Can VoiceScapes Assist in Menu Navigation?

Steffen Werner, Christopher Hauck, Nicholas Roome, Connor Hoover and Daniel Choates
Cognition & Usability Lab, University of Idaho

Providing better information access to blind users is an important goal in the context of accessible interface design. Similarly, designers of user interfaces benefit from alternative interface techniques for usage scenarios in which visual (graphical) interfaces are either not possible or suboptimal. In our study we compared a traditional serial aural presentation of menu items to a new simultaneous aural presentation of up to seven menu items. These continuously present VoiceScapes allow the user to actively scan the auditory display to find the most appropriate command. While VoiceScapes are more difficult and attentionally more demanding than other formats of presentation, extended use might allow experienced users to more efficiently navigate complex menu hierarchies. A first pilot experiment with 13 sighted participants presented here tested the basic viability of this approach.

INTRODUCTION

Information access is one of the hallmarks of our modern society. However, access to an abundance of digital information has been largely designed for use via visual/graphical user interfaces. Blind and visually impaired users often have to rely on assistive technologies, such as screen readers, to access information that is designed for a sighted audience. In addition, visual presentation of information might not be advisable, feasible, or desired in certain usage scenarios even for sighted users – for example while driving, under situations of visual glare, for telephony applications, or conditions of dark-adaptation. In this paper we introduce VoiceScapes as an extension of existing research on auditory interfaces and start to explore whether VoiceScapes show potential for an efficient navigation of an information space through multiple levels of menus.

Currently screen readers are a crucial technology used to access digital information for non-sighted users (Borodin, Bigham, Dausch, & Ramakrishnan, 2010; WebAIM, 2014). Assistive technologies like iOS's Voice–Over, or independent MS Windows-based software tools such as JAWS (Freedom Scientific, n.d.) and NVDA (NV Access, 2013) represent indispensable general solutions to the problem of accessing web-based content, as well as interacting with largely graphical user interfaces. However, anyone who has ever tried to use these web-browsers can attest to the slow and frustrating user experience these tools provide for novice users. After extensive practice users can develop strategies to employ screen readers more efficiently – by speeding up synthesized sound output (up to 500 wpm), keyboard shortcuts to access known 'landmarks' within a site or program, or specific task-based strategies (Borodin et al., 2010).

Inherent shortcomings of current screen readers have been recognized by many researchers. Strain, McAllister, Murphy, Kuber, and Yu (2010), for example, focused on the lack of spatial awareness in traditional screen readers and proposed the extension of screen readers through a multimodal augmentation that informs the user about their current location within a grid-like structure. In addition, specific mark-up standards like W3C Accessible Rich Internet Applications (ARIA) (W3C, 2014) attempt to provide screen readers with more information about the structure of documents. A much more dramatic proposal was envisioned by IBM's extreme blue group. Their concept of a 'conversational internet' would relieve blind users from having to navigate a visually biased, graphical interface and instead rely on natural language interfaces to create a better user experience that can be applied to a broad range of websites (Gomez, Lane, & Dailly, 2013).

In conjunction with these general solutions, an alternative path for non-visual access, both on the web as well as for desktop or mobile computing, relies on specific aural interfaces optimized for particular applications, infor–mation displays, or content domains. An early example of this approach is Emacspeak, a speech output interface to Emacs, the UNIX family of text editors (Raman, 1996). It was designed to give blind or visually impaired users the ability to receive feedback via synthesized speech. By being directly integrated into the program, Emacspeak can take advantage of context-specific information to optimize its output.

Content-specific solutions like the provision of an auditory glance at algebraic expressions (Stevens, Brewster, Wright, & Edwards, 1994) and information-specific solutions, like the sonification of graphs (Walker & Mauney, 2010) have been proposed to improve access of blind users to specific content. Even in the gaming world new approaches to audio-only, 3D game engines have been successfully employed (e.g., www.papasangre.com).

In contrast to the approaches mentioned above, our project attempts to improve the selection and navigation performance of aural interfaces. One of the main impedements of speech-based interfaces is the *serial nature of information display*. One way to solve this problem, as mentioned above, rests on the acceleration or compression of speech (Borodin et al., 2010) or the introduction of sometimes unintelligible, highly compressed speech icons (so called spearcons, Walker, Nance, & Lindsay, 2006). While speeding up presentation is one possible solution, the presentation of multiple auditory objects or streams that can actively be attended to is a second way to present more information in the same time frame. Beginning with Cherry

(1953), researchers have investigated the ability of human listeners to attend to more than one auditory (or speech) input at a time. Schmandt and Mullins (1995) presented AudioStreamer, a system presenting multiple audio recordings simultaneously. More recently, Guerreiro and Goncalves (2014) tested the ability of blind users to search for relevant content in 2, 3, or 4 concurrent talkers that were presented through spatially separated channels. Their results showed that approximately 87% of their blind users were able to identify all relevant information from two simultaneous talkers, while a majority of users was still able to listen for information across three talkers. Using spatial separation was reported as one of the most important features of their display.

A slightly different approach was taken by Parente (2006) who used four distinct assistants' voices spatially distributed around the listener to offer a new type of task-oriented audio display. Users would interact with his Clique system either through keyboard or voice input.

Sodnik, Jakus, and Tomazic (2011) created an auditory display to assist in hierarchical menu navigation, using both spatial location of the spoken menu items and pitch of background music as navigational aids (aligned horizontally or vertically for menu navigation). To assure intelligibility they limited the number of simultaneous sound sources to three. Their main finding was that participants found it difficult to attend to multiple audio sources and that the auditory interface required more attention than traditional interfaces. Prior to their study, Frauenberger, Putz, and Holdrich (2004) had used auditory icons in a virtual room to study navigation with up to six 'menu' items.

Building on these previous findings, the purpose of this study is to investigate whether a simultaneous presentation of up to seven menu items can provide a viable auditory interface alternative to serially presented menu lists. We conceive of a 'VoiceScape' as the simultaneous presentation of multiple streams of speech, each spoken by a separate voice. To aid in the discriminability of the different streams we envision VoiceScapes to be spatialized, with a particular voice (speaker) being associated with a particular spatial location or direction/azimuth. For an enduring auditory presentation of short elements such as menu items we simply repeat the specific item multiple times by looping the recorded speech elements.

One of the potential benefits of simultaneous presentation lies in the possibility of active browsing of the VoiceScape instead of a passive reception of serially presented items and waiting for the right one to come along. However, based on previous results, we expect the cacophonic presentation of up to seven audio streams to present a demanding task for the listener. The main question for our pilot study, therefore, is to establish whether listeners are able to distinguish such a large number of auditory streams and to elicit user feedback about the viability of such a demanding display type.

Given that one of the main motivations for VoiceScapes is the use for blind or low-vision individuals it is of course critical to eventually test this interface concept on the target population – especially given the vast experience of blind users with auditory displays. However, for this pilot study we had to rely on a convenience sample of undergraduate participants.

METHOD

Participants.

A total of 15 sighted participants took part in this study. Two participant's data was lost due to recording failure for a total of 13 complete data sets. Participants were undergraduate students at a university in the Northwest. The age range of participants was 18-23 years of age.

Materials

For the construction of the auditory interface we selected eight top-level menus from MS Word and Adobe Photoshop and reduced the number of items within each menu to seven items (see Figure 1). We took care that both often-used and infrequent menu items as well as single-word and multi-word menu items were present.

Sample Menus and Menu Items Used in the Study

MS Word Insert Menu	Adobe PS Edit Menu
1. Page numbers	1. Content-aware scale
2. Date and time	2. Perspective warp
3. Autotext	3. Free transform
4. New comment	4. Auto-align layer
5. Footnote	5. Define pattern
6. Cross reference	6. Color settings
7. Watermark	7. Convert to profile
MS Word Format Menu	**Adobe PS Image Menu**
1. Paragraph	1. Reveal all
2. Document	2. Duplicate
3. Bullets and numbering	3. Apply image
4. Borders and Shading	4. Calculations
5. Text direction	5. Variables
6. Change case	6. Apply Data set
7. Autoformat	7. Analysis

Figure 1: Sample materials.

Following Brungart, Simpson, Ericson, and Scott (2001) we had seven speakers read the menu items (three female and four male speakers, age range 13-43 years) to increase the discriminability of the audio stimuli. We approximately matched each speaker's cadence by compressing or stretching their sound files without altering their pitch. We also attempted to match speakers' overall volume level. For VoiceScape presentation, each menu item was then assigned a specific location and speaker (1 through 7, see Figure 3, left panel). Location and speaker were perfectly correlated so that each speaker's voice was always presented from the same location. Each menu item was accompanied by the corresponding number (1-7) of the direction the sound was

presented from. The numbers were spoken by the same voice as the corresponding menu item. To accommodate the variable durations of different menu items and to introduce phase shifts between the item presentations we copied each sound multiple times to create VoiceScape sound files approximately 15 seconds in duration that were looped during the experiment. Different menu items were thus repeated at different frequencies (see Figure 2).

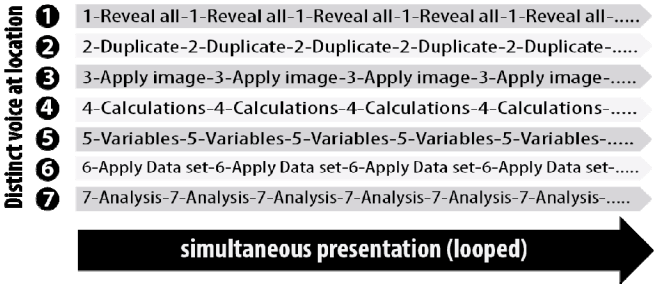

Figure 2. Example of a VoiceScape menu presentation. Each row shows the repeated presentation of the menu item and the associated number. Due to the different durations of the spoken digits and menu items the temporal presentation across different menu items drifts over time.

In addition to spatialized VoiceScape stimuli we created two alternative serial audio presentations based on the recordings of an additional speaker to avoid confusion about the assignment of speakers to location. Regular speed serial presentation stimuli included the concatenation of the seven auditory stimuli used in the corresponding VoiceScape presentation, together with their respective speaker number. The average duration of these serial presentations was 13.3 seconds. Fast serial presentations were created by compressing these audio files to 50% duration while leaving pitch constant (average duration 6.6 seconds).

The sounds were presented via seven Bose Companion 2 Series III Multimedia speakers arranged in a semi circle around the participant (30° separation, see Figure 3 left Panel) to maximize spatial separation and with it intelligibility (Drullman & Bronkhorst, 2000). In the serial sound presentation conditions only the center speaker was used. Audio files were prepared in Audacity.

Procedure

Participants were informed about the purpose of the study and were instructed to verbally identify all the menu commands and their respective numbers while listening to the looped audio stimuli. This somewhat unusual task allowed us to test the ability of participants to discriminate and identify the menu items in a straightforward manner. The initially planned recording of manual response times proved impossible given the technical limitations of our experimental software's sound controller.

Identification times and accuracy were measured based on voice recordings of the participants. There were four different conditions. Participants either listened to VoiceScapes with increasing complexity (starting with only three speakers turned on, then five, then seven) or in decreasing complexity (VoiceScapes with seven speakers and reducing to five and three). The purpose for this manipulation was to test the intelligibility of VoiceScapes at different levels of difficulty and to allow participants in the 3-7 condition to identify three of the sounds before moving on to a higher-order VoiceScape.

In the serial presentation conditions participants either listened to the regular speed or fast serial order stimuli. The specific menu to be tested during a particular trial was randomly assigned to sound condition (without replacement). The test was broken into two blocks with the first block exposing the participants to each condition once, and the second block repeating the same order of conditions but with a new set of menus.

Participants were asked between blocks about their subjective assessment of the difficulty of the task, followed by a systematic subjective evaluation of the task after the conclusion of the second block. A modified version of the System Usability Scale (SUS, Brooke, 1996) was presented at the end of the study to assess general usability assessments from the participant's perspective and their judgment of feasibility for a non-sighted user.

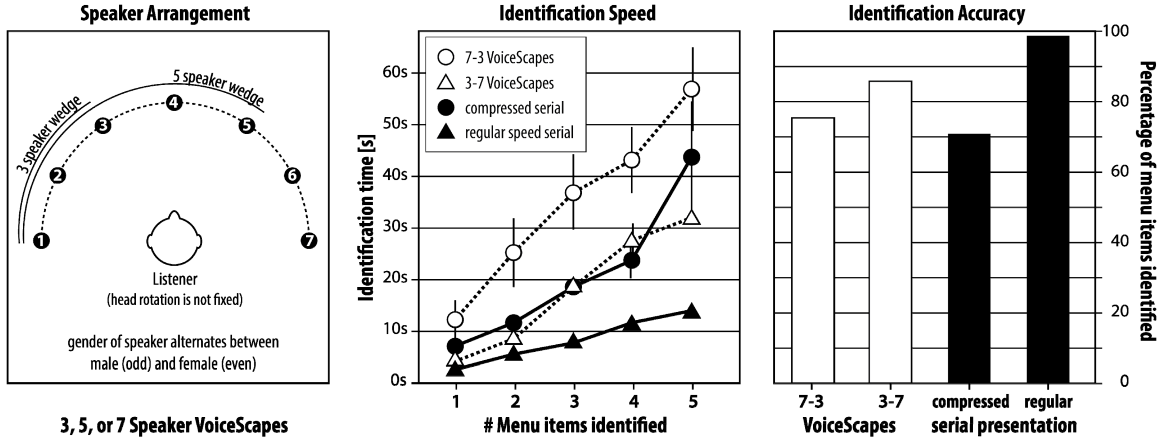

Figure 3: Speaker arrangement (left), identification times (center), and identification accuracy (right). Error bars (center) indicate SE_{mean} whenever the error is larger than the symbol depicting the mean.

RESULTS AND DISCUSSION

Overall accuracy data showed a clear benefit of the normal speed serial presentation over all other conditions (98.0% of menu items correctly identified, see Figure 3, right panel). Condition 3-7 with increasing VoiceScapes (from three to seven total speakers) performed better (87.6%) than the decreasing VoiceScape condition (76.5%), which is not surprising because of the reduced auditory interference for conditions with only three auditory channels compared to seven channels. Surprisingly, the compressed serial presentation performed worst, with only 70.3% of menu items being accurately identified.

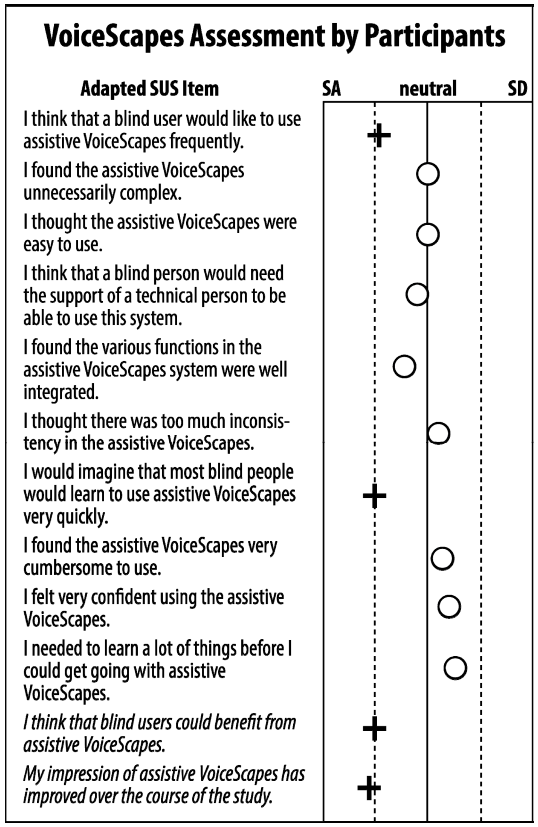

Figure 4: User evaluation of VoiceScapes

To compare identification times across conditions we plotted the identification time for the first identified menu item, the second, and so on, up to the fifth item (see Figure 3, central panel). Since some participants did not identify more than five items we did not include higher-order items in this analysis. From the graph it is quite clear that the menu items are identified quickest in the normal serial presentation format, while it takes four times as long to identify the items through the seven-channel (7-3) VoiceScape. The 3-7 increasing VoiceScapes and compressed serial presentation were very similar in their respective identification speeds.

While participants in our study found the VoiceScapes challenging, the adapted SUS scale suggests that most users found the level of difficulty acceptable and thought that the system would be of benefit to non-sighted users, indicated by the positive evaluations on these items (see figure 4). While we obviously don't take our sighted participants' evaluation of the usefulness of VoiceScapes for blind users at face value, the overall positive assessment might indicate that the system is not too difficult to interact with.

Qualitative data was collected to assess the perceived usefulness of and attitude towards the VoiceScape system. Overall, participants perceived the VoiceScapes as a positive experience. One participant stated that performance could improve with practice, and indeed, even within these short trials users began developing creative strategies for navigating the VoiceScape. For instance, focusing on one speaker or one voice at a time were both utilized. Whether or not these are optimal navigation strategies or whether there are significant long-term practice effects needs to be further investigated.

Compressed serial presentation was almost unanimously disliked, which is not surprising given the poor performance (70.3% accuracy) of this condition. The regular speed serial presentation and three speaker VoiceScape were most liked and perceived as the most useful and practical. Accuracy data also supports these claims. However, given the non-existing experience of our sighted participants with compressed speech it is unclear whether this is a valid comparison – as Borodin et al. (2010) reported, many blind users of speech readers are quite comfortable with compressed speech. The use of compressed natural speech (vs. synthetic speech) might also raise issues of external validity of our results regarding compressed speech.

GENERAL DISCUSSION

As is clear from these results, menu items in Voice-Scapes are harder to understand and slower to identify than menu items presented in a traditional serial manner at normal presentation speeds. This was not surprising. Participants also reported, consistent with the findings by Guerreiro and Goncalves (2014), that auditory VoiceScapes demanded a lot of their attention and were initially perceived as overwhelming. However, participants also reported that the process of identifying menu items became easier as the study progressed.

An important consideration for future research is the familiarity of users with a particular menu and the VoiceScape setup. While novice users will have a difficult time identifing menu items when being presented seven options simultaneously, an experienced user might be able to direct their attention to the relevant spatial segment and familiar voices of the VoiceScape to confirm menu options and their associated actions (e.g., key presses). This should provide VoiceScapes with a temporal advantage, given that the information is constantly present (unlike serial presentation). In this way, VoiceScapes would serve as a memory aide in a confirmatory role after sufficient experience with an information space has been established.

A limitation of this study lies in the preliminary nature of the speech stimuli used in this study. Some participants felt

that the volume of the speakers was not matched well enough and that some speakers were hard to understand. Improvements of these issues could make VoiceScapes easier to understand and use. Similarly, the choice of using 50% compressed natural speech in the fast serial presentation conditions, while motivated by previous research, could shine too negative a light on accelerated serial speech. More in-depth research would be required to validate these findings.

Two promising lines of future research with VoiceScapes are the incorporation of an active spatial modulation of sound by the listener – so that while the entire VoiceScape is continually present, certain speakers/directions can be amplified – similar to the flashlight model of attention. A second line of research that we are currently undertaking is to investigate more fully how repeated exposure to the same VoiceScapes will enable users to quickly identify the relevant information based on recognition memory within a VoiceScape.

Finally, it is critical that after an initial pilot testing phase a prototype of a VoiceScape environment is tested within the target population of blind and visually impaired users. Only then can we make a valid assessment of the shortcomings of compressed serial presentation vs. the potential benefits (and drawbacks like the increased attentional demand) of simultaneous, spatialized presentation of information.

REFERENCES

Borodin , Y., Bigham, J.P., Dausch, G., & Ramakrishnan, I. V. (2010). More than meets the eye: A survey of screen-reader browsing strategies. *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A), April 26-27, 2010, Raleigh, North Carolina.*

Brooke, J. (1996). SUS: a "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. *Usability Evaluation in Industry*. London: Taylor and Francis.

Brungart, D.S., Simpson, B.D., Ericson, M.A., & Scott, K.R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J. Acoust. Soc. Am. 110(5),* 2527-2538.

Cherry, C.E. (1953). Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America 25(5),* 975-979.

Drullman, R. & Bronkhorst, A.W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. J. Acoust. Soc. Am. 107(4), 2224-2235.

Frauenberger, C., Putz, V., & Holdrich, R. (2004). Spatial auditory displays: A study on the use of virtual audio environments as interfaces for users with visual disabilities. In *Proceedings of the Seventh International Conference on Digital Audio Effects (DAFx'04), Naples, Italy,* 384-389.

Freedom Scientific (n.d.). BLINDNESS SOLUTIONS: JAWS®. Retrieved March 9, 2015, from http://www.freedomscientific.com/Products/Blindness/JAWS

Gomez, N., Lane, D., & Dailly, J. (2013). The conversational internet: Creating a natural language interface for visually impaired people to converse with the web. *W4A2013 – Communications, May 13–15, 2013, Rio de Janeiro, Brazil*

Guerreiro, J. & Goncalves, D. (2014). Text-to-speeches: Evaluating the perception of concurrent speech by blind people. *ASSETS'14, October 20-22, 2014.*

NV Access. (2013, January 11). Retrieved March 9, 2015, from http://www.nvaccess.org/

Parente, P. (2006) Clique: a conversant, task-based audio display for GUI applications. *ACM SIGACCESS Accessibility and Computing 84:* 34-37.

Raman, T.V., Emacspeak - direct speech access (1996). *Proceedings of the second annual ACM conference on Assistive technologies. ACM: Vancouver, British Columbia, Canada.*

Schmandt, C. & Mullins, A. (1995). AudioStreamer: Exploiting simultaneity for listening. *CHI'95 short paper, May 7-11, 1995,* 218-219.

Sodnik, J., Jakus, G., & S Tomažič (2011). Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. International Journal of Human-Computer Studies 69 (1), 100-112.

Stevens, R.D., Brewster, S.A., Wright, P.C., & Edwards, A.D.N. (1994). Design and evaluation of an auditory glance at algebra for blind readers. *Auditor Display: The Proceedings of the Second International Conference on Auditory Display, 1994*, 21-30.

Strain, P., McAllister, G., Murphy, E., Kuber, R. & Yu, W. (2007). A grid based extension to an assistive multimodal interface. In *Extended Abstracts of on Human Factors in Computing Systems, CHI'07, San Jose, USA,* 2675 – 2680.

W3C (2014). Accessible Rich Internet Applications (WAI-ARIA) 1.0. Retrieved March 9, 2015, from http://www.w3.org/TR/wai-aria/

Walker, B. N. & Mauney, L. (2010). Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. *ACM Transactions on Accessible Computing 2, 3*. 12:1-16.

Walker, B. N., Nance, A., & Lindsay, J. (2006). Spearcons: Speech-based earcons improve navigation performance in auditory menus. *Proceedings of the International Conference on Auditory Display (ICAD 2006)*, London, England (20-24 June). pp. 63-68.

WebAIM (2014). Screen reader user survey #5 results. Retrieved March 9, 2015, from http://webaim.org/projects/screenreadersurvey5