SUPPORTING NAVIGATION IN AUDITORY INTERFACES USING PERSONALIZATION AND MULTIPLE SYNTHETIC VOICES

by

PEER SHAJAHAN

A Thesis submitted to
the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements for the Degree of
MASTER OF SCIENCE

Department of Computer Science
University of Manitoba
Winnipeg, Manitoba

Abstract

Auditory interfaces commonly use synthetic speech for conveying information. In many instances the information being conveyed is hierarchically structured, such as menus. However, using the auditory channel the structure of the information is not "visible". In particular, the hierarchical nature of auditory interfaces is not revealed explicitly to users. As a result, navigation can be complex and it is considered to be one of the critical issues in auditory interfaces. In an effort to address the issue of navigation in auditory interfaces, two solutions were suggested and investigated in this thesis.

The first solution utilizes a hierarchy re-structuring approach. It aims at reducing the navigation time in auditory interfaces by allowing users to customize/personalize their favorite menus in the interface. The system developed allows users to bookmark any given node in the menu-based system. Bookmarks provide a method for skipping the entire hierarchy structure to access only a node of interest. By means of this approach, users can access personalized information more efficiently and without spending time listening to prompts before making decisions. The results of an evaluation of this technique show that users can quickly access items of interest. However, if the hierarchy is significantly re-structured by the user, performance results degrade.

The second solution described in this thesis focuses on reducing the error rates in auditory interfaces by providing navigation cues to the users, using multiple synthetic voices. In this study, different synthetic voices were created, in a structured manner, by manipulating synthetic parameters for representing the nodes in the hierarchy. This study

consists of two experiments. The first experiment focuses on representing small hierarchies using multiple synthetic voices, whereas the second experiment focuses on representing complex hierarchies using multiple synthetic voices. The experimental results suggest that multiple synthetic voices can be used to represent and provide navigation cues in hierarchies.

Acknowledgements

I would like to thank my supervisor Dr. Pourang Irani for his friendship, motivation, insightful guidance throughout the research. I regard myself as being extremely grateful and thankful to him, because he has been more than just a supervisor.

I would like to thank Jennifer Lai, IBM Watson research labs, for her initial ideas and suggestions in this project. I would also like to thank Fonix Corporation for granting the DECtalk development toolkit for this research. I would also extend my thankfulness to Dean Slonowsky for assisting in performing the statistical analysis.

I would also like to extend my thankfulness to Dr. Christel Kemke and Nivedita Kadaba for proofreading my papers. Finally, I would also like to thank my friends and family for their unconditional encouragement and support.

This research is supported by a NSERC grant.

Declaration:

The experiment described in Chapter 4 has been published in the proceedings of Human Factors in Telecommunication (HFT-03) [SI03]. Substantial parts of the experiment described in Chapter 5.1 have been published in the 8th IEEE international conference on Information Visualisation (IV-04) [SI04]. The experiment described in Chapter 5.2 has been published in the proceedings of the International Conference on Auditory Display (ICAD-05) [SI05b]. The experiments described in Chapter 5 have been submitted for publication in the International Journal of Speech Technology [SI05a].

LIST OF FIGURES

LIST OF TABLES

Chapter 1 : Introduction

1.1 Introduction

In our day-to-day activities, most of the information we process is perceived through visual and auditory means. Over 80% of the information processed is through visual perception. As a result, we typically utilize heavily our visual channels. Auditory perception can be used in situations, in which our peripheral vision is not active. Often, these two senses are combined to perceive and interpret information in our daily life. In human-computer interaction, extending visual and auditory capabilities to harness information from our environment is referred to as multimodal interactions. Some of the examples include, listening to a song while editing a document, and listening to radio while driving a car.

We rely heavily on auditory information, when visual perception is overloaded or cannot be used to perceive information. For example, all the computer systems require our visual sensors to interpret information. Hence, users can miss out some important information, because their visual system may already be overloaded by some other information. As interfaces become more pervasive and visually rich, the human visual system becomes overloaded with other information that users can miss. For example, while playing a computer game, the visual sense is preoccupied by a large amount of moving images on the screen. Hence, the visual system has to be dedicated to the task at hand and cannot be used to gather any other information. Also, there are some situations where visual displays cannot be used to provide information (such as while assisting visually-impaired

users). In these circumstances, the information can be effectively conveyed through auditory interfaces.

1.2 Motivation for Research in Auditory Interfaces

Most auditory interfaces use speech as a medium of input and/or output to provide information to users. In such instances, the speech is produced by a computer (also called computer generated speech or synthetic speech). Speech not only conveys information from the speaker to the listener, but is also considered as the primary means of human communication. In recent years, the use of synthetic speech in auditory interfaces has grown rapidly due to the introduction of Interactive Voice Response (IVR) Systems. Using IVR systems, end-users are able to perform their day-to-day banking transactions, pay their bills, and listen to movie listings at their own convenience.

While auditory interfaces provide information to users, navigating these interfaces is often considered to be one of the most common problems. In such systems, the user queries the application based on a predetermined set of options. Typically, options are presented hierarchically, with the user entering the system through the root node. For example, to pay a bill, a user has to navigate a hierarchy, which branches into sub-hierarchies until the final task is accomplished. Information structured as hierarchies has an implicit ordering, organized into what is commonly referred to as levels. The hierarchy begins with a component at the topmost level, and all other components are related to their preceding entities. The hierarchy is further decomposed into sub-

hierarchies with similar characteristics and has very little correlation between its components at any given level.

Hierarchies are commonly used for structuring and organizing complex information. For example, file structures, disk trees, and many interface elements, such as menus, are organized into hierarchical structures. In addition, many forms of data can be structured into hierarchies. However, to extract information organized as hierarchies, interfaces have to adequately facilitate navigational tasks. For example, Wolf et al. [WKK95] says that one of the most important problems in current telephony applications is "*navigation"*. This is because of the limitations in these devices (i.e. options/menus are presented in a serial mode and are arranged in an invisible hierarchical fashion). As a result, the error rate for obtaining the required information is high. Also, a considerable amount of users' time is spent navigating these hierarchies, particularly if the hierarchy is wide and deep [MS96, Ros85, WKK95]. Hence, in order to improve navigation in auditory interfaces, interfaces should provide navigational cues to the users. In general, the problem can be stated as "how can we improve navigation in auditory interfaces".

1.3 Research Overview

The high-level goal of the research described in this thesis was to determine an effective way of improving navigation in hierarchies (such as file systems and cellular phone menus) and to evaluate users' experience in navigating these hierarchies. The secondary goal was to analyze the observations obtained from the users to design future

interfaces. In this research, the issue of navigation in auditory interfaces had been addressed using two different and novel solutions:

1. Personalizing menus in touch-tone interfaces
2. Using multiple synthetic voices to represent hierarchies.

Personalizing Menus in Touch-Tone Interfaces

This approach has been developed with the intention of improving navigation (i.e. reducing the navigation time) in hierarchies, such as telephone-based interfaces. In this technique, the system will allow users to personalize their favorite menus in the menu-based exemplar. The aims of this part of our investigation are:

- To determine if personalization improves navigation in auditory interfaces.
- To evaluate and compare the effect of personalization with traditional interfaces.

Using Multiple Synthetic voices to represent Hierarchies

This solution has been proposed with the intention of improving navigation in a wide range of auditory interfaces, such as

- Assisting visually disabled users
- Assisting users, whose eyes and hands are busy, such as while driving and playing games

The main motivation behind using multiple synthetic voices to represent hierarchies is attributed to the belief that if multiple synthetic voices can be used to provide sufficient information and navigation cues to the users, then a significant amount of users' navigation time can be reduced. The aims of this part of the thesis are:

- To investigate whether multiple synthetic voices can be used to present hierarchical information.
- To evaluate the effect of multiple synthetic voices in locating items in hierarchies.
- To determine if the rules that are used to create the hierarchies are easy to remember and recall.

1.4 Structure of the Thesis

Figure 1.1 shows the structure of the thesis and how the chapters contribute to address the issue of navigation in auditory interfaces. Chapters 2 and 3 provide the background work, Chapters 4 and 5 describe the suggested solutions, followed by the limitations, conclusions and future work in Chapter 6.

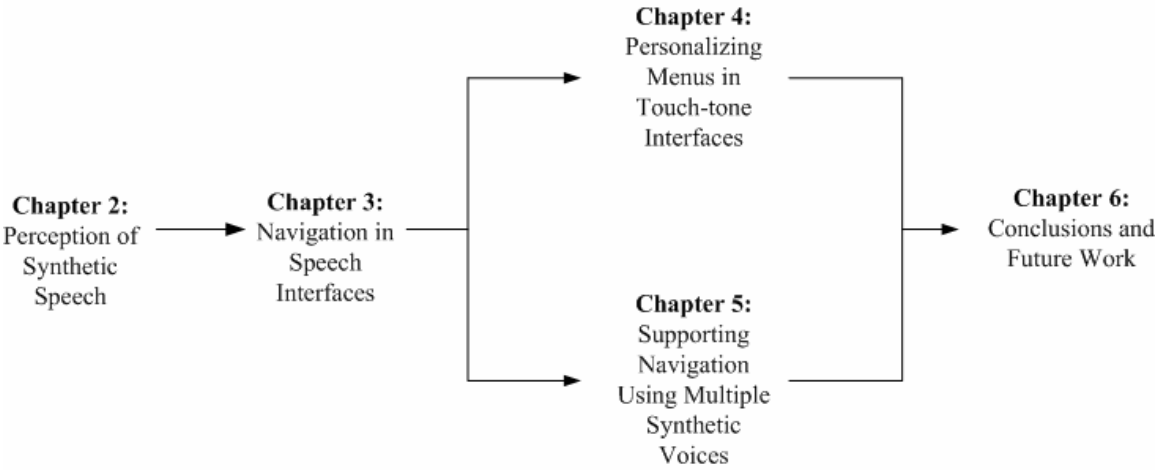


Figure 1.1: Structure of the thesis

Chapter 2 gives an introduction to the perception of synthetic speech. This chapter provides a review of the literature, which explains the effect of synthetic speech in conveying auditory information. The effect of synthetic speech was analyzed on various

perpetual factors (such as comprehensibility of messages, listening to synthetic speech while driving, mixing synthetic speech and human speech, etc.). This chapter then gives detailed information about studies that describe the social variables (such as emotions and personality) that can be included in synthetic speech. From results in this literature guidelines are created for different synthetic voices by manipulating speech parameters, such as pitch, speech rate and volume.

Chapter 3 discusses why navigation is considered as an important issue in auditory interfaces. This chapter provides a background of existing research in the field of improving navigation in auditory interfaces. This chapter also highlights the limitations and the disadvantages with the existing research and concludes with the proposed alternative solutions to improve navigation in auditory interfaces.

Chapter 4 describes the architecture and inner workings of an implemented system, which has been designed to reduce the navigation time in auditory interfaces, by restructuring menus in the interface. Using this system users can bookmark their favorite options and later access these bookmarked options, easily and quickly, by bypassing the various layers of prompts and options. This chapter also evaluates the effect of personalization in auditory interfaces, by comparing the personalized auditory interface with the traditional auditory interface.

Chapter 5 defines a design framework, which aims at improving navigation in auditory interfaces by reducing the error rates. This chapter suggests that error rates for obtaining

the required information can be reduced by providing navigational cues, using multiple synthetic voices, to the users. This chapter describes two experiments, which focus on determining the effect of multiple synthetic voices for conveying hierarchical information. These experiments were conducted to analyze the recall rates of multiple synthetic voices. An initial experiment suggests that multiple synthetic voices can be used to represent small hierarchies. The results of this study show that participants recalled the nodes in the hierarchy, better with multiple synthetic voices than with single synthetic voice. The second experiment described in this chapter was conducted to analyze the effect of multiple synthetic voices on complex hierarchies. This experiment also focused on finding out the effect of training on recall rates. The results of the experiment show that that multiple synthetic voices provide navigation cues to recall components in a complex hierarchy. The results also suggest that the type of training does not have any significant effect on recall rates.

Chapter 6 summarizes the contributions of the thesis, discusses its limitation, followed by future work.

Chapter 2 : Perception of Synthetic Speech

Speech is a gift to humans, since humans are the only species who can understand and process speech [Slo79]. Research suggests that even four day old babies can distinguish and respond differently to their native language [Slo79]. Speech is more than just a carrier of words from the speaker to the listener, i.e. speech also conveys a wide range of information (such as information about gender, emotions, etc.) about the speaker. This information is produced based on the variations in the vocal cues of the speaker. The perception of these vocal cues is measured based on the vibrations of the vocal cord. The vocal cord, located at the larynx, is considered as the primary source for speech production. The speech sounds are produced when there is a vibration in the vocal cord, which normally occurs when the air is exhaled from the lungs. The vocal cord is also responsible for producing louder and higher voices. For example, the increase in the frequency of the vocal cord vibration will amplify the pitch of the voice, whereas the increase in the amplitude will intensify the loudness of the voice.

Speech is composed of several factors, such as speech parameters, language, environment, accent, etc [Slo79]. However, speech parameters are considered to be the most important factor that influences the perception of speech. Although, speech is composed of several speech parameters, fundamental frequency, intensity and speech rate are considered as the most important speech parameters that affect the perception of speech [Slo79].

Fundamental frequency: Fundamental frequency (F_0) is often perceived as the pitch of the speech. Pitch is determined based on the number of vibrations of the vocal cords. Pitch is measured in Hertz (Hz).

Intensity: Intensity is perceived as the loudness of speech. Loudness is considered as one of the major components in identifying the emotions in speech. For example, voice intensity increases for anger and decreases for sadness [CDCS$^+$01].

Speech Rate: Speech rate is defined as the number of words spoken per minute. Speech rate helps to identify the personalities (such as introverts and extroverts) and emotions of the speakers. For example, extroverts tend to speak faster than introverts. Also, speech rate is faster for anger and happiness than for sadness [CDCS$^+$01]. Speech rate is measured in words/minute.

Various studies suggest that a variation in the above mentioned speech parameters will allow listeners to identify the gender, emotion [Bac99, PDR96], social identity [LHGF60], and personalities [BGS67] of the speaker. For example, humans can easily identify the gender (male or female) of the speaker based on the pitch. Ramsay [Ram66, Ram68] suggests that humans easily differentiate various speech characteristics (such as personality and emotion), using the speech parameters, such as speech rate and volume. For example, human personalities can be identified using the speech rate, i.e., extroverts speak faster than the average speaking rate, where as introverts tend to speak slower than the average speaking rate [SBSR75].

Although human speech is easy to understand, it cannot be used to convey dynamic information, such as news, weather forecast, etc. Hence, synthetic speech is introduced to convey dynamic auditory information. The next section described in this chapter introduces synthetic speech and explains how synthetic speech is evaluated, followed by the various lines of research on synthetic speech. Figure 2.1 describes the structure of the chapter.

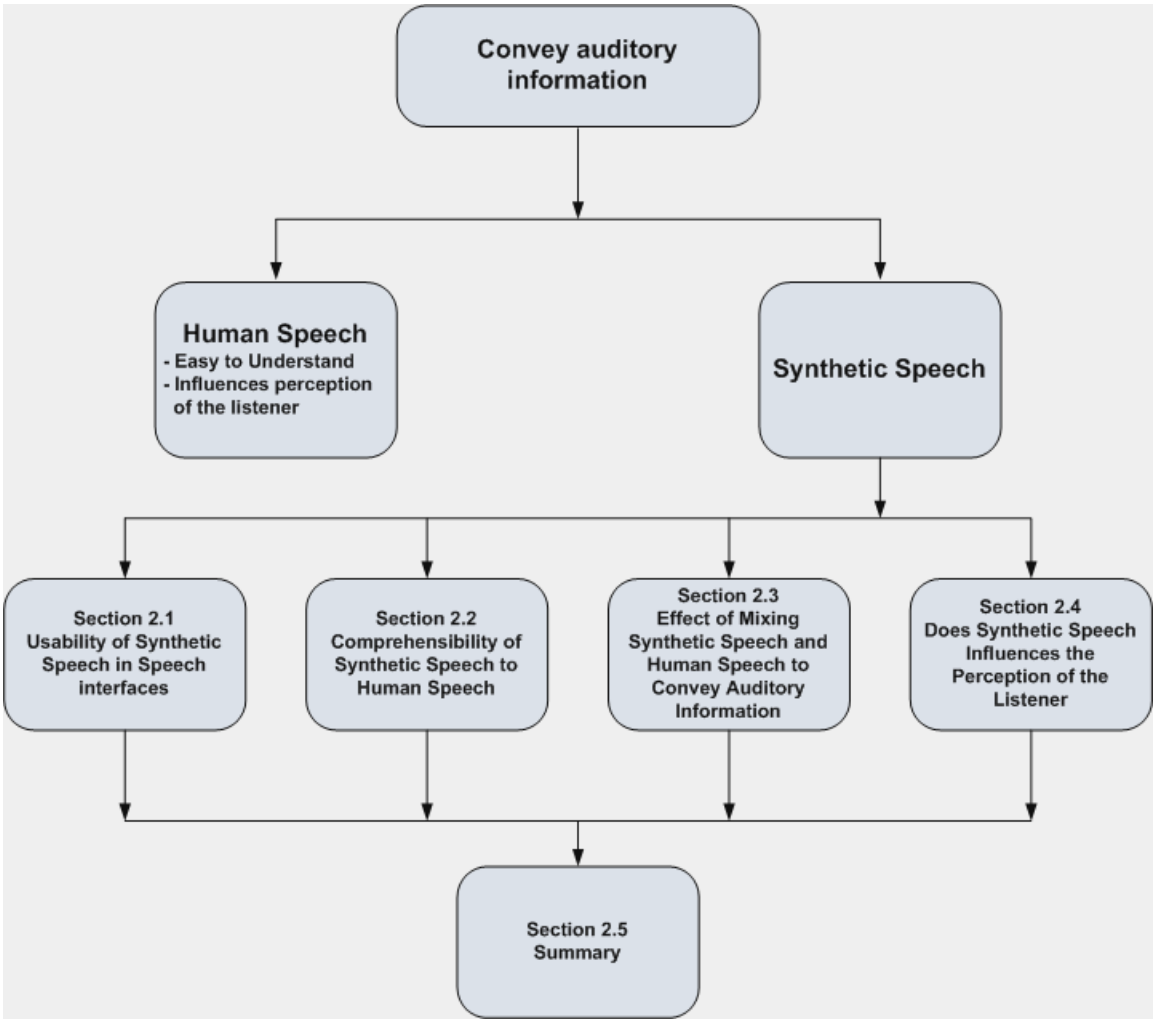


Figure 2.1: Structure of Chapter-2

2.1 Usability of Synthetic Speech in Speech Interfaces

In human-computer interaction, auditory information is generally conveyed through synthetic speech, which is evaluated based on intelligibility (understandability) and naturalness [BCS+99, SN85]. *Intelligibility* is measured in terms of how well Para-linguistic information (such as words and sentences) in computer generated speech is understandable to the users [FN99]. Currently, the best available synthetic speech systems, also called as Text-To-Speech (TTS), systems offer 97% intelligibility. This is close to the intelligibility of human speech (intelligibility of human speech is 99%) [BCS+99]. *Naturalness* is measured in terms of how similar synthetic speech is with respect to natural human speech. Lack of naturalness in TTS clearly tells the users that the speech is produced by a machine and not by a human. Though the best TTS systems offer 97% intelligibility, a significant amount of users' time is consumed and a greater effort is involved in understanding and recognizing synthetic voice parameters when compared to understanding and recognizing human speech. This is due to the lack of naturalness in synthetic speech [Oli97, SN85].

In recent years, the popularity of synthetic speech has increased due to the growing demand for Voice User Interfaces (VUIs). For instance, VUIs are used where access to high information bandwidth (i.e. visual displays) is impractical and the task of querying and delivering information is complex. VUIs primarily support tasks where the eyes and hands are busy (such as in driving and playing video games), and where verbal interaction is the most effective medium of communication (such as in assisting the

visually disabled). Some of the potential applications, where synthetic speech is commonly used are:

Automated Telephone Services

One of the most important areas that benefited from the use of synthetic voices is telecommunication services. In 2000, A report by the White House [Hou00] says that only two-third of the homes in the United States have access to the internet, which is believed to be more than that of any other country in the world. However, even with the advent of internet technology, the most common means of communication around the world is still through the telephone or cellular phones. This popularity is attributed to the fact that users perceive information, from the automated systems, based on the intelligibility of the synthetic speech and not on the naturalness of the speech [Dut97]. In addition, the increase in popularity of synthetic speech systems is due to the fact that 70% of users require very little interactivity with the system in order to obtain required information, some examples include cinema listings, road directions, railway reservations, etc.

Assisting Visually Impaired Users

A vast majority of applications (such as newspapers, email, document, web, etc.) make extensive use of graphics and text. Hence, accessing these applications is almost impossible for visually impaired users. James and Winograd [JW98] suggest that there are 11 million visually impaired users in the United States and many more in other countries. These users can access computers using TTS systems. Using TTS systems,

users can browse a computer and obtain the required information, such as listening to the news, email, and interacting with the system vocally.

The results from the above discussed studies describe how synthetic speech is evaluated and where synthetic speech is used. However, further research has been conducted to investigate the comprehensibility of synthetic speech. The next section described in this chapter compares the comprehensibility of synthetic speech to human speech.

2.2 Comprehensibility of Synthetic Speech

Various studies have suggested that humans recognize and remember words and sentences in synthetic speech in the same way as they recognize and remember these in human speech. In 2000, Lai et al. [LWC00] compared the comprehensibility of synthetic speech and human speech. In their study, the participants listened to messages, consisting of short, medium and long passages, in both pre-recorded human speech and synthetic speech. The synthetic speech was produced using one of the five commercial TTS systems; DECtalk for Windows 95 v 4.4, Acu-Voice AV1700 text reader, IBM Via Voice Outloud (from VV ’98), L&H TTS engine version 6.03, and Lucent Release 2. At the end of playing each set of messages, participants’ comprehension accuracy for both synthetic and human speech was measured. The results show that the comprehension rate for the messages is slightly lower for synthetic speech when compared to the comprehension rate of human speech (67% for synthetic speech and 73% for human speech). The main reason for the decrease in comprehensibility of synthetic speech is that it lacks the naturalness of human speech. The next section describes a detailed evaluation of a study,

which suggests that the naturalness in speech-based interfaces could be improved by mixing synthetic speech and human speech.

2.3 Mixing Synthetic speech and Human Speech

Some existing commercial systems (such as www.conita.com [con]) propose to mix male TTS voice with female TTS voice, while others proposed to balance the use of both human and synthetic speech (i.e., use synthetic voice for presenting dynamic information and use pre-recorded voice to present static information) in auditory interfaces. In today's world most of the information (such as news, weather, and email) is dynamic. Therefore, producing the pre-recorded human speech for presenting dynamic information is not only unrealistic but also time consuming. Hence, synthetic speech is used to present the dynamic information. Let us consider an example, "you got a mail from Mr. X", in this case the message "you got a mail from" is static, so pre-recorded human speech can be used to present the information, whereas "Mr. X" is dynamic and varies depending on the sender, hence synthetic voice can be used to present this information. Though the effect of the mixing the male TTS voice with female TTS voice was not formally evaluated, the later (mixing human speech with synthetic speech) was evaluated in various viewpoints (such as task performance [GL03] and the user's attitudinal response towards the system [MAE+99, Spe97].

Studies suggest that mixing synthetic speech and human speech in auditory interfaces leads to inconsistency. In 1999, McInnes et al. [MAE+99] conducted a study to compare the effect of mixing human speech with TTS, to a TTS-only condition. In this study, the

users were asked to listen to the messages using both the conditions. The users' attitudinal response towards the system was measured and evaluated. The results of the study show that users' attitude is more positive towards the mixing approach than with the TTS only approach. However, this study fails to assess the performance of the tasks (such as time taken to perform the tasks). In order to address the above issue, Gong and Lai [GL03] conducted a study to evaluate the effect of mixing synthetic speech and human speech in performing tasks, such as managing emails. In their study, they compared the effect of a pure synthetic voice condition to the combination of synthetic voice and human voice condition, in managing emails and calendar tasks. The participants were divided into two groups. In the first group, the participants performed the email and calendar managing tasks with pure synthetic voice, whereas in the second group, the participants performed the same tasks with the combination of synthetic voice and human voice. The results of the study suggest that user performance decreases with the mixing (TTS with human speech) approach when compared to using pure synthetic voice. The results also suggest that a decrease in task performance using the mixing approach is due to the lack of consistency in the interface. This is because, in the mixing approach, the users have to shift their processing modalities between the human speech and the TTS.

The studies discussed above primarily focus on evaluating the intelligibility and the understandability of synthetic speech. However, some studies also suggest that synthetic speech can influence the listeners' perception, as described in the next section.

2.4 Influencing User's Perception Using Synthetic Speech

In human-human interaction, the gender of the speaker plays an important role in influencing users' perception [Eag83], ranging from willingness to conformity [FSMKW80]. For example, praise or comments that are delivered from the male voice is considered to be more influential than the same praise or comments delivered from the female voice. To examine the above implication in human computer interaction, Lee et al. [LMB00] conducted a study, in which the participants were presented with a dilemma-situation scenario on the computer screen. Two possible options for that situation were also presented to them. The participants' task was to choose the correct option. Once the participants read the scenario, the computer orally presented its suggestions for one of the two possible options. After listening to the computer's suggestion, participants were asked to fill out a paper-and-pencil questionnaire to find out if they can easily identify the gender of the computer generated speech. The results show that participants did not have any problems in identifying the gender of the TTS voice. Consistent with the literature [Eag83] that gender plays an important role in influencing users' perception, their results also demonstrate that a male voice is more influential on user's decisions than a female voice, whereas the female voice is regarded as being more socially attractive and trustworthy than a male voice.

Apart from conveying gender, vocal characteristics of human speech also manifest personality, such as introversion and extroversion. To find out the influence of synthetic speech on a listener's interpretation of personality, various studies [BGS67, IN00, NL01, NMF+95] were conducted to investigate whether people could identify and respond to

computer personalities in the same way as they would recognize and respond to human personalities. Nass and Lee [NL01] created personalities (introvert and extrovert) by manipulating synthetic voice parameters such as pitch, volume, pitch rate and pitch range. Prior to the start of the experiment, the participants' personalities were assessed using a paper-pencil questionnaire. After assessing their personalities, the participants were equally divided into 2 groups; the introvert group and the extrovert group. In the experiment, the participants listened to the computer generated speech (either introverted speech or extroverted speech) to assess reviews about some books available on the Internet. After listening to the reviews, participants were asked to identify the personality of the computer generated speech. They were also asked to rate the quality of the book, and the credibility of the reviewer. The results illustrate that the participants convincingly distinguished between the introvert and extrovert voices. The results show that personality dimensions such as introversion and extroversion can be created by manipulating synthetic speech parameters and also have an influence on the user.

In recent years, various studies [BNS02, Cah89, SG03] have been conducted to find out the degree of recognizing emotions in synthetic speech. Cahn [Cah89] says that six types of emotions: angry, disgusted, glad, sad, scared and surprised, can be produced using synthetic speech. In this study, five sentences were synthesized for each of the emotions. The voices were then played to the users, to find out if they can identify emotions in these voices. The results of this study suggest that a significant amount of emotions were identifiable by the users.

The results from the above discussed studies suggest that users could identify the social variables (such as emotions, personalities and gender) in synthetic speech, in the same way as they identify these variables in human speech. This implies that synthetic voice is robust enough to allow a large variation of distinctions/discriminations by manipulating the synthetic voice parameters.

2.5 Summary

A number of studies have been conducted to analyze the effect of synthetic speech in various dimensions. One line of research focused on analyzing the effect of synthetic speech in comprehending auditory messages. The results from this research suggest that auditory information can be successfully conveyed to the users using synthetic speech. Another line of research focused on finding out the effect of identifying social variables in synthetic speech, such as recognizing personalities, gender and emotions. Results from these studies suggest that social variables can be created using synthetic speech by manipulating synthetic speech parameters, such as pitch, speech rate and intensity. In other words, these studies suggest that various synthetic voices can be created by manipulating synthetic speech parameters. These results are an integral part of this thesis. Since discriminating between synthetic voices will facilitate creation of different synthetic voices.

This thesis describes two solutions, which focus on improving the navigation in auditory interfaces. The second solution, improving navigation using multiple synthetic voices, described in this thesis focuses on manipulating synthetic speech parameters in order to

create various synthetic voices for providing navigational cues to the users. This is explained in Chapter 5.

Chapter 3 : Navigation in Speech Interfaces

Advances in speech synthesis and speech recognition techniques have led to a proliferation of voice based applications in recent years. Additionally, higher accuracy levels in speech recognition algorithms have facilitated the creation of interaction techniques that rely heavily on speech input and output. Through the use of voice commands, users can give directives to an application, navigate between different modes of an application and provide input to fill out a form or a document. Applications are also designed to give output in the form of voice based prompts using TTS algorithms. For instance, in IVR systems, menus are presented acoustically to the users. Similarly, interfaces (such as those that assist the visually impaired users) are also completely dependent on auditory interactions for conveying information to the users. In such interfaces, the graphical display cannot be used to provide navigational cues to the users. Likewise, in situations where multitasking (such as in driving and getting directions) is critical, applications are dependent on auditory interactions (input/output) to facilitate users' goals.

The remainder of this chapter is organized as follows. The next section discusses some of the common problems with auditory interfaces, followed by some of the current solutions to these problems, in the subsequent sections. Figure 3.1 shows the structure of this chapter.

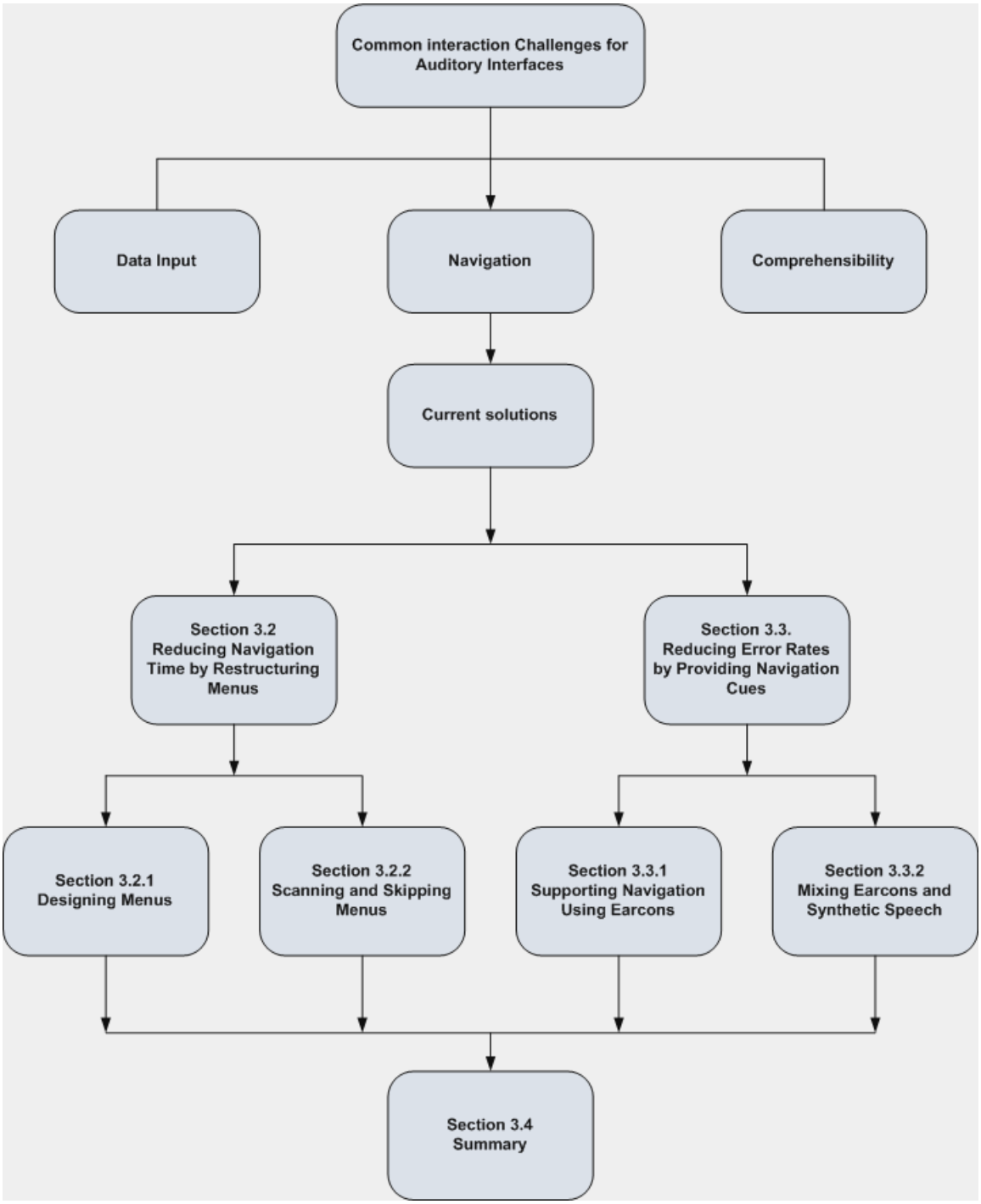


Figure 3.1: Structure of Chapter-3

3.1 Interaction Problems with Auditory Interfaces

Several challenges are apparent in auditory interfaces: comprehensibility, navigation, data input, etc. Navigation is one of the most critical issues. Navigation is defined as searching, browsing, and/or scanning for information in a system. Many users consider navigating the set of menus in TBIs (such as finding the information about movie listings and addresses of different malls in the city), a tedious process. This is because, a significant amount of the interaction in voice based applications is dedicated towards navigation tasks. Navigation is funneled through menus, buttons or commands. Options for executing a given command are presented hierarchically with the user entering the system through the root node. By listening to the appropriate prompts (or menu items) the user can navigate the hierarchy that branches the users' calling path into sub-hierarchies until the final task has been achieved. Using sub-hierarchies for directing a user in a system has direct implications on users' performance in navigating the system. In general, options/menus in these interfaces are presented serially using synthetic voice (sometimes pre-recorded human speech is also used) and no additional aid is provided to the listeners' short-term memory [HN89] about the position of a node in the hierarchy. This narrow bandwidth available for interaction results in degraded usability performances in speech-based interfaces, such as TBIs, [KHHK99, SHS95]. In systems with large hierarchical structures, where many levels are necessary, users can lose context, can go astray, or can increase their error rates in selecting options. One of the major obstacles in promoting speech-based interfaces has been the high level of user frustration in navigating hierarchical menus [Yan95]. One of the most important reasons for the above mentioned problem is that these interfaces are not explicit about the parent-

child relationship between the menu items. Hence, in order to improve navigation, Rosson [Ros85] suggested that the interfaces should constantly update the users' on their current position in the hierarchy.

In an effort to address the above mentioned problem, various studies had been conducted and evaluated. These studies suggest that navigation in auditory interfaces can be improved in two different ways: either by reducing the navigation time (by restructuring menus in the hierarchy) or by reducing the error rates (by providing navigational cues to the users). The following sections provide a review of the literature on these studies.

3.2 Restructuring Menus in the Hierarchy

The studies described in this section focus on reducing the navigation time in auditory interfaces by restructuring menus in the hierarchy. Some of the widely accepted approaches include effective menu design for locating desired items (Menu Design) and allowing users to skip uninterested menus (Scan and Skip).

3.2.1 Menu Design

One of the major factors that influence the navigation in auditory interfaces (such as TBIs) is the structure and size of the hierarchy. Searching particular information is considered as the fundamental task in auditory interfaces (e.g. TBIs). Therefore, the usability of these applications was measured based on the effectiveness of obtaining the requested information. Martin et al. [MWW90] conducted a study to evaluate the effect of menu design (structure) in TBIs. In their study, a deep and narrow structure was

compared to a wide and shallow structure. The deep and narrow structure contained six levels with two menus/items per node, whereas the wide and shallow structure contained two levels with eight menus/items at each node. Participant's tasks involve identifying particular information in the interface. The results of this study suggest that the participants performed the tasks better with wide and shallow structures than with deep and narrow structures. One of the main reasons for the decrease in the performance with the deep and narrow design is that the users are forced to make many selections in order to obtain the information. However, the results also implied that the other factors that may affect the user's performance are: the rate of synthetic speech, the time taken by the user to enter the input and the time taken by the interface to make the menus available to the users.

On the other hand, results from several studies suggest limits on the number of menu choices ranging from no more than four [RE89, SH93] or in some cases up to a maximum of nine [BM99]. Suhm et al. [SFG01] compared long menus containing items with well-defined functions to shorter menus consisting of prompts that compressed several functions into one menu item. Their results indicate that long menus with specific and clearly defined categories, can route users more efficiently than systems containing short menus with items consisting of broad categories. Since long menus consisting of items with succinct functions result in fewer layers than short menus, Suhm et al. [SFG01] suggest using long menus with fewer levels if possible. Another study by McInnes et al. [MNA^{+}99] suggested that navigation in TBIs is improved by confirming the caller’s actions. Confirmation reinforces the user’s previous action and therefore

facilitates further progress in their task. Other studies have also suggested some solutions to improve the navigation in auditory interfaces, such as: optimizing the potential path that a user takes to travel in a system [Bal99], finding a balance between the length of the menus and their associated prompts [SGF01]. However, all these studies have failed to convey to the users that the menus in the auditory interfaces are arranged as hierarchies.

Results from the above discussed studies suggest that a considerable amount of navigation time can be reduced by designing the best possible layout for menus items in the hierarchy. In addition, navigation time can further be reduced by allowing users to scan and skip the uninterested menus. The next section highlights some of the advantages of implementing “scan and skip” options in auditory interfaces.

3.2.2 Scan and Skip

Resnick and Virzi [RV92] have suggested an alternative method to reduce navigation time in hierarchical TBIs, called skip and scan. In this study, they have suggested that one of the main problems with the TBIs is that users are forced to listen to all the prompts/menus before they make a selection. In some cases, the size of the prompts may be very long. Therefore, users may get frustrated easily. In order to address the above problem, a new technique, skip and scan, was introduced. This technique will allow users to skip the uninterested menus. Figure 3.2 shows the menu selection in the skip and scan interface. Users can press “9” in the telephone keypad to skip the current prompt or select “7” to listen to the previous prompt. In this study, the skip and scan interface was compared to the traditional telephony interfaces. User’s tasks involved

finding a particular option in the hierarchy. The results of this study show that users performed the tasks faster with skip and scan interfaces than with the traditional interface. However, the authors suggest that skip and scan should be used only when there are more than four menus/options are used at one level. For example, if there are 15 movies playing in a city; skip and scan technique can be used to skip the irrelevant movies and find the desired movie quickly and easily.

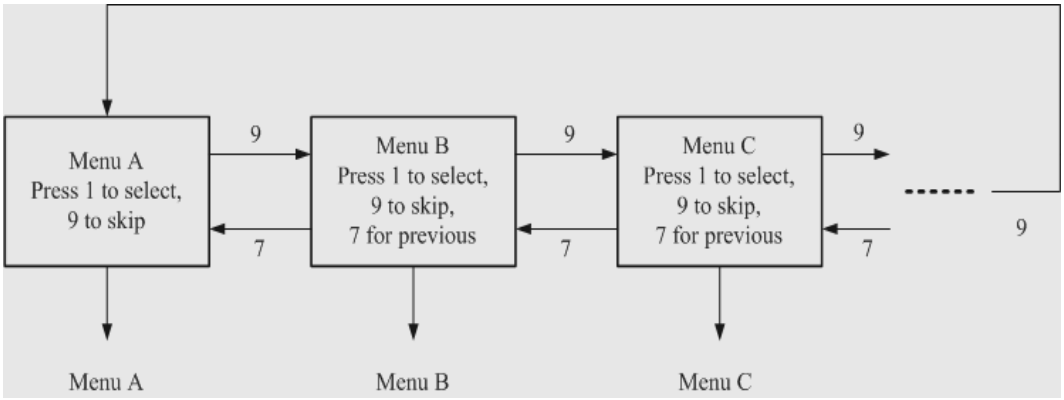


Figure 3.2: Menu selection in scan and skip interface

Although scan and skip is effective in reducing the navigation time in auditory interfaces, even this technique failed to convey that the menu items are arranged in a hierarchical order. Therefore, the chances of losing context in the system are more. In order to overcome the above problem, another line of research suggests that providing navigational cues about the structure of the hierarchy might reduce error rates. The next section describes a new line of research, which focuses on improving the navigation in auditory interfaces by reducing the error rates.

3.3 Providing Navigation Cues to Reduce Error Rates

This section focuses on improving the navigation in auditory interfaces by reducing the error rates of obtaining information. The studies described in this section suggest that a significant amount of error rates can be reduced by providing navigational cues to the users. Some of the widely used approaches include supporting navigation using earcons and reducing error rates by combining synthetic speech and earcons.

3.3.1 Supporting Navigation in Hierarchies using Earcons

Studies suggest that the issue of navigation in auditory interfaces has been addressed through the implementation and evaluation of "earcons". Earcons were developed by Blattner et al. [BSG89], Sumikawa et al. [SBJG86], and Sumikawa [Sum85]. Blattner et al. [BSG89] defined earcons as "abstract, synthetic tones that can be used in structured combinations to create sound messages for representing parts of an interface". Brewster [Bre98] says that earcons are constructed from motives (motives are short musical tones that can be combined in different ways to represent parts of an interface). For example, a simple sound motive can be used to represent a file or a folder and it may be played whenever that particular file or folder is opened or closed.

Brewster et al. [BWE92, BWE93, BWE95] show that earcons can be used to represent a small hierarchical structure of about 5 to 9 items. Participants' tasks involve identifying the position of the associated node in the hierarchy by listening to the played earcon. The results illustrate that participants could identify 80% of the nodes in the hierarchical structure accurately by listening to the earcons. However, the ease of remembering 9

nodes has lead researchers to evaluate the effect of earcons on more complex hierarchical structures. Blattner et al. [BSG89] suggest that hierarchical structures can be represented by creating hierarchical earcons, and hierarchical earcons can be created by manipulating various audio parameters such as pitch (frequency of the tone), register (position of the motive in the musical scale), intensity (loudness of the sound), timber (quality of the sound), and rhythm.

The idea of representing earcons in a hierarchical manner is mainly inspired from Sumikawa's [Sum85] three golden guidelines. Sumikawa [Sum85] suggests that hierarchical earcons can be created by manipulating motives in three different ways. They are as follows:

- ❶ **Repetition:** Repeating the preceding motive and all the parameters that are associated with that motive.
- ❷ **Variation:** Varying one or more audio parameters from the preceding motive.
- ❸ **Contrast:** Altering the pitch and/or rhythm to make an earcon's sound contrast with the preceding one.

By following the above discussed guidelines, Brewster [Bre98] created a hierarchy of 27 nodes. This study was conducted to evaluate the effect of earcons in complex hierarchical structures. Different audio parameters, such as pitch, register, intensity, timber, and/or rhythm, were manipulated to create the earcons that are related in a hierarchical manner. The earcons were used to represent nodes in the hierarchy. The participant's tasks involve identifying the nodes in the hierarchy by listening to the earcons. The accuracy rate and

the effect of training for locating the nodes in the hierarchy were measured and analyzed. The results of this study suggested that participants could recall 81.5% of the earcons. The results also indicated that the type of training also affected the recall rates of earcons. However, one of the major problems with the hierarchical earcons is that increasing the size of the hierarchy would increase the complexity of the earcons. Thus, compound earcons were suggested to represent nodes in the hierarchy.

Brewster et al. [BCH98] designed another study to evaluate the use of compound earcons for representing large hierarchies. Compound earcons were constructed by initially creating a set of sounds and then concatenating these sounds (i.e. creating various combinations of these sounds) according to the number of levels and nodes in the hierarchy. The results of this study suggest that 97% of the earcons were recalled accurately. Although, compound earcons are effective in conveying hierarchical information, they possess some limitations. First, the length of the earcons (in time) increases with respect to the number of levels in the hierarchy. Second, the users have to listen to the entire earcon in order to locate the position in the hierarchy. Hence, when these earcons are implemented in speech-based interfaces (where speech is considered as the dominant mode of interaction) users have to listen to the entire earcon in order to navigate the hierarchy. This in-turn increases the overall users' navigation time.

Results from the above discussed studies suggest that the error rates can be significantly reduced by providing navigational cues about the structure of the hierarchy. However, these studies did not evaluate the effect of earcons in speech-based interfaces, where

speech is considered as the dominant mode of interaction. The next section described in this chapter describes the effect of earcons in speech-based interfaces.

3.3.2 Combining Speech and Earcons to Improve Navigation in Auditory Interfaces

Vargas and Anderson [VA03] suggest that navigation in speech-based interfaces could be improved by combining synthetic speech and earcons. In this study, the researchers compared pure synthetic speech to a combination of synthetic speech and earcons. During the training session, participants were shown a graphical representation of hierarchical menus and were allowed to choose a menu and listen to its associated speech. The participants' tasks involved locating the specific menus items in the hierarchy. In this study, the recall rates and the time taken for each task were measured. Results from this experiment showed that the participants performed the node finding tasks slightly better with the combination of synthetic speech and earcons, than with the speech only condition. Their results also suggest that though participants performed the tasks better with speech and earcons condition, the participants took 18% more time to perform each task than the participants who performed the same tasks with speech only condition.

Although combination of speech and earcons provide navigation cues in hierarchies, they possess several disadvantages. When earcons are combined with speech in speech-based interfaces, the earcons are generally played in the background. This mode of interaction forces users to listen to both the speech and the earcons, in parallel, in order to identify

their position in the hierarchy. This might impose a higher cognitive load on the users' short-term memory and hence will increase navigation time. These results suggest that non-speech audio can be used but at the cost of overloading the user with additional information.

3.4 Summary

The issue of navigation in auditory interfaces has been addressed in different ways. One line of research focused on reducing the navigation time by restructuring the navigation hierarchy, whereas the other line of research focused on reducing the error rates by providing navigational cues to the users. Although, restructuring the hierarchy menu reduces navigation time, the error rates for obtaining the requested information were never reduced. In other words, the major drawback with these studies is that they failed to convey to the users the structure of the information in the auditory interfaces. Therefore, earcons were introduced to convey the arrangement of nodes in auditory interfaces, to the users, in order to reduce the error rates. Results from various studies suggest that earcons were effective in conveying the hierarchical information to the users. The results also suggest that the navigation time for obtaining the required information is increased when earcons are used in conjunction with synthetic speech in speech-based interfaces. This is due to fact that users have to listen to both speech and earcons in order to locate the position of the node in the hierarchy. The studies described in this chapter provide a strong background that a significant amount of work has been done to improve

navigation in auditory interfaces. However, there are many aspects that need to be further investigated.

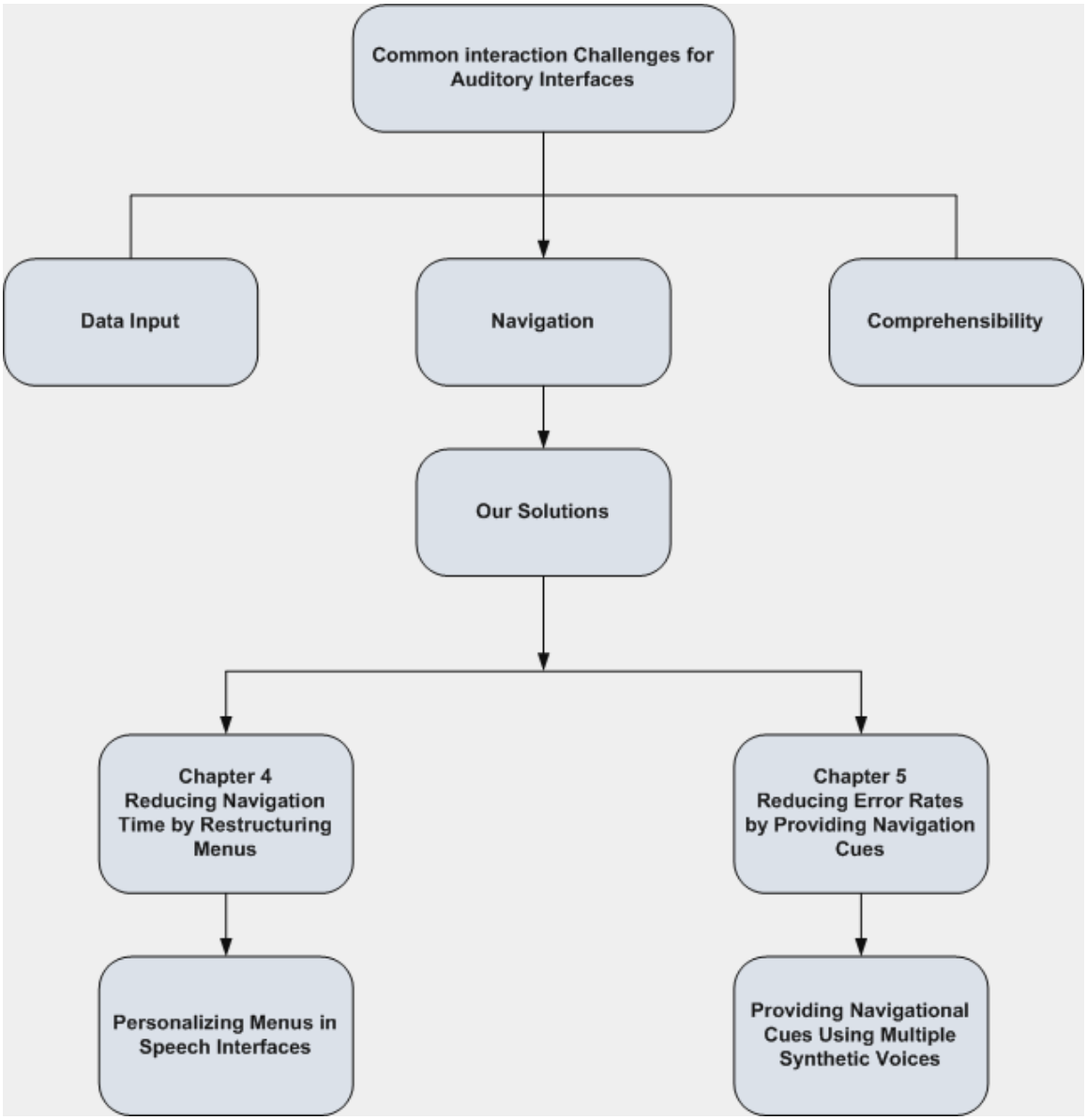


Figure 3.3 Proposed solutions

In an attempt to explore the issue of navigation in voice based interfaces, two novel solutions (Figure 3.3): personalizing speech-interfaces and representing hierarchies using

multiple synthetic voices, are proposed and evaluated in this research. The first solution, personalizing menus in auditory interfaces, focuses on reducing the navigation time by restructuring the menu hierarchy whereas the second solution focuses on reducing the error rates by providing navigation cues. The following chapters describe the implementation and the experimental analysis of the above proposed solutions in detail.

Chapter 4 : Restructuring Menus - Personalizing menus in Speech Interfaces

In recent years, touch-tone interfaces (or telephony interfaces) have been widely deployed for accessing various types of information. Although, some studies [Tat96] have argued that TBIs may not continue to exist, touch-tone voice interfaces continue to represent a significant component of our electronic day-to-day transactions. The popularity of touch-tone interfaces has recently grown due to the introduction of automated call-centers. This growth is a result of the improvement in the quality of synthetic speech. Using touch-tone systems, end-users are able to perform their day-to-day banking transactions, get automated directory assistance services [LM03], pay their bills, retrieve ticketing information, lookup city directions, and listen to movie listings at their convenience. More generally, touch-tone interfaces have been used in instances where access to high information bandwidth (i.e., visual displays) is impractical and the task of querying and delivering information is complex. In most cases touch-tone interfaces are coupled with voice input to support tasks where verbal interaction is the most effective medium of communication, such as while driving and assisting the disabled.

Recently, there has been considerable interest in exploring methods for reducing the amount of time a user necessitates for navigating a hierarchical menu structure, such as TBIs. Some of the major solutions that have been provided involve either optimizing the potential paths a user can undertake in a system [Bal99], finding a balance between the

length of menus and their associated prompts [SFG01], inserting additional location cues in the menus [Bre98], enhancing the system with features such as barging-in [Lar02], and determining the most efficient balance between wide and deep hierarchies in telephony systems [Bon99]. The major drawback to the studies described above is that users are forced into following all the prompts before arriving at their final destination. In certain cases, frequent callers memorize the sequence of actions but nevertheless have to follow the prompts in order to avoid erroneous routing paths.

This chapter describes a new method for improving navigation in touch-tone interfaces by re-arranging the menu structure. The core concept relies on allowing users to personalize their menu options by giving them direct access to nodes of interest in the hierarchy without going through the various layers of prompts.

4.1 Personalization/Customization

Customization is defined as the changes that are made by the user to the default system/interface for efficiently accessing some of the desired options [McG02]. Traditional software applications offer all the functionalities to the users, regardless of their tasks and experience. However, users typically use only a few set of options in an interface [CC84, LJSC00, MM00]. Therefore, personalizing these sets of options in the interface will help users access their preferred options easily and quickly. For example, bookmarked options in a web browser allow users to save the address (node on the Internet) of a favorite page and then refer to that page later. Another similar feature is that of hidden menu items under the Windows operating system. Menus can be extended or

hidden. The most common implementation presents only menu items that have been most recently used. Customization not only allows users to skip over the entire set of uninterested options but also significantly reduces the amount of time users take for accessing their interested features.

The idea of personalizing menus in touch-tone interfaces is mainly inspired from McGrenere's solution to complex software [McG02]. In recent years, many similar types of softwares have been introduced in the marketplace. These software tools initially start with a few set of options, and grow with every new release. Hence, options in these applications have become visually complex. Although, toolbars have been introduced to solve this problem, even toolbars grow in a similar fashion. Recently, McGrenere [MBB02, McG02] created a system for allowing users to customize favorite options in Microsoft Word™. Using this system, users can save their favorite options/menus in the customized interface. This system also allowed users to toggle between the customized interface (MSWord Personal) and the default Microsoft Word interface. In their study, MSWord Personal was compared and evaluated with the default Microsoft Word application. Initially, the personal interface contained a very few functions. However, the users were allowed to modify the interface using the modify function, according to their needs. In a user study, users' satisfaction, and their ability to learn and navigate the software were tested [McG02]. Overall, the results show that the participants preferred using the new customized system better than the default Microsoft Word™. Participants also, suggested that they were able to access their favorite options easily with-out the need to navigate the entire set of menus in order to find their favorite menu/option.

4.2 Experiment 1 - Personalizing Menus for Navigation in Touch-Tone Voice Interfaces

In this part of the thesis, an interface was implemented, for bookmarking or saving the callers favorite nodes in the menu hierarchy. The bookmarks later appear as options higher up in the menu tree. Users can then bypass the various layers of prompts and access directly their bookmarked information. The set of bookmarks created by the user is provided in a different sub-tree of the application.

Figure 4.1.a shows a sample hierarchy before the creation or insertion of personalized bookmarks. In this example, the user chooses to save two nodes in the hierarchy. The routing information in the hierarchy is then stored in the database for that user. Figure 4.1.b depicts the new hierarchy once the user has saved nodes of the hierarchy. A new level gets created from which the user has either the choice to follow the default menu options (sub-tree labeled “Main”) or to choose the saved options in the personalized list (sub-tree labeled “Personal”). The personalized list will dynamically create prompts for each saved node that exists in the database. By selecting an item in the list of personal options, the user is directly routed to the corresponding node in the original hierarchy. From that point onward in the session, control in the application is passed to the destination node. In essence this solution “flattens” the hierarchy of prompts into a linear list with connections to nodes in the original list.

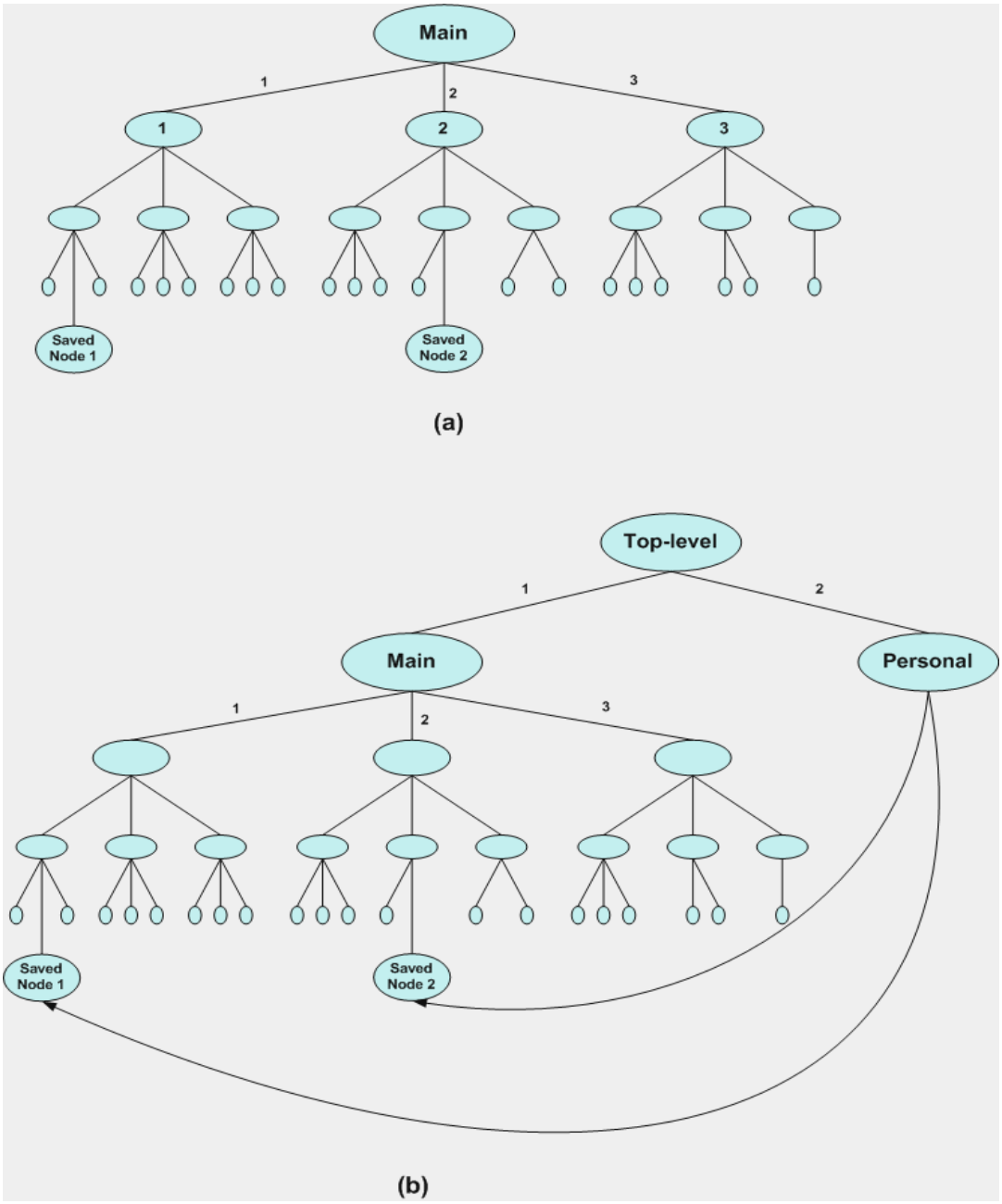


Figure 4.1: Diagram showing the linking of bookmarked nodes in the list of personal options to the nodes in the original tree. The user can avoid traversing the higher level sub-trees by selecting a node in their personal options.

4.2.1 Interface

At the interface level, each touch-tone menu item consists primarily of two elements, a function label and an associated action [Bal99]. For example, in the menu item “To review your appointments, press one”; the function label “review your appointments” is coupled with the user action of “pressing one”. Associating a node that is deep in the tree to a personalized option becomes problematic as the user would request more information about the node, in particular about its function label. For example, saving the node that provides information about the user’s appointments may have taken a route that gets narrower as the user traverses deeper into the hierarchy. Higher level nodes serve the purpose of filtering the flow of the user’s actions. In the example given above, the user may reach a level at which the choice would be to select “business related tasks” or “personal tasks”. The next level in the tree would then list the appointments specific to the type of choice made at the level above. Therefore using the function name associated with the node being saved is insufficient to identify the menu item, i.e. business appointments are different than personal appointments and therefore simply saving the title “review your appointments” in the list of bookmarked nodes will not clearly identify the menu function. Therefore, in saving menu nodes it is necessary to distinguish the label of the node that has been saved from the label of the node in the original hierarchy.

To overcome the ambiguities that may arise from simply associating the menu item’s label in the list of personalized options, the implementation allows users to record a message using their voice to identify the node in the hierarchy that is bookmarked. For

instance, if the user is interested in book-marking the node that accesses information about their personal appointments, then the user could potentially save a recorded message that identifies the level in the hierarchy related to that information by saying “personal appointments”. The recorded message gets saved with the list of bookmarks. When the user wants to later access the saved personal options, the system replays the recorded message. The duration of a user’s recording is fixed to a maximum of ten seconds.

To save nodes in the hierarchy the user presses the “9” key at which time the system will request that they save a recorded message to label the prompt in their personal list. Figure 4.2 shows the scenario for saving “personal appointments” in the personalized list. The user also has the choice to delete the entire list of saved prompts. When the system replays the list of personal prompts, the order of the menu items in the list of bookmarks is dictated by the order in which the menu items were saved. Therefore, the first node saved will be the first that will be replayed in the list of prompts in the personal options. Figure 4.3 shows how the control is transferred from the personalized interface to the default interface.

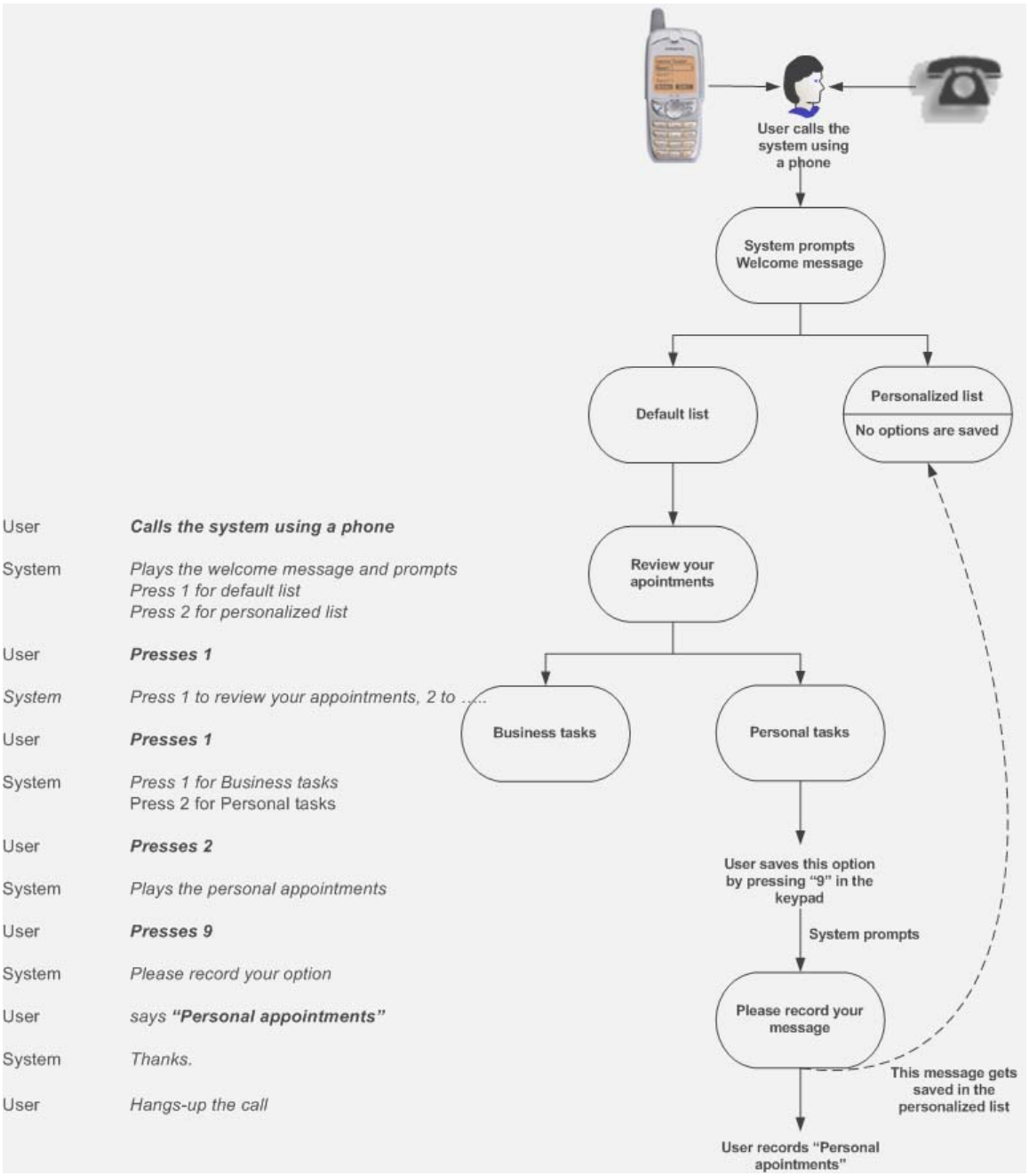


Figure 4.2: Scenario for saving "Personal appointments" in the personalized list

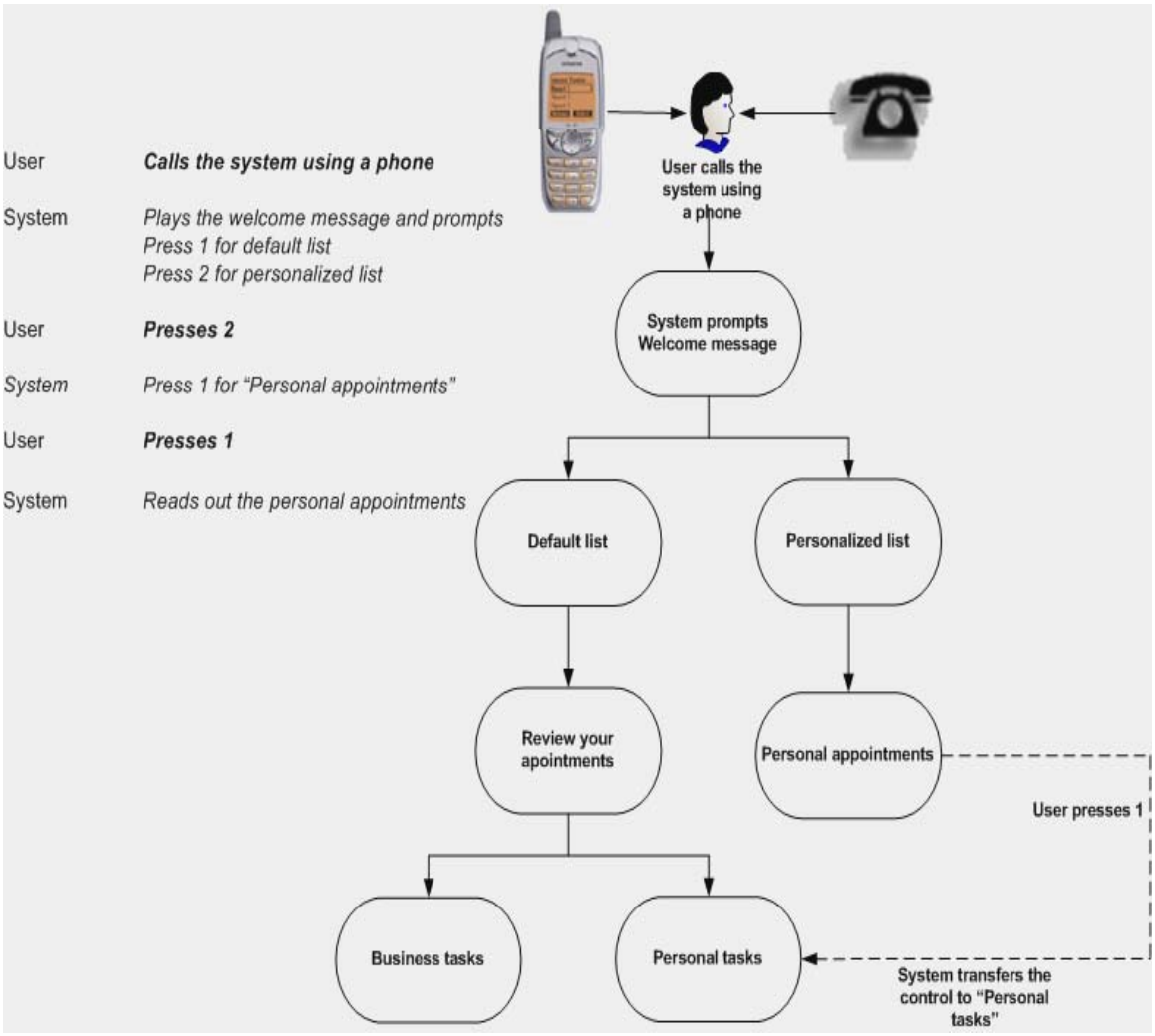


Figure 4.3: Transfer of control from the personalized list to the default route in the main menu structure

4.2.2 System Architecture

The prototype is built by extending the VoiceXML (VXML) standard for touch-tone applications. While many telephony platforms exist, Voicegenie [voi] was selected for testing purposes, because Voicegenie (located in Toronto, Canada) is considered as the world's leading VXML platform provider. At the core of the system is a JSP (Java Server Pages) application that dynamically creates the VXML pages for the user (Figure

4.4). The JSP application creates the list of prompts from the data residing in the database. Each prompt is associated with the appropriate node in the original hierarchy. As users browse the system, bookmarked menus are saved in the database with associated pointers to the nodes in the hierarchy. Each bookmarked menu is also attached to a recording that identifies the function of the menu item. All state variables that have been created during a session also get stored in the database and are linked with the node that is bookmarked. Upon entering into a new session, the JSP application re-creates the entire list of prompts including those saved under the bookmarks.

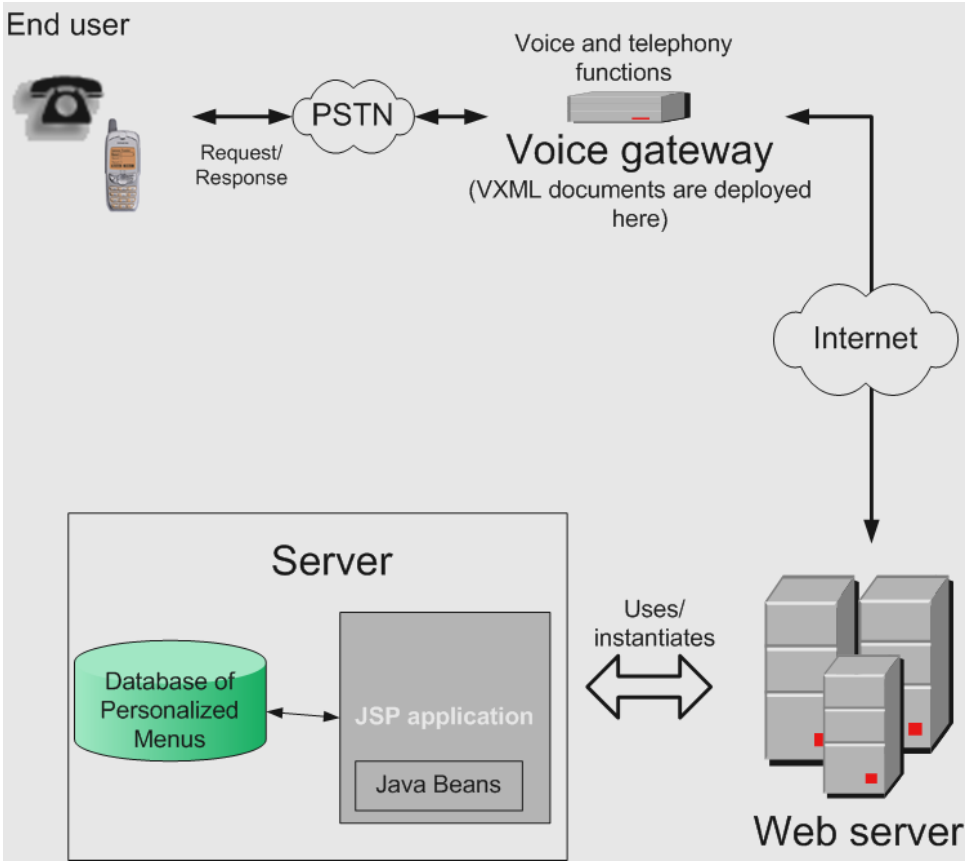


Figure 4.4: High-level system components

4.2.4 User Interaction Dialogs

The proof of concept system was developed by implementing two touch-tone applications. The first application was created with the intention of having a simple hierarchy (local bus transit information) and the other is structured using a complex hierarchy (cinema listings which require that the user traverse several layers).

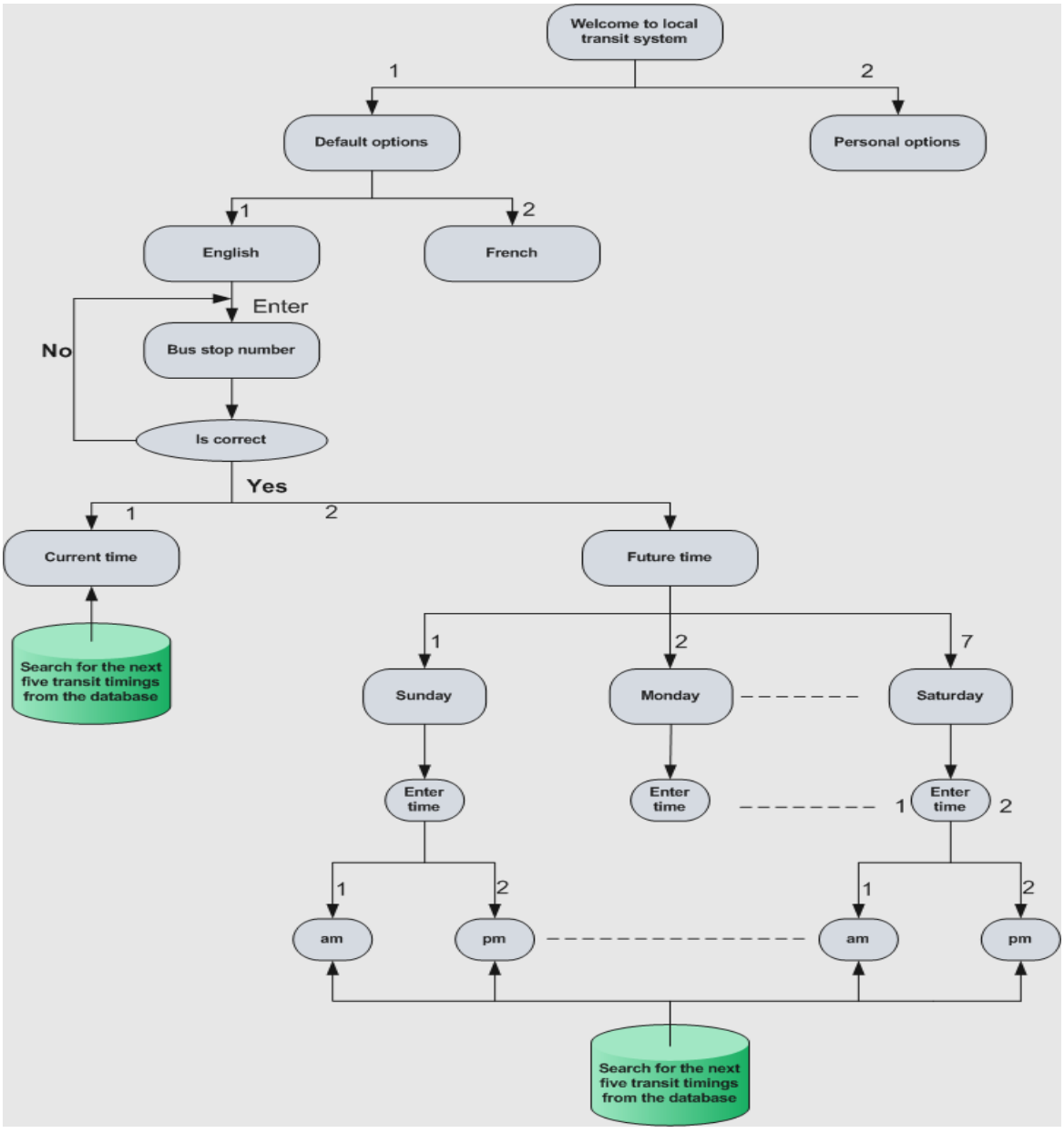


Figure 4.5: Interface with only default options (Local transit system)

The implemented bus transit application is identical to the one that is currently available by the local city transit system with the exception of the additional features of saving personal favorites. The structure of this application is shown in Figure 4.5. The levels of users' interaction with the system were illustrated through a scenario for the transit system. The sequence of direct dialog interactions adopted by the user is illustrated in Table 4.1. In this scenario, the user saves the times for the current schedule for the bus stop 60613. The caller dials the system at 12:30 pm and therefore all current timings after 12:30 pm are presented. The user then opts to save the current timings for the selected bus stop for future reference. When the user bookmarks the menu node, the system requests a recording from the user that will later get replayed.

Table 4.1: Interaction dialog for retrieving information about current bus schedules and book-marking the relevant node in the menu

System	Welcome to the local transit system. Press 1 for default options or 2 for personal options.
Caller	Presses 1
System	Welcome to the local transit system. To save information at any level please press "9". Press 1 for English or 2 for French.
Caller	Presses 1
System	Please enter the bus stop number.
Caller	Enters 60613
System	Press 1 for current schedule or 2 for future schedule.
Caller	Presses 1
System	The current schedule is: - Bus 75 at 1:00 pm, 1:35 pm and 2:20 pm - Bus 60 at 1:40 pm, 2:15 pm and 2:45 pm

Caller	Presses 9
System	If you wish to save current times please record an associated prompt for this function.
Caller	Says “Current schedule at stop number 60613, in English”
System	Item saved.

Table 4.2 shows the sequence of prompts once a node has been saved in the system. The caller accesses the system at 3:30 pm and selects the previously saved node for the current timings. The system is routed to the node in the original tree that retrieves the list of the current bus timings, i.e. timings after 3:30 pm. In this example, the user is able to avoid one layer in the menu tree by directly accessing a previously saved node.

Table 4.2: Interaction dialog for retrieving information about current bus schedules using the list of book-marks

System	Welcome to the local transit system. Press 1 for default options or 2 for personal options.
Caller	Presses 2
System	Press 1 for “Current schedule at stop number 60613, in English”
Caller	Presses 1
System	The current schedule is: - Bus 75 at 3:40 pm, 4:15 pm and 5:10 pm - Bus 60 at 4:10 pm, 4:55 pm and 5:35 pm

Figure 4.6 shows the interface, which includes both the default options and the personal options. The lists of bookmarks that are saved in the “personalized list” are:

Bookmark 1: English language -> “current timings” for the bus stop number “60613”

Bookmark 2: English language -> “future timings” for the bus stop number “26532” for “Sunday” at “12:00 pm”

Bookmark 3: English language -> “future timings” for the bus stop number “23422”

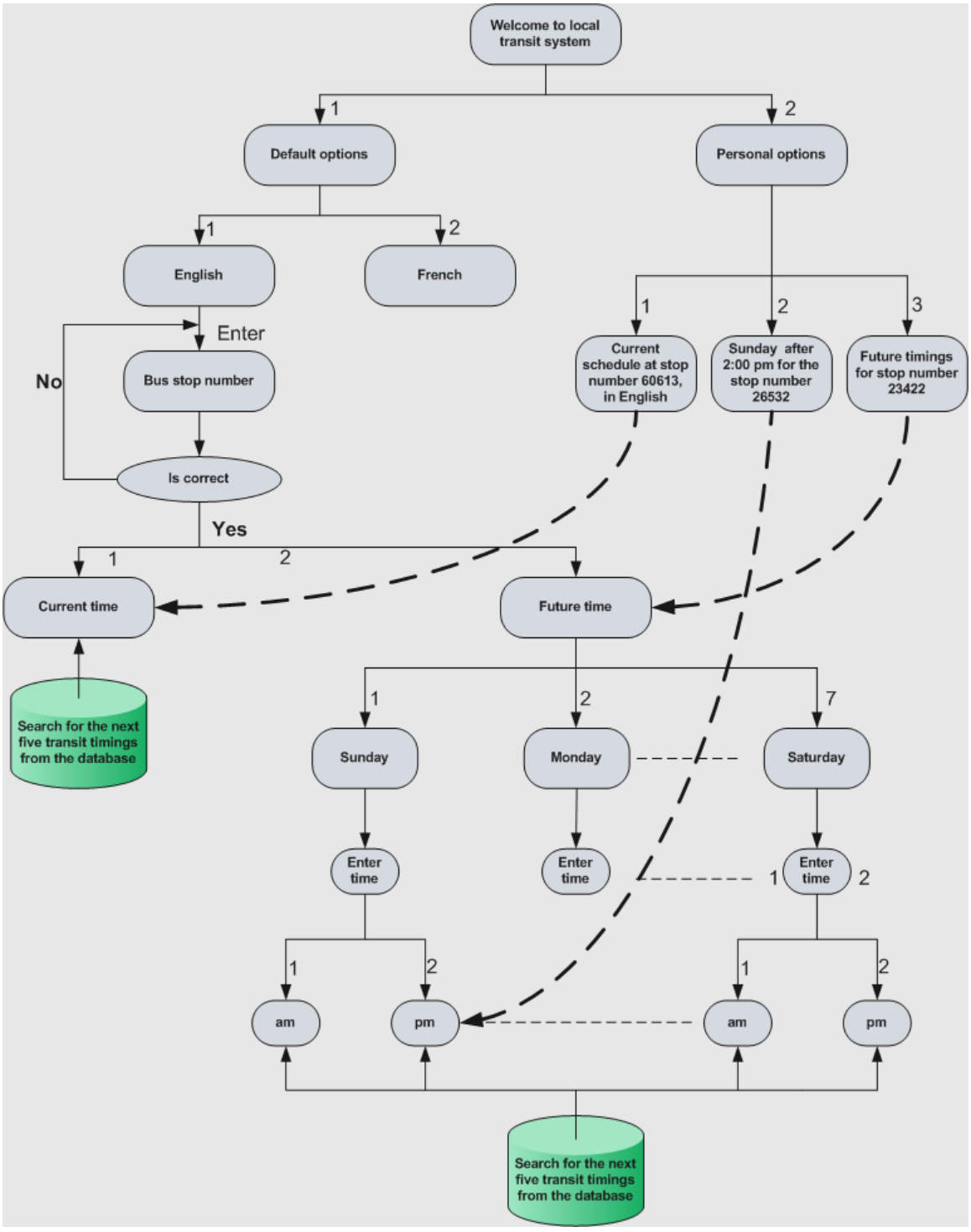


Figure 4.6: Interface with both default options and personal options (local transit system)

The number of dialog interactions between the user and the system is significantly reduced when more complex queries are involved. Table 4.3 illustrates the sequence of actions for a user requesting timings for a future bus schedule. In this scenario the user goes to work at 1:00 pm on Sundays and therefore wishes to save the schedule for 12:00 pm.

Table 4.3: Interaction dialog for accessing bus schedules for Sundays at 12:00 pm and for saving the node of interest in the list of book-marks (this sequence is replicated to match the dialog provided by our city transit system)

System	Welcome to the local transit system. Press 1 for default options or 2 for personal options.
Caller	Presses 2
System	Welcome to the local transit system. To save information at any level please press “9”. Press 1 for English or 2 for French.
Caller	Presses 1
System	Please enter the bus stop number.
Caller	Enters 60613
System	Press 1 for current schedule or 2 for future schedule.
Caller	Presses 2
System	Enter the day, 1 for Sunday and 7 for Saturday.
Caller	Presses 1 (Sunday)
System	Enter the time.
Caller	Enters 12:00
System	The bus timings are: - Bus 75 at 12:15 pm, 12:50 pm and 1:35 pm - Bus 60 at 12:20 pm, 1:40 pm and 2:15 pm
Caller	Presses 9
System	If you wish to save this schedule please record an associated prompt for this

	function.
Caller	Says “Sunday after 12:00 pm for the bus stop number 26532”
System	Item saved.

Table 4.4 demonstrates the reduced amount of interaction necessary for accessing information that requires multiple layers of user input. Upon entering a new session users can directly access their saved information. The system will therefore replay the bus schedules for the previously saved date and time and allow the user to bypass four layers of prompts. By reducing the amount of criteria inputted into the system (see Table 4.3), possible user errors are avoided in requesting the needed information.

Table 4.4: Retrieving future timings from the list of book-mark options

System	Welcome to the local transit system. Press 1 for default options or 2 for personal options.
Caller	Presses 2
System	Press 1 for “Current schedule at stop number 60613, in English” Press 2 for “Sunday after 12:00 pm for the bus stop number 26532” ……
Caller	Presses 2
System	The schedule is: - Bus 75 at 12:15 pm, 12:50 pm and 1:35 pm - Bus 60 at 12:20 pm, 1:40 pm and 2:15 pm

The second application (cinema listings) is similar to the first application (local bus transit information), except that the structure of the hierarchy is more complex in the second application. The structure of the application, before saving any option in the

personalized list, is shown in Figure 4.7. Figure 4.8 shows the structure of the interface after saving some options in the personalized list.

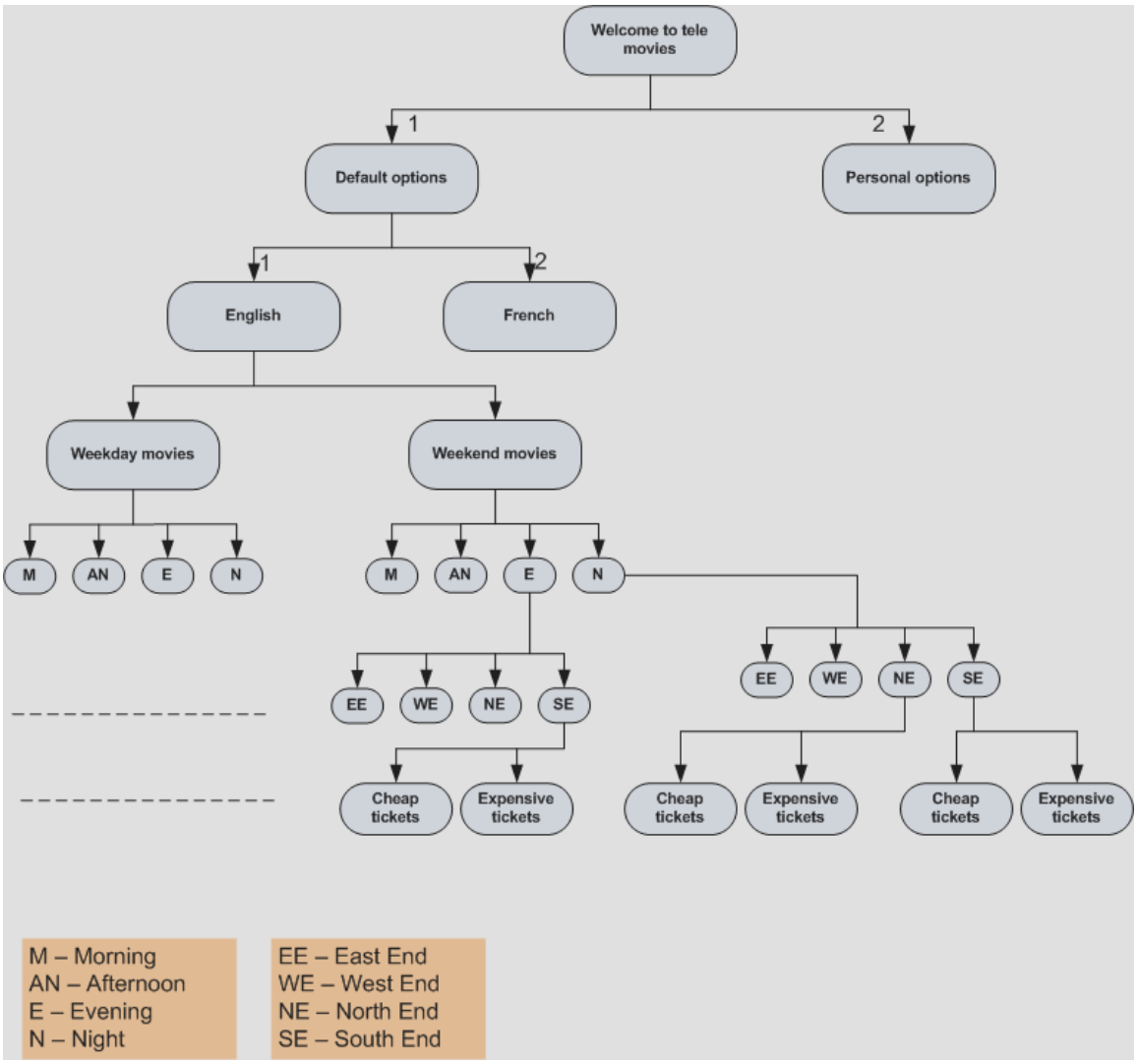


Figure 4.7: Structure of the second application - before saving any option in the personalized list

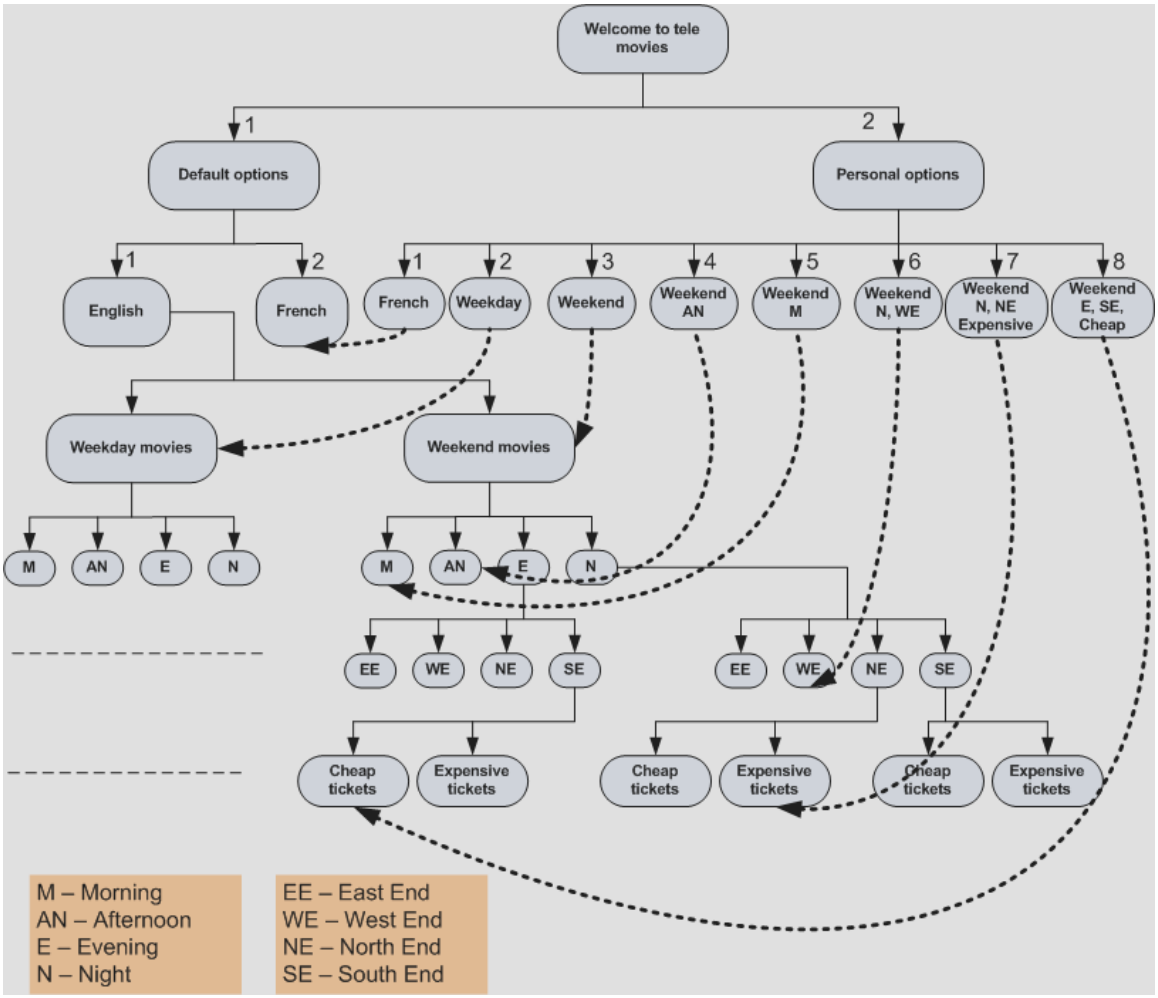


Figure 4.8: Structure of the interface - after saving some options in the personalized list

4.3 Evaluation

The main goal of this experiment is to determine whether personalization improves navigation in auditory interfaces. In order to evaluate the effect of personalization in auditory interfaces, personalized touch-tone interfaces were compared with conventional touch-tone interfaces. Two sets of applications were used for testing

the new interface. In the first application the users interacted with a simple hierarchy and in the second application the users managed a complex hierarchy.

4.3.1 Hypothesis

Based on the results of earlier studies on personalizing options/menus at the interface, the following effect was anticipated:

❶ If there are fewer options in the personalized list, subjects will perform the tasks faster with the personalized interface than with the traditional interface. Because, fewer options in the personalized list will help the users to listen to the prompts (menus) quickly and easily.

❷ If there are more options in the personalized list, subjects will perform the tasks faster with the traditional interface than with the personalized interface. Because, users are forced to listen to all the prompts before they obtain the required information. Therefore, the navigation time increases with the increase in the personal options.

4.3.2 Method

4.3.2.1 Design

A within group experiment was conducted to test the two conditions described earlier in this experiment: personalized touch-tone interface and conventional touch-tone interface. A fully balanced Latin-square design was used to reduce any learning effects. In this experiment, the subjects were randomly assigned to one of the above mentioned

two conditions. The experiment focused on measuring the time taken to obtain the requested information.

4.3.2.2 Materials

The experiment was tested on the voice genie platform to simulate the effect of using the telephone to access information.

4.3.2.3 Participants

10 students from a local university volunteered to participate in this study. None of the participants reported a history of auditory disorders. All the participants reported that they had heard about or used touch-tone interface systems at least once. The participants also stated that they had previous experience in listening to synthetic speech and spoke English without any evident accent.

4.3.2.4 Procedure

Participants took the test one at a time and were told that the intention of the evaluation was to test the system and not their abilities. Prior to the start of the study, the subjects were given a brief introduction about IVR systems. The subjects were informed of the main motivation behind building the personalized touchtone interface. The participants were then given a brief explanation of the tasks to be performed in the experiment and then asked to sign a consent form. Detailed information about

personalized touch-tone interface and a demo was also provided to all participants. The demo included examples of accessing information using conventional touch-tone interface and personalized touch-tone interface. Participants were shown how to bookmark a node and access the bookmarked node using the personalized touch-tone interface. When the users indicated that they were comfortable with accessing information using both the touch-tone interface systems, the real experiment was conducted. Users were given a booklet that contained the various options they had to bookmark. After saving all the options that were listed in the booklet, the users performed various sets of tasks using both interfaces.

The lists of tasks that were performed by the participants are shown in Table 4.5. After completing all the tasks, the users were asked to evaluate the conventional touch-tone interface and the personalized touch-tone interface systems in terms of speed, navigation, ease of use, and adaptability.

Table 4.5: Lists of tasks performed by the participants

Application 1 (Local transit system)	
Task number	***Tasks***
T1	Find the current bus timings for the bus stop number “60613”

T2	Find the bus timings for Sunday after 12:00 pm, for the bus stop number "60613"
T3	Find transit timings for Monday after 22:25 for the bus stop number "23422"

Application 1 (Local transit system)	
Task number	***Tasks***
T1	Find the list of English language expensive Hollywood movies that are played at nights on weekends in the north end
T2	Find the list of English language cheap Hollywood movies that are played in the evenings on weekends in south end
T2	Find the list of English language cheap Hollywood movies that are played at nights on weekends in the north end
T4	Locate English language, weekends, night, west end
T5	Locate English language, weekends, night, west end
T6	Locate English language, weekdays
T7	Locate English language, weekends, afternoon
T8	Locate French language

4.3.3 Results and Discussion

To measure the difference between the conventional touch-tone interface and personalized touch-tone interface, the time taken by each participant to complete the tasks in the given list was recorded. The task completion times were then tabulated and compared qualitatively. After the completion of the given tasks, the users were requested to complete a subjective questionnaire to analyze the degree of satisfaction while using the conventional touch-tone interface and personalized touch-tone interfaces.

4.2.5.1 Task Completion Time

The participants performed the tasks using the personalized touchtone interface and conventional touch-tone interface, in both the applications. In the first application, the subjects managed three bookmarks in the personalized list (Figure 4.6). The subjects' tasks involved locating the option (node) in the hierarchy, using both the interfaces. The time taken by the subjects to locate the nodes was recorded. The results of this experiment suggest that the subjects performed all the tasks comparatively better (faster) with the personalized interface than with the traditional interface. The average task completion times for the first application are summarized in Figure 4.9.

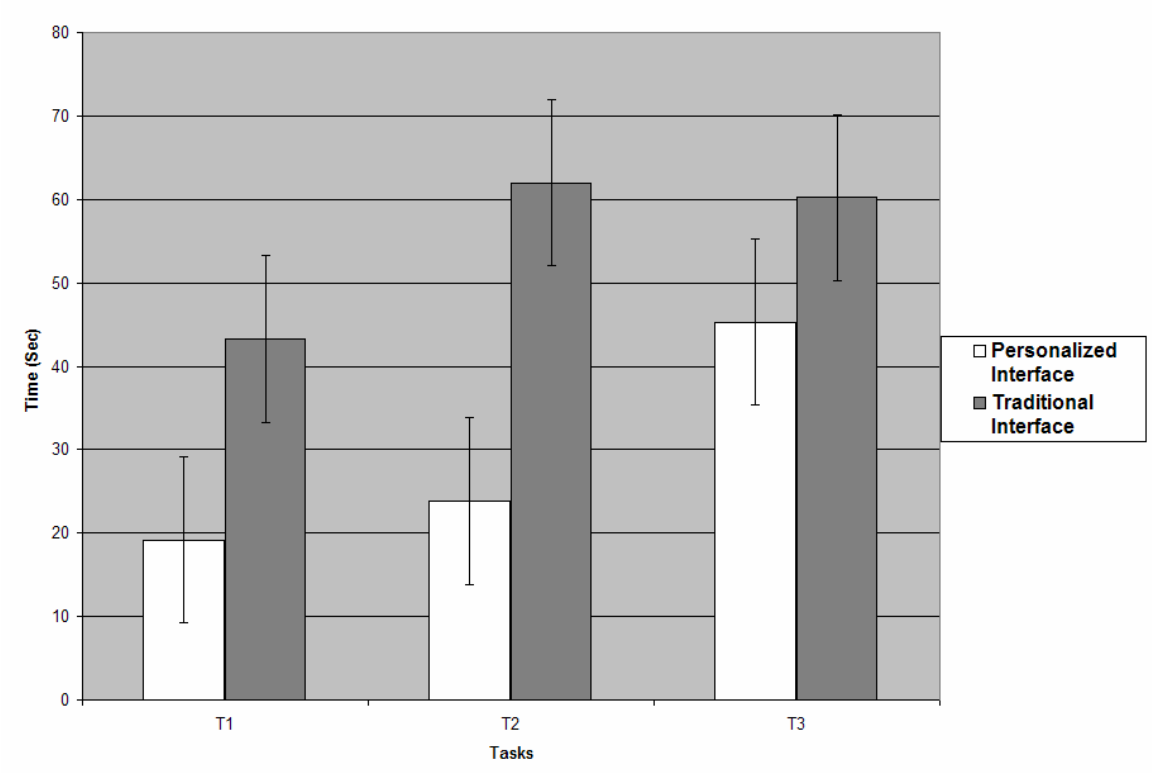


Figure 4.9: Average task completion time for Application-1(local bus transit system)

The overall task completion times for both the interfaces were averaged in order to find the average time taken by each subject to perform the tasks. The overall results show that the subjects performed the tasks 1.9 times faster with the personalized interface than with the traditional interface (Figure 4.10). A T-Test shows that there is a significant difference in performance between the personalized interface and the default interface, $T(9, 0.05)=-7.781$, degree of freedom=9, $p<.001$. The improvement in performance while using personalized touch-tone interface is attributed to the fewer number of options in the personalized menu list. Overall, the results confirm hypothesis-1 that the subjects will perform the tasks faster with the personalized interface than with the traditional interface.

In the second application the subjects managed to save eight options in the personalized list (Figure 4.8). The subjects' tasks involved locating the position of the node in the hierarchy. The time taken to complete each task was measured. The results of this study suggest that subjects performed the tasks (except T6 and T8) considerably faster with the default interface than with the personalized interface. The overall results were averaged for both the interfaces in order to find the average time taken by each subject to perform the given tasks. A T-Test shows that there is a significant difference in performance between the personalized interface and the default interface, $T(9, 0.05)=4.949$, degree of freedom=9, $p<.001$. Consistent with hypothesis-2, the overall results show that the subjects performed the tasks 1.15 (Figure 4.10) times faster with the traditional interface than with the personalized interface. Also, the reason for improved performance using the conventional touch-tone interface in the second application is the increase in navigation time, due to the increase in the number of options in the personalized menu list.

One of the major problems with the personalized list is that the options are saved sequentially. Hence, if the user wants to select the sixth option in the personalized list, then the user has to listen to the previous five options (atleast) before selecting the sixth option, unless the user knows beforehand that the option is saved in the sixth place. For example, in this application (Cinema listings), in order to perform T2, the subjects have to listen to all the eight options in the personalized list before selecting the eighth option (see Figure 4.12). However, in a traditional interface, the subjects have to drill through only five levels in order to locate T2 (see Figure 4.12). This can be improved by requesting shorter prompts to be recorded and by allowing users to skip between prompts.

Another factor that facilitated the navigation in traditional interfaces was training. Since the users had some previous training with the default interface, they selected the options/menus more quickly with that interface. For example, in traditional interfaces, when the users reach the fourth level, the system prompts "Press 1 for weekdays, Press 2 for weekends". As soon as the users hear "Press 1 for weekdays", they interrupted the system by "Pressing 2" in the keypad, because they know that they have to "Press 2 for weekends". Thereby, a significant amount of navigation time was saved in traditional interfaces.

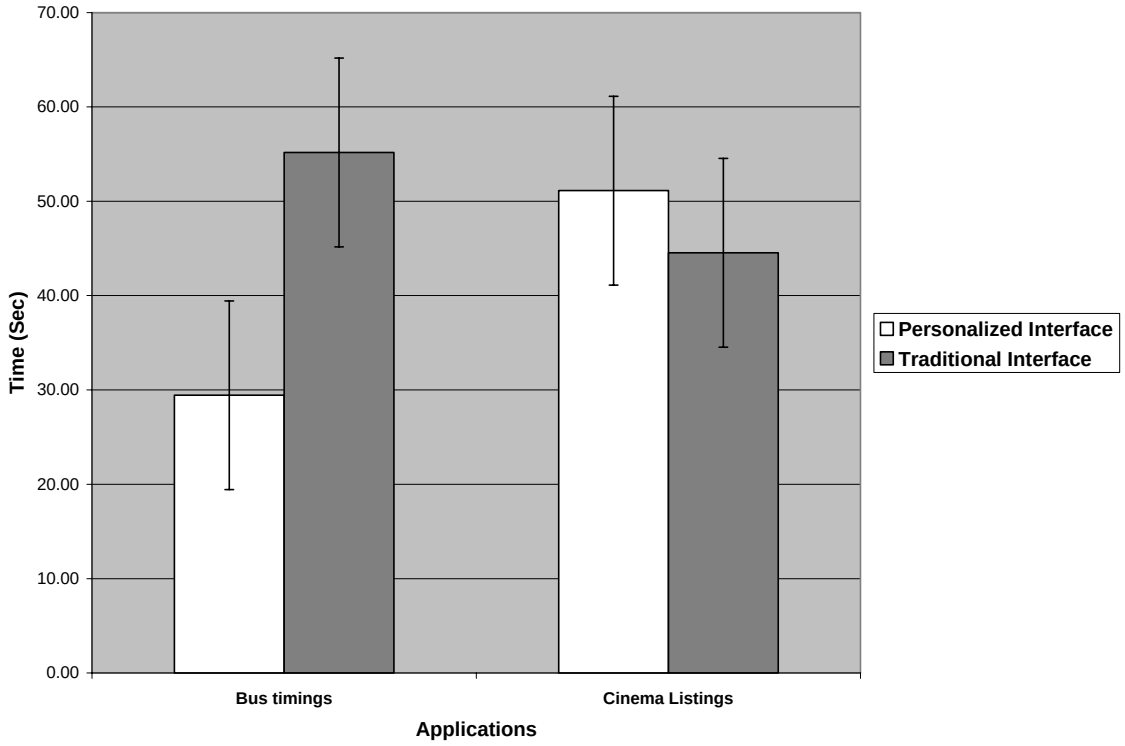


Figure 4.10: Average task completion times for both the applications

It is inferred from the experiment that as the options in the personalized menu increase, the navigation times also increase and hence the performance reduces. However, it was also noted that the performance of the participants while using the personalized touch-tone interface improved when the required options were placed closer to the beginning of the menu. Hence, ordering of the bookmarks in the menu plays a very important role in the performance of the personalized touch-tone interface.

4.2.5.2 Keypad Selections

The average number of keypad selections that are required to complete the tasks is shown in Figure 4.11.

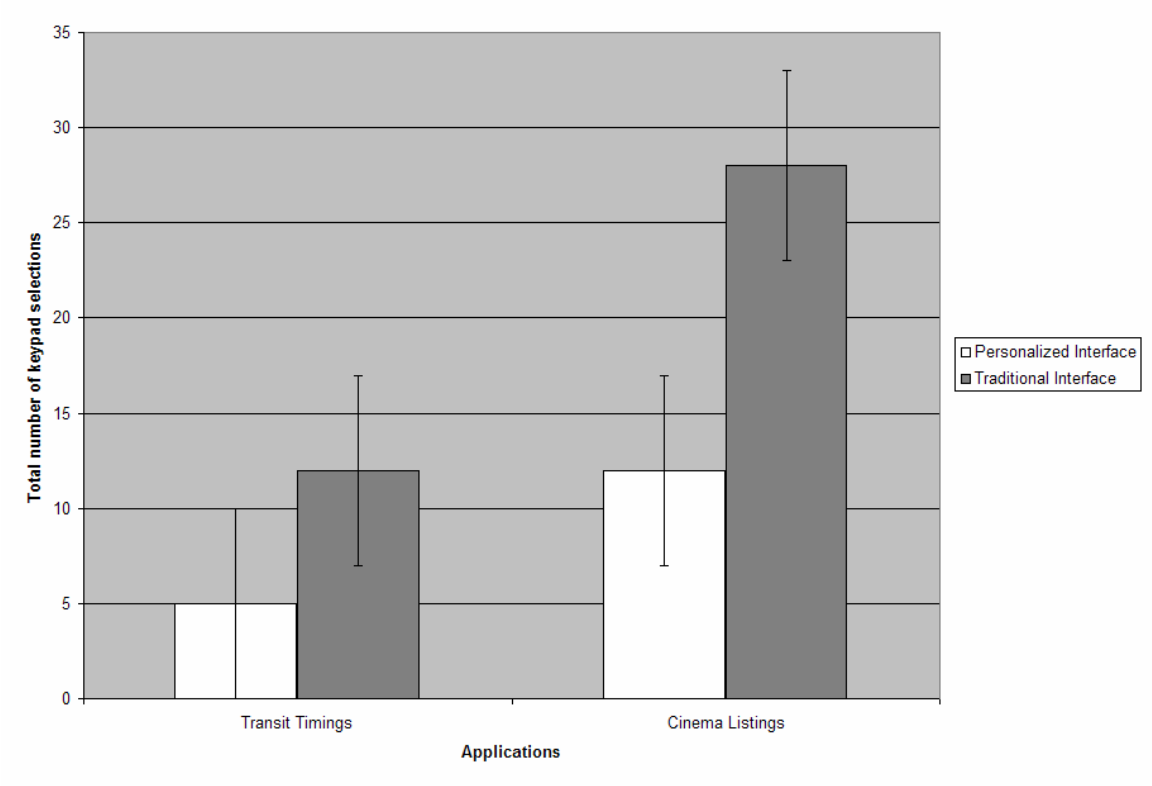


Figure 4.11: Average number of keypad selections required to complete the tasks

The results of this study suggest that the number of keypad selections is reduced significantly, with the personalized interface than with the default interface, for both the applications. This is because the default interface forces the users to select an option at each level in the hierarchy, before the user moves from one level to another. For example, in order to reach Q2 (Figure 4.12), the users have to press the key in the keypad, at least 5 times (1 selection for each level). However, the same destination can be reached using the personalized interface by just a single click on the keypad.

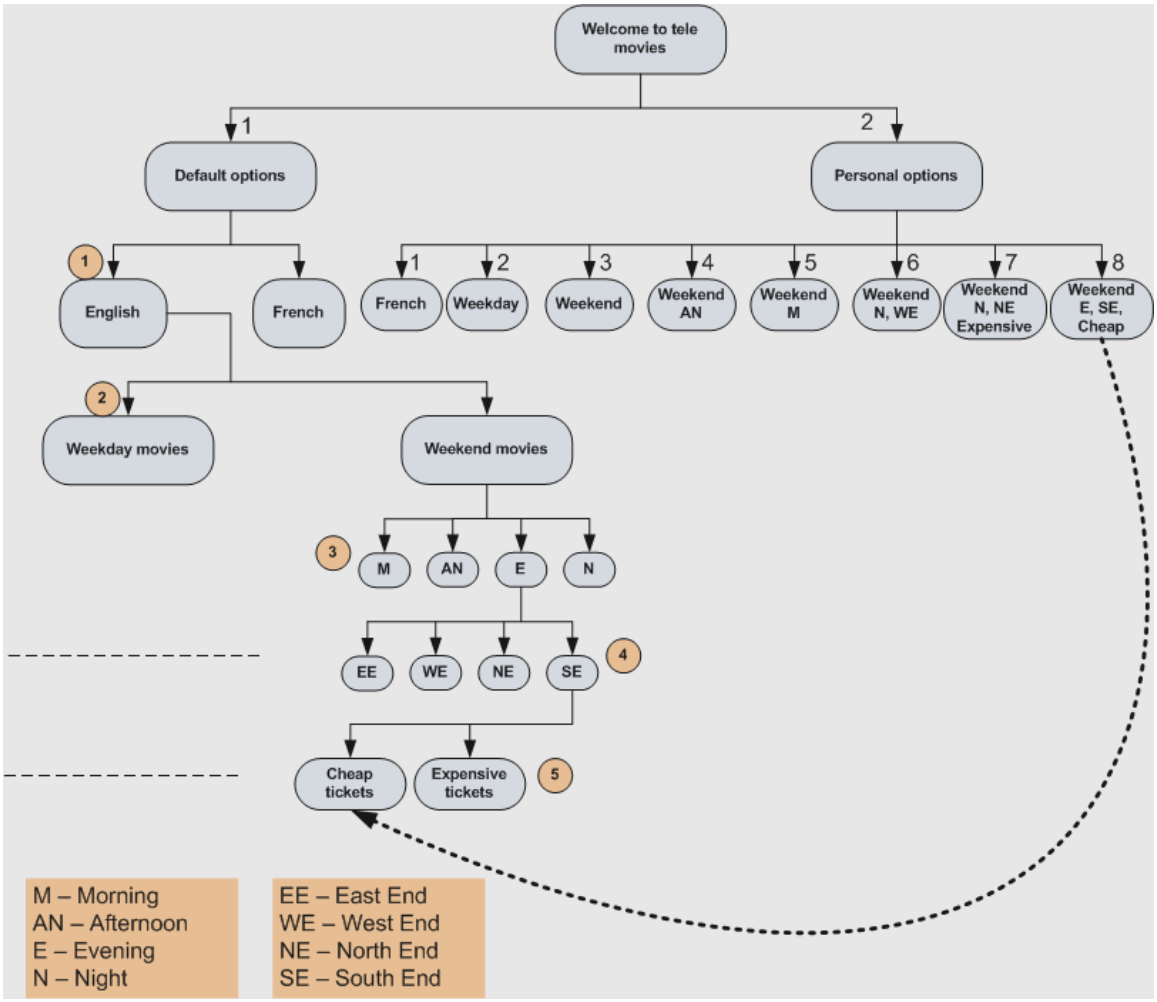


Figure 4.12: Number of keypad selections required to locate a node in the hierarchy

Results show that the participants are able to select the correct option in the personalized menu list and move to the appropriate level in the hierarchy in both the experiments. Though some of the participants found the system confusing at first, all of them were able to eventually understand the system and reach the required targets.

4.2.5.3 User's Experience Measure

After the completion of both the experiments, the participants were asked to fill-out two post-experiment questionnaires. The first post-experiment questionnaire compares personalized touch-tone interface and conventional touch-tone interface in terms of speed, navigation, ease of use and adaptability. The first post-experiment questionnaire contained questions comparing the performance of personalized touch-tone interface and conventional touch-tone interface for the same set of tasks. For each task, the participants chose the interface that they felt was better suited for that task. The analysis of the first questionnaire has been categorized below:

- ***Efficiency:***

  Eighty percent of the users favored personalized touch-tone interface for speed. The general opinion was that they found targeted information quicker when they used personalized touchtone interface as they had bookmarked the information earlier. All the users also confirmed that they would like to use personalized touch-tone interface for accessing information in their daily transactions.

❶ ***Navigation***

Sixty percent of the users favored personalized touch-tone interface to choose the right menu in order to reach the next level efficiently and effectively (faster and better).

❸ ***Adaptability***

Fifty percent of the users said that personalized touch-tone interface is easy to learn. All the users said that language is not a major barrier in accessing information using personalized touchtone interface, since the information can be stored in the local language of the users. Seventy percent of the participants said that they prefer personalized touch-tone interface, because they know exactly which option they have to choose in order to reach the targeted information.

❹ ***Ease of Use***

Ninety percent of the users said that they liked using the personalized touch-tone interface system for the tenth attempt as they understood the system well and were comfortable using and navigating through the options. Eighty percent of the users said that they would like to use personalized touch-tone interface as they could skip levels and options that slow down their navigation speed. All the users also highly recommended personalized touchtone interface for flexibility and ease of use.

The second post-experiment questionnaire evaluated the attitudinal response of the users of the personalized touch-tone interface. Participants' suggestions were collected at the end of the questionnaire. All the questions, except suggestions, were scaled between 1 and 5 ("1"="strongly disagree", "5"="strongly agree"). The results are shown in the Table 4.6.

Table 4.6: Summary of the second post experiment questionnaire

Statement	**Mean** (1-5)
Q1: Difficult to locate menu options using the personalized touch-tone interface.	2.4
Q2: I am able to complete my tasks efficiently using personalized touch-tone interface.	4
Q3: The speed of personalized voice interface is good for accessing information that I regularly like to query the system.	4.5
Q4: I am satisfied with the speed of accessing information using personalized voice interface.	3.5
Q5: I like the technique of recording my voice for providing information about personalized menus in the "Personalized Menus list".	4.1
Q6: Recorded menus in the "Personalized Menus List" improve navigation.	4.1
Q7: Replay of recorded human voice is pleasant.	4.1
Q8: Personalized touch-tone interface is easy to learn.	3.5
Q9: I know how to use the system after the trial session.	3.3
Q10: Personalized touch-tone interface is easy to use.	4.5

Q11: I can remember the list of menus that I have saved in the "Personalized Menus List".	2.5
Q12: Recording option is a necessary element of the interface.	4.3
Q13: Personalized touch-tone interface is the preferred choice for daily transactions.	4.4
Q14: Overall satisfaction of personalized touch-tone interface.	4.1

From the second post-experiment questionnaire, it is inferred that most of the users supported the idea of personalizing the menus at the interfaces (Q12). The questionnaire also reveals that most of the users preferred to use the personalized interface for daily transactions (Q3, Q13) because it is easy to learn (Q8), intuitive and easy to use (Q10). Users also favored the idea of recording human voice for saving an option in the personalized list (Q5, Q12). They have also suggested that listening to their own voice is pleasant (Q7). However, one of the major problems with the personalized list is that the users found it very difficult to remember the options that they have saved in the personalized list (Q1 and Q11). One way to solve the above problem is by allowing the users to scan and skip the uninterested menus in the list. This way a significant amount of users' navigation time can be saved. Overall, the users supported the idea of personalizing menus in the interface to improve navigation in auditory interfaces.

4.2.6 Conclusion and Future work

In this chapter, a method for personalizing menus in touch-tone interfaces was described. The method consists of creating bookmarks for nodes in the tree of menu options that the user can then access immediately. Results of the evaluation indicate that navigation time is reduced significantly when subjects use the personalized touch-tone interface for accessing information. Users mentioned that they were pleased with the effectiveness and efficiency of the personalized touch-tone interface and that the feature would be most useful in transactions performed on a regular basis.

Navigation time in the personalized touch-tone interface can be further decreased by allowing the user the choice of ordering the bookmarks in the personalized menu list. As personal options increase, users have to wait until the entire sequence of prompts gets replayed in order to obtain the information they require. Future work will consist of allowing users to perform maintenance operations on their list of options. For example, users may want to re-record a particular bookmarked option, change the order of the bookmarks in the entire list, and be given the flexibility of skipping non-interested bookmarks in the middle of their operations. Overall, current advances in technology will permit the implementation of personalized menus in touch-tone interfaces. The results of our preliminary investigation suggest that users would be willing to adopt the technology if it were made available.

Chapter 5 : Navigation Cues - Using Multiple Synthetic Voices to Improve Navigation in Hierarchical Structures

The issue of navigation in auditory interfaces has been addressed using two different ways: reducing the navigation time and reducing the error rates. However, both navigation time and the error rates are very closely interconnected. In other words, if a significant amount of error rates can be reduced, then a considerable amount of navigation time can also be reduced. One of the main reasons for the increase in the error rates, in auditory interfaces, is that the underlying navigation structure is not explicit or visible to the user. For example, in many applications users enter the system from the root node and branch into various dialogs (nodes and levels) in order to extract the necessary information. Since, voice-based applications do not explicitly depict the arrangement of nodes (i.e. the parent-child relationship between dialogs), users can lose track of their position in the navigation hierarchy.

In some respects, navigation in auditory interfaces can be generalized as the task that requires locating and branching to the appropriate path that leads to the object of interest. Navigation can be facilitated by providing cues about the user's location in the navigation structure. In this chapter, a method for assisting a user in identifying the location of nodes in the underlying navigation structure of an auditory interface is presented. This chapter focuses on analyzing the richness of multiple synthetic voices for representing hierarchies. Multiple synthetic voices were chosen to provide navigation cues in hierarchies, because studies [NL01, NMF+95] suggest that social, psychological, and

emotional impressions can be created by manipulating synthetic voice parameters such as pitch, pause, volume, and speech rate. Other studies by Furui [Fur86], Johnson et al. [JHH84], Sambur [Sam75], and Stylianou et al. [SCEM98] also suggest that voice characteristics, such as speech rate, pitch contour, and duration of pauses, have a great influence on speaker identification. These studies suggest that different synthetic voices can be created by manipulating synthetic voice parameters. However, the ability to manipulate synthetic speech parameters, to produce voices that are related in a hierarchical manner, has not yet been investigated.

This work is primarily inspired from the literature on earcons and their application to auditory interfaces. The issue of navigation in auditory interfaces has been addressed primarily through the implementation and evaluation of non-speech audio, called "earcons". Although, earcons are effective in conveying hierarchical information, they have several limitations (please see Chapter 3.3.1 and Chapter 3.3.2 for more details). In general, studies by Brewster [Bre98], Brewster et al. [BCH98], and Vargas and Anderson [VA03] suggest that speech-based interfaces should, in addition to providing information to the users, also assist with the task of navigation. This can be achieved by providing additional cues in these interfaces. Hence, the idea of representing hierarchies using multiple synthetic voices was proposed. The motivation behind this approach is in the belief that if synthetic speech can by itself provide both information and navigational cues to users, then a significant amount of navigational time may be reduced.

In this chapter, the results of two experiments to analyze the effect of multiple synthetic voices in representing general hierarchies are described. The first experiment in this chapter describes an initial study that was undertaken to determine whether multiple synthetic voices improve navigation in small hierarchies. The second experiment illustrates the effect of multiple synthetic voices in complex hierarchies, such as those found in real-world applications. The second experiment also investigates the effect of training on recall rates. Effect of training is an important factor that has to be considered if we want to employ multiple synthetic voices in real-world interfaces, such as telephone banking where training is normally not provided to users.

5.1 Representing Small Hierarchies using Multiple Synthetic Voices

The main goal of this experiment was to determine whether hierarchical relationships can be created from multiple synthetic voices. In this experiment, the effect of synthetic voices was evaluated on small hierarchies (9 nodes). This evaluation is based on the methodology similar to Brewster et al. [BWE92, BWE93] in which they initially evaluated the effectiveness of earcons on hierarchies with a small set of nodes (9 nodes).

The basic structure of a tree that was used in this experiment is shown in Figure 5.1. A different synthetic voice was constructed (with the rules described in section 5.1.3) for each node in the hierarchy. Similarly, each node in the hierarchy was assigned a unique phrase. These phrases did not include any information or hint about the parent-child relationship. All the phrases also contained the same number of syllables and were not

repeated in any of the hierarchies. Some of the sample text phrases were “the big dog slept on the floor last night”, “this place is too cold to go out”. Words in the selected phrases did not contain related words.

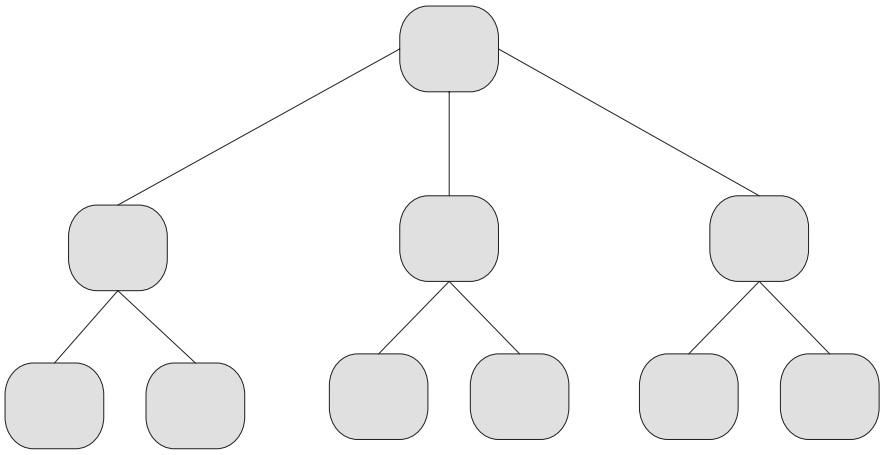


Figure 5.1: Structure of the hierarchy used for analyzing the effect of single synthetic speech versus multiple synthetic voices

5.1.1 Construction of Multiple Synthetic Voices

Multiple synthetic voices were created using DECtalk version 4.61 (Fonix Corp., www.fonix.com). DECtalk facilitates the manipulation of the various parameters of synthetic voices to create new voices. DECtalk is a formant-based speech engine, which uses linguistic rules for converting text to speech. In this study, a formant-based speech engine was used instead of a concatenative speech engine, such as AT&T Next-Gen TTS system, because concatenative speech engines do not facilitate easy manipulation of some of the speech parameters, such as breathiness, average pitch, and pitch range.

As a starting point the same guidelines proposed in [BSG89, BWE93] for the creation of hierarchical earcons was applied to create multiple synthetic voices. Sumikawa [Sum85] proposed three guidelines (Repetition, Variation, and Contrast) for creating earcons, to represent them in a hierarchical manner. In this experiment, the rules proposed by Sumikawa was adopted and modified to devise the following guidelines for creating various structured combinations of synthetic voices. The guidelines are as follows (MSSG stands for Multiple Synthetic Speech Guideline):

- **MSSG1 - Duplication:** Duplicate all the synthetic voice parameters and their values from the parent node. For instance, if a parent node is created with two parameters, speech rate (with value = 110 words-per-minute) and average pitch (with value = 120 Hz), and if the voice for the child node is created with the same parameter values that are used in the parent node, then it is referred to as duplication.
- **MSSG2 - Variation:** Alter the values of one or more synthetic speech parameters between two related nodes. For example, if a parent node is assigned a speech rate of 110 words-per-minute, then its child can be assigned a speech rate of 150 words-per-minute.

5.1.2 Parameters Used for Creating Multiple Synthetic Voices

The various speech parameters that were selected to create multiple synthetic voices are:

- **Average Pitch (AP):** defines the variation in the pitch contour for a given synthetic voice. For example, increasing the average pitch may result in agitation,

where as reducing the pitch will result in calmness of the speaker. Average pitch is measured in Hertz (Hz).

- **Pitch Range (PR):** is used to expand or shrink the swings in pitch. Different emotions in a voice can be perceived by varying the pitch range. For example, increasing the pitch range will increase the level of dynamism projected by the voice. In turn this could lead to a perception of happiness in the voice. Reducing the pitch range will project the image of sadness in the speaker. Pitch range is represented in %.
- **Speech Rate (SR):** is defined as the number of words that a system can speak in one minute. It is measured in terms of words-per-minute (wpm).

The main reason for choosing the above mentioned voice parameters to create hierarchical relationships, is as follows:

- Various studies [Cah89, NMF+95] suggest that average pitch, pitch range, and speech rate are the three main voice parameters that have a significant impact on the users' perception of the voice personalities and emotions in speech interfaces.

The values for the above mentioned voice parameters were chosen to achieve the strongest possible contrast between two different voices. The values to these parameters were assigned in a structured and systematic manner. The values were also carefully chosen such that they would not degrade users' comprehension of the text being vocalized (comprehension range). Comprehension of the text was tested during several pilot runs. In the case of speech rate, the upper limit was set to 250 wpm. This is based on

results by Slowiaczek et al. [SN85], which suggested that a users' comprehension of synthetic speech decreases for speech rates greater than 250 wpm. Two extreme nodes in a sub-tree (i.e. two siblings at the extremity of a sub-tree) were assigned the lowest and the highest value of the parameters chosen to represent that sub-tree. For example, if the speech rate for the left-most node was assigned 90 wpm, then the speech rate for the right-most node was assigned 210 wpm. The nodes in between the two extremity nodes were assigned values ranging between these two extreme values.

5.1.3 Conditions Tested

In order to test the validity of the hypothesis, three conditions are created: Single Synthetic Voice (SSV), Multiple Synthetic Voice-1 (MSV-1) and Multiple Synthetic Voice-2 (MSV-2). The following rules are used for creating the hierarchies:

Single Synthetic Voice (SSV):

- All the nodes have the same pitch, pitch range and speech rate (AP=110, PR=135%, SR=170 wpm). The only perceivable difference between nodes was the text labels.

Multiple Synthetic Voice-1 (MSV-1):

The rules that were used to create the MSV-1 hierarchy are as follows (see Figure 5.2):

- The root node is assigned a "neutral" synthetic voice. The following values are used to represent this node: AP=306 Hz, PR=210% and SR=160 wpm.

- The nodes at the second level were all created with the same pitch. The pitch used in this level is AP=110 Hz and the pitch range PR=135%. Different values of speech rate are assigned to each node in the second level to create a sufficient contrast between them. For example, the left node was created with a low speech rate (SR=90 wpm), the middle node with a medium speech rate (SR=150 wpm), and the right node with a high speech rate (SR=210 wpm).

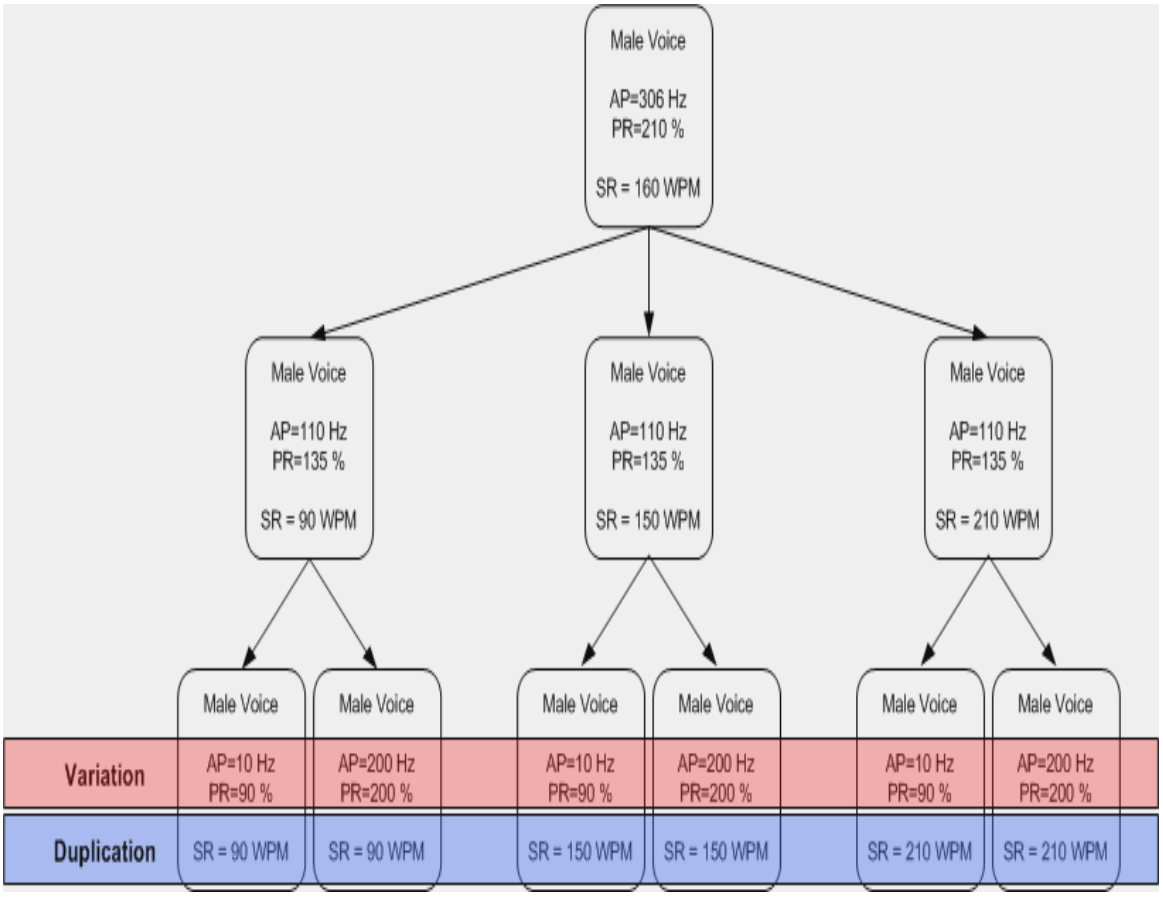


Figure 5.2: Synthetic voice parameters and the values that are used to create the hierarchy using MSV-1

- The nodes at the third level were created with the same speech rate as that of their parent node. In this case the speech rate was inherited, i.e. children nodes of the

left most sub-tree in the hierarchy inherit a SR=90 wpm. The contrasting parameter at the leaf nodes was the pitch. The left child had a low pitch (AP=10 Hz, PR=90%), and the right child a high pitch (AP=200 Hz, PR=200%).

Multiple Synthetic Voice-2 (MSV-2):

The rules that were used for creating this hierarchy were the same as those used in creating the previous hierarchy (MSV-1) with the exception that speech rate is varied across levels and pitch is varied within a level (see Figure 5.3).

- Similar to MSV-1, the root node was assigned a "neutral" synthetic voice. We used the following values to represent this node: AP=306 Hz, PR=210% and SR=160 wpm.
- The nodes in the second level were all created using the same speech rate. The speech rate selected for this level was SR=160 wpm. The differentiating parameter for the nodes in the second level was pitch. For example, the left node had a low pitch (AP=10 Hz, PR=90%), the middle node had medium pitch (AP=110 Hz and PR=135%), and the right node had high pitch (AP=200 Hz, PR=200%).
- The nodes in the third level were created by using the same pitch as in their parent node (i.e. they inherited pitch). The leaf nodes were differentiated by changing the speech rate. For example, the left child of the sub-tree had a low speech rate (SR=90 wpm), and the right child of the same sub-tree had a high speech rate (SR=210 wpm).

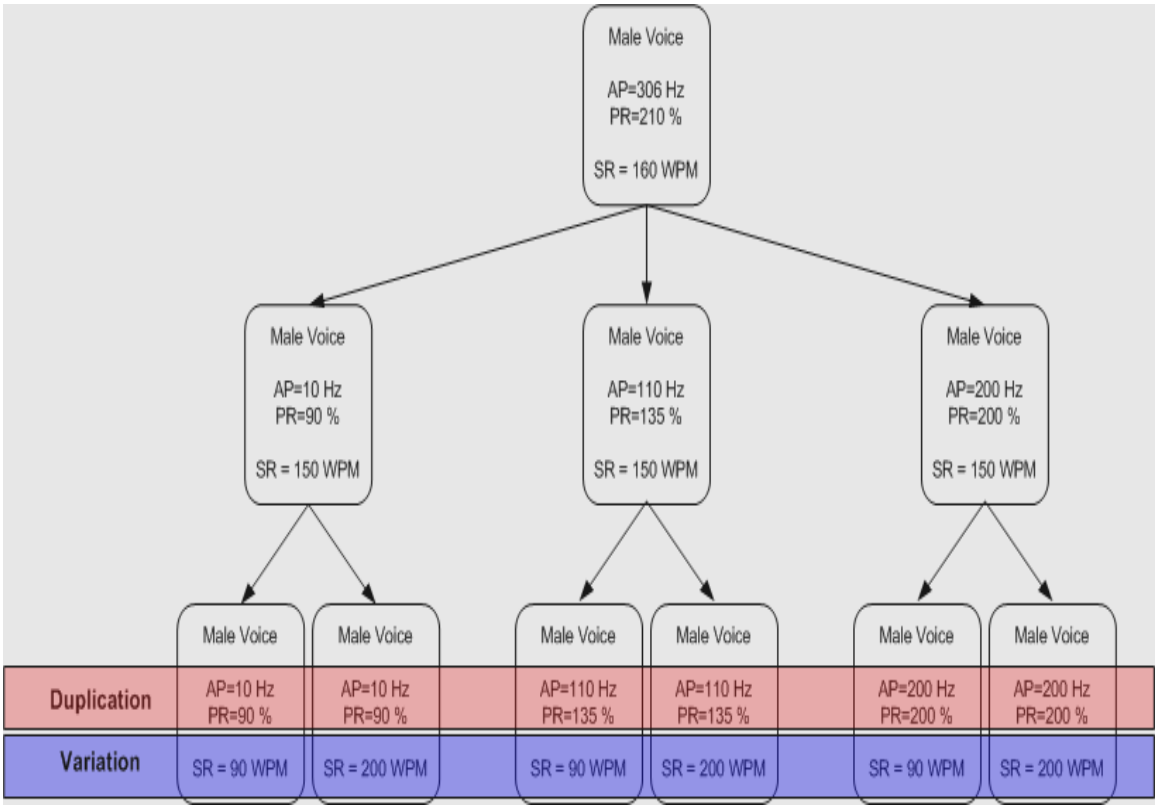


Figure 5.3: Synthetic voice parameters and the values that are used to create the hierarchy using MSV-2

5.1.4 Hypothesis

Based on the results of earlier studies on the use of non-speech audio for representing hierarchies and on the perception of synthetic speech, the following effect was anticipated:

❶ Multiple synthetic voices will result in higher performance, for tasks requiring identification of node positions in a hierarchy than single synthetic voice.

5.1.5 Method

5.1.5.1 Design

A within group experiment was conducted to test the three conditions described earlier in this experiment: Single Synthetic Voice (SSV), Multiple Synthetic Voice-1 (MSV-1), and Multiple Synthetic Voice-2 (MSV-2). A fully balanced Latin-square design was used to reduce any learning effects. In this experiment, the subjects were randomly assigned to one of the above mentioned three conditions. The experiment focused on measuring the participants' accuracy of locating a node in the hierarchy.

5.1.5.2 Materials

In this experiment, the hierarchies were presented using PowerPoint files, with appropriate .wav files (sample rate = 11.025 KHz, 16-bit Mono) on the various nodes. The slides were shown using a Dell Inspiron 8600 laptop with a 15.4" display using Intel® Integrated laptop audio speakers. Lower quality voice resolution was used as it simulates the output in several different types of environments, such as telephony interfaces.

5.1.5.3 Participants

12 undergraduate students from a local university volunteered for this experiment. None of the participants reported a history of auditory disorder or exhibited any hearing problems. All the subjects also stated having previous experience in listening to synthetic

speech. In order to avoid any potential language difficulties, only native English speakers were recruited.

5.1.5.4 Procedure

Before running the experiment, the participants were presented with a write-up, which described the rules that were used for creating the nodes in the hierarchies (presentation session). The participants were then given the hierarchy (Figure 5.1) and were asked to click on the nodes and listen to the associated phrases. For each participant, the order in which the nodes were clicked was randomized. Participants were also allowed to click on any particular node to re-listen to the voice at that node (up to a maximum of three clicks were allowed). Overall, the participants received very limited help from the experimenter. The experiment began after asking the subjects whether they felt comfortable about their understanding of the rules that were used for creating the hierarchies.

During the experiment, six voices were randomly selected from the set of 10 voices that were previously presented to the subjects. From the list of six voices, 2 were from level-2 and 4 were from level-3. The participants were then asked to locate the node that was played, by writing down on a given sheet (the sheet contained a figure similar to Figure 5.1) its position in the hierarchy. The accuracy in identifying the correct position of the node in the hierarchy was computed.

5.1.6 Results and Discussion

The results of the above experiment are summarized in Table 5.1, which reports error rates based on the type of mapping (i.e. type of voices used). The results were obtained by averaging all subjects' scores. A Kruskall-Wallis test (non-parametric ANOVA) conducted on the results statistically shows that the subjects performed differently with all three types of hierarchies ($p < 0.0001$). The mean error rate shows that the subjects were four times more accurate with MSV-1 than with SSV and 2.8 times more accurate with MSV-2 than with SSV. A Mann-Whitney test statistically shows that the subjects performed significantly better with both multiple synthetic voice conditions ($p < 0.0001$ and $p < 0.005$ respectively for MSV-1 and MSV-2) than with the single synthetic voice. However, the results do not show that there is a statistically significant difference between MSV-1 and MSV-2 (Mann-Whitney test, $p = 0.277$).

Figure 5.4 shows the error rates per voice type per level in the hierarchy. From the entire set of six nodes that the users were required to locate in each hierarchy, two nodes were on level-2 (level-1 is the root) and four nodes were on level-3. The observation shows that the users' accuracy with the single synthetic voice condition degraded with increase in the depth of the hierarchy. A Kruskall-Wallis test on level-2 of the hierarchy statistically shows that there is no significant difference in performance across all three conditions ($p=0.297$). Pair wise Mann-Withney tests show that there is no significant difference in performance between SSV and MSV-1 ($p=0.219$), SSV and MSV-2 ($p=0.932$), and MSV-1 and MSV-2 ($p=0.266$), on level-2.

Table 5.1: Average error rates of finding the appropriate node in the hierarchy

	SSV	***MSV-1***	***MSV-2***
Error Rate	62.5%	15.3%	22.2%

On level-3 of the hierarchy a Kruskall-Wallis test shows that there is a significant difference in performance across all the three conditions ($p<0.0001$). Pair wise Mann-Whitney test shows that there is a significant difference between SSV and MSV-1 ($p < 0.0001$) and SSV and MSV-2 ($p < 0.0001$). However there is not any statistical difference between MSV-1 and MSV-2 ($p=0.799$). Inspection of the mean error rate indicates that the subjects performed better with MSV-1 and MSV-2 than with SSV, on level-3.

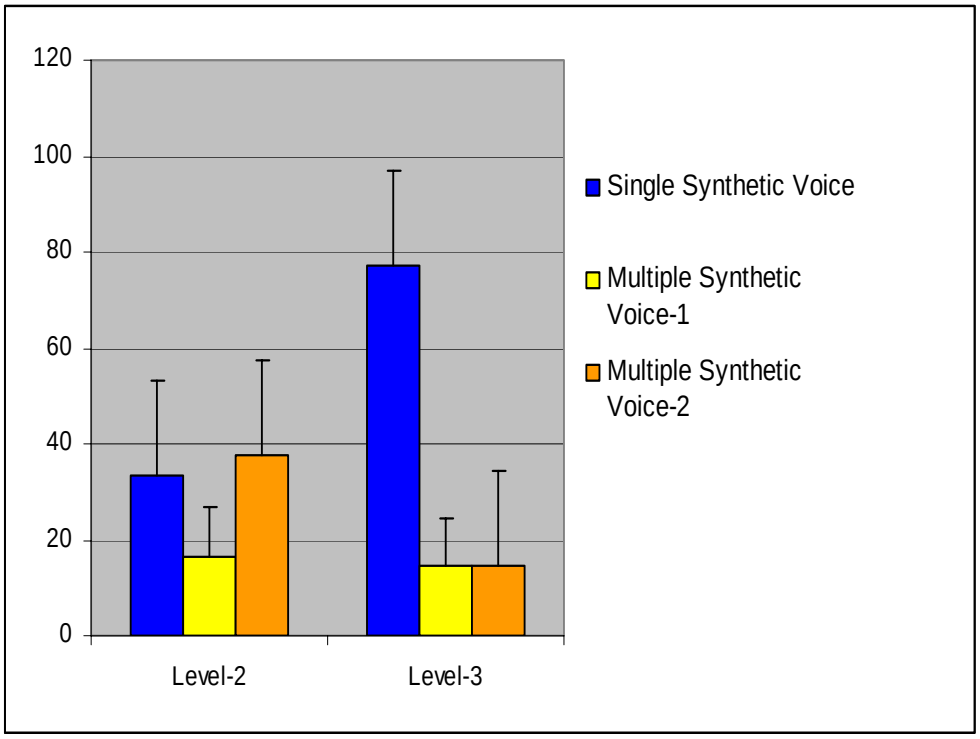


Figure 5.4: Average error rates of finding the appropriate node in the hierarchy

Overall, the results confirm the hypothesis that users are capable of locating the position of nodes in hierarchies created with multiple synthetic voices. The results also confirm that altering the mapping of the multiple synthetic voices on the hierarchy (i.e. for creating MSV-1 and MSV-2) does not make any significant difference. However, by looking at the mean error rates, the results suggest that the subjects performed slightly better with MSV-1 than with MSV-2. This may suggest that pitch is a better dimension than speech rate for identifying depth in the hierarchy, and identifying nodes in a level may be better achieved by varying the speech rate. Further experimentation is required to determine which mappings are best suited for representing deep and wide hierarchies.

5.2 Representing Complex Hierarchies using Multiple Synthetic Voices

The results of the previous experiment suggested that multiple synthetic voices could be used to improve navigation in small hierarchies. However, the experiment did not analyze the effect of multiple synthetic voices in complex hierarchies, such as those found in file systems and TBIs. The purpose of the second experiment was to analyze the effect of multiple synthetic voices in representing complex hierarchies. In this experiment, a hierarchy of 27 nodes with 4 levels was created. Figure 5.5 shows the hierarchy that was used in the experiment. This hierarchy was a diagrammatic representation of the files and folders in a computer system.

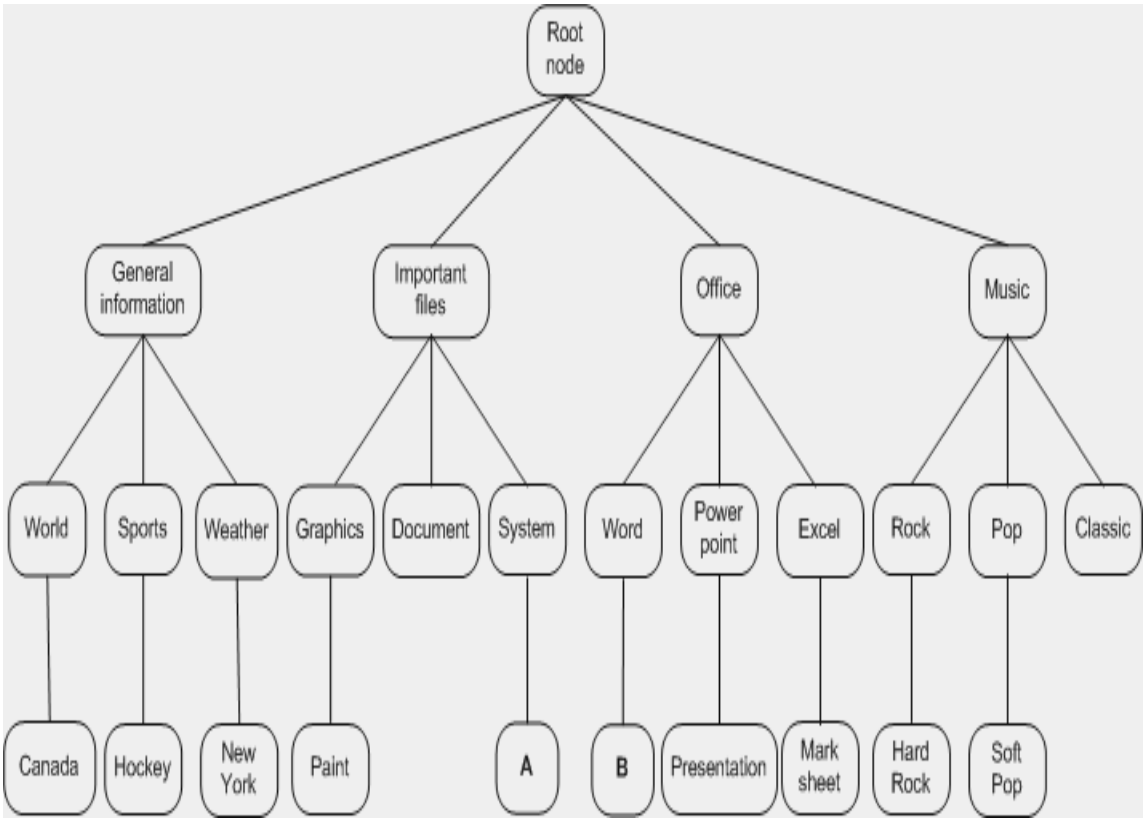


Figure 5.5: Hierarchy used for testing the effect of multiple synthetic voices on a larger hierarchy

5.2.1 Manipulation

In this experiment, the two rules MSSG1 and MSSG2 that were used in the previous experiment was incorporated. However, to construct related voices in a complex hierarchy an additional rule was devised:

- ***MSSG3 - Inclusion:*** Include one or more speech parameters to the preceding voice in order to create unique voices. For example, if the child node duplicates all the parameters and the exact values from its parent node (such as speech rate and pitch), then other voice parameters such as laryngealization and breathiness

can be added (inclusion rule) to the child node, to make the child node sound different from its parent node.

5.2.2 Parameters Used for Creating the Multiple Synthetic Voices

For this experiment more voice parameters were added to create the synthetic voices. In addition to using average pitch, speech rate and pitch range, the voices were created using laryngealization, breathiness and volume gain to create voices for representing hierarchies.

- **Laryngealization (LA):** At the beginning and end of sentences, many speakers turn their voice on and off irregularly. This gives a querulous tone to the voice. This departure from perfect periodicity is called creaky voice quality, which is often referred to as laryngealization. The LA option specifies the amount of laryngealization, in the voice. The value LA=0 specifies (no laryngealization) and LA=100 specifies (maximum laryngealization).
- **Breathiness (BR):** Some voices can be described as breathy. The vocal folds vibrate to generate a breathy noise along with the voice. BR option ranges from 0 dB (no breathiness) to 70 dB (strong breathiness).
- **Gain in Volume (GV)**: GV option gives information on the intensity (volume) of the voice. GV ranges from 0 dB (no volume) to 70 dB (full volume)

Laryngealization, breathiness and gain in volume are often considered as important parameters that contribute to measure the quality of the voice [Cah98] and hence form an integral part of the rules.

5.2.3 Rules for Creating the Hierarchy

The rules that were used to create the hierarchy are described below (Figure 5.6).

- The root node was assigned a "neutral" synthetic voice. The various speech parameter values that were used to represent this node are: AP=306 Hz, PR=210% and SR=160 wpm.
- All the nodes in the second level were created with the same speech rate. The speech rate selected for this level was SR=220 wpm. Four families (sub-trees) of nodes (Family-1, Family-2, Family-3 and Family-4) were created in this level. These families were differentiated from each other by varying the pitch. The pitch parameters used for each family are described as follows :
  - For Family-1: male voice with normal pitch (AP=110 Hz, PR=135%).
  - For Family-2: male voice with low pitch (AP=10 Hz, PR=95%).
  - For Family-3: female voice with normal pitch (AP=200 Hz, PR=175%).
  - For Family-4: female voice with low pitch (AP=100 Hz, PR=135%).
- Nodes belonging to the third level were created by inheriting the same pitch (AP and PR) from their respective parent nodes. The distinctive feature between the second level nodes and the third level nodes was the difference in the speech rate, i.e. for the second level nodes SR=220 wpm was used and for the third level nodes SR=160 wpm was used. At this level, the nodes that belonged to the same family were differentiated using speech rate (SR), Laryngealization (LA) and Breathiness (BR). For all the nodes in the third level, all the parameters and their values, except the speech rate, were inherited from the parent node. The speech rate was set at SR=160 wpm. In addition to these features,

- For the middle node, a new parameter LA=50 was also added (inclusion rule).
- For the right node, new parameters (BR=55 dB and GV = 55 dB) were added (inclusion rule) to distinguish this node from the other nodes.

- Nodes that belonged to the forth level were created by inheriting all the parameters and their values, except the speech rate, from the parent node. The speech rate used in this level was SR = 100 wpm.

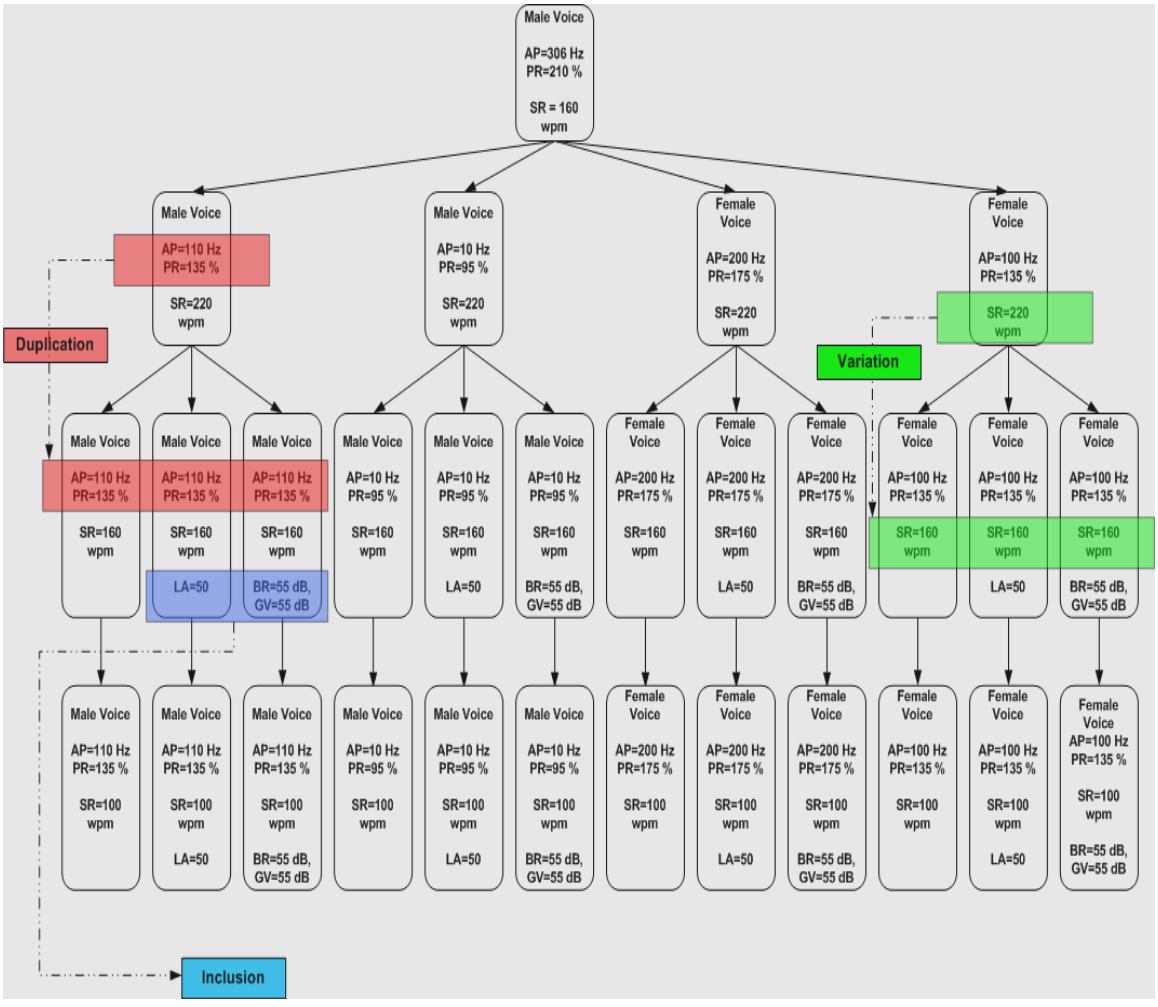


Figure 5.6: Synthetic voice parameters and the values that are used to create the complex hierarchy using Multiple Synthetic Voices

5.2.4 Hypotheses

Based on the results of Brewster [Bre98] and on the studies of perception of synthetic speech in speech-based interfaces, the following results are anticipated:

❶ ***Hypothesis 1:*** Participants should be able to perform the node finding tasks with high accuracy by listening to the multiple synthetic voices than without.

❷ ***Hypothesis 2:*** Participants should be able to locate the position of the unheard voices (nodes A and B) in the hierarchy by recalling the rules that were used to construct the multiple synthetic voices.

❸ ***Hypothesis 3:*** If the rules that were used to create the multiple synthetic voices are easy to remember, then the type of training should not have an effect on the recall rates of the node identification tasks.

5.2.5 Method

5.2.5.1 Design

A two-condition (group-1 and group-2) between-subject experiment was conducted, where the type of training given to the participants was varied between the groups. The participants were randomly assigned to one of the two groups. In this study, the accuracy rate for locating nodes in a hierarchy was measured and analyzed.

5.2.5.2 Materials

The materials used on this experiment were the same as those used in the first experiment.

5.2.5.3 Participants

16 students from a local university volunteered for this study. Only native English speakers who did not exhibit any auditory disorder were selected. Though all the participants admitted to have heard synthetic speech on various occasions, all of them claimed that they had no prior experience in working with synthetic speech.

5.2.5.4 Training

At the start of the experiment, all the participants were given a write-up, which explained the rules that were used to create the hierarchy of multiple synthetic voices. However, the type of training received by the participants varied between the groups. For group-1, the experimenter showed the hierarchy and explained the rules that were used to represent the hierarchical information. Participants belonging to group-2 were given three minutes to learn the rules by themselves, and did not receive any help from the experimenter during the training session. During the training session, the participants were shown the structure of the hierarchy (Figure 5.5) and were asked to click on a node to listen to its associated synthetic voice.

5.2.5.5 Procedure

The experiment began when participants claimed to be comfortable with the system. During this experiment, fourteen voices were randomly selected from the set of 29 possible voices in the hierarchy, and were played to the participants. Three of the voices were from level 2 (root is at level 1), five of the voices were from level 3 and the remainder were from leaf nodes (level 4). Out of the fourteen voices, two voices (A and B) were new voices that had not been played to the participants during the training session (see Figure 5.5).

Table 5.2: List of voices played to the participants during the experiment

Questions	Location	Where is the node labeled…
Q1	*Level 2*	*General information*
Q2	*Level 4*	*Soft pop*
Q3	*Level 4*	*New York*
Q4	*Level 2*	*Important flies*
Q5	*Level 4*	*Canada*
Q6	*Level 3*	*Word*
Q7	*Level 4*	*Mark sheet*
Q8	*Level 2*	*Music*
Q9	*Level 3*	*Sports*
Q10	*Level 3*	*Classic*
Q11	*Level 3*	*Graphics*
Q12	*Level 3*	*Weather*
Q13	*Level 4*	*A*
Q14	*Level 4*	*B*

After playing each voice, the participants were asked to locate the position of the node in the hierarchy. The participants were given a sheet of paper with a hierarchy containing unlabelled nodes (similar to Figure 5.5 with the labels removed). They were asked to label the nodes in the hierarchy based on the order of presentation (i.e. the first node played was labeled 1, etc.), as in experiment 1. The labeling provided by each subject was used to compute their accuracy in identifying the position of the nodes in the hierarchy. Upon completion of the first twelve questions (Q1-Q12), the participants were informed that Q13 and Q14 are new voices and had not been shown to them during the training session. They were asked to identify the position of these two voices in the hierarchy. The lists of voices that are played during the experiment are shown in the Table 5.2.

5.2.6 Results and Discussion

The overall results are summarized in Table 5.3, which reports the overall accuracy rate for each question. Consistent with hypothesis 1, the overall result shows that the participants could recall 84.38% of the voices accurately. In order to analyze the occurrence of the errors in the tasks, the questions were grouped based on the different levels in the hierarchy, i.e. out of 14 questions, 3 are from level-2, 5 from level-3 and 6 from level-4.

Table 5.3: Overall recall rates for each question

Questions	Recall rates (%)
Q1	93.75
Q2	68.75
Q3	70
Q4	100
Q5	100
Q6	87.5
Q7	68.75
Q8	68.75
Q9	81.25
Q10	43.75
Q11	100
Q12	100
Q13	100
Q14	93.75

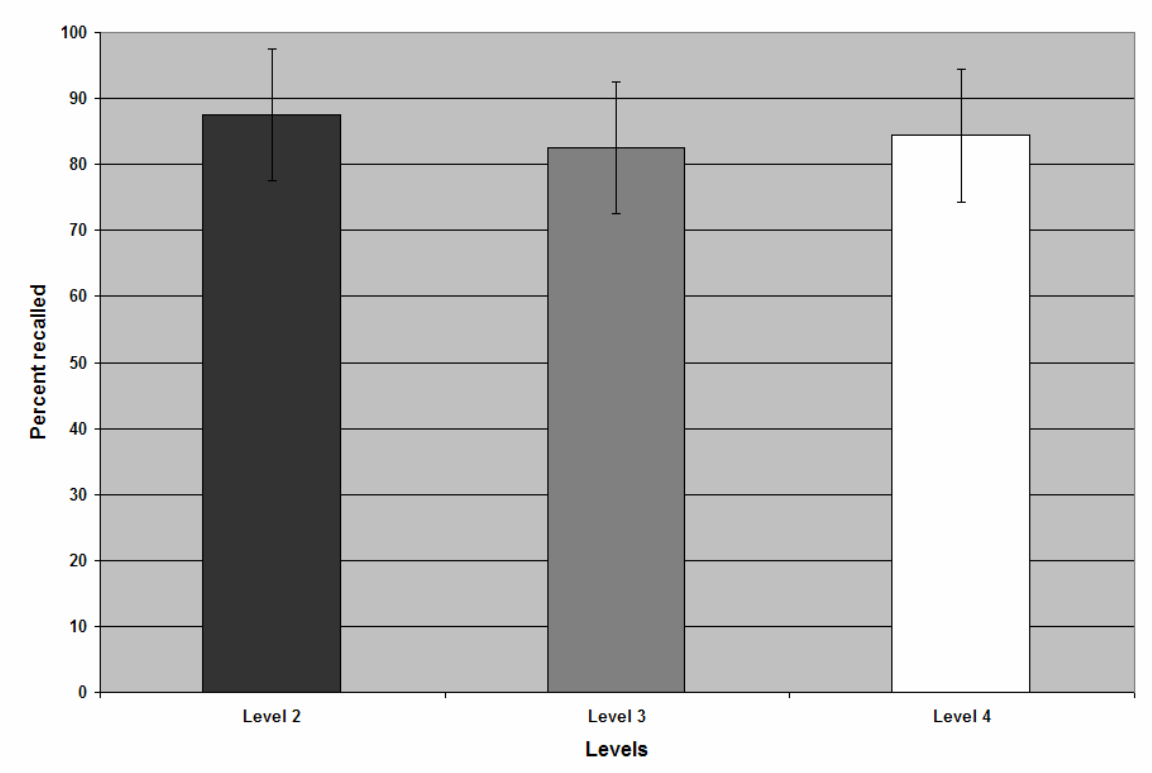


Figure 5.7: Overall recall rates at each level

Figure 5.7 shows the overall recall rates of multiple synthetic voices at each level. Since the overall recall rates at each level are almost the same, the recall rates for each family were analyzed, where Family-1 consists of nodes inclusive and belonging to "General Information" (Figure 5.5). Figure 5.8 shows the overall recall rates for each family. The results show that the highest number of errors occurred in Family-4. This shows that the voices that were used to create this family were not distinguishable from Family-3. Hence, the results conclude that the parameter values should be chosen such that the voices are easily distinguishable.

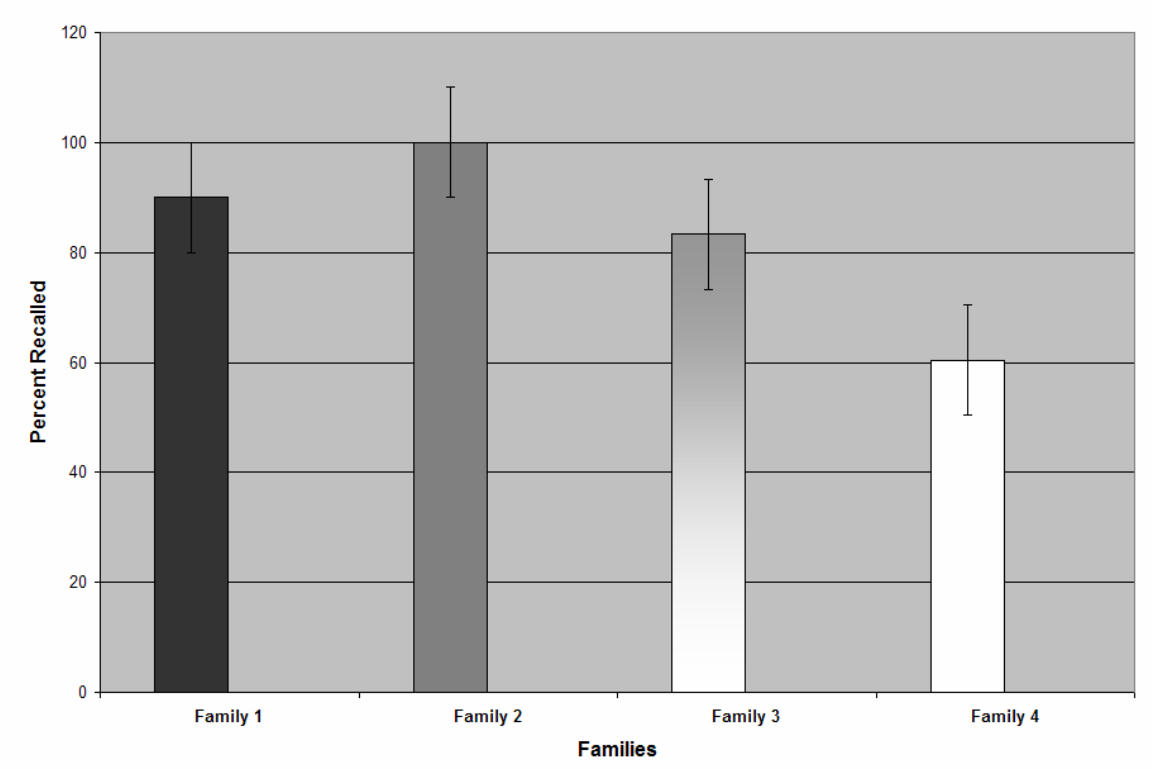


Figure 5.8: Overall recall rates for each family

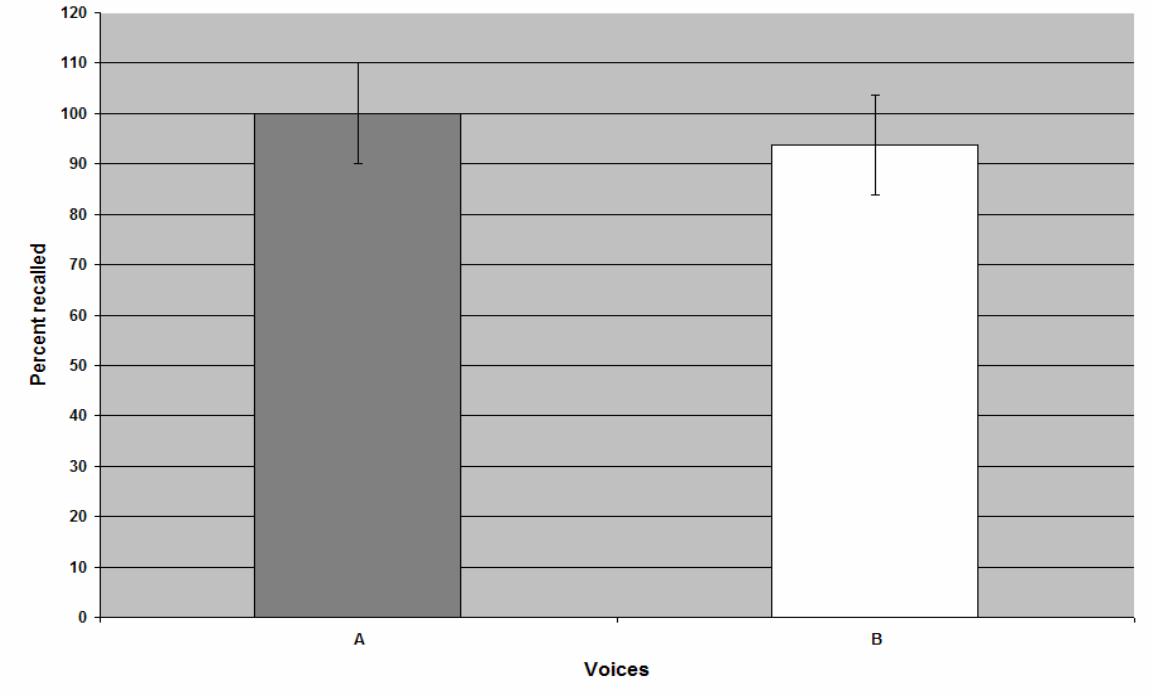


Figure 5.9: Overall recall rates for the two new voices

Figure 5.9 shows the recall rates of unheard voices. The results show that 96.88% of the voices were recalled accurately. Out of 16 participants, all the participants recalled “A” and 15 participants recalled “B” correctly. Hence, consistent with hypothesis-2, the results show that the rules that were used for creating multiple synthetic voices were easy to remember and recall. In order to analyze the effect of training, the task completion rates were compared (the number of accurately identified nodes) of each group. A one-way ANOVA F-test (p=.8387, degree of freedom = 15, F ratio = 0.043), showed that the type of training had no significant effect on recall rates. Consistent with hypothesis-3, the overall results suggest that extensive training is not required to extract hierarchical relationships from the multiple synthetic voices.

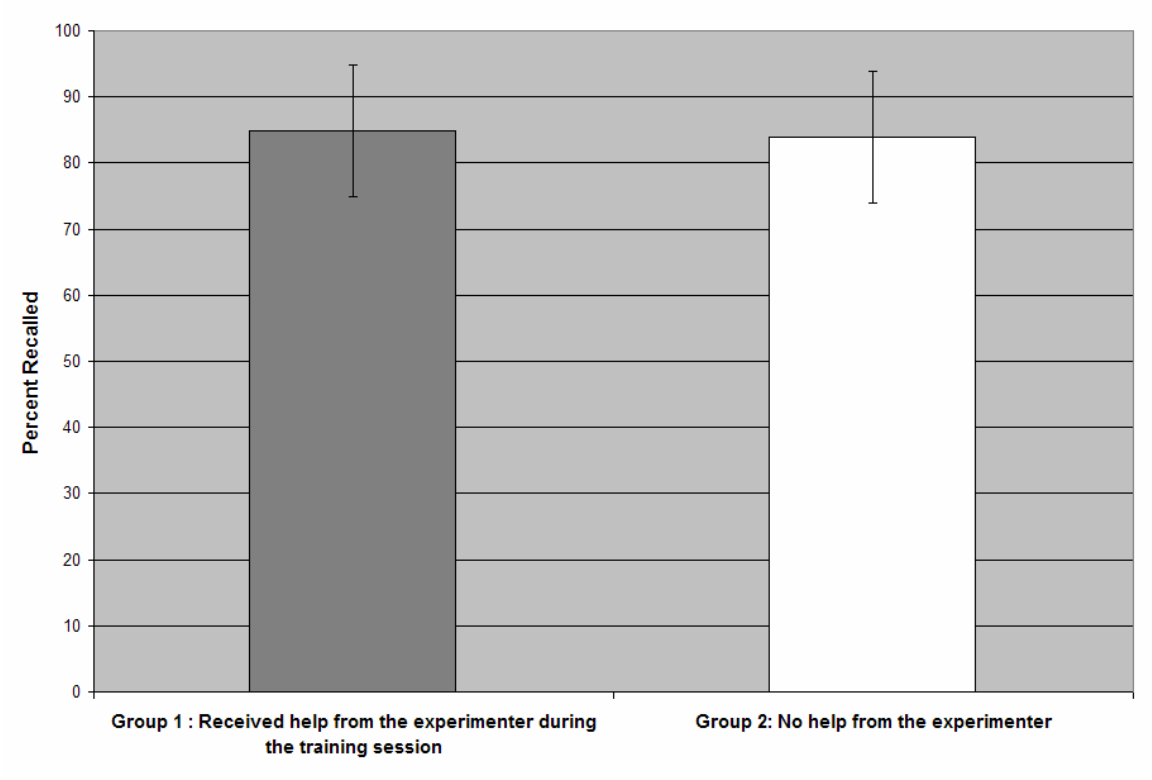


Figure 5.10: Overall recall rates for the two groups (Group-1 and Group-2)

Overall, the results of this experiment confirmed all the three hypotheses, that multiple synthetic voices provide navigation cues to recall components in a complex hierarchy. In this experiment, pitch was used to differentiate between the families (sub-trees); however, further experimentation is required to decide on the best combination of parameters that facilitate identifying hierarchical relationships.

5.3 Conclusion

Two experiments were conducted to analyze the effect of multiple synthetic voices on small and complex hierarchies. In the first experiment, multiple synthetic voices were created by manipulating speech parameters such as pitch, pitch range and speech rate. Multiple synthetic voices were created in a manner that facilitates identifying hierarchical relationships between the nodes in a tree, using two basic design guidelines: Duplication (duplicate exactly the value of a preceding synthetic voice parameter) and Variation (alter the values of one or more synthetic speech parameters between two related nodes). The participants' tasks involved identifying the position of the node in the hierarchy by listening to a played voice. Participants performed node finding tasks 4 times better with MSV-1 than with SSV (Error rate: MSV-1=15.3%, SSV=62.5%) and 2.8 times better with MSV-2 than with SSV (Error rate: MSV-1=22.2%, SSV=62.5%). Hence, consistent with the hypothesis, the results of this experiment suggest that multiple synthetic voices can be constructed to facilitate locating nodes in hierarchies.

In the second experiment, along with as pitch, pitch range and speech rate, other speech parameters such as laryngealization, breathiness, and gain in volume were also

manipulated. Multiple synthetic voices were created using the following three guidelines: Duplication, Variation and Inclusion (include one or more speech parameters to the preceding voice in order to create unique voices). In this experiment, participants' accuracy rate and the effect of training were measured. Participant's tasks involve locating the node in the hierarchy. Consistent with hypothesis 1, the overall results suggest that participants identified 84.38% of the nodes accurately. The results also suggest participants recalled the unheard voices with high accuracy (96.88%) without requiring extensive training (hypothesis 2). Consistent with the hypothesis 3, the results suggest that the type of training did not have any impact on recall rates (p=.8387), which is considered as the one of the major milestones in this research. Hence, the results recommended that the rules that were used to create multiple synthetic voices were easy to remember and recall.

Chapter 6 : Conclusions

The main aim of this thesis is to improve navigation in auditory interfaces. A summary of the research and the results that are obtained in this research are given below. This chapter also discusses some of the limitations of the work described in this thesis and also suggests some solutions to overcome these limitations. The chapter concludes by suggesting some of the areas of future work in auditory interfaces.

6.2 Summary of the Contribution

The main focus of this thesis is to analyze the different approaches that have been taken to deal with the issue of navigation in auditory interfaces. The main motivation for this study is the following question, “How can navigation be improved in hierarchical auditory interface structures?” In an endeavor to answer this question, this research addresses the issue of navigation in auditory interfaces using two different novel solutions, as mentioned below.

6.2.1 Personalizing menus in speech interfaces

Chapter 4 describes the detailed implementation and evaluation of a new interface design (referred to as a personalized interface) that has emerged from this research. A combination of the following elements defines the novelty of this design:

❶ A Personalized interface is a multiple interface (an integration of a default interface and a personalized interface). This interface allows users to toggle between two interfaces efficiently.

❷ A Personalized interface is easy to use and adapts to the users effortlessly.

❸ A Personalized interface starts without any options in the interface and grows as the user adds more functions to it. In other words, a personalized interface is the minimal favorite interface when compared to the full interface.

In general, voice-based applications rely heavily on hierarchically structured menus (or prompts) for creating dialogs between the user and the information that is being retrieved. The highly hierarchical structure of this method of retrieving information leads to spending significant amount of time in navigating through the various layers for obtaining the requested information. However, the implemented system described in chapter 4 allows users to bookmark any given node in the menu-based system. This provides a method of bypassing the entire hierarchy structure to access only the node of interest in the menu tree. Results of a user satisfaction study show that users prefer using the personalized system for accessing data in touch-tone applications over conventional methods for navigating touch-tone interfaces.

6.2.2 Improving Navigation Using Multiple Synthetic Voices

Many commercial applications use synthetic speech for conveying information. In many cases the structure of the information is hierarchical (e.g. menus). Chapter 5 describes the results of two experiments that examine the possibility of conveying hierarchies using multiple synthetic voices. These studies focus on reducing error rates in auditory interfaces by providing navigational cues about the user's location in the navigation structure. Chapter 5 presents a method that assists users in identifying the

location of nodes in the underlying navigation structure of an auditory interface. Some of the main contributions to this novel framework are outlined below:

❶ Results of Chapter 5 suggest that a higher accuracy rate was obtained with multiple synthetic voices. This accuracy rate is slightly better than the hierarchical earcons suggested by Brewster [Bre98].

❷ The effect of training does not have a significant impact on the recall rates of synthetic voices, which is considered as a major milestone in this research, because users typically cannot spend significant training times with an auditory interface.

This chapter is mainly based on the belief that if hierarchical structures can be highlighted using synthetic speech, then navigation in these hierarchies can be improved. In the first experiment, hierarchies containing 10 nodes, with a depth of 3 levels, were created. Synthetic voices were used to represent nodes in these hierarchies. A within-subjects study (N=12) was conducted to compare the effect of multiple synthetic voices with single synthetic voices, to locate the positions of nodes in a hierarchy. Multiple synthetic voices were created by manipulating synthetic voice parameters according to a set of design rules. Subjects were trained using the set of design rules. Results of the first experiment showed that the subjects performed the tasks significantly better with multiple synthetic voices than with single synthetic voices. To investigate the effect of multiple synthetic voices on complex hierarchies and the effect of training on recall rates second experiment was conducted. A hierarchy of 27 nodes was created and a between

subjects study (N=16) was carried out, in which the type of training varied between the groups. The results of this experiment showed that the participants recalled 84.38% of the nodes accurately and that the type of training did not have any significant effect on recall rates. Results from these studies imply that multiple synthetic voices can be used to represent and provide navigation cues in hierarchies, such as Telephone Based Interfaces (TBIs).

6.3 Limitations and Perspectives

Whilst this thesis has contributed to the general knowledge of supporting navigation in auditory interfaces, there are some limitations to this work which have been outlined below.

6.3.1 Personalizing menus in speech interfaces

Notwithstanding its popularity, a personalized touch-tone interface also has its limitations. Although, the user satisfaction questionnaire suggests that listening to the recorded voice is pleasing, accessing information using personalized menus depends upon the information that the users have recorded or stored. For instance, the chances of going astray could increase if users had improperly recorded their prompts.

The experiment described in Chapter 4 says that the options in the personalized list are saved sequentially. As a result, the time taken to navigate the personalized list increases, with increase in the number of options in the list. Hence, further investigation is required to improve the navigation in the personalized list. Further design may improve the current designs and also may become a valuable component to the work described in Chapter 4.

For instance, allowing users to skip the uninterested menus may reduce further navigation time.

6.3.2 Improving Navigation Using Multiple Synthetic Voices

The work described in Chapter 5 suggests that manipulating synthetic voice parameters would facilitate navigation in auditory interfaces. However, the work did not evaluate the recognition of multiple synthetic voices over a period of time. The results explained in Chapter 5 suggest that users were able to remember/recall multiple synthetic voices after the training session, i.e. after a very short period of time (approximately 10 minutes). But in the real-world scenario, the frequency of accessing the auditory interface may vary between the users. For example, some users might use the interface once in a week, or once in every two weeks. Therefore, additional investigation is required to analyze the usefulness of multiple synthetic voices over a long period of time.

The experiments described in Chapter 5 focused on analyzing only the recognition rates of multiple synthetic voices. However, these experiments did not evaluate the effect of multiple synthetic voices on cognitive workloads. If higher workload is required to recognize synthetic voice parameters, then a significant amount users' cognitive resources has to be dedicated to extract the information from multiple synthetic voices. Therefore, an in-depth study is required to ensure that the results obtained from the studies described in Chapter 5 are valid in the highly demanding environment.

The major limitation with the research presented in Chapter 5 is directly linked to the novelty of the design framework. Since the design framework and the principles are novel, the guidelines provided in this thesis were very general, and have been applied only by the designer. Therefore, more knowledge on speech and psychology is required to derive the complete set of guidelines for creating multiple synthetic voices for hierarchies. At this point of time, deriving a set of guidelines for creating multiple synthetic voices for hierarchies is difficult, because only a couple of experiments have been conducted to investigate the effect of multiple synthetic voices on hierarchies.

Another major limitation with multiple synthetic voices is increasing the size of the hierarchy. Increasing the size of the hierarchy would be a challenge, because there are not too many synthetic speech parameters left over to manipulate. However, the size of the hierarchy could be increased by mixing multiple synthetic voices and earcons to represent the nodes in the hierarchy.

6.4 Future work

In the previous sections, some of the main contributions of the thesis and the limitations to these contributions have been summarized. This section discusses some of the areas in which the work described in this thesis may be expanded to further improve navigation in auditory interfaces.

6.4.1 Personalizing menus in speech interfaces

Whilst personalizing menus at the interfaces study has shown some improvements in reducing the navigation time in auditory interfaces, some improvements to this model do remain. Future work in the personalization research will consist of allowing users to perform maintenance operations in the personalized list. These operations include, allowing users to re-record a particular bookmark, changing the order of menus in the personalized list, allowing users to skip uninterested bookmarks, and deleting a particular option in the personalized list. By implementing these options in the personalized interface, a considerable amount of navigation time can be saved.

It would be interesting to maintain a small hierarchy in the personalized list, i.e., saving the bookmarked options in a hierarchical order (in the same way as menus are saved in the default interface). This way, the users can avoid listening to some of the uninterested options. Also, a more rigorous study is required to compare the sequential personalized list and the hierarchical personalized list.

6.4.2 Improving Navigation Using Multiple Synthetic Voices

It is interesting to compare the difference in recall rates for multiple synthetic voices to those for earcons in a similar study [Bre98]. In the study designed by Brewster [Bre98], the recall rate for hierarchical earcons was approximately 81.5% and for compound earcons was 97%. The results of the experiment described in the Chapter 5 fall in between these two values (84%). However, for unheard earcons, users were able to locate their position in the hierarchy with an accuracy of 91.5% in comparison to 97%

with multiple synthetic voices. A more rigorous experiment could be designed to compare earcons to multiple synthetic voices on dimensions other than locating elements in hierarchies, such as whether users are able to comprehend information better or quicker with one system over the other.

The results also reveal that the rules that were used to create the multiple synthetic voices were easy to remember and recall. However, this study did not establish the set of manipulation rules that facilitate the best performance. An additional study will be designed to evaluate the effectiveness of multiple parameter configurations for representing depth and width in hierarchies. The results of the experiments, have not clarified the effect of multiple synthetic voices on more complex hierarchies, such as those found in file systems. This could be done either by integrating earcons and multiple synthetic voices to represent the nodes in the hierarchy or by making use of other vocal cues such as richness, smoothness, etc. A study will also be performed to compare the effect of mixing multiple synthetic voices with earcons.

In the future, the effect of multiple synthetic voices will be evaluated on real-time applications that contain hierarchical structures, such as in video games and TBIs. Hence, this approach will provide significant insight to designers once we are able to devise a firm set of guidelines for manipulating the various synthetic voice parameters.

6.5 A Final Word

In conclusion, the work described in this thesis constitutes an important contribution for supporting interaction in auditory interfaces. Further investigation is required to determine the application of the results derived from this thesis.

Appendix A: Questionnaire for Evaluating Personalizing Menus for Navigation in Touch-Tone Voice Interfaces

We are building a new Interactive Voice Response (IVR) system, called "Personalized Voice Interface". As a part of the research we are evaluating our system to find the elements that need to be changed to enhance the performance of our system. Please note that this is an exercise to evaluate our system not your abilities. If you find any difficulty in accessing the information or finding it hard to understand, it is our work to rectify the problems to improve the system. The evaluation process will last for approximately 45 minutes. You are allowed to take a break at any time you deem necessary. All information provided will be treated as confidential and will not be redistributed.

Thank you for spending your precious time in evaluating our system.

Peer M. Shajahan

Dr. Pourang P. Irani

Details about Personalized Voice Interface:

Personalized Voice Interface (PVI) is an Interactive Voice Response (IVR) system in which users use the telephone as a primary medium to access information, such as, transit information. PVI allows users to bookmark their preferences in the "Personalized Menus List". The "Personalized Menus List" saves bookmarked items and transfers control of the system to that menu item when requested by the user.

The goals of building PVI are to reduce users' navigation time in telephony-applications and to build a system where language is not the main obstacle in accessing information. PVI allows users to record their preferences in their own languages and replays users' voices for the bookmarked menus, when accessed via the "Personalized Menus List".

Personal Details

1. Contact Information

1. First Name : _____
2. Last name : _____
3. Phone number : _____ (Optional)
4. E-mail : _____

2. Age range

a) Less than 20

b) 21 – 35

e) 36 and above

3. Sex

a) Male

b) Female

4. Do you have any auditory problem?

a) Yes

b) No

5. Have you ever listened to computer generated speech?

a) Yes

b) No

If yes, please answer the following question.

Did you find it hard to understand computer generated speech?

a) Yes, please specify the reason

b) No

6. Do you know how to use Interactive Voice Response systems for accessing information, such as telephone banking, and finding transit information?

a) Yes

b) No

If yes, please answer the following questions.

1. What is your experience with Interactive Voice Response systems?

a) Never used b) Used a few times c) Used regularly

d) Other, please specify: _____

2. How long have you been using Interactive Voice Response systems?

a) 0 – 2 years b) 3 – 5 years c) 6 – 10 years d) Greater than 10 years, please specify: _____

3. How often do you use Interactive Voice Response systems in a month?

a) Never b) 1 – 5 times c) 6 – 10 times d) 11 – 15 times

e) Other, please specify: _____

Subjective Evaluation Questionnaire

Speed

Questions	***Personalized Voice Interface (A)***	***Conventional Voice Interfaces (B)***
Which of the following systems would you choose to obtain targeted information quicker: A or B?		
Which system would you recommend for accessing information that you regularly use: A or B?		

Navigation

Questions	*Personalized Voice Interface (A)*	*Conventional Voice Interfaces (B)*
Which of the following systems would you use to choose the right menu in order to reach the next level efficiently and effectively (faster and better): A or B?		
In which system do you feel lost when you go into deeper levels in accessing particular information about the system: A or B?		
Which of the following systems would you choose to know better where you are by often listening to the prompt/voice of menu options: A or B?		

Learnability / Adaptability

Questions	*Personalized Voice Interface (A)*	*Conventional Voice Interfaces (B)*
Which of the following systems is easy to learn: A or B?		
In which system, language is not a major barrier in accessing information: A or B?		
In which system, I know exactly where I am or under which branch the targeted option is located: A or B?		

Ease of use

Questions	*Personalized Voice Interface (A)*	*Conventional Voice Interfaces (B)*
Which system would you use for your tenth attempt: A or B?		
In which of the following systems would you choose to skip some of the options that do not like to hear on every use: A or B?		
Which of the following systems provides more flexibility to reach the targeted goal: A or B?		

Users' Satisfaction in using Personalized Voice Interface

Questions	*Strongly disagree*	*Disagree*	*Neutral*	*Agree*	*Strongly Agree*
Q1: Difficult to locate menu options using the personalized touch-tone interface.					
Q2: I am able to complete my tasks efficiently using personalized touch-tone interface.					
Q3: The speed of personalized voice interface is good for accessing information that I regularly like to query the system.					
Q4: I am satisfied with the speed of accessing information using personalized voice interface.					
Q5: I like the technique of recording my voice for providing information about personalized menus in the "Personalized Menus list".					

Q6: Recorded menus in the “Personalized Menus List” improve navigation.					
Q7: Replay of recorded human voice is pleasant.					
Q8: Personalized touch-tone interface is easy to learn.					
Q9: I know how to use the system after the trial session.					
Q10: Personalized touch-tone interface is easy to use.					
Q11: I can remember the list of menus that I have saved in the “Personalized Menus List”.					
Q12: Recording option is a necessary element of the interface.					
Q13: Personalized touch-tone interface is the preferred choice for daily transactions.					
Q14: Overall satisfaction of personalized touch-tone interface.					

Suggestions and Comments:

1. In general how do you feel working with Personalized Voice Interface?

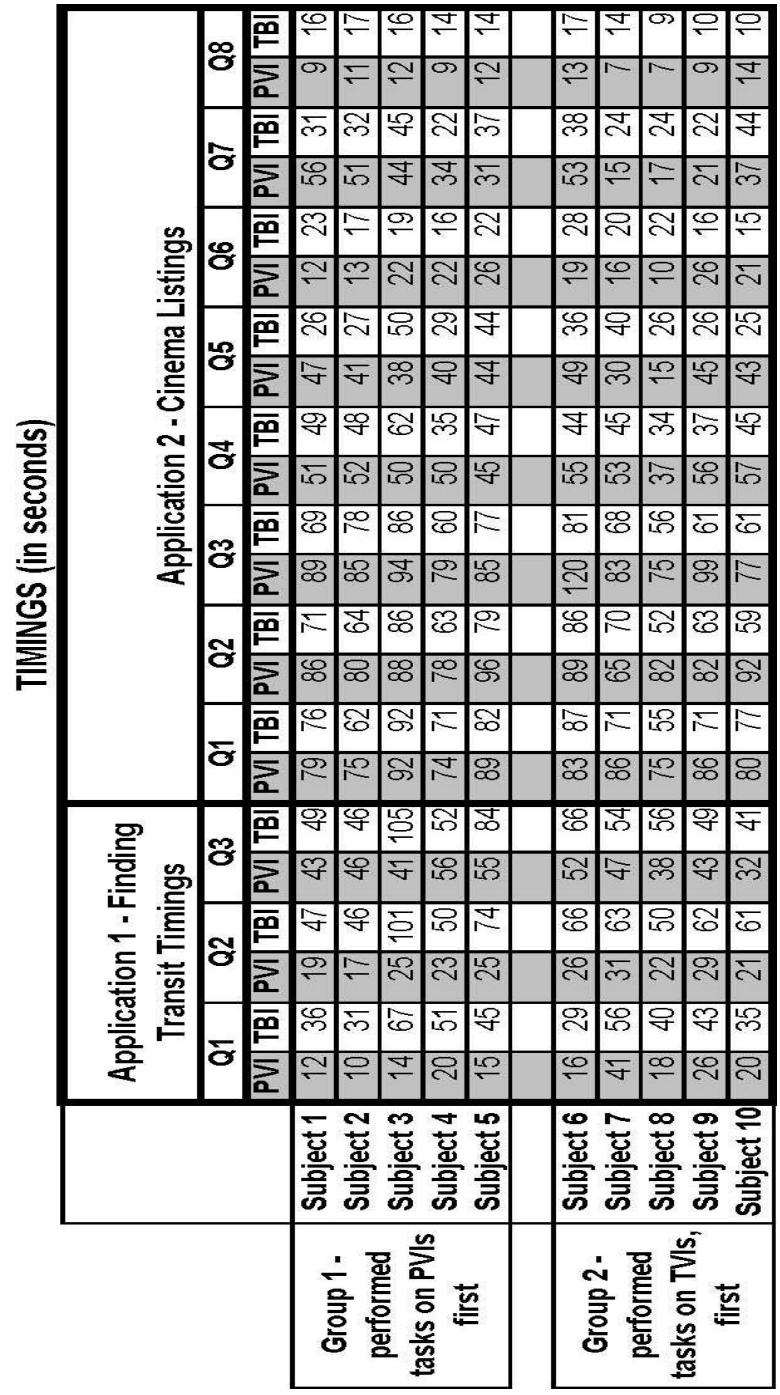
2. What would you like most about Personalized Voice Interface?

3. What would you dislike most about Personalized Voice Interface?

4. Please add any additional comments that you wish to make about Personalized Voice Interface

Thank you for your assistance.

Appendix B: Raw data from the Experiment: Personalizing Menus for Navigation in Touch-Tone Voice Interfaces



TIMINGS (in seconds)

		Application 1 - Finding Transit Timings						Application 2 - Cinema Listings													
		Q1		Q2		Q3		Q1		Q2		Q3		Q4		Q5		Q6		Q7		Q8	
		PVI	TBI	PVI	TBI	PVI	TBI	PVI	TBI	PVI	TBI	PVI	TBI	PVI	TBI	PVI	TBI	PVI	TBI	PVI	TBI
Group 1 - performed tasks on PVIs first	Subject 1	12	36	19	47	43	49	79	76	86	71	89	69	51	49	47	26	12	23	56	31	9	16
	Subject 2	10	31	17	46	46	46	75	62	80	64	85	78	52	48	41	27	13	17	51	32	11	17
	Subject 3	14	67	25	101	41	105	92	92	88	86	94	86	50	62	38	50	22	19	44	45	12	16
	Subject 4	20	51	23	50	56	52	74	71	78	63	79	60	50	35	40	29	22	16	34	22	9	14
	Subject 5	15	45	25	74	55	84	89	82	96	79	85	77	45	47	44	44	26	22	31	37	12	14
Group 2 - performed tasks on TVIs, first	Subject 6	16	29	26	66	52	66	83	87	89	86	120	81	55	44	49	36	19	28	53	38	13	17
	Subject 7	41	56	31	63	47	54	86	71	65	70	83	68	53	45	30	40	16	20	15	24	7	14
	Subject 8	18	40	22	50	38	56	75	55	82	52	75	56	37	34	15	26	10	22	17	24	7	9
	Subject 9	26	43	29	62	43	49	86	71	82	63	99	61	56	37	45	26	26	16	21	22	9	10
	Subject 10	20	35	21	61	32	41	80	77	92	59	77	61	57	45	43	25	21	15	37	44	14	10

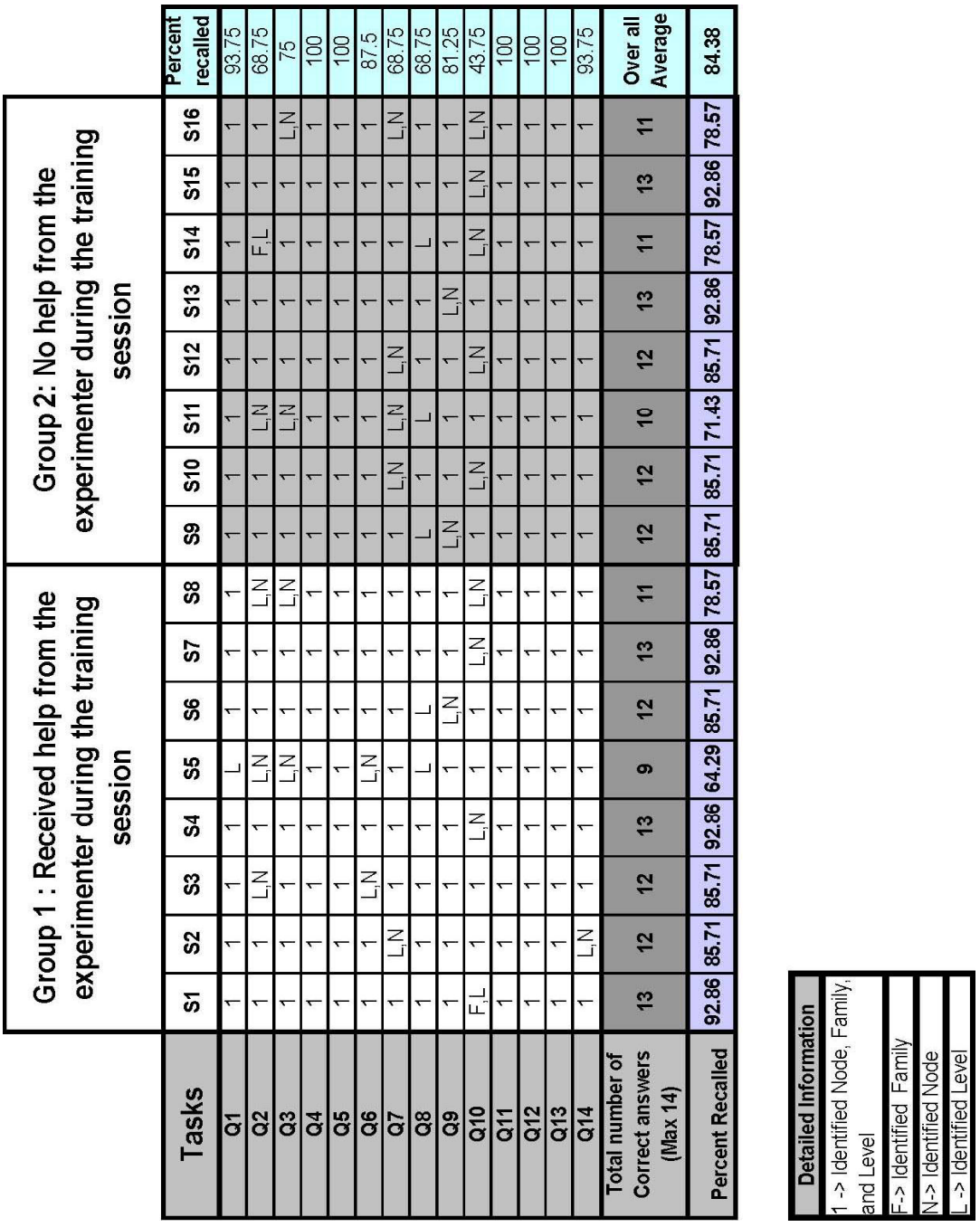
Appendix C: Raw data from the Experiment: Representing Small Hierarchies Using Multiple Synthetic Voices

This appendix contains the raw data from the experiment: Representing small hierarchies using multiple synthetic voices, described in Chapter 5.1.

	Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12		
SSV	Level 2	1	2	2	2	1	2	1	1	1	2	1	0	Total Identified	27
	Level 3	1	1	0	0	2	2	1	1	1	0	2	0		
	Total	2	3	2	2	3	4	2	2	2	2	3	0	Total Incorrect	45
	Total Tasks	6	6	6	6	6	6	6	6	6	6	6	6		
	In Correct	4	3	4	4	3	2	4	4	4	4	3	6		
MSV-1	Level 2	2	2	2	1	2	2	2	2	0	2	2	1		
	Level 3	4	4	3	3	4	3	4	4	3	3	3	3	Total Identified	61
	Total	6	6	5	4	6	5	6	6	3	5	5	4		
	Total Tasks	6	6	6	6	6	6	6	6	6	6	6	6	Total Incorrect	11
	In Correct	0	0	1	2	0	1	0	0	3	1	1	2		
MSV-2	Level 2	1	2	2	2	2	0	1	2	0	2	0	1		
	Level 3	4	3	4	4	4	4	3	3	4	2	4	2	Total Identified	56
	Total	5	5	6	6	6	4	4	5	4	4	4	3		
	Total Tasks	6	6	6	6	6	6	6	6	6	6	6	6	Total Incorrect	16
	In Correct	1	1	0	0	0	2	2	1	2	2	2	3		

Appendix D: Raw data from the Experiment: Representing Complex Hierarchies Using Multiple Synthetic Voices

This appendix contains the raw data from the experiment: Representing complex hierarchies using multiple synthetic voices, described in Chapter 5.2.



	Group 1 : Received help from the experimenter during the training session								Group 2: No help from the experimenter during the training session								Percent recalled
Tasks	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	
Q1	1	1	1	1	L	1	1	1	1	1	1	1	1	1	1	1	93.75
Q2	1	1	L,N	1	L,N	1	1	L,N	1	1	L,N	1	1	F,L	1	1	68.75
Q3	1	1	1	1	L,N	1	1	L,N	1	1	L,N	1	1	1	1	L,N	75
Q4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100
Q5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100
Q6	1	1	L,N	1	L,N	1	1	1	1	1	1	1	1	1	1	1	87.5
Q7	1	L,N	1	1	1	1	1	1	1	L,N	L,N	L,N	1	1	1	L,N	68.75
Q8	1	1	1	1	L	L	1	1	L	1	L	1	1	L	1	1	68.75
Q9	1	1	1	1	1	L,N	1	1	L,N	1	1	1	L,N	1	1	1	81.25
Q10	F,L	1	1	L,N	1	1	L,N	L,N	1	L,N	1	L,N	1	L,N	L,N	L,N	43.75
Q11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100
Q12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100
Q13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	100
Q14	1	L,N	1	1	1	1	1	1	1	1	1	1	1	1	1	1	93.75
Total number of Correct answers (Max 14)	13	12	12	13	9	12	13	11	12	12	10	12	13	11	13	11	Over all Average
Percent Recalled	92.86	85.71	85.71	92.86	64.29	85.71	92.86	78.57	85.71	85.71	71.43	85.71	92.86	78.57	92.86	78.57	84.38

Detailed Information
1 -> Identified Node, Family, and Level
F-> Identified Family
N-> Identified Node
L -> Identified Level

Appendix E: Statistical Tests Used

T-test:

T-test is a one-way ANOVA (Analysis of Variance) test, used to evaluate whether the means of two groups are statistically different.

Kruskal-Wallis Test:

Kruskal-Wallis Test is a non-parametric test, alternative to two sample t-test. Kruskal-Wallis test is commonly used to compare three or more samples and when the sample data is not normally distributed.

Mann-Whitney test:

Mann-Whitney test is the non-parametric test alternative to independent samples t-test. Mann-Whitney test is commonly used when the sample data is not normally distributed. In order to perform the Mann-Whitney test, the data from two samples (S1 and S2) are combined and ranked from lowest to highest. The difference between the sum of these ranks from S1 and S2 were then measured to analyze if the results were statistically different.

F-test:

F-test, sometimes, also called as ANOVA, is very closely related to t-test. In t-test the difference between the means of two groups were measured, whereas in ANOVA, the difference between the means of two or more groups was analyzed. A one-way ANOVA

(or single factor ANOVA) f-test is conducted to analyze the differences between the groups that are distinguished on one independent variable.

References

[Bac99] J. A. Bachorowski. Vocal expression and perception of emotion. Current Directions in Psychological Science, 8(2):53–57, 1999.

[Bal99] B. Balentine. Human Factors and Voice Interactive Systems, chapter Re-Engineering the speech menu, pages 205–235. Kluwer Academic, Massachusetts, United States, 1999.

[BCH98] S. A. Brewster, A. Capriotti, and C. V. Hall. Using compound earcons to represent hierarchies. HCI Letters, 1(1):6–8, 1998.

[BCS+99] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T next-gen TTS system. In Joint Meeting of ASA, EAA, and DAGA, pages 15–19, Berlin, Germany, 1999.

[BGS67] D. Byrne, W. Griffitt, and D. Stefaniak. Attraction and similarity of personality characteristics. Journal of Personality and Social Psychology, 5:82–90, 1967.

[BM99] B. Balentine and D.P. Morgan. How to Build a Speech Recognition Application. Enterprise Integration Group, Inc, 1999.

[BNS02] M. Bulut, S. Narayanan, and A. Syrdal. Expressive speech synthesis using a concatenative synthesizer. In International Conference on Spoken Language Processing, pages 1265–1269, Colorado, USA, 2002.

[Bon99] D. G. Bonneau. Human Factors and Voice Interactive Systems, chapter Guidelines for speech-enabled IVR application design, pages 147–162. Kluwer Academic Publishers, 1999.

[Bre98] S. A. Brewster. Using nonspeech sounds to provide navigation cues. ACM Transactions on Computer-Human Interaction (TOCHI), 5(3):224–259, 1998.

[BSG89] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg. Earcons and icons: Their structure and common design principles. Human Computer Interaction, 1(4):11– 44, 1989.

[BWE92] S. A. Brewster, P. C. Wright, and A. D. N. Edwards. A detailed investigation into the effectiveness of earcons. In Proceedings of the First International Conference on Auditory Display, pages 471–498. Addison-Wesley, 1992.

[BWE93] S. A. Brewster, P. C. Wright, and A. D. N. Edwards. An evaluation of earcons for use in auditory human-computer interfaces. In INTERCHI Conference Proceedings, pages 222–227. ACM Press, 1993.

[BWE95] S. A. Brewster, P. C. Wright, and A. D. N. Edwards. Parallel earcons: Reducing the length of audio messages. International Journal of Human-Computer Studies, 43(2):153–179, 1995.

[Cah89] J. Cahn. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, May 1989.

[CC84] J. Carroll and C. Carrithers. Blocking learner error states in a training-wheels system. Human Factors, 4(26):377–389, 1984.

[CDCS+01] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder. Emotion recognition in human-computer intercation. IEEE Signal Processing Magazine, 18(1):32–80, 2001.

[con] http://www.conita.com.

[Dut97] T. Dutoit. An Introduction to Text-to-Speech Synthesis, volume 3. Kluwer Academic Publishers, 1997.

[Eag83] A. H. Eagly. Gender and social influence: A social psychological analysis. American Psychologist, 38:971–981, 1983.

[FN99] A. L. Francis and H. C. Nusbaum. Human Factors and Voice Interactive Systems, chapter 3, pages 63–97. Kluwer Academic Publishers, Norwell, Massachusetts, United States, 1999.

[FSMKW80] S. Feldman-Summers, D. E. Montano, D. Kasprzyk, and B. Wagner. Influence attempts when competing views are gender-related: Sex as credibility. Psychology of Women Quarterly, 5(2):311–320, 1980.

[Fur86] S. Furui. Research on individuality features in speech waves and automatic speaker recognition techniques. Speech Communications, 2(5):183–197, 1986.

[GL03] L. Gong and J. Lai. To mix or not to mix synthetic speech and human speech? contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses. International Journal of Speech Technology, 6:123–131, 2003.

[HN89] R. Halstead-Nussloch. the design of phone-based interfaces for consumers. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 347–352. ACM Press, 1989.

[Hou00] The White House. Information technology research and development: Information technology for the 21’st century, Jan 2000.

[IN00] K. Isbister and C. Nass. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. International Journal of Human-Computer Studies, 53(2):251–267, 2000.

[JHH84] C.C. Johnson, H.F. Hollien, and J.W. Hicks. Speaker identification utilizing selected temporal speech features. Journal of Phonetics, 12:319–326, 1984.

[JW98] F. H. James and T. Winograd. Representing structured information in audio interfaces: a framework for selecting audio marking techniques to represent document structures. PhD thesis, 1998.

[KHHK99] C. M. Karat, M. Halverson, D. Horn, and J. Karat. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 568–575. ACM Press, 1999.

[Lar02] J. A. Larson. VoiceXML: Introduction to Developing Speech Applications. Prentice Hall, 2002.

[LHGF60] W. E. Lambert, R. C. Hodgson, R. C. Gardner, and S. Fillenbaum. Evaluational reactions to spoken language. Journal of Abnormal and Social Psychology, 60:44–51, 1960.

[LJSC00] F. Linton, D. Joy, P. Schaefer, and A. Charron. Owl: A recommender system for organization-wide learning. Educational Technology & Society, 1(3):62–76, 2000.

[LM03] K. Larson and D. Mowatt. A speech-based human-computer interaction system for automating directory assistance services. International Journal of Speech Technology, 62(6):154–159, 2003.

[LNB00] E. J. Lee, C. Nass, and S. Brave. Can computer-generated speech have gender?: An experimental test of gender stereotype. In CHI'00 extended abstracts on Human factors in computer systems, pages 289–290. ACM Press, 2000.

[LWC00] J. Lai, D. Wood, and M. Considine. The effect of task conditions on the comprehensibility of synthetic speech. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 321–328. ACM Press, 2000.

[MAE+99] F. R. McInnes, D. J. Attwater, M. D. Edgington, M. S. Schmidt, and M. A. Jack. User attitudes to concatenated natural speech and text-to-speech synthesis in an automated information service. In Proceedings of Eurospeech 99 (European Conference on Speech Communication and Technology), pages 831–834, Budapest, Hungary, 1999.

[MBB02] J. McGrenere, R. M. Baecker, and K. S. Booth. An evaluation of a multiple interface design solution for bloated software. In CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems, pages 164–170. ACM Press, 2002.

[McG02] J. McGrenere. The design and evaluation of multiple interfaces: A solution for complex software. PhD thesis, University of Toronto, Canada, 2002.

[MM00] J. McGrenere and G. Moore. Are we all in the same "bloat"? In Graphics Interface, pages 187–186, 2000.

[MNA+99] F. R. McInnes, I. A. Nairn, D. J. Attwater, M. D. Edgington, and M. A. Jack. A comparison of confirmation strategies for fluent telephone dialogues. In Human Factors in Telecommunication, 1999.

[MS96] M. Marx and C. Schmandt. Mailcall: message presentation and navigation in a nonvisual environment. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 165–172. ACM Press, 1996.

[MWW90] M. M. Martin, B. H. Williges, and R. C. Williges. Improving the design of telephone-based information systems. In Proceedings of the Human Factors Society 34th Annual Meeting, volume 1, pages 198–202, 1990.

[NL01] C. Nass and K. Lee. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied, 3(7):171–181, 2001.

[NMF+95] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and D. C. Dryer. Can computer personalities be human personalities? International Journal of Human-Computer Studies, 43(2):223–239, 1995.

[Oli97] J. P. Olive. HAL's Legacy: 2001's Computer as Dream and Reality, chapter 6 : "The Talking Computer": Text to Speech Synthesis, pages 101–131. MIT Press, 2nd edition, 1997.

[PDR96] S. Planalp, V. DeFrancisco, and D. Rutherford. Varieties of cues to emotion in naturally occurring situations. Cognition and Emotion, 10(2):137–153, 1996.

[Ram66] R. W. Ramsay. Personality and speech. Journal of Personality and Social Psychology, 64(1):116–118, 1966.

[Ram68] R. W. Ramsay. Speech patterns and personality. Language and Speech, 11(2):52–63, 1968.

[RE89] T. L. Roberts and G. Engelbeck. The effects of device technology on the usability of advanced telephone functions. In Proceedings of ACM CHI'89 Conference on human factors in computing systems, pages 331–337. ACM Press, 1989.

[Ros85] M. B. Rosson. Using synthetic speech for remote access to information. Behavioral Research Methods and Instrumentation, 17(2):250–252, 1985.

[RV92] P. Resnick and R. A. Virzi. Skip and scan: Cleaning up telephone interfaces. In Proceedings of ACM CHI'92, pages 419–426. ACM Press, 1992.

[Sam75] M.R. Sambur. Selection of acoustic features for speaker identification. IEEE Transactions on Acoustics, Speech, and Signal Processing, 2(23):176–182, 1975.

[SBJG86] D. Sumikawa, M. Blattner, K. Joy, and R. Greenberg. Guidelines for the syntactic design of audio cues in computer interfaces. Technical Report UCRL 92925, Lawrence Livermore National Laboratory, Livermore, California, United States, 1986.

[SBSR75] B. L. Smith, B. L. Brown, W. J. Strong, and A. C. Rencher. Effects of speech rate on personality perceptions. Language and Speech, 18(2):145–152, 1975.

[SCEM98] Y. Stylianou, O. Cappe, and E. E. Moulines. Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing, 2(6):131–142, 1998.

[SFG01] B. Suhm, B. Freeman, and D. Getty. Curing the menu blues in touch-tone voice interfaces. In Extended Abstracts on Human Factors in Computer Systems, pages 131–132. ACM Press, 2001.

[SG03] M. Schroeder and M. Grice. Expressing vocal effort in concatenative synthesis. In 15th International Congress of Phonetic Sciences, pages 2589–2592, 2003.

[SH93] A. L. Schwartz and M. L. Hardzinski. Ameritech phone-based user interface standards and design guidelines. Technical report, Ameritech Services, Inc., 1993.

[SHS95] R. M. Schumacher, M. L. Hardzinski, and A. L. Schwartz. Increasing the usability of interactive voice response systems. Human Factors, 2(35):251–264, 1995.

[SI03] P. Shajahan and P. Irani. Improving navigation in touch-tone interfaces. In Human Factors in Telecommunication (HFT), pages 145–152, 2003.

[SI04] P. Shajahan and P. Irani. Representing hierarchies using multiple synthetic voices. In 8th International Conference on Information Visualisation, pages 885–891. IEEE Computer Society, 2004.

[SI05a] P. Shajahan and P. Irani. One Family, Many Voices: Navigation Cues in Complex Hierarchies Using Multiple Synthetic Voices. International Journal of Speech Technology (Submitted), 2005.

[SI05b] P. Shajahan and P. Irani. Manipulating Synthetic Voice Parameters for Navigation in Hierarchical Structures. In International Conference on Auditory Display. Addison-Wesley, 2005.

[Slo79] D. I. Slobin. Psycholinguistics, chapter 3: Psycholinguistic constraints on the form of grammar, pages 63–72. Scott, Foresman and company, second edition, 1979.

[SN85] L. M. Slowiaczek and H. C. Nusbaum. Effects of speech rate and pitch contour on the perception of synthetic speech. Human Factors, 27(6):701–712, 1985.

[Spi97] M. F. Spiegel. Advanced database preprocessing and preparation that enable telecommunication services based on speech synthesis. Speech Communication, 23(1-2):51–62, 1997.

[Sum85] D. A. Sumikawa. Guidelines for the integration of audio cues into computer user interfaces. Technical Report UCRL 53656, Lawrence Livermore National Laboratory, Livermore, California, United States, 1985.

[Tat96] G. R. Tatchell. Problems with the existing telephony customer interface: The pending eclipse of touch-tone and dial-tone. In Conference

Companion on Human Factors in Computing Systems, pages 242–243. ACM Press, 1996.

[VA03] M. L. M. Vargas and S. Anderson. Combining speech and earcons to assist menu navigation. In Proceedings of the 2003 International Conference on Auditory Display. Addison-Wesley, 2003.

[voi] http://www.voicegenie.com.

[WKK95] C. Wolf, L. Koved, and E. Kunzinger. Ubiquitous mail: Speech and graphical user interfaces to an integrated voice/e-mail mailbox. In Interact, pages 247–252, 1995.

[YLM95] N. Yankelovich, G. Levow, and M. Marx. Designing speechacts: issues in speech user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 369–376. ACM Press/Addison-Wesley Publishing Co, 1995.