EFFECTS OF INTERFACE MODALITY ON SPEAKER PROSODICS

Julie Baca **Joseph Picone**

ABSTRACT

Displayless interface technology must address challenges similar to those presented by the problem of providing GUI access to visually impaired users. Both must address the issue of providing nonvisual access to spatial data. This research examines the hypothesis that such access places a cognitive burden on the user, which in turn will impact the prosodics, i.e. nonverbal aspects, of the user speech.

KEYWORDS: GUI access, displayless interfaces, prosodics

1.INTRODUCTION

In the past decade, the introduction of the graphical user interface (GUI) has profoundly altered the nature of human interactions with computer systems. While sighted users may prefer the more direct interaction of the GUI, this type of interface has presented significant limitations for users with visual impairments [1,2]. The explosion of the World Wide Web (WWW), of which the GUI is an integral component, coupled with the 1990 passage of the Americans with Disabilities Act (ADA), heightens the imperative to find better methods of accessing GUIs for users with visual impairments. Although commercial solutions continue to improve, certain fundamental problems remain unsolved.

Ironically, concurrent to the increasing prevalence of the GUI has been the development and use of ìdisplaylessî interface technology. This technology provides speech-only access for applications in which a visual interface cannot be used, such as telephone-based or mobile applications. Interestingly, this technology must address certain underlying issues common to that of GUI access for users with visual impairments, specifically how to verbally present data that is either inherently spatial in nature, such as geographical maps, or data that is presented through a visuospatial display metaphor. The research presented in this paper assumes that presentation of inherently spatial data through a verbal interface modality significantly increases the cognitive load for the user. Research surveyed in [3], conducted in diverse disciplines, including psychology, education, and human-computer interaction, supports this assumption.

Nonetheless, speech provides a viable alternative for many applications in which spatial data must be presented non-visually, particularly those requiring mobility. For example, Back Seat Driver, a navigational system developed at the MIT Media Lab[4] for taxi drivers in the Boston area, accepts voice commands from the driver and provides directional assistance through synthesized speech. The use of similar technology in a mobile navigational aid for visually impaired travelers in unfamiliar environments is investigated in [5]. The Soldierís Computer, a component of the Digital Battlefield designed to meet the needs of the modern soldier, offers another example of a speech interface which provides rapid portable access to map and directional information in time-critical situations [6].

GUI access researchers also argue that voice input should be included along with other input and output modalities as an option in providing nonvisual access to graphical and spatial data [2]. Used in a multimodal environment, it offers certain advantages by freeing the hands to be used for other tasks, such as accessing a tactile output device.

Regardless of the potential advantages of speech interfaces, widespread use of this technology will require the resolution of many human factors issues, as discussed in [7]. The study detailed in this paper addressed one issue in particular, speaker prosodics. Research, reviewed in [8], has examined the impact of various psychological and cognitive burdens on the prosodics of human speech, e.g. fundamental frequency (F0), the length and location of pauses, and speaking rate. This study examined the possible correlation between the increased cognitive load produced by a strictly verbal presentation of spatial data and the effects of this burden on the prosodics of the userís speech. A better understanding of this issue could contribute to the development of more robust interfaces for applications requiring verbal access to spatial data. Knowledge gained from investigating this issue could be used, for example, to improve prosodic pattern detection algorithms. The limitations of existing algorithms which use only limited acoustic cues, i.e., primarily F0 features, are noted in [9]. It is argued in [9] that additional acoustic cues including pauses and duration, should be used for more robust prosodic pattern detection. Any correlation found between the additional cognitive load induced by displayless navigational interfaces and the prosodics of the userís speech could lend support to this argument.

2.RESEARCH APPROACH

This investigation tested the hypothesis that the prosodics of the userís speech produced during interactions with a displayless

navigational system would differ significantly from that produced in interactions with a multimodal navigational system.

Testing the hypothesis entailed analyzing recordings of user speech interactions with a prototype speech-based interface to a map database. The database contains details of the physical and spatial layout of the U.S. Army Corps of Engineers (USAE) Waterways Experiment Station (WES). Two experiments were conducted in which subjects were given a series of increasingly complex navigational tasks. Both subjects with and without visual impairments participated in the study, although no formal comparisons were made between the two categories since the conditions in the second experiment differed for each. In the first experiment, subjects used only speech to perform the tasks, while in the second, subjects also used either a graphical or tactile interface, depending on their visual capacity. All user speech was recorded during both experiments, post-processed for prosodic content and then analyzed to determine differences between the two interface conditions.

Before discussing results of the experiments, it is necessary to provide more details on the design of the prototype due to its potential effects on the outcome of the study. Also, further details of the prosodic labeling scheme as well as subject selection are given.

2.2Speech-Based Prototype

The speech interface provides access to the program, ëWES Auto Travelí, which uses the information in the map database to assist first-time visitors in navigating the station via spoken instructions. In both experiments, subjects were asked to play the role of first-time visitors to the station and to use the program for assistance in getting from one location on the station to another. After hearing a verbal description of the overall station layout, subjects were given a starting point and a destination for each task and then asked to use the program to determine how to get to the destination. Although the program defaults to giving instructions along a precomputed driving route, in both experiments, subjects were asked to customize the route for walking by issuing various queries and commands.

As stated, in the first experiment, subjects used only a speech interface with no additional modalities, visual or tactile. All interactions between the user and the system were conducted through speech as shown in Figure 1.

The primary hardware and software modules of the system are also shown in Figure 1. To summarize the flow of control between the individual modules, the Speech Input Module receives the spoken request from the user, produces a hypothesis of the utterance, and passes this to the Natural Language (NL) Input Module. The NL Input Module then parses the request and converts it to a database query which it presents to the GIS database. The NL Output Module accepts the result of the query, phrases it in natural language and transmits this to the Speech Output module. Finally, this module presents the natural language response to the user via synthesized speech.

In the second experiment, subjects used an interactive touch screen display of a map of the station in addition to speech. Certain key areas were identified as selectable on the map. The map was presented visually for sighted users and tactilely for users with sight loss; thus the selectable areas were highlighted with visual and tactile markings respectively. Users could touch the selectable areas on the map and hear short descriptions of the areas. They could also query in the same manner as in the first experiment since the enhanced prototype only added multimodal access to the base prototype, but did not replace any existing functionality.

The NL and Speech Modules were designed and developed Iteratively beginning with a technique similar to the ìWizard of Ozî approach defined in [10] and continuing through a series of usability tests for refinement. The wizard sessions differed from the standard technique in that participants typed their interactions with the mock application rather than using a speech surrogate. Nonetheless, the sessions provided a means of eliciting some initial user input in constructing the application grammar and vocabulary.

2.2.1NL Module

The NL parser uses a semantic grammar and some limited contextual knowledge of previous queries to parse and translate requests into database queries. The grammar allows both fixed commands and freely formed natural language queries. Fixed commands are used for such actions as moving along the precomputed route, e.g., ìContinue Forwardî, which causes the program to continue forward one segment in the route, as well as asking for help, e.g., ìWhat can I ask?î. Any type of query to the database can be phrased in natural language, e.g., ìHow far is it to Headquarters?îor ìIs there a sidewalk on this road and is the

traffic heavy here?î

Though initially determined from the wizard sessions, the phrasing of the fixed commands was refined through usability tests. For example, users initially preferred the command, ìGo Forwardî to continue forward one segment on the route. However, this utterance was consistently misrecognized for several speakers in the usability tests due to their lack of articulation of the ìfî in ìforwardî. This produced a sound similar to ìgo-owardî, which the recognizer matched with ìroadî in the phonetic dictionary. The phrase ìcontinue forwardî reduced the effects of coarticulation and was accepted by users as a reasonably intuitive substitute.

2.2.2Speech Input Module

The Speech Input Module required speaker-independent, continuous speech recognition. The research question dictated the need for continuous recognition since no meaningful prosodic analysis could be conducted on isolated words. Speaker-independence was necessary due to the size of the sample population; training to each speaker was not feasible.

The Entropic HTK speech recognition software met these requirements. Implementing the recognizer for 'WES Auto Travel', a small vocabulary application, entailed assembling the language model, phonetic dictionary, and acoustic models in the HTK environment. Development of the language model and phonetic dictionary proved a straightforward translation from the semantic grammar and lexicon of the NL Module. Producing the phonetic dictionary involved a more extensive process of translating phonetic models taken from the Carnegie Mellon University (CMU) triphone pronunciation dictionary to the DARPA phonetic models used by HTK. This process is described in [11]. The final language model contained approximately 150 production rules with a perplexity of ...., the vocabulary less than 200 words.

Other Design Issues

Misrecognition errors present a disadvantage in the use of speech interfaces, causing frustration for users and making it difficult for them to form mental models of the system. Since this could impact the results of the investigation, error-handling strategies were important. As recommended in [7], a minimal confirmation strategy was used, i.e., the program confirmed the userís request only when the consequences of an error could cause significant inconvenience to the user, for example, when requesting an alternate route.

Although response time is also an important issue for speech interfaces, the nature of the tasks in the experiments, i.e., tasks requiring long periods of silence for cognitive planning rather than dictation-type tasks, lessened its significance. In usability tests conducted prior to the experiments, users most often described their reaction to the speed of interface as ìdidnít noticeî or ìslow, but acceptableî. Subjects participating in the experiments reacted similarly.

2.2.3 Speech Output Module

Avoiding auditory overload presented perhaps the most significant issue in the design of the Speech Output Module. Measures taken to achieve this design goal included minimizing the use of auditory lists and speaking directions in brief segments which the user could request to be repeated by either pressing a key or saying ìRepeat Instructionî. It was also important to give users control of the synthesizer speaking rate since users with visual impairments are typically more experienced in listening to synthesized speech than sighted users. The synthesizer used for the prototype, CentigramTruVoice, provided this capability.

2.2.4Multimodal Interface

Although implementation of the graphical and tactile display differed, design of the underlying interface adhered to the same objectives of offering completeness while maintaining simplicity. Most importantly, these objectives motivated the selection of the map for the display, which was designed by a graphic artist for station visitors, rather than a detailed drawing produced from the original database for WES engineers and maintenance personnel. This did not significantly reduce the number of selectable areas on the map and it provided a much more intuitive view for users unfamiliar with the station. Further details of the graphical-audio display are given in [11]. Design of the tactile display proceeded from that of the graphical, with the main distinction that the principle of simplicity was even more critical. Tactile maps cannot provide the same level of detail as visual maps in a meaningful way. Thus, guidelines given in [12] were followed to produce the map for the tactile display.

.3Prosodic Analysis

After the experiments were completed, the speech data collected was transcribed and labeled using the Tones and Break Indices (TOBI) transcription system [13]. Prosodic features analyzed included pauses, breaths, boundary tones, preboundary lengthening, and speaking rate changes. TOBI was selected for the transcription since it provides coverage of multiple aspects of prosody, containing four transcription tiers, an orthographic tier, a tonal tier, a break index tier, based on the ìbreak indicesî defined in [14], and a miscellaneous tier. More details on TOBI can be found in [13].

2.4Subject Selection

While data collected from subjects with and without sight loss was treated separately, certain criteria applied to all subjects, regardless of visual capability, including age, education, and amount of prior computer usage. Subjects were required to be of adult age, i.e., at least 18 years of age or older, and possess the equivalent of a high school education, i.e., high school diploma or General Equivalency Diploma. Also, subjects were required to be using computer software to perform tasks on a regular, i.e., weekly or monthly basis. This ensured subjects possessed at least a minimal level of comfort in computer usage.

3.RESULTS

The experiments were conducted over the course of three months at various universities and rehabilitation agencies, located in Mississippi, Arkansas, and Louisiana. Over 90 subjects participated in the experiments, including 33 sighted subjects and over 60 subjects with visual impairments, with an approximate 50/50 distribution between those with congenital and adventitious vision loss.

Results for sighted subjects showed significantly more pauses, i.e., alpha less than 0.05, which did not occur at an intonational or intermediate phrase boundary (denoted by the symbol ë2pí in TOBI) in the displayless condition than the multimodal condition. These types of pauses were also significantly longer in the displayless condition. In addition, average frequencies were significantly higher and the number of intonational phrase boundaries ending in a low tone (denoted 'L%' in TOBI) were significantly greater in the displayless condition.

Results for subjects with sight loss were similar with certain exceptions. Results for subjects with congenital sight loss showed not only significantly more hesitation pauses, but also significantly more and significantly longer pauses occurring at phrase boundaries (denoted '3p' in TOBI). Also, the number of intermediate phrase boundaries ending in a high tone (denoted 'H-' in TOBI) were significantly greater while the number of low intonational phrase boundaries were not. More details on the tonal results as well as pauses are given in [5].

...In interpreting the results,..there is strong evidence in the data that hesitation pauses are increased, for all categories of users, in the displayless condition. This ...indicates a likely increase in the amount of cognitive effort and planning required to use the displayless navigational interface. Further, hesitation pauses tend to increase the number of misrecognition errors made by the system, which in turn negatively impacts the level of user satisfaction with the interface. These issues must be addressed for these interfaces to gain user acceptance.

Also, while the data from users with and without sight loss was treated separately, differences found in the tonal data, particularly for users with congenital vision loss, are interesting. This aspect of the data should be examined further to determine if it reflects differences in how this category of users adapts to these interfaces.

Finally, both the data regarding hesitation pauses and the tonal data indicate the possible need for multiple acoustic cues in prosodic pattern detection, lending support to the arguments in [9].

To conclude, this study has demonstrated significance differences in the prosodics of user speech when using a displayless versus a multimodal navigational system. ...More generally, it revealed potential problems in the use of displayless navigational systems, particularly in the recognition component. It has also shown possible differences in the way in which these interfaces are accepted by sighted users and users with visual impairments. ....All of these issues warrant further examination....for the successful...deployment...of displayless navigational interfaces.....

ACKNOWLEDGMENTS

This research was partially funded by the USAE WES in Vicksburg, MS.

REFERENCES

1.Boyd, L.H., W.L. Boyd, J. Berliss, M. Sutton, and G.C. Vanderheiden. The paradox of the graphical user interface: Unprecedented computer power for blind people. *Closing the Gap 14* (October):24-25, 60-61, 1992.

2.Vanderheiden, G.C., and D.C. Kunz. Systems 3: An interface to graphic computers for blind users. In *Proceedings of the 13th Annual Conference of RESNA held in Washington, D.C. 20-24 June 1990*, 150-200, 1990.

3.Baca, J. Displayless access to spatial data: Effects on speaker prosodics. An unpublished dissertation, Mississippi State University, iDepartment of Computer Science, to be published May 1997.

4.Davis, J.R. and C. Schmandt. The back seat driver: Real time spoken driving instructions. In *Vehicle Navigation and Information Systems*, 146-150, 1989.

5.Loomis, J.M., R.G. Golledge, R.L. Klatzky, J. Speigle, and J. Tietz. 1994. Personal guidance system for the visually impaired. In ASSETS ë94, *The First Annual ACM Conference on Assistive Technologies, October 31-November 1, 1994, Los Angeles, CA*, 85-91, 1994.

6.Weinstein, C.J. Applications of voice processing technology. *Voice Communication Between Humans and Machines*, National Academy Press, Washington, D.C., 1994.

7.Kamm, C. User interfaces for voice applications. *Voice Communication Between Humans and Machines*. National Academy Press, Washington, D.C., 1994.

8.Scherer, K.R. Speech and emotional states. *Speech Evaluation in Psychiatry*, 189-220, New York:Grune-Stratton, 1981.

9.Wightman, C.W. and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing* 2(4):469-481, 1994.

10.Nakatsu, R. And Y. Suzuki. What does voice-processing technology support today? *Voice Communication Between Humans and Machines.* National Academy Press, Washington, D.C., 1994.

11.Ngan, J. and J. Picone. Issues in generating pronunciation dictionaries for voice interfaces to spatial databases. In *Proceedings of IEEE Southeastcon ë97, April 12-14, 1997, Blacksburg, VA*, 97-99, 1997.

12.Barth, J. *Tactile Graphics Guidebook*, American Printing House for the Blind, Louisville, Kentucky, 1983.

13. K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, TOBI: A standard for labelling English prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP), Banff, Alberta, Canada*, 867-870, October 1992.

14.Price, P.J., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *Journal of Acoustical Society of America* 90:2956-2970, 1991.