, rather than solve a specific problem (Collis & Hussey, 2009). It is suited to prototype development and evaluation of its prototype's speech interfaces. That is, the research approach taken in this study is exploratory research to look at the user's self-perception of task performance, experience, perception and attitude of different types of speech interfaces of personal devices.

Regarding the research design, according to Mohan, McGregor, Saunders, and Archee, (2008) "Surveys can be used to collect information and opinions from a number of respondents in order to conduct research eliciting social, economic or political trends. Questionnaires can be administered as part of a research interview and can help you to develop a structure for the interview" (p. 229). In addition, a methodological triangulation, as suggested by Collis and Hussey (2009), is using more than one method to collect and/or analyse the data (e.g., evaluative rating scales to identify key issues and provide insights into the issues before conducting an interview), which should then lead to greater validity and reliability rather than a single method approach. Therefore, regarding this methodology, several research methods and techniques for gathering and analyzing the data have been used including demographic questionnaires, evaluative questionnaires, and semi-structured interviews. These research methods used in this study of recruiting the participants, and collecting and analysing the data are illustrated in Figure 3.1.

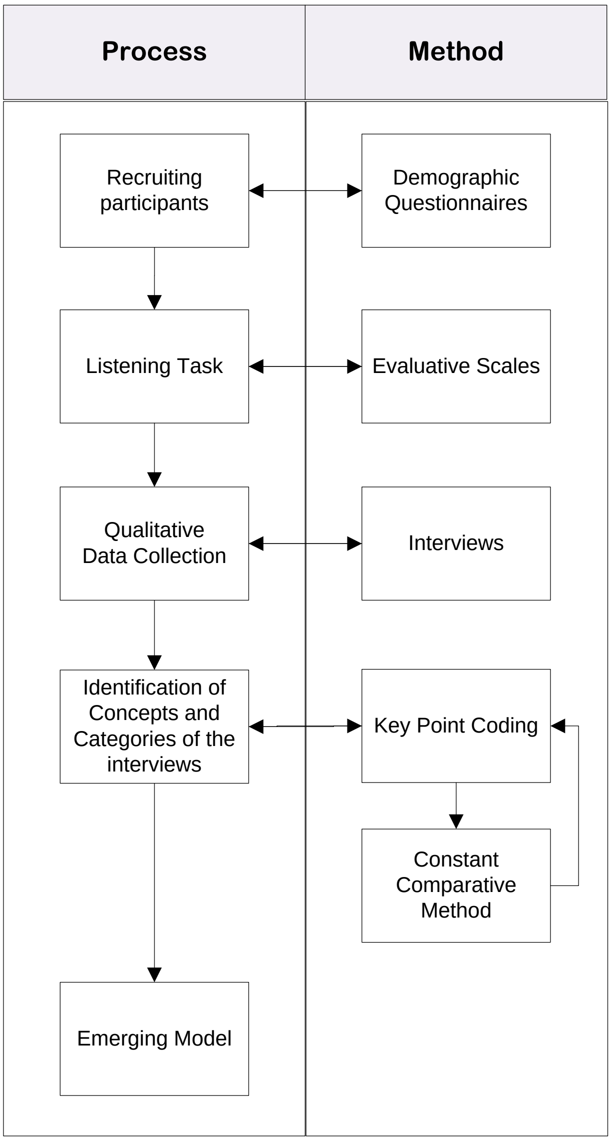

Process
Method
Recruiting participants
Demographic Questionnaires
Listening Task
Evaluative Scales
Qualitative Data Collection
Interviews
Identification of Concepts and Categories of the interviews
Key Point Coding
Constant Comparative Method
Emerging Model

Figure 3.1 - Research Methods Flowchart (source: author)

The first stage is recruiting the participants in which presentation is used to recruit six participants (see Section 3.5). The second stage is the listening task that involves data collection from evaluative questionnaire; this evaluative data that will be used to inform the participants and the researcher about the study and to facilitate the interview process. The third stage, which it is interviewing each of the participants by basing on their statistical analysis of the evaluative scales (see Section 3.7). The fourth stage is the interview data analysis, from which the concepts and categories are identified by the key point analysis approach, then defining the relationships between concepts and categories so far established with the other couples (see Section 3.8). The last stage consists of forming a model, which will be covered in the next chapter.

3.4 Design, Development, and Testing of the Equipment

3.4.1 Environment

Similar to previous studies, (Lee et al., 2000; Sánchez & Aguayo, 2007; Ståhl, Gambäck, Hansen, Turunen, & Hakulinen, 2008; Walker et al., 2006), used a PC to present the stimuli and collect responses. The design of Lee et al.'s study was that each participant had to response verbally to one of two options provided vocally by the computer on 8-point scales, whereas 1 is *definitely doing A* and 8 is *definitely doing B*. After completing the tasks with the computer, the participants filled out a post-experiment questionnaire as the decisions were related to both TTS gender (male and female) by both participants' gender (male and female). Whereas, Walker et al. have used the TTS engine in order to run the experiment, including randomization, response collection, and data recording. Furthermore, Sánchez and Aguayo integrated a TTS engine with their delivered mobile messenger application for the blind, as the TTS system comes with a speech manager to synthesize the speech modification of tone and pace similar to almost all TTS systems. Whereas the speech manager is an application programming interface (API) that allows developers to use speech synthesis engine within websites and mobile applications in order to convert written content into speech. With regards to the developed mobile application, Ståhl et al., (2008) developed a mobile application for the health and fitness that runs on a PDA that can be used during the physical exercise. The mobile companion prototype shows a rabbit image and a bubble text with a TTS system. The TTS system speaks the bubble text. In addition, the application displays information about the exercise status as (duration, distance, pace, and calories). This application is based on the use of speech. It uses a synthesized speech to address the user commands. The user has the ability to control the application during the exercise by giving commands for asking the summary of the exercise, playing music and stopping the exercise. After finishing each exercise, the application gives a summary of the exercise and asks for permission to upload the results to the server in order to be analysed later.

Regarding this study, a customized audio player application has been developed on the iPhone to facilitate running the listening task and to demonstrate an environment of TTS interfaces of personal devices. A mobile device has been used, since this research targets the applications' uses of personal devices. The iPhone has been chosen on other mobile devices because of its proven acceptability, simplicity, speakers' loudness and clearness, and also its multi-touch screen. The application was developed on an iPhone Operating System (iOS) platform to provide faster interactivity as it is better, faster, and easier than other platforms at the time of writing codes. The next sections are more detailed about the specification of the developed application prototype.

3.4.2 Software Specification

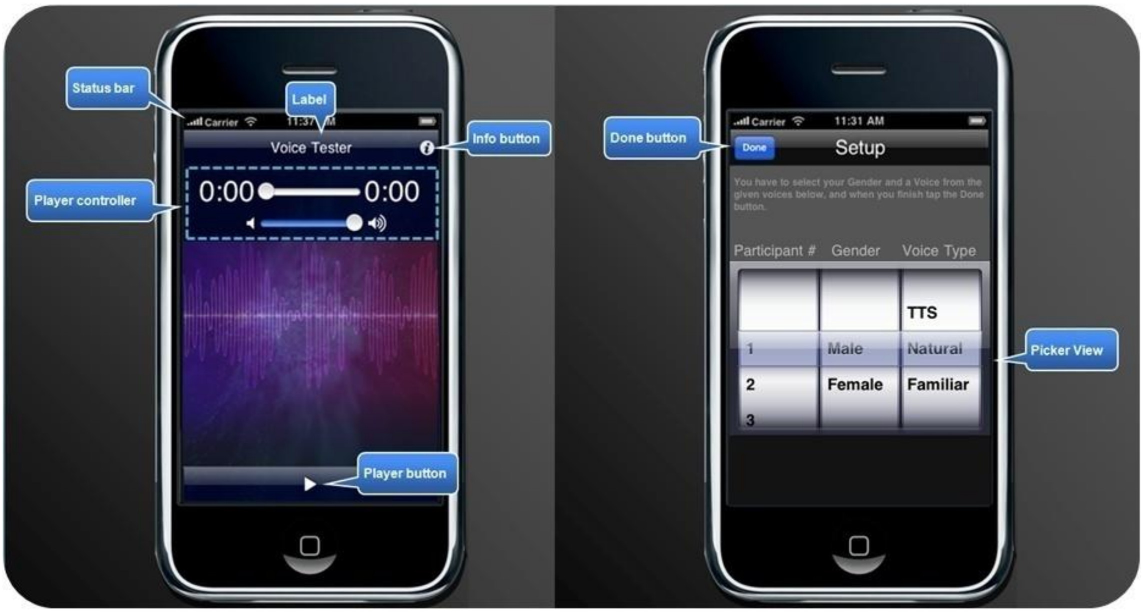
The VoiceTester application software was developed by using an Apple desktop computer (iMac) running MAC OS X as it is, according to Apple Inc. (2011), the world's most advanced operating system. The application was programmed in Xcode and iOS Software Development Kit (SDK), whereas Xcode is software used for building Mac and iOS applications through Object C computer language, also contains Interface Builder to design user interface elements, and simulator to test the functionality, utility, and accessibility of the application currently running on the iPhone (Appendix L – The VoiceTester Code Introduction).

The VoiceTester functions similarly to an audio player. It maintains a database of voices, also, the user can play/pause and forward/backward a selected voice. Interaction with the application is via a combination of speech synthesis and an on-screen display. The background of the application has gradually varying colours to give the live feeling to the users, similar to the sound waves that change according to tone and rhythm of the audio file. Speech synthesis is currently supported on the iOS platform and can be ported to the iPhone. For less programming and because the used computerized speech samples were static only, the Cepstral TTS engine was used to generate computerized voices for both genders.

3.4.3 VoiceTester Prototype: Functional Module

This module has been evaluated using a pilot test for bugs and having colleagues' complete set tasks with the application in a laboratory setting to reassure the application's interface is easy to understand and follow. However, the delivered interface is very simple and modern. Additionally, this multifunction application consists of two views, which are the Main view and the Setup view.

The main view has a Player Controller component integrated into the music player function of the software application. This component supports volume control and progress tracking control, as well as updating the current time and duration of the selected audio file. The setup view has the most important feature of this system, the Picker View component. This has the participant number, gender and type of speech placed at the Setup view with more detailed instructions on the interface. This reduces the confusion generated by users with having extra button/view and delivering as much as possible a consistent and simple interface that almost doesn't require training.



You have to select your Gender and a Voice from the given voices below, and when you finish tap the Done button.

-Main View- **-Setup View-**

Figure 3.2 – VoiceTester prototype showing the frontend and the backend (source: author)

The application software prototype was developed to produce multiple voices under a personal device environment. The first two types are supported for both genders, however, the gender of the computerized and natural voices were controlled and assigned accordingly to the participants' gender, based on evidence from the literature review chapter. In the case of a couple, the familiar voice would be each other's voice (see generating and recording the voices subsection for more details). Figure **3.2** shows the user interfaces (UI) elements of Status bar, Player controller, Player button, Label, Info button, Done button, and Picker View. The functionality of these UI elements is described briefly below:

- **Status bar:** appearing at the very top of the screen. It shows important information about time, signal strength, current network connection, and battery charge.
- **Player controller:** consists of a volume slider, a progress bar slider and an update timer. A volume slider is supported to control and adjust the loudness of the speaker volume; the volume is assigned to the maximum loudness by default, as the listeners are supposed to hear the voices clearly and loudly. The progress bar shows progress of the task. An update timer of the audio file consists of current time and duration. The update timer is developed to give feedback about the current time of the audio file, to show audio progressing time and to give the user the ability to move the audio file forward/backward if the listener miss-understands any word, as well as, providing information about the length of the audio file.
- **Player button:** it is the Play/Pause button of the audio player controlling the playing or pausing of the audio file, responding to user touches. By tapping this button the selected voice from the picker view will be played, accordingly, the update timer will start counting by showing the current time, progress and duration of the audio file.
- **Label:** used for displaying text as title and that depends on the selected values from the Picker view component; the title bar shows the selected options from the picker view of gender and voice type.
- **Info button:** flips the screen to the Setup view of the application, most probably will be used by the participants to select a different voice.
- **Done button:** this button simply flips back from the Setup view to the Main view.
- **Picker View:** this is the spinning wheel of values. It is used here to assign a participant number, a gender and a type of speech.

3.4.4 VoiceTester Protocol

The purpose of this section is just to look at the research protocol of the delivered application. The application was called 'VoiceTester' because its purpose is to test and evaluate different voices through listening tests. Whereas, a listening test is an approach of testing specific speech sample or recording through human listening experiments. The aim of the listening test is to gather important information about users' perceptions and experiences of the speech samples to inform the speech interface design.

The researcher would prepare the environment for each participant to practise the task on a silent mode before starting the test to reduce the bias. A participant's number, gender and a voice type preferred to be selected first by the researcher. And this process has to be repeated twice more as there are another two speech samples have to be played and listened to by each participant. After that, each participant would begin listening to each of the given voices and then evaluate the performance of each voice on a paper-and-pencil evaluation form which helps focus the interview.

The task is introduced and demonstrated, later in this chapter, as basically it is listening to the instructions of each speech samples and the process is exited whenever a step is finished. There are however, three listening tests to complete the listening task (see Appendix D – Listening Task), in addition, up to 5-minute rest after finishing each listening test for the evaluation purpose and to prepare the candidate for the next step (see Appendix E – Evaluative Questionnaire), and at the end of the listening task each participant will be interviewed about his/her responses, experience and perception (see Appendix F – Interview Questions). The following subsection describes the auditory stimuli in more details.

3.4.5 Auditory Stimuli

All the speech samples have to be pre-integrated with the VoiceTester application for ease of use apart from the familiar speech samples which were different from one participant to another. The duration of the audio file of the natural voice was slightly different for each gender, and the familiar voice was different for each subject. This difference in the length of the human (natural and familiar) speech samples due to controlling the speed and duration of the spoken words of the human voice is almost impossible. By comparison, the length of the computer-generated speech samples was the same in both genders because it was read by TTS engine.

To avoid memory recall issues for the participants and to reduce the bias, a delay was added to separate questions in an audio file. A similar procedure has been done with previous research (Walker, Nance, & Lindsay, 2006), where E-Prime software has been used for combining the audio cues and TTS segments in one file separated by a few seconds of silence. Pilot testing was conducted to make sure that all the sounds and words were clearly understandable. The length of the passage was also controlled against being too long or too short. If too short, the participants would not have had enough information to evaluate the speech perception. If too long, it could have affected the quality of the reading adding to the involvement of the listener. Therefore, the duration of each audio file is around one minute including the 5-second silent delay to separate the prompts (this is described in more details in the following subsection).

Currently, there are a few types of commercial TTS engines and many researchers used different types upon availability and comprehension of each. For example, Walker et al., (2006) chose Cepstral TTS engine, Sánchez and Aguayo (2007) chose ACAPELA TTS engine, and Gong and Lai (2001) used an IBM via Voice Outloud TTS engine. However, as pointed out by (Lai, Wood, & Considine, 2000 cited from Gong & Lai, 2001), using a different TTS engine wouldn't have a significant impact on the voice comprehension, as the authors found no major impact on the perception of synthetic speech through evaluating five types of commercial TTS engines.

Regarding this study, the computerized voices were generated by using Cepstral TTS engine. The male and female natural voices were recorded by Native English NZ speakers with local accent. Participants had to nominate three familiar names whose voices they would like to use in their personal applications, and then one, giving the reason for the preference of that voice over others. A 5-point rating scales has been used to measure the familiarity (1 = *I don't know him/her that much* and 5 = *I know him/her very well;* see Appendix A – Demographic Questionnaire).

The following subsections explain the instruments used in generating and recording the speech samples, and illustrate the speech samples integration with the developed application.

3.4.5.1 Generating and Recording the Speech Samples

Generating the TTS samples was by Swift Talker software of Cepstral engine. David and Callie were selected because of their clear British accent; they have been used as computerized voices for male and female genders respectively. The SwiftTalker software comes with the Cepstral package that has the ability to generate voices simply by entering text and controlling the characteristics of voice pitch, speech rate, and volume by its editing tools. The speech rate is 170 WPM (word per minute). The samples rate for David is 16000 Hz and for Callie is 22050 Hz by default.

Regarding the natural and familiar speech samples, most importantly, all speakers were recorded separately and were instructed to read the short passage (see Appendix G – Stimulus Passage for Readers). They were also given the opportunity to practice reading the passage prior to the actual recording. This precaution was taken to guarantee a relaxed, clear and uninterrupted delivery of the reading. Recording the natural and familiar speech samples was via Voice Memos application. Voice Memos is a built-in voice recorder application of the iPhone. This application was also used to record the interview sessions. Voice Memos has the ability to record and mange the recorded audio file by its editing tools of naming, sharing and trimming (see Figure **3.3**).

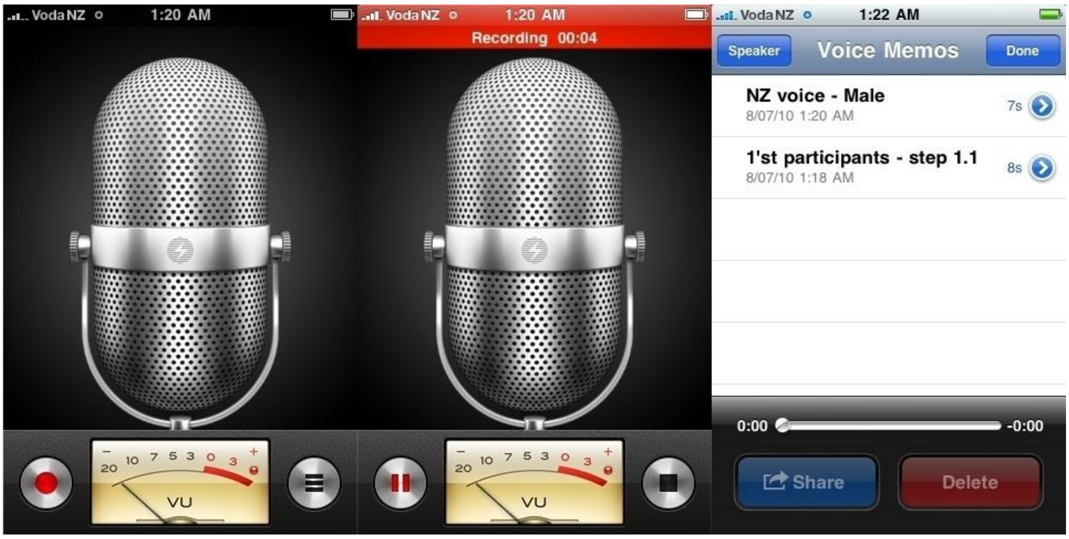
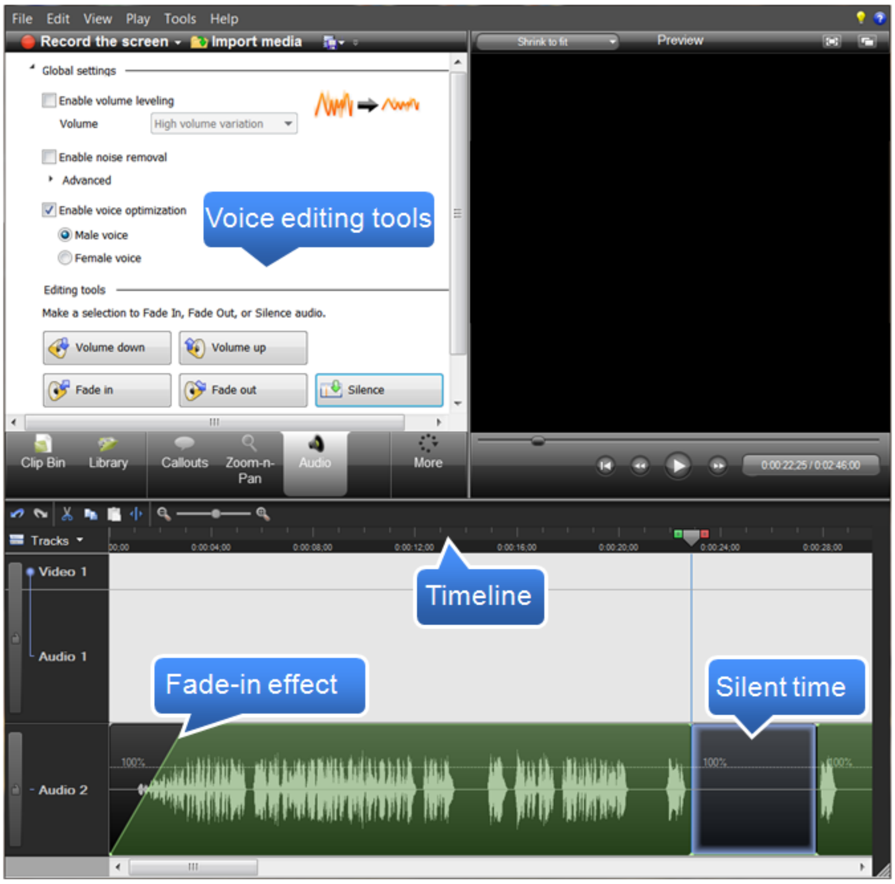


Figure 3.3 - Voice Memos Application

Camtasia software (Version 7) has been used to edit the used voices and re-sort the questions randomly in each voice by separating them with a silent delay. Camtasia is a computer software that enhances and improves audio quality by editing tools of volume levelling, noise removal, voice gender optimization, fade in/out the beginning and the end of the audio clip respectively, and adding up to 5-second silent delay to separate questions in each audio file. The enhancements could be applied to the duration of a clip, a timeline section, or entire timeline. Moreover, all the voices were high quality and customized with one volume level. Figure 3.4 shows the effects taken on the input signal of an audio file.

Assigning the Generated TTS voices and the recorded natural voices is static for all the participants with similar gender, while assigning the familiar voices to its related participants is different for each participant (see the following section for more details).



Global settings

□ Enable volume leveling

Volume: High volume variation

□ Enable noise removal

Advanced

☑ Enable voice optimization

● Male voice

○ Female voice

Editing tools

Make a selection to Fade In, Fade Out, or Silence audio.

Figure 3.4 - Shows the actions taken on the signal output by Camtasia

3.4.5.2 Speech Samples Integration

All the speech samples have to be integrated with the application first through Xcode, which is, the researcher responsible for this step and because the familiar voice is unique among participants. Therefore, the researcher has to assign the participant number to its related familiar speech sample from the Setup view of the delivered application before starting the test.

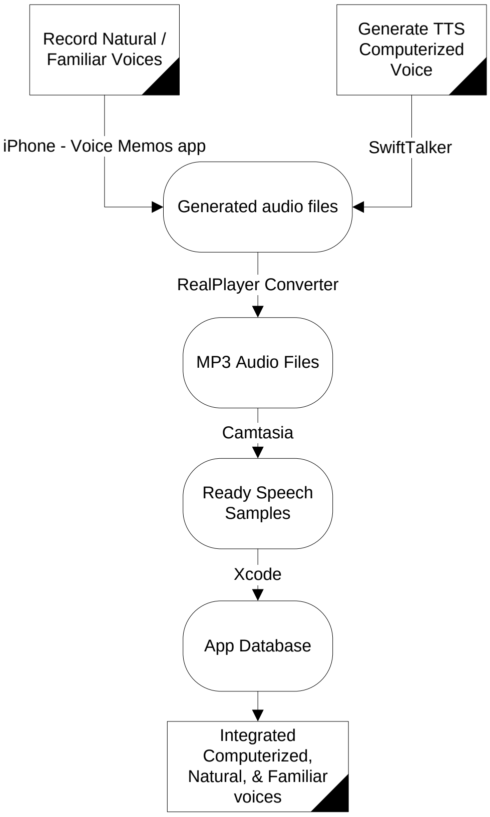


Figure 3.5 - Speech Samples Integration Process (source: author)

The integration process of the audio treatments is completed in four steps (Figure 3.5). The first step is recording the human voices (natural voices and familiar voices) by the Voice Memo app of the iPhone in which the recorded file emailed to a computer, and generates the computerized speech samples by the SwiftTalker software. The second step involves converting the audio files with different formats via RealPlayer converter to MP3 format as this format has less size and can be played by the VoiceTester. Treating the audio files by Camtasia is the third step. Step number four uses Xcode to integrate the treated speech samples within the application database. Now, the VoiceTester app should be ready for the listening task.

3.4.6 Practice the Listening Task

Regarding the task procedure, the participants firstly have to follow the process of selecting their number, gender and a voice from the given voices at the Setup view. Secondly, they have to navigate to the Main view and then tap the play button to start listening to the selected voice (see Section 3.6 for more details). However, they have the ability to pause the played voice by tapping the pause button, and also they are able to control the volume and the playing track through the volume slider and the progress bar slider respectively. The task has to be done via a VoiceTester app, and this application could be launched for the participants before starting the task for practicing purposes. The application can be launched by tapping on its icon (Figure 3.6).

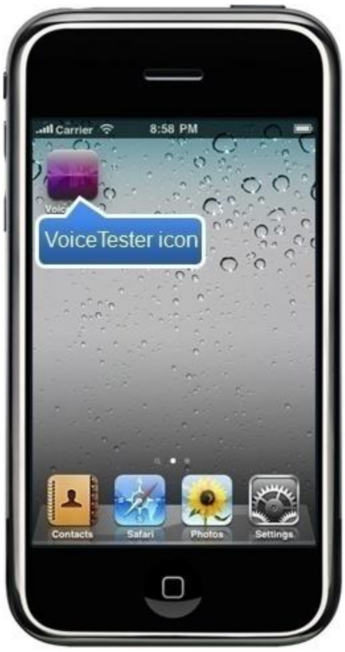


Figure 3.6 - Screenshot showing the VoiceTester icon (source: author)

Once the application is launched, the title displays the name of the application, then, after selecting a voice the label text would display the chosen user voice. Either the researcher or the participant has to assign the participant's number[1], gender and the voice type from the setup view of the developed application. The number preferably is given by the researcher. However, the hint used at the setup view encourages the user to select his/her number, gender and voice type (Figure 3.2, Setup view). Now, the speech samples should be prepared for the specified participant, the play button should be taped to start the listening. An alert is used at the end of each step to instruct the candidate which actions are next required (Figure 3.8). Moreover, the user has the ability to pause anytime. Accordingly, the update timer will be paused. In the example of Figure 3.7, the screenshot shows different actions of the play/pause button. All the details for the voice selection process are placed at setup view. However, the participants have the chance to practice on the application of the iPhone before the test began as the device turned on a silent mode illustrated here. In addition, the instructions were placed at the setup view for unpractised well users.

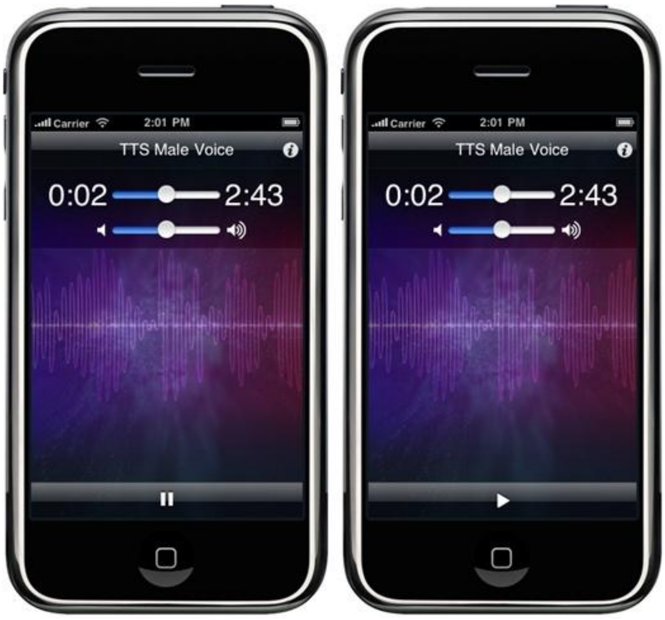
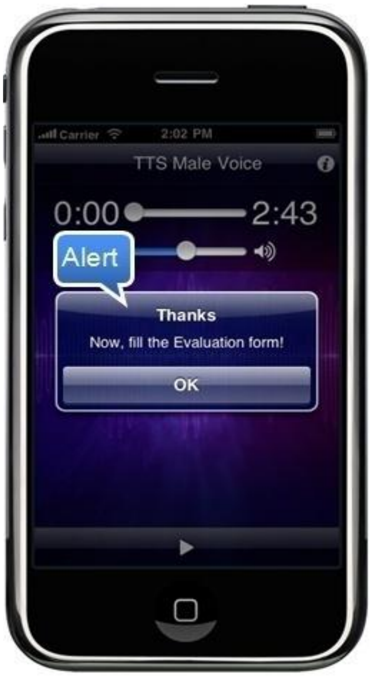


Figure 3.7 - Play and Pause interactive screens (source: author)

[1] Participant's number is used for confidentiality purpose, as well as, to allocate the familiar audio file of each targeted participant.

Once a step is completed, the device vibrates and the application presents an alert window. According to the mobile user experience of the *iOS Human Interface Guidelines* (Apple Inc., 2011), alerts should be used only for situations that require immediate user attention. Here it is used at the end of listening of each speech sample to inform the participants that this step has been ended and now it's time to complete the evaluation form (see Figure 3.8). In addition, this process is necessary for user reassurance and also to reduce the impact of bias. Otherwise, the participant would start with another voice without evaluating the voice performance or could be interrupted by the researcher for information about the process.



Thanks

Now, fill the Evaluation form!

Figure 3.8 - The application alerts whenever a step is completed (source: author)

3.5 Research Design and Preparation of Data Collection

3.5.1 Design of Research Instrument

The design of the research instrument is based on the research question described in Chapter 1 and the framework of analysis developed in Chapter 2 (Figure 2.3). The analysis framework identifies significant factors related to the impact of voice characteristics of the types of speech towards users' perceptions and experiences. This has influenced the scope of the study, the structure of the designed interview, the analysis of the data, and the report structure.

Several data collection methods of demographic questionnaires, evaluative scales, and semi-structured interviews will be used to collect data in this research, and the purpose of each is briefly discussed below:

- The demographic questionnaire will be used to demonstrate participants' demographics data with regard to age, gender, favoured gender and familiar voice.
- The evaluative scales were designed based on the Gong and Lai (2001) study. A similar structure was designed to inform the participants and the researcher about the listening task and to facilitate the process of the interview data. Thus, the participants completed the questionnaire after each test regarding self-perception of task performance.
- The *interview* term refers to a method for collecting primary data in which a sample of interviewees are asked questions to find out what they think, do or feel (Collis & Hussey, 2009). Semi-structured interview is an interview guide which structures the questions that have to be explored in the interview (Patton, 2002). The interview guide, most importantly gives the interviewer awareness to design the questions according to the limited time of an interview. The interviewer can freely explore, probe, and ask questions to explain and clarify that particular subject. Hence, semi-structured interview questions were selected and designed through a series of brain storming and discussions with the project supervisor. The final version of the interview questions can be seen in Appendix F – Interview Questions. The interview questions started with an open question followed by more directed questions. Table 3.1 shows how the interview questions relate to the research question and objectives developed in Chapter 1. And also the purpose of the last question '*Are you able to efficiently complete your task using this application?*' is to see if the developed application contributed in any bias.

Table 3.1 - Research question and interview questions relationship

Research question / objective	**Interview questions**
What are the perceptions of different speech types of TTS interfaces?	1-2, 4-5
How has each of the examined voice characteristics contributed in the acceptability of each type of speech?	3

In time with the design of the research instrument, the methods of recruiting and identifying participants were also conducted. The following section describes the participants' recruitment method.

3.5.2 Recruitment Method

Participants for this study were recruited from a postgraduate class and a public library through presentations. The presentation was related to the participants' confidentiality, risk, and time; as the participants will be known to the researcher only. However, detailed information will remain confidential to the research, and their identity will remain confidential to the researchers. The volunteers completed a demographic questionnaire, which is required approximately five minutes to be filled (see Appendix A – Demographic Questionnaire). In the demographic questionnaire, volunteers could indicate their willingness to participate further in the study. Potential volunteers were matched with criteria for inclusion in the study. The demographic questionnaire used as a selection tool for six participants consisting of three couples, being those who are about to complete the listening task and interview. The six participants need approximately 45-60 minutes to do the listening task and the interview. The participants are not at risk of physical, psychological or social discomfort when they take part in the study. The details of the research included in the study were provided to the participants via a participant information sheet, (see Appendix B – Participant Invitation Letter). After that, written consent was gained from those who would like to participate in the study, (see Appendix C - Participants Consent). The six participants have to be invited to participate via a phone call using the contact information that they provided in the demographic questionnaire. The next section describes the participant selection process.

3.5.3 Participants Selection

In a qualitative inquiry, there are no rules for sample size. Extensive valuable information can be obtained from a small number of people, particularly if the selected cases are information-rich (i.e., what the purposeful sampling focus on; as claimed by Patton, (2002). In purposeful sampling the size of the sample is determined by informational considerations. More or less participants depend on the level of in-depth information as Patton (2002) stated "If the purpose is to maximize information, the sampling is terminated when no new information is forthcoming from new sampled units; thus redundancy is the primary criterion" (p. 202). Numerous, qualitative studies used small samples, such as Walker et al., (2006) used in their study a small sample size consisting of nine undergraduate students. Both Gong and Lai's (2001) and Walker et al.'s (2006) participants reported have no hearing problems. Participants has been chosen from both genders equally to reduce the bias of participants' gender, since this research is targeted for the public users of mobile applications so participants preferred to be selected from different age groups. A similar approach has been used by Gong and Lai (2001) who recruited 24 participants (12 males, 12 females). Since, the specification for the design of qualitative samples has to be with minimum samples based on expected reasonable coverage of the phenomenon given, the purpose of the study however, a qualitative sample design should be understood to be flexible and emergent (Patton, 2002). Thus, reasonably, six participants (three males, three females) had to be recruited, preferably three couples, to facilitate the testing process and to simplify the familiar voices selection and recording.

Regarding this study, volunteers have to be recruited through presentations. At least six participants in the form of three couples have to be chosen through applying specific criteria. Participants must be over 20 years old, healthy with no hearing problems, using English as their first language in order to avoid any potential difficulty in understanding the synthetic speech, and each couple must consist of both genders and be from different age groups (21-35, 36-50, and over 51 years).

3.6 Listening Task Conditions and Procedure

All the speech samples contained almost the same questions of the demographic questionnaire filled by the participants apart from the open-ended questions, to let them focus on the speech samples, however sorted in a different order to avoid bias by any effects of a participant's memory. Similar procedures have been applied by Walker et al. (2006), whereas the questions of each speech sample were sorted differently, and participants were instructed to answer the questions as quickly as possible while still being accurate. The speech samples of this research were computerized, natural, and familiar voices; whereas, each participant performed the task once with each speech sample, where gender controlled according to participant’s gender, regardless the familiar voice were the preferred gender was the opposite gender to form a couple. This resulted in a total of three listening tests for each participant. The total duration of the three speech samples together is around three minutes. It is estimated it would take in total up to 30 minutes for the participants to accomplish the listening task including the evaluation process and the break time after each listening test. The evaluation form encourages the participants to give feedback about the task performance of the heard voice and to reflect that feeling on ranking scales.

The participants gave their informed and voluntary consent to participate in the research. In addition, they were encouraged to express ideas that could be applied to the research process. After that, the researcher will record the participants’ familiar voices to assign it to the developed application. The three types of speech samples were evaluated according to users’ self-perception of task performance through evaluative scales, as well as, explored according to users’ perception, experiences, and attitudes towards the speech interfaces through interviews. The listening task has to be done by the developed iPhone application, with each participant working through the task by listening to the speech samples (CV, NV, and FV) to investigate performance and perception of how types of speech are preferred to be used in TTS interfaces. After each listening test each participant has to complete paper-and-pencil evaluative scales and after listening to the three speech samples he/she will be interviewed. The participants would then follow the sequential order of the speech samples. This process is simply to give the participants all the same experience.

Simliar approaches were used in previous studies by Gong and Lai (2001) and Lee et al. (2000). Gong and Lai who investigated the impact of a male gender in both conditions (mixing human voice and TTS vs. TTS-only); their task was managing some emails and calendar tasks of the virtual assistant system of the mobile phone, and their responses used to measure each of voice performance, perception and attitude. Whereas, Lee et al. applied a usability test on the participants in a usability lab and consent was obtained for the videotaping. The participants told that the purpose was to test the prototype of the virtual-assistant application, where the actual purpose was to measure the impact of using different types of voices on the speech output.

With regards to this study, the participants told that the purpose of the listening task is just to measure the performance of each speech sample, where the evaluation process of the listened voices is just to facilitate focus the interview and the interview process contributes by exploring answers that refer specifically to the perception of the listened speech samples. Eventually, the description of the participants' perceptions discussed via an interview, and taken into consideration for exploring user perceptions and experiences of different voices (CV, NV, and FV) of a better speech interface in the next chapter. In this study a listening task was used to examine the impact of different types of speech on the user's perceptions and experiences. The following subsections describe the listening task procedure.

3.6.1 Prior to the Listening Task

Prior to the actual listening task, volunteers filled out a demographic questionnaire to provide information regarding age, gender, TTS applications used, and their favoured familiar voices (see Appendix A – Demographic Questionnaire). On the questionnaire, participants nominated three familiar people that they use or would consider using their voices within the personal applications. Each of these familiar names was rated by likeness and familiarity of the particular person. Afterwards, each participant had to select one of the familiar names provided in order to record his/her favoured voice by the research and then uses that voice as a preferred familiar voice for that particular participant. A reason for selecting this person has to be collected at the demographic questionnaire, and whether the familiar person would be likely to be available for recording his/her voice has to be indicated as well. After that, for each participant three speech samples were designed and integrated to the VoiceTester application as the above-mentioned speech samples integration process (Subsection 3.4.5.2).

3.6.2 Execution of the Listening Task

Participants were instructed to interact with the VoiceTester application to listen to the three speech samples. Each listening test was performed individually. The participants moved through the process firstly with the computerized voice, secondly with the natural voice and finally with the familiar voice. This sequential order of the speech samples has been followed to give the participants all the same experience. After each listening test, general self-perception of task performance of the speech samples in the listening task was measured on a ranking scale, as well as, feedback was collected from the participants to rate quality, general feeling, and any failures due to the experiment's setup. The scale represents a subjective feedback of the listening task. The three listening tests were performed on the same day for each participant (with at least an interval of 5-minute in between). The participants were able to repeat/pause the voices as required. The participants had the choice to stop performing at any time during the task's schedule. After completing the listening task with the VoiceTester application, each participant has been interviewed by the researcher for 30 minutes about his/her body language, feelings, and reactions. Similar procedures have been previously used successfully in audio research studies (Gong & Lai, 2001; Lee, et al., 2000).

3.7 Conducting the Data Collection Phase

Several data collection methods of demographic questionnaires, evaluative scales, and semi-structured interviews will be used to collect data in this research. Because the model is based on user participation, the research design is focused on techniques based on users' perception and experiences. A pilot test was conducted with one of postgraduate colleagues prior to starting the data collection phase. This was done because pilots are useful for testing and refining the research instrument, as well as the data gathering protocol (Seidman, 2006). The pilot testing was an important step for the researcher to gain confidence and experience in interviewing. During the pilot test, the project supervisor who also attended the interview session provided some support by probing some responses that needed more clarification and gave suggestions for improving the interview technique of the researcher. The pilot test suggested some minor modifications of the words and sentence structures to improve interviewee understanding.

The same protocol and research instruments were repeated in each interview. This should improve the reliability of the research, as well as maintaining uniformity of the data, simplifying data analysis (Patton, 2002). Most of the interviews were conducted at the participants' work places to minimize the interruption of participants' schedule. After confirming the interviewee's consent to record the interview, the structure of the interview and the management of privacy were restated. This was followed by a review of the research background and motivation, as well as a high-level conceptual description of the phenomenon to be investigated, with an opportunity for the participant to clarify any points. This ensured all participants had a basic understanding of the concepts and terminology used, but minimised the influence of their future answers to questions.

A good interview depends on open thoughts, feelings, knowledge, and experience of both the interviewer and the interviewee (Collis & Hussey, 2009; Patton, 2002). However, the reason for interviewing starts with the hypothesis that the perspective of someone else's is meaningful, knowable, and able to be made comprehensible. The principle of interviewing is to find out what the other is thinking and, to gather their perception which is called the 'inner perspective' (Patton, 2002). Hence, a sample of interviewees was asked questions about their perceptions. In addition, participants were encouraged to share their perceptions and values in a non-threatening environment. The interviews were held in the participants preferred place and with their consent. While these establish a rapport between the interviewer and the interviewee, rapport depends on the ability to express compassion and sympathy without judgment by conveying to the interviewee that his/her knowledge, experience, attitude, and feeling are important. However, Patton supposed it must not undermine the interviewer neutrality concerning what the interviewee tells. Interviews of the six participants were recorded using the Voice Memos app of the iPhone. After that, a key point analysis approach, according to Patton (2002), was used by the researcher to transcribe the interview data. The results and findings of the collected data are presented in the next chapter.

3.8 Data Analysis

Data analysis follows the data collection stage and in this section, the methods of analysis used to analyse the demographics data, the evaluative data and the interview data are all discussed. The demographics data will be analysed in the next chapter to demonstrate participants' gender, age, and also to show their preferred gender, accent and familiar voices. Qualitative analysis transforms data into findings, and the challenge is that there is no rule for that transformation. This is because each qualitative study is unique and depends on the skills, training, insights, and analytical intelligence and style of the analyst (Patton, 2002). The collected data from each participant are analysed, and qualitative findings are emphasised because of the small sample size. The purpose of this analysis is to organize the description in a manageable way. The method of qualitative analysis employed in this research is a key point analysis of the interview data, which includes the notes taken during the listening test with a full transcript of the interview. The recorded interview from each participant will be transcribed. The intellectual and mechanical work of analysis according to Patton (2002) is coding data, finding patterns, labelling themes, and developing category systems. The interview data will be reduced into categories based on the identification of key ideas. Throughout the qualitative analysis, interesting individual quotes provided by participants were identified for inclusion in the following chapter.

One of the common issues discussed in literature is the generalization of the findings from the participants. This has important implications for this thesis and so several points of view are presented and discussed in the following section, and related to the research in this study.

3.9 Generalizability, Reliability and Validity of the Findings

Generalizability is the degree to which the research findings (often based on a sample) can be extended to other settings. Since this study is interpretive, it uses small samples to produce rich, subjective, qualitative data and its findings will be allowed to be generalized from one setting to another similar setting. The findings obtained from a sample of six participants can be used to represent and make inferences to a model (Collis & Hussey, 2009). This research has a sample size of six participants in the form of three couples chosen from different age groups (21-35, 36-50, and over 51 years) to increase the credibility by reducing the participants' age bias as the study targeted the public users of mobile devices. All the participants preferred to be in couples to facilitate the listening test process and to simplify the familiar voices selection. As in the case of a couple, their familiar voice would be each other's voice. Generalizability can also largely depend on a reader's interpretation based on their experience, in that the findings of the research may match the reader's experience as defined by Patton (2002). Thus, other researchers can use the findings of this study as a base to conduct further research in this area.

Reliability, the research is reliable if it can be conducted again in the future and to produce the same results and inferences. For a research to be reliable, the sample size must be representative, that is, not too small or too large. In this case, an initial sample size of six participants (three males, three females) is enough for qualitative data to reach the saturation level if not then more participants will be needed to draw findings about the three speech types. According to interpretivism paradigm, the produced findings are with low reliability as it would not produce the same findings if this study is replicated (Collis & Hussey, 2009). However, the focus of the developed application is to provide the same experience to all the participants so that the collected data will be valid and accurate, so as not to rely on recall and to remove the potential for the participants to report different things which can help add credibility to the research (Patton, 2002). In addition, adopting a methodological triangulation (i.e., using the evaluative rating scales to improve the interview process) would add validity to the study. This means the findings of the research study are very much reliable.

Validity is the degree to which the finding of a research accurately reflects the phenomena under study (Collis & Hussey, 2009). However, the validity of a sample size depends on the in-depth information which will illuminate the research question, described by Patton (2002) as “the validity, meaningfulness, and insights generated from qualitative inquiry having more to do with the information richness of the cases selected and the observation/analytical capabilities of the researcher than with samples size” (p. 243). However, being an interpretive study, it will embark extracting data which will provide detailed explanations as well as capture the essence of the phenomena being studied. The aim of interpretivism is to provide full insight and understanding of those involved in the phenomena which adding to the validity. Also, Myers and Avison (2002) stated that a research is valid if it demonstrates what the researcher claims it does. However, the validity of a research can be undermined if there are research errors such as poor samples, faulty research procedures and inaccurate measurements. The main reason of conducting a research study is to try and find solutions to the research question. The research findings must therefore have valid relevance to the research topic. It is also important getting each participant to experience the speech samples and identify points in the listening task in order to ask about it in the interview using the observation (audio recording) thus adding validity to the study. In addition, according to Collis and Hussey (2009), adopting a methodological triangulation (i.e., using the evaluative rating scales to guide the interview questions) would increase validity of the study. Therefore, the adopted research design including the research paradigm, methodology, methods and techniques are valid hence the findings of this study are valid as well.

3.10 Conclusion

This chapter justifies a specific approach and methodology that would be useful for this study as the data collection and data analysis methods to be used have been discussed and justified, verifying that these will enable the aims and the research question to be realized and solved. With regards to the research methodology, the use of the listening task approach is vindicated; the use of demographic questionnaire, evaluative scales, and semi-structured interview as data collection method is verified; using demographic analysis and key point analysis method as data analysis for the demographic data and the interviews respectively is also logical. This chapter also includes details on the implementation of the selected research methods and techniques, showing how the used protocols contribute to generalizability, reliability and validity.

Regarding the developed mobile application, Xcode software of an iMac was used to develop the application, and an iPhone 3Gs was used to run the listening task to present different speech samples. In terms of generating and recording the stimuli, the Cepstral engine was used for generating the CV samples, Voice Memo was used to record the human voices (NV and FV), Camtasia software was used for fixing the audio files, and the Xcode was used for integrating the speech samples with the developed application.

In the next chapter, the findings of the interviews data are presented, along with some relevant discussions relating to the research question.

CHAPTER IV
RESEARCH FINDINGS AND DISCUSSIONS

4.1 Introduction

This chapter presents an analysis of the results and the findings. The results are discussed together with the appropriate findings. Moreover, the key themes are discussed through conducting an analysis to the collected data. In this chapter, the results and the findings of this analysis progression are presented in the form of research analysis results and findings, with discussion of important aspects of these findings. Developing the conclusions of this research, answering the research question, and presenting possible future research directions and limitations are based from the analysis and discussions of the results and findings of this chapter.

To achieve the depth of understanding of current practices and issues in evaluating suitable TTS interfaces, in addition to their association with existing literature, the data obtained from the interviews were coded, categorized, and analysed upon the key point coding of Patton (2002). And the data obtained from the speech samples evaluation forms were analysed by Excel and descriptive figures of the means opinions were exported to find the differences in the perceived quality of the voices. Developing the structure of the data analysis was supported from the literature review to identify patterns of practice in selecting acceptable TTS interfaces. The analysis of the data comprises a basic demographic analysis, results from listening activity tests, interviews, comparison of practice, as well as comparison with current literature.

In the following sections, the results and findings of this research are analysed and discussed. Section 4.2 presents and discusses the background of the participants. Section 4.3 describes the research results, findings, analysis and discussion. Section 4.4 includes several subsections constructed around the dimensions of the framework of analysis presented at the end of Chapter 2 (Figure 2.3), discusses the analyses of the interview data with key point coding and presents the findings of the interviews. Section 4.5 discusses the findings, as well as the resultant model from the analysis of the finding. Finally, Section 4.6 summarises the main findings and conclusions from this chapter.

4.2 Background of the Participants

In a key point coding technique there should be some similar characteristics among the interviewees in order to ensure that the comparison of the interviewees along these similar dimensions is valid (Patton, 2002). For example, the aim was to select three healthy couples, over 20 years of age with English as their first language. The three couples have been named X, Y, and Z, accordingly. Also, the preferred gender of their current TTS applications and their favoured familiar voices were included in the demographic data gathered. Moreover, there are also specific demographic questions asked in the survey questionnaire.

Another aim was to see if there were variations in practices and issues among different participants, as well as participants selected demographically. From literature it is not clear if preferred familiar voice gender depend on the user gender. Because some studies showed that the preferred gender depended on some factors and especially on the type and purpose of the application itself (Subsection 2.5.1), this therefore was also collected as part of the demographics. The experience, familiarity with the relationships and likeness, and then nominating a favourable familiar person may also affect the data gathered about speech interfaces preference, so this was included in the alongside collected data partially for the participant background data. Some of these demographic data are summarized in Table 4.1.

Table 4.1 - Participants background

Couple	Participant	Age	Gender	Preferred Gender	Familiar voice
X	1	57	Male	Female	Wife
	2	54	Female	Doesn’t matter	Husband
Y	1	37	Male	Doesn’t matter	Sister
	2	36	Female	Male	Brother
Z	1	23	Male	Female	Friend
	2	21	Female	Male	Friend

From the table above, information about the preferred participant gender and voice preferences is exposed. From the interviewees' transcripts, two out of six participants can be categorized as experienced TTS users since they are iPhone users and they are using GPS navigation systems with a TTS interface in weekly bases. Two from the remaining four participants are considered as motivational participants, as they stated 'they would like their familiar voice to be integrated within some of their personal application', while the first two participants (Couple X) were unfamiliar with applications based on speech synthesis interfaces, as they commented that 'it's hard to accept new things easily in that age'.

Background data was collected from the participants to determine whether preferred familiar voice gender depends on the user's gender, experience, and age or on other factors and especially on the purpose of the application itself. From the background data presented in the above table, note that four out the six participants preferred the TTS voice to be from the opposite gender while two out of six participants did not have preference of any gender. This indicates that the users' gender could influence their preference of the familiar voices. Based on participants' responses to a pre-experiment survey, participants are aged from 21-57 years old, and had been chosen from different age groups; two participants from 21-35 age range, two from 36-50 age range, and two over 51 years old. Regarding prior exposure to TTS, four participants reported to have heard TTS once or twice, while three reported listening to it with some regularity but less than a few times a week. None of the participants reported working with TTS. All participants were native English speakers with no reported hearing problems. Table 4.2 describes in some detail couples' occupation and lifestyle attributes. As this shows, these attributes figure importantly in understanding certain features in familiar voice selection practice.

Table 4.2 - Subjects description

Couple	Age	Occupation & Important Lifestyle Attributes
X	51+	Manager and homemaker, husband and wife, aged 57 and 54 respectively. They have been married for more than 30 years. For that reason they chose each others' voice over other nominated voices as theirr familiar voice, explaining their feelings of being more comfortable and trusting in this choice. They use a GPS navigation system with female gender and English accent, voicemail of their own voices, and answer machine phone with his wife's voice. However, she doesn't like that commenting 'I do not actually enjoy hearing my voice'. The husband preferred an opposite gender as either his wife's voice or his daughters' voice. However, the other half preferred her spouse to be a first priority and her best friend with a similar gender as second priority, as she doesn't care about the voice gender. Additionally, both of them ranked a family member's voice higher than a friend's voice.
Y	36-50	Dentist and pharmacist, brother and sister, aged 37 and 36 respectively. They use a GPS with a female gender and UK English accent. However, the brother doesn't mind the voice gender while his sister prefers a male voice instead. He has selected his sister's voice to be his familiar voice; however, he prefers Norah Jones voice (Singer). His sister preferred a friend to be her familiar voice as she said 'he has a clear deep voice' but she chose her brother's voice for reasons of availability. However, both of them considered the maximum familiarity for each other and less for the other nominated people who are not family members.
Z	21-35	Web designer and accountant, male and female, aged 23 and 21 respectively. This couple have been best friends for more than 10 years. They preferred each other's voice as they are familiar with these voice characteristics of accent, tone and pace. Moreover, he used an opposite gender for his GPS navigation system and she thinks it depends on the language style as she would like hearing a female voice for an American accent and male voice for a British accent. However, both of them favoured opposite gender to be their familiar voice. Also,, a higher ranking went to their family and relative members than to each other's friends.

The differences in a listening task performance, voice perception and voice characteristics of gender and accent are discussed in the next section.

4.3 Research Results and Findings

In this section, the research analyses, results and findings are presented. The aim of this study is to understand speech perception in relation to the examined speech samples. This research specifically implicates speech interface in the development of GMT application for TBI patients helping them to undertake activities with the necessary support and preventing them from making errors while gaining insight and understanding before more rigorous investigation. The research findings of this study are based on the interviews of the six participants. These were the six participants who agreed to participate in this research. The interviews were conducted with three males and three females in order to control gender bias, aged ranged from 21 to 61 years old.

The following subsections are based on the aspects of the framework of analysis identified in Chapter 2 (Figure 2.3). The interview transcripts are analysed and coded to develop themes and patterns. Their implications for research and existing research are discussed. In Subsection 4.3.1 the results of the self-perception of task performance evaluation are analysed. Four adjectives have emerged into themes related to speech interfaces perception. In Subsection 4.3.2 the interviewees' perception is analysed by the key point analysis, which further emerged from codes, concepts and categories to structure a model.

4.3.1 Results of the Listening Task

The results of evaluating the speech samples are measured on four semantic adjectives. Each adjective has been measured on rating scales which are intended to elicit the participants' perception of task performance in the context of the different speech samples. Appendix H – The Evaluative Data (Table H.1) shows the overall self-perception of task performance for the CV, NV, and FV conditions of each participant in the form of a table. And, Table H.2 summarises and shows, accordingly, the collected evaluative data of the six participants. The results show that there is significant variety in the preferences of the speech samples represented by the participants and the highest ranked are shown in bold text. By exploring these data to Excel sheet bar charts generated, showing the self-perception of task performance according to demographic data of the participants. However, all these results were not included in this study.

4.3.2 Findings of the Interviews

This section shows the analytical framework approach for organizing and reporting qualitative data, as the responses of the six participants' interviews can be organized question by question, especially where a standardized interviewing format is used. Interesting individual direct-quotes provided by participants were identified for clarifying the situation and thoughts of each participant. As suggested by Patton (2002), the basis of qualitative reporting findings relies on the balance between adequate description and direct quotations. And reporting these findings considered as the final step in data reduction. Moreover, during data collection and analysis of the interview data, each two participants of a couple have been grouped into X, Y, and Z, as they share many things in common (e.g., age, relationship, familiar voice, etc.; see Table 4.2). The followed number represents the participant number of that couple (i.e., X_1 represents the first participant of Couple X).

Additionally, during the debriefing at the interview after all the listening tests, some participants at the FV condition commented that it would be best to record the FV with better instruments making sure it's professionally recorded as the NV. Others were forced to choose a second priority familiar person instead of their most preferred person, because of some reasons (e.g., they are not available for recording their voices or privacy issues). Some participants at the NV condition also commented that the recorded speech was quite easy to understand, and commented in the CV condition that the fixed prompts were preferred to be slower. Thus, all these credible comments and notes have been included in the analysis of the interviews in the next section.

4.3.2.1 Couple X: Participant 1 (X_1). Preferred the familiar voice as 'it is warmer' and recommended using familiar voice with his personal applications mentioning he welcomed hearing his wife's voice on the answer phone at home. Additionally, he preferred using his personal voice for his own mobile phone. Regarding the natural voice, he perceived it as real and nice and liked the native accent. He judged that however by depending on the New Zealand accent. Interestingly, he recommended using a computer-generated voice in the case of electrical devices (microwave, heater, TV, etc). And also he expressed the computer-generated voice as impersonal; however he perceived the quality as good and he didn't mind it even for it being computerized.

Moreover, he expressed his preference for the opposite gender by stating 'if my familiar voice is male then I wouldn't like it that much', and also obviously as he using a GPS navigation system with a TTS female gender and UK accent for availability as he would prefer native accent.

4.3.2.2 Couple X: Participant 2 (X_2). Also preferred the familiar voice because of ease of understanding, comfortableness and relaxation stating, 'Hearing my husband's voice is easier to understand, makes me feel more comfortable when you hear someone you more familiar with and relax more.' She also recommended using her personal voice with her personal application, answer machine or voice mail, as she doesn't want her friends to hear her husband's voice. However, she said 'I do not enjoy hearing my voice' so probably a computer-generated voice is recommended in this case as an alternative solution. Regarding the natural voice, she perceived it easy to understand the nice native accent. However, she doesn't mind a variety as she is used to different accents from TV, and also she doesn't mind an accent in the case of a computerized voice as a worldly accent would be preferred. Surprisingly, she recommended TTS interfaces of electrical devices (microwave, heater, TV, etc) to be more computerized soundings. And also she preferred the person to be slower and clearer in the case of a computer-generated voice. However she could understand the CG voice easily and she didn't mind the quality.

Moreover, she doesn't have preference for the gender; being obvious from using GPS navigation system with a similar gender.

4.3.2.3 Couple Y: Participant 3 (Y_1). He described the familiar voice as having a great impact on his perception as well as being more understandable and suitable for him. However, he preferred the computerized voice over the other voices because of their familiarity and interactivity, but preferred his familiar voice to be a preferred celebrity's voice instead of his sister's voice. Also, he understood the natural voice easily, although he is used to hearing a computerized voice as TTS interface, however, he didn't mind the idea of using natural voice as a speech interface. Moreover, the accent of the CV is not a matter for him, but he found the UK accent interesting so he preferred the natural voice to be with UK accent instead of with native accent. In addition, he perceived the CV as it is less fluid than the natural voice articulating each word more slowly as claimed, 'it can tell one spoken word at a time'.

A familiar voice is preferred in case of availability because of it being more comfortable to hear someone you know. However, if it is difficult to produce individualized familiar voices then natural voices would be better for others as it is more understandable especially if it was with a likeable accent. For general use probably a natural voice would simplify the use of its applications, otherwise if it is impossible then a computerized voice is preferred because it is something that he is more used to.

He didn't mind the quality, moreover, he also doesn't have preference for the gender, although, he is using a GPS navigation system with an opposite gender and UK accent.

4.3.2.4 Couple Y: Participant 4 (Y_2). She didn't like hearing her brother's voice, although she preferred her friend to be her familiar voice for the reason of availability issues. She also recommended a professional recording of the familiar voice especially with the speaker practicing the passage for recording to fit the intonations of a natural voice. On the other hand, the natural voice was her preferred voice as it is recorded professionally with good intonation, paced speech and accent. She found the natural voice is the most appropriate voice as it is clear and a lot easier to listen to, and she liked the native accent. However, she perceived the computer-generated voice as 'robotic, just like the GPS, and I have to concentrate little harder to understand'; moreover, she thinks that the natural voice 'simplifies the use of the application because it's natural'.

Furthermore, she prefers male gender with UK accent as she thinks it is the clearest accent. She likes her GPS navigation system with a similar gender and UK accent.

4.3.2.5 Couple Z: Participant 5 (Z_1). He thinks he did well in the listening task, stating it was easy to distinguish between the given voices and the way that they were recorded or generated. He understood the familiar voice but expressed his perception about the recording quality and the speaker.

He recommended the familiar speech samples to be tested with the end user to make sure it is understandable and can then be applied to the application; moreover, he would prefer the familiar voice if more time was spent on it and recorded with professional equipment to eliminate distortion. The natural voice was perceived as the favourite one for its professional speaking as he expressed his perception about this voice by 'probably recorded in a better environment with better equipment as well'. Regardless of the advancement of technology of the computer-generated voice, the voice was hard to understand by reasons of pace and pronunciation of words. In addition, he thinks 'the Computer Generated technology has still not developed enough to match a natural voice', which is true as discussed in Chapter 2, and he personally thinks having a female voice could make it easier to understand.

He recommended using a natural voice as a common solution by stating 'I chose a female voice because I know the girl more, so I'm used to the way she speaks'. However, he would prefer the natural voice even more if it was a female voice as generally a female voice has a nicer tone to enjoy with use of applications based TTS interfaces. He thought that by having an option to choose between the male and female voice in each category, probably would help him to distinguish which gender to prefer.

4.3.2.6 Couple Z: Participant 6 (Z_2). She chose her familiar person for the reason that 'he has a clear New Zealand accent, easy on the ear and clear enough to understand'. She thinks she did well with both the natural voice and the familiar voice, stating it was easy to understand the given speech samples and the way that they were recorded or generated. She understood the CSS sample but expressed negative perception towards the synthesized speech samples by her unenthusiastic ranking of the computer synthesized voice performance.

She recommended to use both voices but each one for its specific uses and purposes, natural voice for a common use application like a GPS navigation system and familiar voice for a personal application and, personally, she thinks people with special needs would much appreciate such interface. She doesn't mind about the gender, however, she used a female gender with her GPS as she thinks it depends on the language and the accent itself so in the case of English she prefers female. She prefers the natural voice on computerized voice with her GPS as it is a non personal device so there is no need for a familiar voice. Also, she would use a native accent if available.

4.4 Analysis of the Findings

This section includes a discussion of some attributes that are structured from the data analysis of interviewing the participants about speech perception. In line with the framework of qualitative analysis according to Patton (2002) is starting with cross-case analysis, meaning the grouping together of interview answers from different participants to common questions, or analyzing different perspectives on central issues. The intellectual and mechanical work of the qualitative analysis is coding data, findings patterns, labelling themes, and developing category systems. This essentially means analysing the core content of interviews to determine what's significant. After completing this initial thematic coding process these themes were then categorised into the factors from the framework of analysis depicted in Figure 2.3. This allowed the usefulness of the framework of analysis to be evaluated by checking that all the identified themes could be categorised reasonably within the dimensions of the framework of analysis. Otherwise, the model represented by the framework of analysis would need to be extended or modified. Also this framework is the basis for comparison of the interview data with literature. For instance: a quote from the transcript which was coded as "understandable" could reasonably be placed in the "easy" concept of the analysis framework. During the coding process the link between a coded extract of the transcript and its location in a participant's original transcript was handled by utilizing the colour coding (i.e., red, blue, and green which these primary colours represent the categories FV, NV, and CV, respectively). The process of categorizing the coded themes was done in iteratively until all of the codes were categorized.

The coding started from the first interview that led to patterns, and then these patterns emerged to generate themes, which then emerged to explore the framework (Figure 2.3). Closed coding was used in the content analysis, which it is the key points identified in the transcripts where compared with patterns and themes so far established and adjustments made to themes to reflect accumulated findings. Additionally, it is considered as qualitative data reduction method. These, were then used in subsequent analysis. Alternatively, this process of searching for patterns and themes is called ‘pattern analysis’ or ‘theme analysis’ as suggested by Patton (2002). However, Myers and Avison (2002) defined this process as the constant comparative method, which requires searching out and checking of contrasts and negative evidence. When no additional data were being collected to develop or add to the set of concepts and categories, a situation noted by (Glaser & Strauss, 1967 as cited from Myers & Avison’s study) referred to as ‘theoretical saturation’ emerged. Emerging the themes is shown diagrammatically in Subsection 4.4.4.

The following subsections illustrate the steps of the intellectual and mechanical work of the key point analysis of coding data, findings patterns, labelling themes, and developing category systems.

4.4.1 Key point coding

The first step of qualitative analysis is developing some manageable classification or coding scheme. The interview data of the three couples (six participants) were analysed, according to Patton (2002), by the key point analysis.

The first letter and the followed number represent the key points made on an interview data of the same participant of that couple. The ‘X’ letter represents Couple X, used to differentiate key points over other key points made by subsequent participants. The followed number identifies the participant number of that couple. That is, X_1 and X_2 refer to participant 1 and 2, respectively, of Couple X; Y1 and Y_2 refer to participant 4 and 5, correspondingly, of Couple Y; Z_1 and Z_2 refer to participants 5 and 6, respectively, of Couple Z. The table below shows the key point analysis and codes from the data in Couple X. The numbering used to identify attributes sequentially, started at the first interview, continuing on through subsequent interviews. The key point $X_1$1a arose on a second subsequent pass of the data to avoid the re-sequencing on every pass (see Table 4.3).

Table 4.3 - Key points analysis and codes from the data in Couple X

Id	Key point	Code
- **Participant 1 (X_1)**		
$X_1$1	*The familiar voice is better, its warmer just more personal*	Warm Personal
$X_1$1a	*With personal applications it's better to have a familiar voice.*	Preferable
$X_1$2	*The person with familiar voice would feel more comfortable than computer-generated voice which would feel too cold.*	Comfy Cold
$X_1$3	*The overall voice quality was good, the computer-generated voice was fine, and I didn't mind it even for it was not real, But I prefer the familiar voice in such applications.*	Fine Preferable
$X_1$4	*The computer generated voice is impersonal, and obviously the familiar voice is personal and feels happier with.*	Impersonal Personal Restful
$X_1$5	*If my familiar voice is male then I wouldn't like it that much, I prefer female familiar voice.*	Dislike similar gender Prefer opposite gender
$X_1$6	*Local natural voice is real and nice but is depending on the NZ accent*	Natural Nice
$X_1$7	*Use familiar voice instead of the computer generated voice, as this voice is real with better quality, more understandable.*	Superior Real Understandable
$X_1$8	*Recommend computer-generated voice in case of microwave, heater, television, etc…*	Not convenient
$X_1$9	*Using GPS navigation system with a TTS female gender and UK accent.*	Opposite gender preferable

$X_1$10	*Welcome hearing my wife voice on the answer phone at home.*	Relaxing
- **Participant 2 (X_2)**		
$X_2$1	*Hearing my husband voice is easier to understand.* *Makes you more comfortable when you hear someone you more familiar with.*	Ease of understand Comfy
$X_2$2	*Relax more if that's comfortable you used to that accent you far more comfortable about it and so you probably relax more.* *Just I would relax more, and maybe take it warmer.*	Relaxing Warmer
$X_2$3	*TTS voices: clear enough that were fine. Computer-generated voice: could understand clearly, didn't mind it.*	Fine
$X_2$4	*Voice Mail or home answer machine: prefer my own voice, a people ringing in that know me, I don't want them to hear my husband answers, and I want them to hear me.* *I do not enjoy hearing my voice.*	Dislike hearing own voice
$X_2$5	*Easier to understand a familiar voice than computer-generated voice.*	Ease of understand
$X_2$6	*Good idea to use familiar voice with my TTS devices.*	Suitable
$X_2$7	*TTS of Electrical devices: better to be more computers sounding really.*	Not convenient
$X_2$8	*My familiar voices just got nice voices and clear.* *I've chose my husband, because he's far more familiar to me I suppose in a sense of trust and in a sense of relaxation and that sort of thing.*	Clear Trustable Relaxing
$X_2$9	*Probably a familiar voice, unless the person would be slightly slower and clear,* *not a real accent, a worldly accent,* *I don't mind about the gender at all.*	Preferable Fast
$X_2$10	*I like my husband voice as I could understand him*	Likable

	clearly.	Ease of understand
$X_2$11	*Natural voice: I actually like the nice accent, the kiwi accent I can hear easier to understand. I do find the kiwi accent is easy.* *Used to many accents from TV so don't really mind a variety.*	Likable Ease of understand

The codes are then analysed and those that relate to a common theme are grouped together. This higher order commonality is called a *concept*. An illustration is given in Subsection 4.4.2. Concepts are then grouped and regrouped to find yet higher order commonalities called *categories*, illustration in Subsection 4.4.3. It is these concepts and categories that lead to the emergence of a *model.*

4.4.2 The emergence of concepts

With regards this study, eight concepts represent users' perception and attitude have been adopted from relevant study (Gong & Lai, 2001). Which are they: difficult/easy, uncomfortable/comfortable, inconvenient/convenient, and inefficient/efficient. The code 'warm' emerged from $X_1$1. The codes from all other key points were compared with this to see if similar codes occurred often. The following codes were considered to have commonality: 'warm' from $X_1$1; 'comfy' from $X_1$2; 'restful' from $X_1$4; 'nice' from $X_1$6; 'relaxing' from $X_1$10; 'easy of understand' from $X_2$1; 'relaxing' from $X_2$2; 'relaxing' from $X_2$8; 'easy of understand' from $X_2$10; and 'easy of understand' from $X_2$11. The common characteristic would be 'comfortable' and this was the first concept to emerge from the data. Other combinations of codes led to further concepts and these were added to Table 4.4.

Begin by looking for recurring regularities in the data. These regularities reveal patterns that can be sorted into categories (Patton, 2002). The process of comparing the codes with each other, to find higher order commonality, produced the concepts from the codes. The eight concepts of Couple X are summarised below.

Table 4.4 - Emergence of concepts from the codes in Couple X data

Difficult	
Easy	$X_1$7, $X_2$1, $X_2$5, $X_2$10, $X_2$11
Uncomfortable	$X_1$2, $X_1$4, $X_2$4
Comfortable	$X_1$1, $X_1$2, $X_1$4, $X_1$6, $X_1$10, $X_2$1, $X_2$2, $X_2$8, $X_2$10, $X_2$11
Inconvenient	$X_1$5, $X_1$8, $X_2$7
Convenient	$X_1$1, $X_1$1a, $X_1$3, $X_1$4, $X_1$5, $X_1$9, $X_2$6, $X_2$8, $X_2$9
Inefficient	$X_2$9
Efficient	$X_1$3, $X_1$6, $X_1$7, $X_2$3, $X_2$8

4.4.3 Categories

By comparing each concept in turn with all other concepts, further commonalities are found which form the even broader categories, to verify the meaningfulness and accuracy of the categories and the placement of data in categories. This method of continually comparing concepts with each other is called constant comparative method (Patton, 2002). However, with regards to this study, three categories have been also emerged based on the framework of analysis (Figure 2.3), which they are: FV, NV, and CV perception.

4.4.3.1 The emergence of categories form Couple X (X_1 and X_2)

The first couple consisting of husband and wife aged 57 and 54 respectively. This couple considered as participants one and two. By applying the constant comparison technique to each concept in turn, common themes were found amongst the following concepts: *easy; uncomfortable; comfortable; inconvenient; convenient; efficient.* These share the theme of FV perception. This was the first category to emerge from the data and is demonstrated diagrammatically in Figure 4.1. However, the iteration between data and concepts ended when enough categories and associated concepts had been defined (Myers & Avison, 2002).

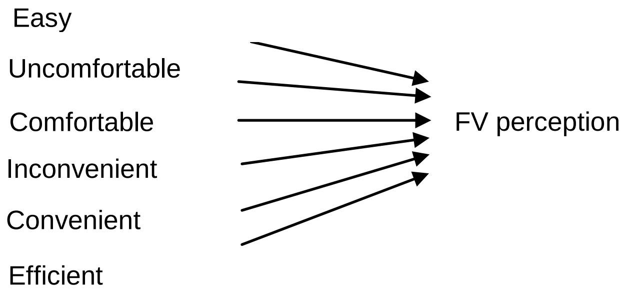


Figure 4.1 - Diagrammatical emergence of the category 'FV perception' for Couple X

By comparing the other concepts and grouping *easy*, *comfortable*, *and efficient* the 'NV perception' category emerged in Figure 4.2.

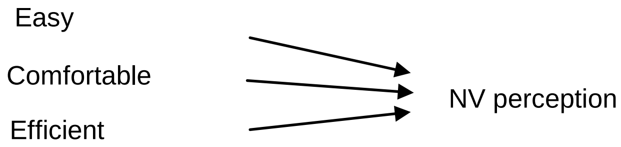


Figure 4.2 - Diagrammatical emergence of the category 'NV perception' for Couple X

Grouping *uncomfortable, inconvenient*, *convenient, inefficient and efficient* in Figure 4.3 gave the category 'CV perception'.

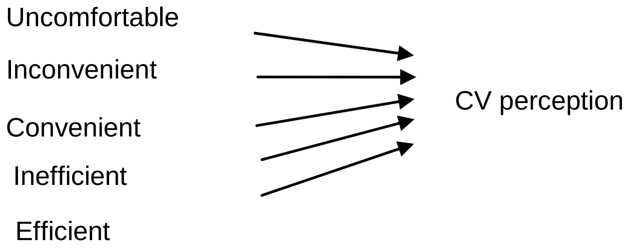


Figure 4.3 - Diagrammatical emergence of the category 'CV perception' for Couple X

Data from two other Couples were analysed to further establish or otherwise define these categories and fill the empty concepts.

4.4.3.2 The emergence of categories from Couple Y (Y_3 and Y_4)

The second couple was brother and sister aged 37 and 36 respectively. This couple considered as participants three and four. The analysis proceeded by comparing the new key points with the concepts and categories established. New concepts will be established. This evidence added further substance of the findings. Since qualitatively there are no statistical tests to tell if an observation or pattern is significant, so the researcher must rely first on his/her own intelligence, experience, and judgment; secondly, the researcher should take seriously the responses of those who were studied or participated in the inquiry (Patton, 2002). The full key point coding and the emergence of concepts can be found at Appendix I – Full Interview Analysis of Couple Y.

The concept 'Inconvenient' of CV perception category, as well as, the concepts 'Uncomfortable' and 'Efficient' of each FV perception and CV perception categories had no support from this data analysis. However, under each category a new concept emerged as:

FV perception category: 'Inefficient' - $Y_1$1, $Y_2$1

NV perception category: 'Convenient' - $Y_1$2, $Y_1$5, $Y_1$8, $Y_1$10, $Y_1$11, $Y_2$2, $Y_2$4, $Y_2$5

CV perception category: 'Difficult' - $Y_2$3

4.4.3.3 The emergence of categories from Couple Z (Z_5 and Z_6)

The third couple were friends, male and female aged 23 and 21 respectively. This couple were considered as participants five and six. The analysis of the key points into their codes continued, by searching for key points in the data and identifying codes. For details about the full analysis showing the key points and codes and emergence of concepts of this couple can be found at Appendix J – Full Interview Analysis of Couple Z.

There isn't a new concept emerging so this can be considered as reaching the saturation level to stop the analysis. As subsequent interviews took place, in any couple, the process of constant comparison continued. Key points identified in the transcripts were compared with the established concepts and categories and adjustments made to categories and reflect accumulated findings. These, in turn, were then used in subsequent analysis. However, the iteration between data and concepts ended when enough categories and associated concepts had been defined (Myers & Avison, 2002). Modelling each of the categories is shown diagrammatically in the following section.

4.4.3.4 The emergence of categories from 3 Couples

These categories and their relevant concepts are displayed in Table 4.5.

Table 4.5 - Emergence of categories from the concepts in the data from 3 Couple

FV perception	
Difficult	
Easy	X_17, X_21, X_25, X_210, Y_19, Z_11, Z_21
Uncomfortable	X_24
Comfortable	X_11, X_12, X_14, X_110, X_21, X_22, X_28, X_210, Y_11, Y_112, Z_17, Z_112, Z_21
Inconvenient	X_15, Y_11, Y_21, Z_11, Z_19
Convenient	X_11, X_11a, X_13, X_14, X_15, X_26, X_28, X_29, Y_14, Y_19, Y_111, Y_112, Z_14, Z_18, Z_22, Z_26
Inefficient	Y_21, Z_18
Efficient	X_17, X_28, Z_110
NV perception	
Difficult	
Easy	X_211, Y_12, Y_25
Uncomfortable	
Comfortable	X_16, X_211, Y_18, Y_22, Y_26
Inconvenient	
Convenient	Y_12, Y_15, Y_18, Y_110, Y_111, Y_22, Y_24, Y_25, Z_12, Z_15, Z_23,
Inefficient	

Efficient	$X_1$6, $Y_1$5, $Y_1$8, $Y_2$2, $Y_2$4, $Y_2$5, $Y_2$6, $Z_1$2,
CV perception	
Difficult	$Y_2$3, $Z_1$3, $Z_1$6, $Z_1$11, $Z_2$4
Easy	
Uncomfortable	$X_1$2, $X_1$4
Comfortable	
Inconvenient	$X_1$8, $X_2$7, $Z_1$11
Convenient	$X_1$9, $Y_1$3, $Y_1$7, $Z_1$4a, $Z_1$6, $Z_1$13 $Z_2$5, $Z_2$7
Inefficient	$X_2$9, $Y_1$6, $Y_2$3, $Z_1$3, $Z_1$6
Efficient	$X_1$3, $X_2$3

4.4.4 Emerging Model

By linking the categories and investigating the connections between concepts the model emerges. As shown below, the model consists of three categories, namely: FV perception, NV perception, and CV perception. Each category has been linked to eight concepts (i.e., difficult/easy, uncomfortable/comfortable, inconvenient/convenient, and inefficient/efficient) to make the data analysis easier and more accurate. Each concept has been represented by a group of codes with similar meanings which has been taken from key points of the interview data.

4.4.4.1 FV Perception

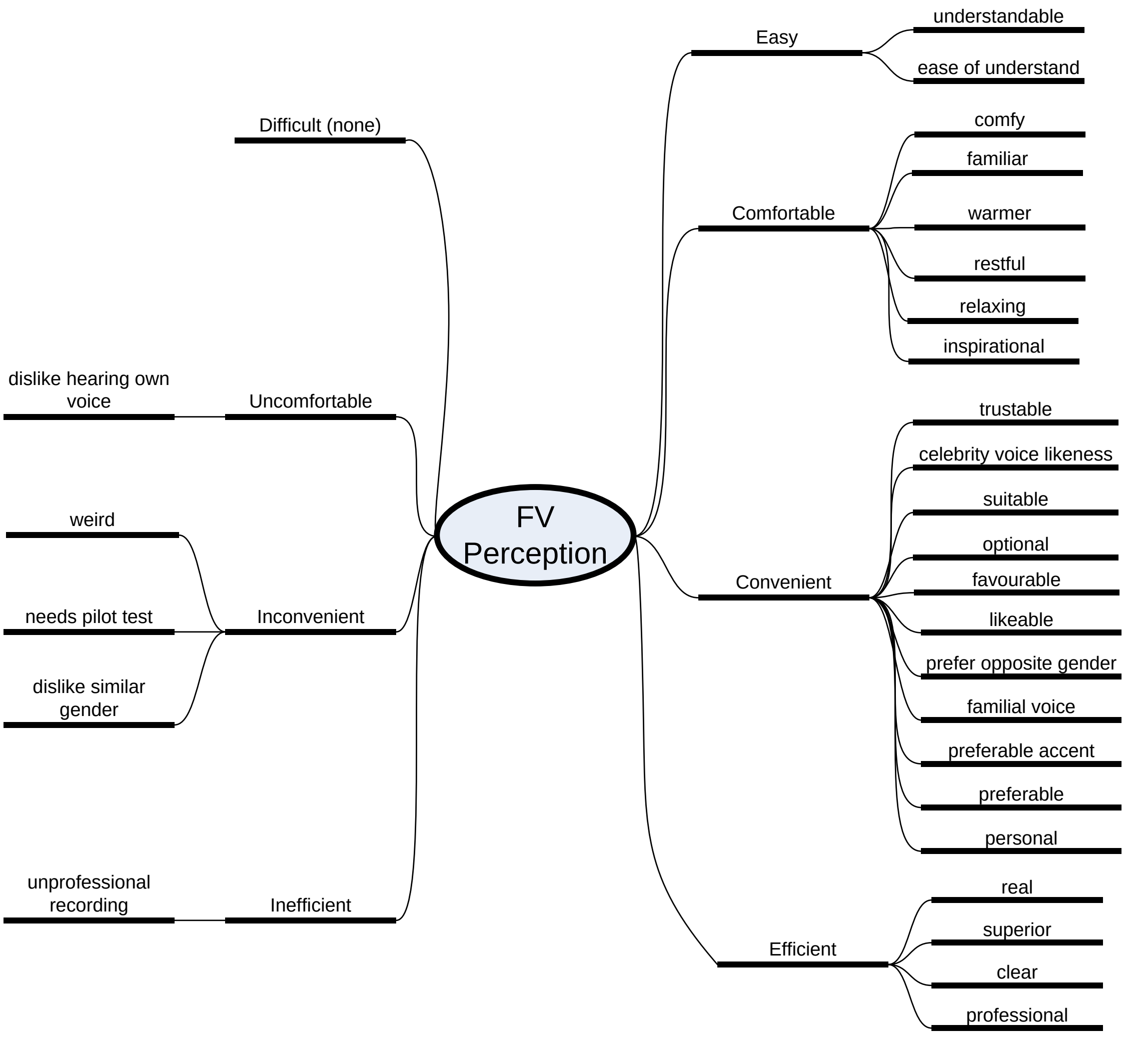


Figure 4.4 - Emergent concepts in 'FV perception' derived from key point analysis of interview data in 3 couples (source: author)

FV interface is familiar therefore it is warm, comfy, relaxing, and restful, however, hearing own voice is uncomfortable. Regarding voice convenience, FV is favourable as it is trustable. Moreover, it is preferable with opposite gender; a celebrity voice would be recommended for such speech interface. This voice is suitable for personal applications based speech interfaces, however, NV interface would be recommended if producing individualized FV interface is not available. Familial voice is recommended to be used in home phone's answer machine. FV needs to be recorded and tested with the end user before it is applied to the application, because of the person's voice is known so it is weird to listen to, and because of the familiar people are not talented speakers. In contrast, NV is more professional and is probably recorded in a better environment, with better equipment and talented speakers. However, this voice is preferred over CV, as it is real with better quality and is more understandable.

4.4.4.2 NV Perception

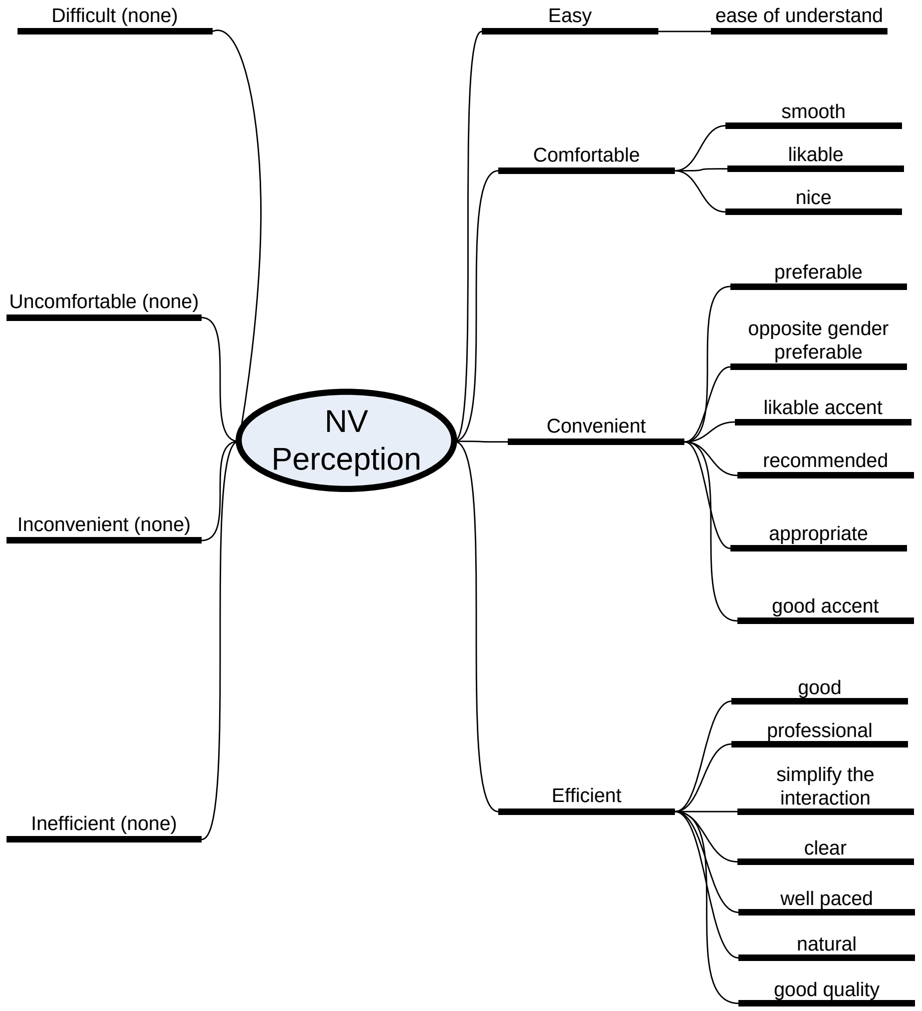


Figure 4.5 - Emergent concepts of 'NV perception' derived from key point analysis of interview data in 3 couples (source: author)

NV interface is perceived as an easy voice to understand. Regarding comfortableness, it is considered smooth and nice. With regard to convenience, it is preferable to use opposite gender with a native accent speaker, and it is preferred and recommended to have such a good accent. In regard to voice efficiency, the NV is good, clear, natural, well paced, and perceived professionally recorded with good quality; therefore, it simplifies the interaction of speech interfaces.

4.4.4.3 CV Perception

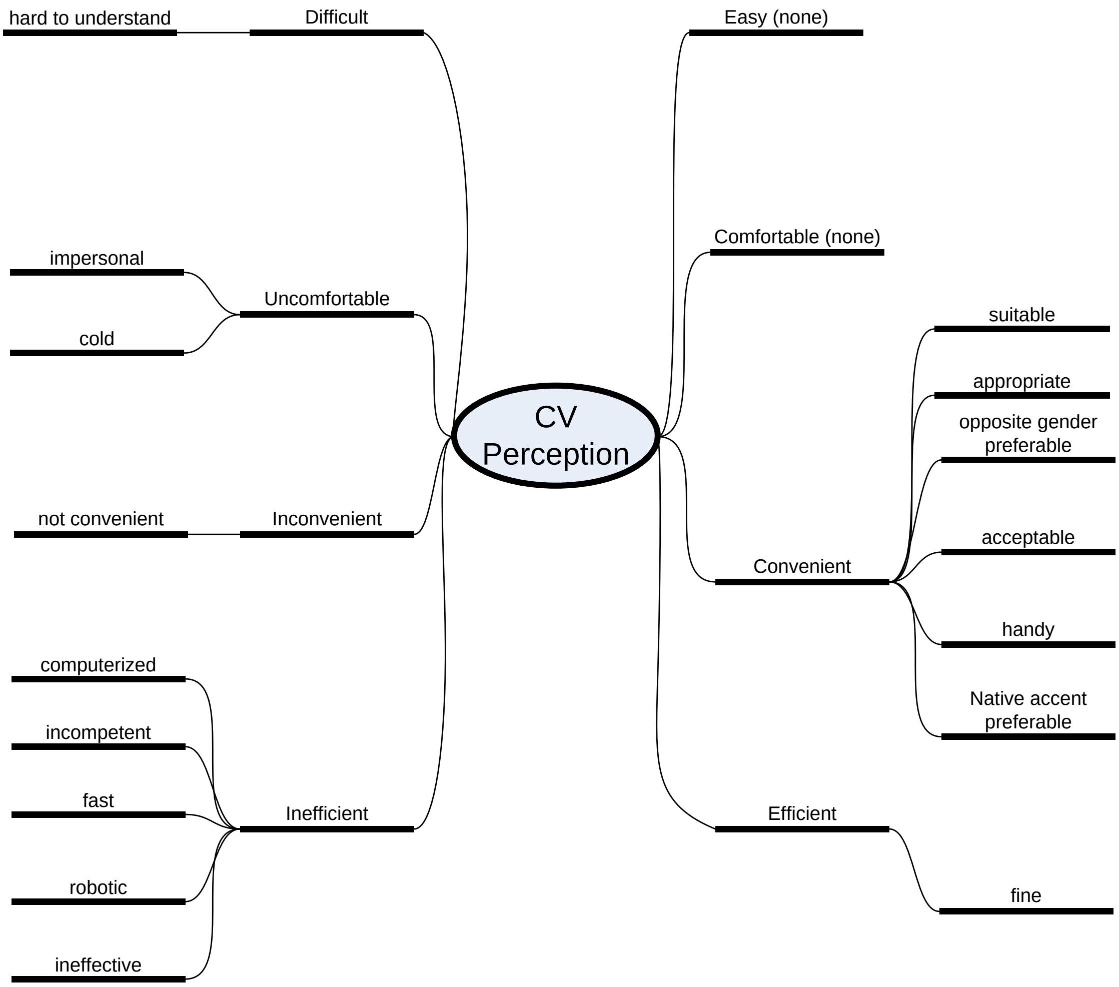


Figure 4.6 - Emergent concepts of 'CV perception' derived from key point analysis of interview data in 3 couples (source: author)

CV interface, regardless of the advancement of TTS technology, the voice is hard to understand due to the synthetic tone, speed and pronunciation. CV is preferable to be used with an opposite gender and a clear accent. Overall, it is understandable, but, isn't easy. CV perceives as uncomfortable because it is impersonal. On the other hand, obviously the familiar voice is personal. CV would suffice because it is something that most people are used to. CV is recommended for impersonal devices, which makes it inconvenient for personal applications.

However, it is the most appropriate, suitable, acceptable and handy for synthesizing dynamic content (e.g., GPS navigation system). But, CV is inefficient because the computer-generated voice technology has still not developed enough to match a natural voice which sounds robotic. Overall, listening to synthetic speech is unpleasant as it is unnatural. With regards to efficiency, it is fine, but considered computerized, incompetent, fast, robotic, and ineffective. With regards the perception of synthetic speech interfaces section in Chapter 2 and the present study, CV interface considered difficult to understand, furthermore, it is impersonal and cold, which makes it difficult and uncomfortable voice to listen to.

4.5 Discussion

Participants had no problems understanding the native speakers and their familiar speakers, but reported they had a hard time in understanding the CV samples. Accordingly, participants felt more comfortable with the NV and their FV. As a result, participants found the quality of recording a talented person was more convenient than recording their FV. In both Gong and Lai's (2001) study and the present study, analyses were performed according to the CV and NV, and showed that the more negative perception in the CV interface may contribute to a worse task performance by the user, although users interacted with the FV interface had less efficiency and convenience than the NV interface. Thus, the explanation that people performed better at the NV interface because their familiar voices weren't recorded professionally, also their familiar people most probably were not as effective and convenient as the native speakers who used their voices for the NV recording. Participants thought the native speakers have the highest efficiency among the other conditions. However, with the FV interface, users need to deal with their familiar voice and thus may be more able to stay focused on the task. This is supported from the notes that taken by the researcher during the listening test. Most of the participants in the familiar condition looked very focused and absorbed most of the time during their interaction with the VoiceTester application. However, users interacting with the FV interface also were more comfortable than the other conditions. Furthermore, participants in the FV condition who had been sitting back in their chairs when listening to their familiar people voices as opposed to the CV condition were participants would often bend forward towards the mobile phone when the CV started to play.

Fascinatingly, users who interacted with either FV or NV interface thought they performed the task better and thought the VoiceTester application was easier to use than interacting with the CV. In line with this fact, the results of self-perception of task performance (see Appendix H – The Evaluative Data, the analyses of the evaluative data) lent some support. On the other hand, provisional evidence suggests that this may have been caused by the strong presence of the pleasant voice of the voice talent that guided the interaction with the user. Such a pleasing natural voice of a native speaker probably makes the users feel better overall. The FV preference explanation is partially supported by the link between liking and relationship of their familiar people and their perceived ease of using the system. Moreover, the contrast with the almost perfect voice of the voice talent probably made the unprofessional recorded FV appear to be not as good as the NV, and even made the CV sound worse.

This important user insight seems to further suggest that the FV may mainly make users feel more comfortable rather than help them in better understand the content of the prompts to carry out the tasks. Their partner voices considered as their most preferred familiar voices, as well almost all of them preferred the familiar voice to be with their opposite gender. Although the explanations proposed here are suggestive rather than conclusive, the findings of the study demonstrate the importance of examining interfaces from other perspectives. Gong and Lai (2001) also suggest that the quality and pleasantness of the voice in speech interface are important and affect users' perception and attitude.

The above-mentioned diagrams show the emergent model of speech interfaces that can be summarised as follows.

Firstly, perception of different voices is a set of feelings to assist the effective operations of speech interfaces acceptability. The usefulness of speech interfaces will be impaired if users' perceptions are not recognised and supported among personal devices/applications. The FV perception is the best voice among the others bringing a sense of trust and relaxation as well as feeling warmer and more personal. Regarding other studies – the celebrity voice is far more enjoyable to run whilst listening to music (Byrne, 2007), as well as, listening to familiar music can have a positive effect on enjoyment of and motivation for performing physical exercise (Wijnalda et al, 2005). Regarding the present study, stated by the participants, the FV is easier to understand and feeling of more comfortable apart from participant four, as she stated 'if I know the person then it's weird!' while participant three (i.e., her familiar person) stated 'makes a big difference hearing someone you know'. The reason for this is that both of them were forced to choose each others' voice for availability issues, whereas participant two stated on selecting her husband she had 'feelings of trustworthy and comfortableness'. The natural local voice perception is real, nice and the local accent is easier to understand over other accents being professionally recorded. The CV perception is not real, is impersonal but still understandable with a high level of naturalness.

Secondly, the characteristics of different voices are a set of rules to assist the effective operations of speech interfaces acceptability. The usefulness of speech interfaces will be impaired if the speech characteristics' preferences for personal devices/applications are not recognised. With personal applications it's better to have a familiar voice, as it is real with better quality than a CV and also the person listening to a FV would feel more comfortable than hearing a NV. The FV preferred to be used with personal electronic devices; the NV is preferred to the impersonal electronic devices as GPS; while the CV is preferable with electrical devices such as machines; Family members and own voice are preferable for answer machines of the home phone and voice mail of the personal mobile phone respectively; the most use for personal speech interface devices/applications are GPS navigation system, voice mail of the mobile phone, and answer machine.

Thirdly, complicated voice characteristics will assist in identifying its impact on the acceptability of speech interfaces of personal devices. These voice characteristics are selecting a proper accent and gender, and selecting a proper familiar person in case of using FV; the FV has to be recorded with care as the familiar person has to have a nice tone and has to be tested with the targeted end user first. However, still selecting the right gender is considered one of the most important characteristics. Although the findings has differed across studies – in the present study, most of the participants preferred an opposite gender in case of speech interfaces, whereas in other study, Lee et al. (2001) found that their participants prefered a similar gender. The accent is one of the tested characteristics to incorporate in the speech interfaces acceptability; the familiar person accent considered as the easiest and most preferred accent followed by the natural local accent, whereas the unfamiliar accent was perceived as hard to understand the human and non human voices.

Finally, currently available speech samples can be difficult to use in the right place. It is therefore important to record the FV with care and pilot test it with the end user to make sure it is nice and easy to understand; the own voice has some difficulties as some people don't like hearing their own voice; the NV has only positive perceptions with no drawbacks; the CV needs to be intelligent and slow with worldly accent in addition to the high naturalness. The results of Gong and Lai's (2001) study lent som support to the last two statements.

4.6 Conclusions

This chapter includes presentation and discussion of the analysis of the most important results and findings of this research. This research focuses on various aspects relating to exploring the impact of different types of speech interfaces of personal devices on users' perception that have been identified from the literature. These various scopes contain perception of importance in developing speech interfaces and identifying speech interfaces barriers. These aspects are explored in more details based on self-perception of task performance of the six participants. Furthermore, comparisons have been made of the findings to current literature during this research. Lastly, the chapter concludes with a discussion of several interesting points related to the impact of different speech interfaces on user's perception, with some probable frameworks recommended for future research.

The following chapter presents the conclusions of this research. It assigns the main findings to the research objectives and research question, in addition to reflecting on the contribution to knowledge. Towards the end of this concluding chapter, there are several discussions about the limitations of the research, possible future work in this area and the implications for research.

CHAPTER V
CONCLUSIONS

5.1 Introduction

This chapter presents the conclusions of this study, the main findings of this investigation which were structured according to the framed research question, and presents the contributions of this study to the areas of the impact of different types of speech interface on users' perception. The major limitations of the study are also outlined and followed with some recommendations for possible future research. And finally, identifies potential implications to the researchers, developers, and public users of mobile application.

In the following sections, the main findings and conclusions of this study are summarized and related to the main research goal, aims and research question. Section 5.2 demonstrates a brief review about the research objective, design and scope. Section 5.35.4 describes a review of the main findings of this study and relates them to the research question. Section 5.4 presents the contribution to the field of this study. Section 5.5 states the limitations of this study and summarizes some mistakes in the research design had been taken. Section 5.6 explores some possible future research that relates to this topic. Finally, Section 5.7 discusses the implications of the findings of this research for practice, as well as, sharing lessons learnt.

5.2 Research Design

The main motivation for conducting this research of speech interfaces of personal devices is based on wide literature review and is outlined as:

- To help the researchers and developers understand the benefits of putting extra effort into the overall quality of speech interfaces of personal applications. Through better understanding the users' perception towards different types of speech.
- The clear importance of speech interfaces development of personal devices as a significant factor in improving the acceptability of software development.

- The lack of empirical research on perception of natural and familiar speech interfaces in the current literature, especially in the context of development speech interfaces of personal devices.

These provide a strong rationale to explore the users' perception of different speech types in order to obtain deeper understanding of this phenomenon. Based on these motivations, the objectives of this research are described as:

- To investigate participants' perception of importance regarding the examined types of speech, and also review their self-perception of task performance that allocated to related activities.
- To identify tools and techniques used that potentially support exploring the perception of speech interfaces, as well as determine their acceptability.
- To identify the voice characteristics used (e.g., gender) to explore its impact towards the different types of speech needing to be shared.
- To identify barriers that would hinder users from accepting the speech interface, as well as the methods for overcoming these obstacles.
- To compare what is being reported in literature and actual practice, in order to identify gaps or provide supporting empirical evidence.

In order to meet these research objectives, a question is formed to investigate the issues embodied in the research objectives. The main research objectives and the research question are summarized below in Table 5.1 along with the relationship between them.

Table 5.1 - Link between the research objectives and research question

Research objectives	Research question
To investigate participants' perception regarding the types of speech, as well as their performance allocated to related activities.	What are the users' perceptions of different types of speech of speech interfaces?
To identify tools and techniques used to potentially support the exploration of exploring the users' perception of speech interfaces, as well as to determine their efficacy.	

The selection of the three couples who participated in this study was limited to healthy New Zealanders aged over 20 years. Although it is acknowledged that the TBI patients' perspectives should provide further insights, this study only considers healthy group's perspective. This is because it is more likely that the healthy group will put more effort in taking the listening tests because of the motivation of satisfied perspectives.

Due to the interest of this study in the perspectives of its participants, an interpretive approach is chosen and an inductive research method is selected as the research method. Semi-structured interview was chosen as the data-collection method, the interviews transcribed and qualitatively analysed using key point coding and categorizing techniques to identify emerging patterns along several significant dimensions as identified from literature. The research design has suited this exploratory research and has provided some insights into the users' perception of speech interfaces of personal devices contributing to the research findings. These findings are the subject of the next section.

5.3 Research Findings

Three main conclusions can be drawn from the findings of this research, structured according to the framed research question.

The reason why users perceive the FV more positively than others can be explained by the principle of familiarity, which in the context of the present study was observed in the following aspects. The context can also apply to a voice that its listeners term as familiar or not. People are comfortable with familiar voices. People are normally interested in a familiar voice because they tend to identify with the voice even in a crowd. Therefore, context influences perception. However, in both Sánchez and Aguayo's (2007) study and the present study, analyses were found that the end users prefer to listen to natural voice instead of synthetic voice, even though the quality of the synthetic voices is good and accepted. Moreover, Mullennix and Stern's (2010a) resuts and the present findings indicate that people view CSS less favourably than natural human speech

What are the users' perceptions of different types of speech of speech interfaces?

In terms of ease of understand the speech samples, the participants in the CV condition found that the CV was hard to understand by comparison to the NV and FV conditions. Participants had no problems understanding the native speakers and their familiar speakers, but reported they had difficulty understanding the computerized prompts.

In terms of the level of feeling comfortable listening to the speech samples, the participants in the CV condition found that this voice had less listening comfortableness than the NV and FV conditions. Accordingly, participants felt more comfortable with their natural and some of them with their familiar voices as well. However, the FV ranked higher.

For the convenience level of the speech samples, the participants in the CV reported that this voice had the least convenience compared to the NV and FV conditions. As a result, participants found the quality of recording a talented person was more convenient than recording their familiar persons' voices.

In terms of efficiency of the speech samples, the participants in the CV condition thought that the CV had the lowest efficiency in intonation of wording in long prompts among the NV condition and the FV condition, as participants thought the native speakers had the highest efficiency among the other conditions.

This section has summarized what has been found out in relation to each investigated research issue. The following section is showing the contribution to knowledge.

5.4 Contribution to Knowledge

Despite the limitations of this study which is addressed in the following Section, this research is one of the first which has explored users' perception towards different speech interfaces (CV, NV, and FV) of personal mobile applications. In addition, it has made a contribution by developing the VoiceTester app of the iPhone that can be used in any speech perception and performance studies, as it simulates an appropriate environment of personal devices.

Another unexplored area this study touched upon was related to the users' perception of the voice characteristics. This study included exploring the users' perception of three different types of speech samples, while, previous studies mainly employed only two types.

5.5 Limitations of the Study

This section talks about summarizing some limitations in the design and implementations of this research.

The first limitation of this study concerns the fact that each interview provides a viewpoint limited to the given representative's perspective. It is possible that other roles would have different points of view that might enhance the richness and validity of the data. That is, all of the data collected during the interviews are the participants' perceptions. It is possible for the data to be interpreted differently from the original thoughts of the researcher, thus threatening the validity of the data.

With larger sampling the findings would be more general, reliable and valid as more participants would represent the models more accurately. Besides, the involved participants were healthy and from New Zealand, therefore, the results can only be valid with respect to this specific group. In addition, the findings concerning these three couples permitted generalization only to another similar setting.

There was no current speech synthesizer with a NZ accent, so a UK accent was used instead. Also, the gender was controlled to a similar gender following evidences from previous studies. However, the similar gender perceived negatively in the case of a CV, where most of the preferred participants were already using an opposite gender within their applications that based speech interfaces (e.g., in-car GPS navigation system). This point might have affected the perception of the NV as the gender was controlled to similar gender as well, where an opposite gender is preferred. Based on these facts, definitely exposing each participant to both genders is necessary to explore deeper perception about gender-based affinities to accent.

Some of the familiar persons weren't available to record their voices so another person was chosen instead. During the interviewing, some participants in the FV condition commented that it would be better to record the FV with better instruments and make sure it's professionally recorded as the natural voice. However, some of them were forced to choose a second priority familiar person instead of their most preferred one, because of availability issues as they were not in the same country or had privacy concerns. In addition, some participants in the CV condition commented that the prompts were preferred to be slower and the participants required some pre-exposure and practice for listening. Based on this, in the case of a CV, participants needed to warm up before the actual test.

As is common in research, this study has raised more questions than it has answered, suggesting some productive areas of further research related to this study for discussion in the next section.

5.6 Recommendation for Future Research

In the process of data collection and analysis, and in interpreting the findings, some contributions were made that could inform future research about areas needing further investigation. The most important of these are outlined below.

The adopted methodological triangulation should lead to greater validity and reliability than a single method approach (Collis & Hussey, 2009). Regarding the research design, survey has been used to collect information and opinions from a number of respondents in order to conduct research, and evaluative questionnaire has been administered as part of a research interview helping in developing a structure for the interview (Mohan et al., 2008). Additionally, since the design of the present research involved qualitative methods of data collection and analysis, it was found to be a very appropriate approach to exploring speech perception. This method allowed a more profound understanding of the problems under investigation. For this reason, it is recommended that future studies should also use similar design. Without the qualitative data, it wouldn't have been possible to find differences between the spontaneous and rationalized responses of the participants.

Reflection on the demographics of this study indicates that this study lacks input from people with cognitive impairments, which is clearly an important role in the development of the GMT speech interface. It is very likely that TBI patients' perspectives would provide further insights to the practice of development of speech interfaces understanding. Further investigation therefore, incorporating wider participation (including the TBI patients) can be considered as one possible future research area.

As mentioned in Chapter 1, this project is a subtask of a future study focusing on exploring the TBI patients' perception to improve the GMT system acceptability. This investigation may be useful in identifying users' perceptions toward different types of speech, thus leading to other relevant insights that could be useful for improving the current practice of speech interfaces development. In this study, all the participants were healthy, and the results could not be generalized to users with TBI or cognitive impairments. A study on speech perception of TBI patients may also open the opportunity to verify users' perception, so that future studies on perception of different speech types may consider comparing these two groups. Regarding the GMT application, Liu et al. (2008) which showed the need to include arrows and outlines, became clear as a result of their pilot testing. However, with small images of indoor landmarks, it can be difficult for a user to know where to focus. Their future work will be focused on user interface improvement and uses wider range of study in order to involve indoor and outdoor way finding. This system has the capability to show the routes to the lost objects, indicating landmarks to the nearest market and to the TBI patients. This study strongly recommends a merger with the current GMT system, as the system will be able to navigate and assist the TBI patients to avoid making errors inside and outside.

Results from the literature review identify several aspects that are relevant to the development of speech interfaces. The relevance of these aspects has been confirmed in this study, thus providing a framework for similar future research in this area. However, the level of influence brought by each of these aspects to the development of personal applications based speech interfaces is still not clear. Investigation on this matter would provide practitioners with insights in determining the effort that has to be spent by the researchers and the developers in order to get an accessible speech interface, particularly in the practice of speech interface development.

There are features required in order to make the VoiceTester application worth releasing to the Apple store to deliver helpful tools for researchers and future research on voices evaluation and perception. That is the ability to customize the UI elements. And the ability to record/generate voices and store them directly on the application internal database for ease of voices integration without the need for external software/device.

In view of the audio stimuli providers, the study involved two native English speakers, but these were controlled by participant's similar gender. For this reason, both genders need to be examined on each participant regardless of the participant gender, to study gender stereotype of the natural voice. Since, NZ is a multi-culture country so different accents for the natural voice have to be examined. However, it might be interesting for future studies to include speech samples of celebrity voices, as it would have a common familiarity for different cultures. Moreover, the familiar voice is dependent on the relationship status, it is therefore recommended for future studies to examine the effect of other demographic variables, such as ethnicity, relation status, etc.

This study made reference to speech perception when interpreting the findings. Since previous research had not taken speech perception of familiar voices into consideration, it was not possible to compare and reference conclusions made in this aspect.

5.7 Implications for Research

From the previously-mentioned suggestions for future research, this section demonstrates some practical implications of the findings of this research. Four main recommendations that could be useful to the acceptability of speech interfaces were identified.

First, it is recommended that researchers offer a number of accents which can be selected for both genders of each examined type of speech to familiarize users with the linguistic and cultural diversity in the world. This will encourage optimistic and unbiased perception to the listeners in order not to introduce psychological barriers in the listeners or users.

Second, developers and researchers should be aware that a computerized voice could be negatively perceived if it impedes the users' ability to provide clear instruction and understandable input. In this case, they should work towards overcoming pronunciation inaccuracies, negatively affecting users' ability to understand.

Third, the researchers should put extra effort into the overall quality of speech interfaces of personal applications, and should also pilot test the output by the end users or listeners to encourage positive and open-minded perceptions. In the case of dynamic content, alternative to the CV, the researchers should put extra effort into the personalized familiar voices, although, in the case of static content, they should consider putting extra care into the overall quality of recording familiar/natural voices.

Fourth, recommending a smarter way of finding two familiar people for each participant, in the case of a future study on the familiar voice or characteristic, by recruiting a group of four friends like students (two males and two females), with each one of them able to select one familiar male voice and one familiar female voice from the group. In this case, the linguistic features of age, relationship, familiarity level, etc will be controlled. And also, the study will focus intensely on perception of familiar voices regarding familiar gender and accent.

Although, the recommendations proposed here are suggestive rather than conclusive, the findings of the study demonstrate the importance of examining speech interfaces from other perspectives. This study also suggests that the quality and pleasantness of speech interfaces are important, affecting users' perception and voices performance. However, the result of this research has provided information on self-perception of task performance for each type of speech. This information is useful when planning similar future research on speech interfaces.

On the other hand, researchers and medical practitioners in the field need to be always up to date with the latest CSS technology in the market. This allows them to get hold of the latest and best technology to assist them in bettering the lives of their patients. Moreover, this allows them to formulate informed decisions about the best possible direction to take and the best research avenues to pursue.

REFERENCES

Apple Inc. (2011). *iOS human interface guidlines: User experience.* Retrieved from http://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/MobileHIG.pdf

Atkinson, R. D. (2008). Why is the digital information revolution so powerful? In R. D. Atkinson & D. D. Castro (Eds.), *Digital quality of life: Understanding the benefits of the information technology revolution* (pp. 1-6): The Information Technology and Innovation Foundation.

Bayard, D., & Green, J. A. (2005). Evaluating English accents worldwide. *Te Reo, 48*, 21-28.

Beskow, J., Granstrom, B., & House, D. (2002). A Multi-Modal speech synthesis tool applied to Audio-Visual prosody. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in speech synthesis: COST 258: The naturalness of synthetic speech* (Vol. Part V Future Challenges, pp. 372-382). New York, NY: J. Wiley.

Bunnell, H. T., & Pennington, C. A. (2010). Advances in computer speech synthesis and implications for assistive technology. In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment* (pp. 71-91). University of Pittsburgh at Johnstown, USA: IGI Global. doi:10.4018/978-1-61520-725-1.ch005

Byrne, R. (2007). *Adapting running pace with music.* Retrieved from http://sucs.org/

Carter, P. (2007). Liberating usability testing. *Interactions, 14*(2), 18-22. doi:10.1145/1229863.1229864

Cepstral Inc. (2011). Cepstral Text-to-Speech (Version 4.0) [Computer software]: Cepstral Corp. Retrieved from http://www.cepstral.com

Collis, J., & Hussey, R. (2009). *Business research: A practical guide for undergraduate & postgraduate students* (3rd ed.). Basingstoke, UK; New York, NY: Palgrave Macmillan.

Definition of perception. (2011). In *Oxford Dictionaries Online*: Oxford University Press. Retrieved from http://oxforddictionaries.com/definition/perception

Duggan, B., & Deegan, M. (2003). *Considerations in the usage of text to speech (TTS) in the creation of natural sounding voice enabled web systems*. Paper presented at the meeting of the Proceedings of the 1st international symposium on Information and communication technologies, Dublin, Ireland. Retrieved from http://portal.acm.org.ezproxy.aut.ac.nz/citation.cfm?id=963600.963686

Dutoit, T., & Stylianou, Y. (2003). Text-to-Speech synthesis. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (Vol. 3, pp. 323-338). Oxford, New York, NY: Oxford University Press.

Flach, G. (2002). Interface design for speech synthesis systems. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in speech synthesis: COST 258: the naturalness of synthetic speech* (Vol. Part V Future Challenges, pp. 383-390). New York, NY: J. Wiley.

Gong, L., & Lai, J. (2001). *Shall we mix synthetic speech and human speech? Impact on users' performance, perception, and attitude*. Paper presented at the meeting of the Proceedings of the SIGCHI conference on human factors in computing systems, Seattle, Washington, United States. doi:10.1145/365024.365090

Keller, E. (2002). Towards greater naturalness: Future directions of research in speech synthesis. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in speech synthesis: COST 258: the naturalness of synthetic speech* (Vol. Part V Future Challenges, pp. 383-390). New York, NY: J. Wiley.

Koul, R., & Dembowski, J. (2010). Synthetic speech perception in individuals with intellectual and communicative disabilities. In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment* (pp. 177-187). University of Pittsburgh at Johnstown, USA: IGI Global. doi:10.4018/978-1-61520-725-1.ch011

Lee, E. J., Nass, C., & Brave, S. (2000). *Can computer-generated speech have gender? An experimental test of gender stereotype*. Paper presented at the meeting of the CHI '00 extended abstracts on human factors in computing systems, The Hague, The Netherlands. doi:10.1145/633292.633461

Liu, A. L., Hile, H., Kautz, H., Borriello, G., Brown, P. A., Harniss, M., & Johnson, K. (2008). Indoor wayfinding: Developing a functional interface for individuals with cognitive impairments. *Disability and Rehabilitation: Assistive Technology, 3*(1), 69 - 81. doi:10.1080/17483100701500173

Lumsden, J. (2008). *Handbook of research on user interface design and evaluation for mobile technology*. doi:10.4018/978-1-59904-871-0

Mahmud, A. A., Mubin, O., & Shahid, S. (2009). *User experience with in-car GPS navigation systems: Comparing the young and elderly drivers*. Paper presented at the meeting of the Proceedings of the 11th international conference on Human-Computer Interaction with mobile devices and services, Bonn, Germany. doi:10.1145/1613858.1613962

McPherson, K. M., Kayes, N., & Weatherall, M. (2009). A pilot study of self-regulation informed goal setting in people with traumatic brain injury. *Clinical Rehabilitation, 23*(4), 296-309. doi:10.1177/0269215509102980

Mohan, T., McGregor, H., Saunders, S., & Archee, R. (2008). Interviewing and negotiating. In *Communicating as professionals* (2nd ed., pp. 217-229). Melbourne, Australia: Cengage Learning.

Moser, L. E., & Melliar-Smith, P. M. (2008). Speech-based UI design for the automobile. In J. Lumsden (Ed.), *Handbook of research on user interface design and evaluation for mobile technology* (pp. 446-460). Aston University, UK: IGI Global. doi:10.4018/978-1-59904-871-0.ch027

Mullennix, J. W., & Stern, S. E. (2010a). Attitudes toward computer synthesized speech. In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment* (pp. 205-218). University of Pittsburgh at Johnstown, USA: IGI Global. doi:10.4018/978-1-61520-725-1.ch013

Mullennix, J. W., & Stern, S. E. (2010b). Overview: Important issues for researchers and practitioners using computer synthesized speech as an assistive aid. In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment* (pp. 1-8). University of Pittsburgh at Johnstown, USA: IGI Global. doi:10.4018/978-1-61520-725-1.ch001

Mullennix, J., & Stern, S. (2010). *Computer synthesized speech technologies: Tools for aiding impairment*. doi:10.4018/978-1-61520-725-1

Myers, M. D., & Avison, D. (2002). Qualitative Research in Information Systems: A reader. In *Introducing Qualitative Methods series* (pp. 312): SAGE. Retrieved from http://www.uk.sagepub.com/booksProdDesc.nav?prodId=Book205159

Pandzic, I. S. (2002). *Facial animation framework for the web and mobile platforms*. Paper presented at the meeting of the Proceedings of the 7th international conference on 3D Web technology, Tempe, Arizona, USA. doi:10.1145/504502.504507

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Los Angeles: Sage.

Patton, M. Q. (2005). Qualitative Research. In *Encyclopedia of Statistics in Behavioral Science*: John Wiley & Sons, Ltd. doi:10.1002/0470013192.bsa514

Sánchez, J., & Aguayo, F. (2007). *Mobile messenger for the blind*. Paper presented at the meeting of the Proceedings of the 9th conference on User Interfaces for all, Konigswinter, Germany.

Sawusch, J. R. (2005). Acoustic analysis and synthesis of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (Vol. 1, pp. 708): Blackwell Publishing Ltd.

Saz, O., & Lleida, E. (2010). The use of synthetic speech in language learning tools: Review and a case study. In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment* (pp. 188-204). University of Pittsburgh at Johnstown, USA: IGI Global. doi:10.4018/978-1-61520-725-1.ch012

Schmidt-Nielsen, B., Harsham, B., Raj, B., & Forlines, C. (2008). Speech-based UI design for the automobile. In J. Lumsden (Ed.), *Handbook of research on user interface design and evaluation for mobile technology* (pp. 237-252). Aston University, UK: IGI Global. doi:10.4018/978-1-59904-871-0.ch015

Schroeter, J., Conkie, A., Syrdal, A., Beutnagel, M., Jilka, M., Strom, V., … Kapilow, D. (2002). A perspective on the next challenges for TTS research. *Proceedings of IEEE Workshop on Speech Synthesis,* 211-214. doi:10.1109/WSS.2002.1224411

Ståhl, O., Gambäck, B., Hansen, P., Turunen, M., & Hakulinen, J. (2010). A mobile fitness companion. *Swedish Institute of Computer Science*. doi:10.1.1.156.1148

Sutherland, D., Sigafoos, J., Schlosser, R. W., & Lancioni, G. E. (2010). Are speech-generating devices viable AAC options for adults with intellectual disabilities? In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment* (pp. 161-176). University of Pittsburgh at Johnstown, USA: IGI Global. doi:10.4018/978-1-61520-725-1.ch010

Symonds, J., Parry, D., & Briggs, J. (2007). An RFID-based system for assisted living: Challenges and solutions. *Studies in Health Technology and Informatics, 127*, 127-138.

Tatham, M., & Morton, K. (2005). *Developments in speech synthesis*. Chichester, West Sussex, England: J. Wiley.

Uchanski, R. M. (2005). Clear speech. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (Vol. 2, pp. 708): Blackwell Publishing Ltd.

Vaissière, J. (2005). Perception of intonation. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (Vol. 2, pp. 708): Blackwell Publishing Ltd.

Wester, M., Dines, J., Gibson, M., Liang, H., Wu, Y., Saheer, L., et al. (2010). Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. *Proceedings of 7th ISCA Speech Synthesis Workshop.* Retrieved from http://hdl.handle.net/1842/3680

Wijnalda, G., Pauws, S., Vignoli, F., & Stuckenschmidt, H. (2005). A personalized music system for motivation in sport performance. *Pervasive Computing*, IEEE, 4(3), 26-32. doi:10.1109/MPRV.2005.47

GLOSSARY

The following definitions refer to the terms used in this thesis. They may differ from definitions used in other publications.

Assistive technology (AT) A general term for assistive, rehabilitative, and adaptive devices for people with disabilities.

Computer-Aided Speech and Language Therapy (CASLT) The use of synthetic speech in language learning tools

Computer-generated voice (CV) A pre-generated voice of synthesized speech by computer software where typed text was converted into spoken voice output

Computer-Synthesized Speech (CSS) or Text-to-Speech (TTS) The technique of speech synthesis where the entered text is converted to speech

Familiar voice (FV) A pre-recorded speech sample of a familiar person, that can be selected preferably by the participants

Mean Opinion Score (MOS) The average score of responses by subjects for specific rated condition

Natural voice (NV) A pre-recorded speech sample of a native speaker

Screen reader The ability of a TTS interface to speak items on the screen by using the TTS synthesis technique

Speech Generating Devices (SGDs) Any type of device that is uses speech synthesizer mechanism to synthesize speech from entered text.

Speech perception The process of hearing, interpreting and understanding sounds of language ("Definition of perception," 2011)

Speech recognition The capability of a device to recognize the voice, which is exclusive such as a fingerprint of an individual

Speech-to-Action The ability of a machine to convert a speech within the voice into action

Speech-to-Speech translation (S2ST) A system involves both speech interface techniques of recognition of a speech by a machine and converts it to a synthesised speech in another language

Speech-to-Text The ability of a machine to recognize words in a given voice and convert it to a text

Visual Text-to-Speech A combination of facial animation with computer-synthesized

(VTTS) speech. It involves lip reading and facial animations into the production of a speech.

Voice Output Communication Aids (VOCAs) Any type of assisted living devices that its functionality based on speech output

Voice Quality Measurement (VQM) The subjects' opinion after perceiving speech signals processed through various distortion conditions

VoiceTester An application has been developed on the iPhone in this study, to facilitate the listening task by simulating environment of speech interfaces on personal devices

Xcode A developer toolset for Mac, iPhone, iPod touch, and iPad to include the interface builder, iOS simulator, and all required tools and frameworks

APPENDICES

Appendix A – Demographic Questionnaire

Demographic Questionnaire

1. Fill your contact details, if you would like to participate in the study:

   Name: ....

   Phone Number: ....

   Age: ....

   Email: ....

   Ethnicity: ....

2. Is English your first language? (please tick one) Yes○ No○

3. Do you have experience with using touch mobile devices? Yes○ No○

4. What is your gender? Male○ Female○

5. Which gender voice do you most prefer when listening to text-to-speech voice generated by a computer? (e.g., the audio heard from an in-car navigation system):

   Male○ Female○ Doesn't matter○

6. Do you use any devices/applications such as GPS navigator system, voice mail, answer machine, etc…? Yes○ No○ (If No, go to Question 6)

   What type of devices and/or applications they are with its current text-to-speech gender? And do you like its text-to-speech voice quality, accent and gender?

   ………………………………………………………………………………………..

   ………………………………………………………………………………………..

   ………………………………………………………………………………………..

7. Can you name at least 3 examples of people you know?

   Please rate by writing your relationship and circling the familiarity of the person in the column.

#	Name of your familiar person	Relationship	Familiarity
1			1 2 3 4 5
2			1 2 3 4 5
3			1 2 3 4 5

*Familiarity (1= I don't know him/her that much, 5 = I know him/her very well)

8. Please choose one person from your list above and describe what do you feel when you listen to this voice within your personal applications, whose voice would you choose and why?

   ………………………………………………………………………………………..

   ………………………………………………………………………………………..

   ………………………………………………………………………………………..

9. Additional comments (please feel free to provide any comments about any of your responses)

Appendix B – Participant Invitation Letter

Participant Information Sheet

AUT
UNIVERSITY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

Date Information Sheet Produced:

15 July 2010

Project Title

Impact of Different Speech Interfaces of Personal Devices on Users' Perception

An Invitation

My name is Mazen. I'm interested in developing mobile applications with a text-to-speech user interface. This study is funded by the Health Research Counsel. You are invited to take part in a survey and then you may wish to participate further in the study. The goal of this project is to determine the impact of text-to-speech voices on the application's acceptability of personal devices. Acceptability indicates the approval level of a technology by the users.

Please remember that your participation in this study is entirely voluntary (your choice). If you do agree to take part you are free to withdraw any time prior to the completion of data collection without giving a reason. This information sheet explains the research study. Please feel free to ask anything you do not understand or if you have a question at anytime.

What is the purpose of this research?

The purpose of this research is to collect data about user preferences for text-to-speech voices of computer-generated voice, natural voice, and familiar voice.

How was I identified and why am I being invited to participate in this research?

People are being invited to take part if they are healthy, not supervisors' students, over 20 years old, and English is their first language.

What will happen in this research?

If you agree to take part, you have to sign a consent form and if you will be selected for the acceptance testing, the researcher will record your familiar voice to integrate it within the application that you are about to use for the listening test. After that, you will be given a task with using the application running on an iPhone. Basically, the application has number of recorded prompts which you need to listen to and follow. The test is about listening to them and you will perform thrice first time with a computer-generated voice, second time with natural voice and then with your familiar voice. Afterwards, you will be interviewed to explore your perception from the listening to the recorded prompts. The interview with the test will take approximately 30-45 min.

What are the discomforts and risks?

This participation has no risk on you. However, you could ask for a break anytime you feel tired and you could withdraw anytime without any consequences to you.

How will these discomforts and risks be alleviated?

The interviewer will interview the participants at a suitable time and location.

What are the benefits?

The information from this study will explore the impact of different types of text-to-speech voices on the acceptability of personal text-to-speech applications. The result will be extremely useful to the text-to-speech interface of the developed software for Traumatic Brain Injury patients, as the recommended software will assist them undertake activities with support to help prevent them from making errors. In general, this information will help the researchers to understand the benefits of putting extra effort into the overall quality of text-to-speech voices in personal applications.

How will my privacy be protected?

All information you give will be kept confidential and your name will not be known to anyone apart from the researchers. We will keep the information locked in a cabinet.

What are the costs of participating in this research?

There will not be any cost to you except your time (approximately 30 to 45 minutes).

What opportunity do I have to consider this invitation?

You will have at least one week in which to consider the invitation.

How do I agree to participate in this research?

If you agree to participate, you have to sign a written consent form provided by me. All your information will be confidential apart to the researcher as the information will be kept in a cabinet, and any reports will make sure that you cannot be identified.

Will I receive feedback on the results of this research?

It’s an optional choice, you could request a summary of the results by indicating it on the consent form; usually, this will take 4 months after the interview.

What do I do if I have concerns about this research?

Any concerns regarding the nature of this project should be notified in the first instance to the Project Supervisor, Judith Symonds, Email: judith.symonds@aut.ac.nz, Phone: (09) 921 9999 x 5879.

Concerns regarding the conduct of the research should be notified to the Executive Secretary, AUTEC, Madeline Banda, *madeline.banda@aut.ac.nz*, 921 9999 ext 8044.

Whom do I contact for further information about this research?

Researcher Contact Details:

Principal Investigator	Mazen Wadea	Phone: (021) 0257-4040

Project Supervisor Contact Details:

Principal Investigator	Judith Symonds	Phone: (09) 921 9999 x 5879

Approved by the Auckland University of Technology Ethics Committee on *6 August 2010*, AUTEC Reference number *10/118*.

Appendix C - Participants Consent

Consent Form

Project title: ***Impact of Different Speech Interfaces of Personal Devices on Users' Perception***

Project Supervisor:	***Dr Judith Symonds***	***Phone (09) 921 9999 x5879***
Researcher:	***Mazen Wadea***	***Phone 021 0257 4040***

○ I have read and understood the information provided about this research project in the Information Sheet dated 15 July 2010.

○ I have had an opportunity to ask questions and to have them answered.

○ I understand that notes will be taken during the interviews and that they will also be audio-taped and transcribed.

○ I understand that I will use a mobile device (iPhone) during the test, if I volunteer and get selected in the acceptance testing.

○ I understand that I may withdraw myself or any information that I have provided for this project at any time prior to completion of data collection, without being disadvantaged in any way.

○ If I withdraw, I understand that all relevant information including tapes and transcripts, or parts thereof, will be destroyed.

○ I agree to take part in this research.

○ I wish to receive a copy of the report from the research (please tick one): Yes○ No○

Participant's signature:.............................………………………………..

Participant's name:.............................………………………………………

Participant's Contact Details (if appropriate):

……………………………………………………………………………………….

……………………………………………………………………………………….

Date:

Approved by the Auckland University of Technology Ethics Committee on 6 August 2010

AUTEC Reference number 10/118

Note: The Participant should retain a copy of this form.

Appendix D – Listening Task

Listening Task

AUT
UNIVERSITY
TE WĀNANGA ARONUI O TAMAKI MAKAU RAU

Your task is about listening to three different types of speech samples (i.e., computerized, natural, and familiar), the test will be up to 30 min, you could ask the researcher any question before the test starts, and after that you will be interviewed for approximately 15 min to express your perception of listening to the three speech samples.

When you are ready launch the 'VoiceTester' application, and then you have to listen and answer set of questions. However, you could ask the researcher to launch the application for you.

Note: the order here is necessarily. After listening to each voice, device gives message to evaluate that voice.

Step 1: Computerized voice

- Starts: listen to the particular computerized voice and answer the questions of that voice
- Ends: rate that voice by filling an evaluation form

Step 2: Natural voice

- Starts: listen to the particular natural voice and answer the questions of that voice
- Ends: rate that voice by filling the same evaluation form

Step 3: Familiar voice

- Starts: listen to your familiar voice and answer the questionnaire of that voice
- Ends: rate your perception of your familiar voice by filling the same evaluation form

Appendix E – Evaluative Questionnaire

Evaluation Form

Use the scales provided below to evaluate the listened speech samples performance.

Note: please rate by circling the number in the column of the listened voice. Whereas 1 = 'describes poor', 5 = 'describes well'. Repeat this process three times to cover all the given voices.

	Computerized voice	**Natural voice**	**Familiar voice**
How well certain adjectives described each listened voice?			
a) difficult / easy	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
b) uncomfortable / comfortable	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
c) inconvenient / convenient	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
d) inefficient / efficient	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5

After finishing the evaluation of the listening task, the researcher will interview you about each of your responses.

Appendix F – Interview Questions

Sample Interview Questions

The following questions are to explore the participants' perception towards the speaker and his/her way of speaking of each of the listened speech samples. The following questions are to explore the participants' self-perception of task performance, experiences, perception and attitudes towards each of the listened speech samples.

1. How well do you think you performed the task? (express of your performance and experience through each of the steps)

   Computer-generated voice:

   Natural voice:

   Familiar voice:

2. What is your perception about each of the following voices? (express your perception towards the speaker and the speaker's way of speaking, as well as, clarity and liking of each of the voices)

   Computer-generated voice:

   Natural voice:

   Familiar voice:

3. Which voice characteristics do you like and don't like with each of the speech samples in point of view of gender and accent?
4. Which speech type do you prefer to use within your personal applications, the computer-generated voice, the natural voice or the familiar voice? Why?
5. Are you unsatisfied with any of the voices that you just listened to? If yes, then Why?
6. Are you able to efficiently complete your task using this application?

   Additional comments

Appendix G – Stimulus Passage for Readers

Auditory Stimulus Content

‘Note for readers: read this short passage and these questions while respecting the punctuation’

Welcome to the listening test of different voices. You will hear set of questions and you have to write your answers while you are listening to the questions. If you don’t feel have enough time then feel free to pause the player. Okay let’s begin.

- What is your gender? Male○ Female○
- Is English your first language? Yes○ No○
- Do you have experience with using touch mobile devices? Yes○ No○
- Which gender voice do you most prefer when listening to text-to-speech voice generated by a computer? Male○ Female○ Doesn’t matter○

Okay, so that brings you to the end of the test. Now, fill the evaluation form.

‘Thank you for your participation, your help was very much appreciate’

Appendix H – The Evaluative Data

This section presents the mean difference in overall self-perception of task performance for each of the six participants. The table below (Table H.1) shows how well certain adjectives described each listened voice of each participant, whereas 1 describes poor and 5 describes well.

Table H.1 – Self-perception of task performance of each participant

- **Participant 1**

	Computerized voice	Natural voice	Familiar voice
difficult / easy	3	5	5
uncomfortable / comfortable	3	4	5
inconvenient / convenient	3	5	5
inefficient / efficient	2	5	5

- **Participant 2**

	Computerized voice	Natural voice	Familiar voice
difficult / easy	4	5	5
uncomfortable / comfortable	3	4	5
inconvenient / convenient	3	5	5
inefficient / efficient	3	5	5

- **Participant 3**

	Computerized voice	Natural voice	Familiar voice
difficult / easy	4	5	5
uncomfortable / comfortable	5	4	5
inconvenient / convenient	4	4	5
inefficient / efficient	4	4	5

- **Participant 4**

	Computerized voice	**Natural voice**	**Familiar voice**
difficult / easy	4	5	5
uncomfortable / comfortable	3	5	3
inconvenient / convenient	3	5	3
inefficient / efficient	3	5	3

- **Participant 5**

	Computerized voice	**Natural voice**	**Familiar voice**
difficult / easy	4	5	5
uncomfortable / comfortable	4	5	4
inconvenient / convenient	3	5	4
inefficient / efficient	3	5	4

- **Participant 6**

	Computerized voice	**Natural voice**	**Familiar voice**
difficult / easy	4	5	5
uncomfortable / comfortable	4	5	3
inconvenient / convenient	4	5	3
inefficient / efficient	4	5	3

Table H.2 – Self-perception of task performance

	Computerized voice	**Natural voice**	**Familiar voice**
How well certain adjectives described each listened voice?			
difficult / easy	3.83	**5**	**5**
uncomfortable / comfortable	3.66	**4.5**	4.16
inconvenient / convenient	3.33	**4.83**	4.16
inefficient / efficient	3.16	**4.83**	4.16

Regarding the analyses of the evaluation data, as the results have shown, the NV had the best effect on task performance. Analysis of the data presented in Figure H.1, indicates that users had poor task performance when they interacted with the FV interface and even poorer when they interacted with the CV than with the NV interface.

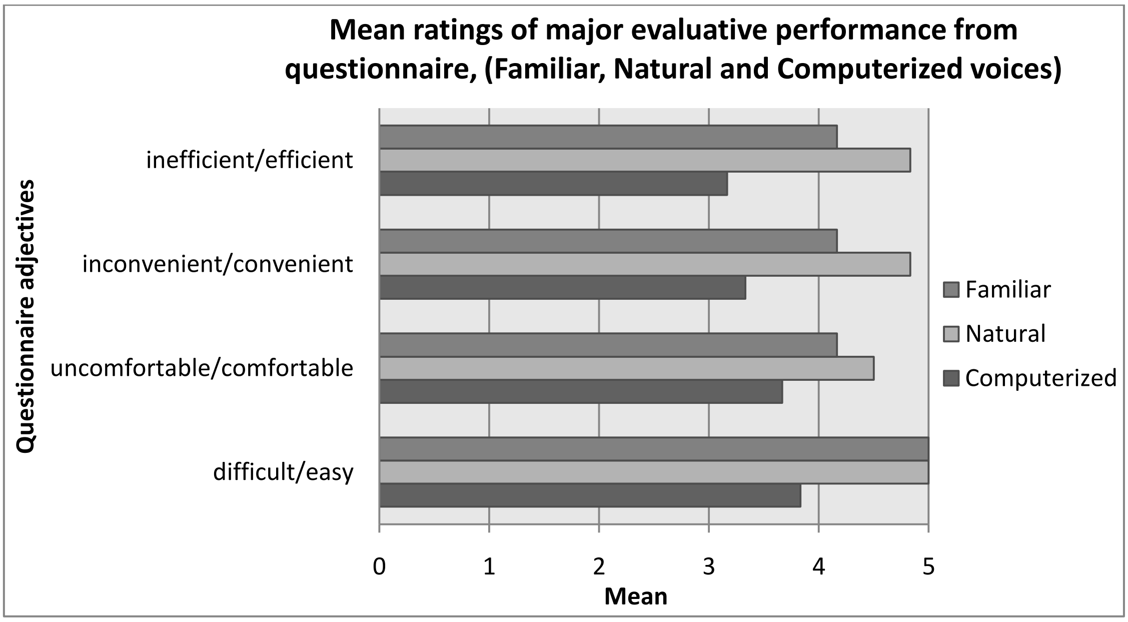


Figure H.1 - Mean difference in overall task performance

Figure H.2 shows the collected demographic data according to the three couples, of the six participants, of the overall task performance for the CV, NV, and FV. Couple X perceived the FV better than the other groups. And they perceived the other voices less than the FV. However, they noted the FV as the best voice, preferred over the other voices. Couple Y perceived the NV better than the other groups. And they perceived the other voices less than the NV. However, they observed the NV as the best voice in preference to the FV. Couple Z perceived the CV better than the other groups. However, they perceived the other voices better than the CV. And, they stated the NV as their preference to the familiar voice.

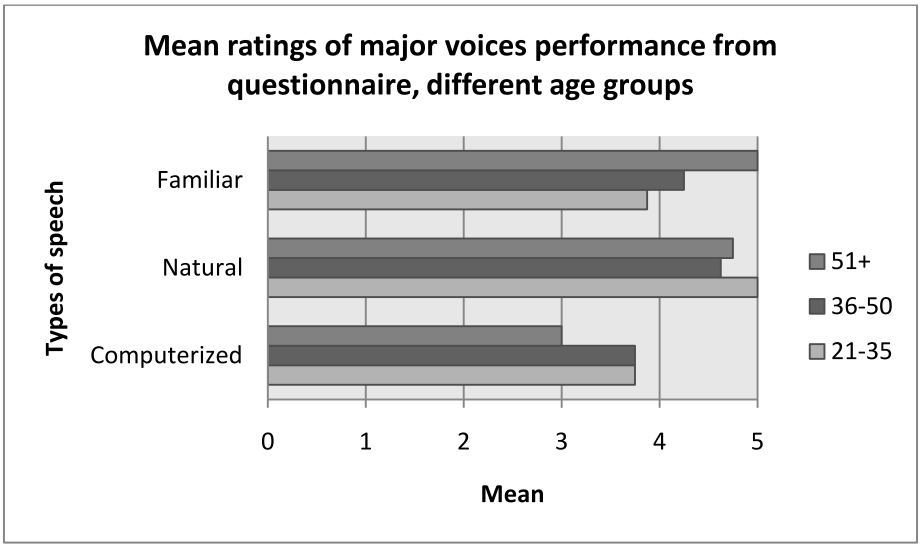


Figure H.2 - Mean difference regarding the three couples in overall task performance

Figure H.3 shows the collected demographic data according to participants' gender of the overall task performance for the CV, NV, and FV. The male participants perceived the familiar voice better than the females. And they perceived the other voices less than the FV. However, they perceived the familiar voice as slightly better than the natural voice. The female participants perceived the NV better than the males. And they perceived the other voices less favourably than the natural voice. However, they perceived the NV as slightly better than the FV.

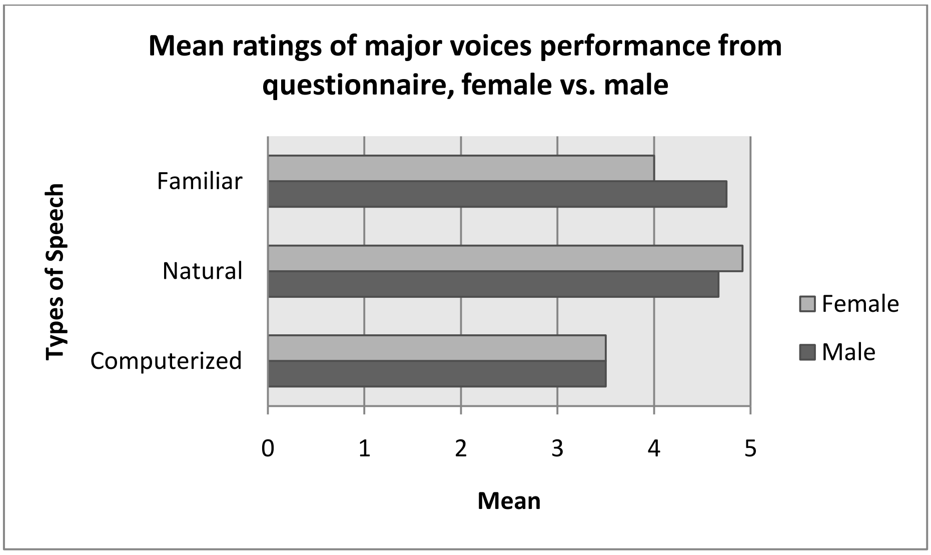


Figure H.3 - Mean difference regarding to participants' gender in overall task performance

The longer average duration in the human voice conditions addressed in Chapter 3 (Subsection 3.4.5.1) is very unlikely to cause the difference in the task performance either because one would expect that a longer duration would give users more time to process and thus lead to better task performance.

This seems to suggest that the CV condition is costly for the user. Processing CV, even though CV is relatively difficult to understand, may also enhance a training effect in that one gets better at understanding CV when he/she hears it more frequently. Unfortunately, because of the design of this study, practising on listening to the voices was avoided to measure the quality and to reduce the bias so the training was only on the use of the VoiceTester application. Moreover, with the CV interface, users also seemed more likely to continue getting the task done. This is supported by the findings that the participants in the CV condition were more willing to repeat part of the synthesis speech and reported to have put more effort in doing the tasks than in the human conditions (NV and FV).

Appendix I – Full Interview Analysis of Couple Y

The second couple was brother and sister aged 37 and 36 respectively. This couple considered as participants three and four. The following tables show the key point analysis of their interview data.

Table I.1 - Key points analysis and codes from the data in Couple Y

Id	Key point	Code
- **Participant 3 (Y_1)**		
$Y_1$1	*FV: makes a big difference hearing someone you know, however, not quite used to this*	Inspirational Weird
$Y_1$2	*NV: quite used to this, was easy though.*	Appropriate Ease of understand
$Y_1$3	*CV: quite familiar, most interactive computer systems have CG voices.*	Suitable Appropriate
$Y_1$4	*I like celebrity voice as my familiar voice.*	Preferable celebrity voice
$Y_1$5	*NV: more fluid. I like the accent.*	Professional Likable accent
$Y_1$6	*CV: can tell one spoken word at a time, less fluid*	Ineffective
$Y_1$7	*CV: female gender is better*	Opposite gender preferable
$Y_1$8	*Quality was good, I didn't mind the natural voice, and I think if it was a different gender it would be nicer.*	Good quality Acceptable Opposite gender preferable
$Y_1$9	*The familiar voice was most understandable and suitable to me.*	Understandable Suitable
$Y_1$10	*For a general use probably a natural voice would simplify the use of its applications,*	Simplify the interaction
$Y_1$11	*If you couldn't produce individualized familiar voices*	Optional

	then natural voice would be recommended.	recommended
$Y_1$12	*I would choose the familiar voice given its availability because provides sense of comfort someone you know,*	favourable Comfy
$Y_1$13	*Otherwise if it is not available I would choose computerize because it is something that I more used to.*	Acceptable Handy
• **Participant 4 (Y_2)**		
$Y_2$1	*FV: because I know the person then it's weird!, and the intonation is not equal*	Weird Unprofessional recording
$Y_2$2	*NV: smooth. Clear well paced speech, accent is good*	Smooth Clear Well paced Good accent
$Y_2$3	*CV: robotic, just like the GPS, I have to concentrate little harder to understand*	Robotic Hard to understand
$Y_2$4	*NV: male gender, this is the clearest voice.*	opposite gender preferable clear
$Y_2$5	*Natural voice, it's a lot easier to listen to and more appropriate and its clear and the accent is very good.*	Ease of understand Appropriate Clear Good accent
$Y_2$6	*I like the natural voice (simplify the use of the application because it's natural)*	Likable Simplify the interaction Natural

Table I.2 - Emergence of concepts from the codes in Couple Y data

Difficult	$Y_2$3
Easy	$Y_1$2, $Y_1$9, $Y_2$5
Uncomfortable	
Comfortable	$Y_1$1, , $Y_1$8, $Y_1$12, , $Y_2$2, $Y_2$6
Inconvenient	$Y_1$1, $Y_2$1
Convenient	$Y_1$2, $Y_1$3, $Y_1$4, $Y_1$5, $Y_1$7, $Y_1$8, $Y_1$9, $Y_1$10, $Y_1$11, $Y_1$11, $Y_1$12, $Y_1$13, $Y_2$2, $Y_2$4, $Y_2$5
Inefficient	$Y_1$6, $Y_2$1, $Y_2$3
Efficient	$Y_1$5, $Y_1$8, $Y_2$2, $Y_2$4, $Y_2$5, $Y_2$6

Appendix J – Full Interview Analysis of Couple Z

The third couple was friends male and female aged 23 and 21 respectively. The couple considered as participants five and six. The following tables show the key point analysis of their interview data.

Table J.1 - Key points analysis and codes from the data in Couple Z

Id	Key point	Code
• **Participant 5 (Z_1)**		
$Z_1$1	*FV: Easy to understand,* *Needs to be recorded and tested with the end user before it's applied to the application.*	Ease of understand Needs pilot test
$Z_1$2	*NV: This is good; using a professional voice would be most preferred. This is my preferred choice.*	Good Professional Preferable
$Z_1$3	*CV: Regardless of the advancement of technology, the voice was hard to understand due to many changes of tone, speed and pronunciation of words.*	Hard to understand Computerized
$Z_1$4	*Home answer machine: my brother, no preference to who records the message, anyone from my family.*	Familial voice
$Z_1$4a	*GPS: CG voice - female*	Opposite gender preferable
$Z_1$5	*The natural voice was my preferred choice, however I would prefer this even more if it were a female voice as this is a nicer tone to enjoy with use of my application.*	Preferable Opposite gender preferable
$Z_1$6	*The Computer-generated technology has still not developed enough to match a natural voice, so at least having a softer female voice can ease the pain of trying to understand*	Incompetent Opposite gender preferred Hard to understand
$Z_1$7	*Had many conversations, very familiar with her speech*	Familiar

	patterns, tone and speed.	
$Z_1$8	*The familiar voice is less professional and if applied, should be recorded with professional equipment to eliminate distortion, etc.* *this accent is preferred*	Unprofessionally recorded Preferable accent
$Z_1$9	*The familiar voice it would be my preferred one, if more time spent on recording*	Unprofessionally recorded
$Z_1$10	*Prefer the natural voice, because its more professional and it was probably recorded in a better environment and better equipment as well*	Professional
$Z_1$11	*The computer generated voice was just shocking Real hard to understand and I'd better of reading.*	Hard to understand Computerized
$Z_1$12	*FV: Chose her because I know her more, so I'm used to the way that she speaks*	Familiar
$Z_1$13	*Generally for a professional voice like what you hear on the news and what you hear on the radio, generally females voices are nicer, they are nicer tone and just easier to listen to.*	Opposite gender preferable
• **Participant 6 (Z_2)**		
$Z_2$1	*FV: has clear New Zealand accent, easy on the ear and clear enough to understand*	Comfy Ease of understand
$Z_2$2	*Preferred this type in case of personal application, and thinks it fits handicapped persons and maybe children and elderly.*	Preferable
$Z_2$3	*NV: performed well with this type, preferred selecting it for my GPS and common use applications instead of the CG.*	Preferable

$Z_{2}4$	*CV: could understand it, however, wasn’t that easy.*	Hard to understand
$Z_{2}5$	*Would choose native accent if available, however, I didn’t mind this one.*	Native accent preferable
$Z_{2}6$	*Home answer machine: My voice, I am the oldest, but wouldn’t mind if anyone of my familiar recorded the message, and definitely my voice for my voice mail.*	Familial voice
$Z_{2}7$	*CV: used a female gender however, and recommended a female in case of English.*	Opposite gender preferable

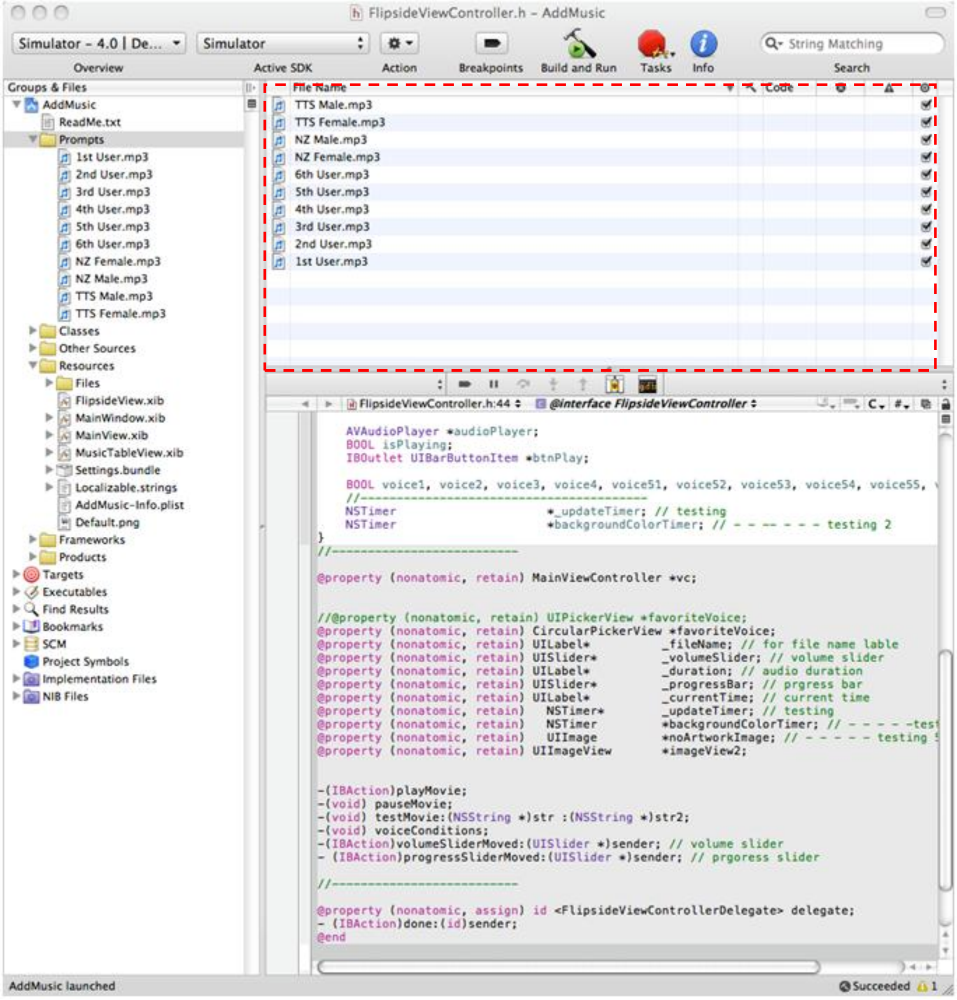
Table J.2 - Emergence of concepts from the codes in Couple Z data

Difficult	$Z_{1}3$, $Z_{1}6$, $Z_{1}11$, $Z_{2}4$
Easy	$Z_{1}1$, $Z_{2}1$
Uncomfortable	
Comfortable	$Z_{1}7$, $Z_{1}12$, $Z_{2}1$
Inconvenient	$Z_{1}1$, $Z_{1}9$, $Z_{1}11$
Convenient	$Z_{1}2$, $Z_{1}4$, $Z_{1}4$a, $Z_{1}5$, $Z_{1}6$, $Z_{1}8$, $Z_{1}13$, $Z_{2}2$, $Z_{2}3$, $Z_{2}5$, $Z_{2}6$, $Z_{2}7$
Inefficient	$Z_{1}3$, $Z_{1}6$, $Z_{1}8$
Efficient	$Z_{1}2$, $Z_{1}10$

Appendix L – The VoiceTester Code Introduction

This section presents the operating environment of the developed application and sample of its coding.

The application can be run on an Apple computer running OS 10.x Snow Leopard platform with software such as Xcode 3.2.5 of SDK 4.2 (Software Development Kit). Xcode is the developer toolset for Mac, iPhone, iPod touch, and iPad to include the Xcode IDE (Integrated Development Environment), iOS Simulator, and all required tools and frameworks. The below screenshot shows the Xcode software and the speech samples integrated into the developed application database.



```
    AVAudioPlayer *audioPlayer;
    BOOL isPlaying;
    IBOutlet UIBarButtonItem *btnPlay;

    BOOL voice1, voice2, voice3, voice4, voice51, voice52, voice53, voice54, voice55, 
    //---------------------------------------
    NSTimer                     *_updateTimer; // testing
    NSTimer                     *backgroundColorTimer; // - - -- - - - testing 2
}
//-------------------------

@property (nonatomic, retain) MainViewController *vc;

//@property (nonatomic, retain) UIPickerView *favoriteVoice;
@property (nonatomic, retain) CircularPickerView *favoriteVoice;
@property (nonatomic, retain) UILabel*          _fileName; // for file name lable
@property (nonatomic, retain) UISlider*         _volumeSlider; // volume slider
@property (nonatomic, retain) UILabel*          _duration; // audio duration
@property (nonatomic, retain) UISlider*         _progressBar; // prgress bar
@property (nonatomic, retain) UILabel*          _currentTime; // current time
@property (nonatomic, retain)   NSTimer*        _updateTimer; // testing
@property (nonatomic, retain)   NSTimer         *backgroundColorTimer; // - - - - -tes
@property (nonatomic, retain)   UIImage         *noArtworkImage; // - - - - - testing 
@property (nonatomic, retain) UIImageView       *imageView2;


-(IBAction)playMovie;
-(void) pauseMovie;
-(void) testMovie:(NSString *)str :(NSString *)str2;
-(void) voiceConditions;
-(IBAction)volumeSliderMoved:(UISlider *)sender; // volume slider
- (IBAction)progressSliderMoved:(UISlider *)sender; // prgoress slider

//-------------------------

@property (nonatomic, assign) id <FlipsideViewControllerDelegate> delegate;
- (IBAction)done:(id)sender;
@end
```

The sample code below is the coding of the ‘FlipsideViewController.h’ class that shows all the headers of the main methods and the initialized variables with a brief description.

```
//  VoiceTester - FlipsideViewController.h
//  Created by Mazen Wadea on 19/07/10.
//  Copyright AUT 2010. All rights reserved.
//
#import <UIKit/UIKit.h>
#import <AVFoundation/AVFoundation.h>
#import <AudioToolbox/AudioToolbox.h> // for vibrating
#import "MainViewController.h"
#import "CircularPickerView.h"

@protocol FlipsideViewControllerDelegate;
@interface FlipsideViewController: UIViewController {
        id <FlipsideViewControllerDelegate> delegate;
#define namesComponent 0
#define voicesComponent 1
#define participantsComponent 2
        IBOutlet CircularPickerView  *favoriteVoice;
        IBOutlet UILabel             *_fileName;            // for file name label
        IBOutlet UISlider            *_volumeSlider;        // volume slider
        IBOutlet UILabel             *_duration;            // audio duration
        IBOutlet UISlider            *_progressBar;         // progress bar
        IBOutlet UILabel             *_currentTime;         // current time
        IBOutlet UIImageView         *imageView2;
        IBOutlet UIBarButtonItem     *btnPlay;

        AVAudioPlayer                        *audioPlayer;
        BOOL                                 isPlaying;
        BOOL voice1, voice2, voice3, voice4, voice51, voice52, voice53, voice54, voice55, voice56;
        NSTimer                              *_updateTimer;
        NSTimer                              *backgroundColorTimer;
}
@property (nonatomic, retain) MainViewController      *vc;
@property (nonatomic, retain) CircularPickerView      *favoriteVoice;
@property (nonatomic, retain) UILabel*                _fileName;          // for file name label
@property (nonatomic, retain) UISlider*               _volumeSlider;     // volume slider
@property (nonatomic, retain) UILabel*                _duration;          // audio duration
@property (nonatomic, retain) UISlider*               _progressBar;      // progress bar
@property (nonatomic, retain) UILabel*                _currentTime;       // current time
@property (nonatomic, retain) NSTimer*                _updateTimer;
@property (nonatomic, retain) NSTimer                 *backgroundColorTimer;
@property (nonatomic, retain) UIImageView             *imageView2;

- (IBAction) playMovie;
- (void) pauseMovie;
- (void) testMovie:(NSString *)str :(NSString *)str2;
- (void) voiceConditions;
- (IBAction) volumeSliderMoved:(UISlider *)sender;     // volume slider
- (IBAction) progressSliderMoved:(UISlider *)sender;    // progress slider

@property (nonatomic, assign) id <FlipsideViewControllerDelegate> delegate;
- (IBAction) done:(id)sender;
@end
@protocol FlipsideViewControllerDelegate;
- (void) flipsideViewControllerDidFinish:(FlipsideViewController *)controller;
@end
```