Research Report

Designing Audio Interfaces for Election Systems:

A review of the research and recommendations

October 1, 2015

Caitlin Rinn
Nancy Frishberg
Whitney Quesenbery
Center for Civic Design

hello@civicdesign.org
civicdesign.org

Designing Audio Interfaces for Election Systems

In 2002, the implementation of the Help America Vote Act resulted in a major shift in the accessibility of elections. Old punch-card and mechanical lever-based machines were replaced with new voting systems, and for the first time, many voters with disabilities were given the option to vote independently through the use of an audio ballot.

Audio ballots greatly increase access, privacy, and independence for voters with visual, dexterity or mobility, and cognitive disabilities. Voters with low vision or blindness are able to use a voting system like a screen reader, so that all instructions and contests on the ballot can be spoken aloud to them. Voters with moderately impaired vision, low literacy, or cognitive disabilities can also benefit from an audio ballot as audio provides supplementary assistance when they have difficulty reading the visual interface. Voters with dexterity or mobility issues may find it helpful to hear the ballot rather than having to position themselves in front of a visual screen.

Yet despite these advantages, there is little research into making audio features of ballots and other election systems accessible, learnable, and usable. As a result, there are few good guidelines for the design of audio features that can help make them as effective and usable as the visual interface, and even fewer specifically applicable to voting.

While some work is being done to allow for people to vote from the convenience of their home, where they have the use of their own personal assistive technology using an electronic system to mark a ballot, the majority of accessible voting in the U.S. still relies on voting systems at a polling place set up as a "self-contained system" or kiosk.

The goal of this review in late 2015 is to gather pertinent audio interface research from a variety of domains of application, in the hopes that this wide-ranging survey can lead to improving the accessibility and usability of audio interfaces in election systems.

Contents

Types of audio used in an interface

A variety of interactive systems use audio either as a primary of supplementary method of communicating information to users. Audio can provide feedback on an action, alert or warn users, or act as an alternative interface to more typical visual interfaces. This review will cover several types of audio, with speech being the primary focus as it is an essential component to the audio interactions of a kiosk-style voting system.

While some audio interfaces have a capacity for speech as both an input and output method, the review will only discuss characteristics of speech output, as speech input is not commonly found in voting systems due to potential privacy risks and interference from ambient sound[1]. In addition, the Election Assistance Commission's Voluntary Voting System Guidelines 1.1 (VVSG) [2] prohibits the use of speech as the only way for a voter to provide input to a voting system operation[3].

Speech output

There are two types of speech that can be used in an audio interface, depending on how the audio is produced: **natural** or **synthetic speech:**

- *natural speech* is speech that is pre-recorded by a human
- *synthetic speech*, also called text-to-speech or TtS, is speech that is artificially generated by a computer program. Synthetic speech can be created using either **formant** or **concatenative** synthesis.

The difference between formant and concatenative synthesis is how the sounds are formed (Neerincx, Cremers, Kessens, van Leeuwen, & Truong, 2008).

Early synthetic speech used formant synthesis, where computers simulate the sounds of a human voice word-by-word. This resulted in robotic sounding speech that was far less intelligible[4] than natural speech (Anderson, 2013: Neerincx et al., 2008). The reduced intelligibility of this type of synthetic speech is caused by its comparatively low fidelity and lack of prosodic cues

---

[1] Speech input has been integrated into the Prime III voting system in a way that does maintain privacy. See section: Voting Systems Designed for Universal Usability, Prime III

[2] The VVSG is issued by the U.S. Elections Assistance Committee in accordance with the Help America Vote Act (HAVA)

[3] VVSG 1.1 - 3.3.9. "Speech not required. The voting system shall not require voter speech for its operation."

4 Speech intelligibility is an assessment of how well users are able to correctly identify individual words that are spoken.

(e.g. changes in word duration and pitch through phrases and sentences) (Papadopoulos, K., & Koustriava, E. 2015; Papadopoulos, Koutsoklenis, Katemidou, & Okalidou, 2009; Roring, Hines, & Charness, 2007).

Newer speech programs use concatenative synthesis[5], which improves the intelligibility and naturalness of synthetic speech by sampling snippets of sounds from human voice recordings and stringing them together to create words and then sentences (Neerincx et al., 2008).

Concatenative synthetic speech has become a popular choice for many audio interfaces, since it is inexpensive (as it doesn't require a voice actor), can easily adjust to changes in menu options or text, and is relatively easy to understand. Recent improvements to this technology has also allowed for increased expression in synthetic speech.

For example, IBM's Expressive TtS Engine can produce synthetic speech in one of four speaking styles: neutral declarative, asking a question, showing contrastive emphasis, conveying good news, and conveying bad news (Eide et al., 2004). Expressive speech is capable of producing a more natural sounding synthetic voice by dynamically altering the style of speech based on the content of the message, and by adding in paralinguistic events such as sighs and breaths. The goals of expressive speech are both to increase user satisfaction with the system, and also to improve user efficiency. Matching the speaking style to the content of the information being communicated can potentially reduce the cognitive burden on the user. IBM provides the example of a customer interacting with a system in order to book a flight:

> *Customer: I'd like a flight from New York to Denver tomorrow **morning**.*
>
> *System: I have a flight from New York to Denver tomorrow **evening**.*

With a tradition TtS program, the user may not notice the slight change between the flight requested and the flight that is available, leading to potential mistakes and longer interaction times. However, an expressive TtS system can use contrasting emphasis to highlight the change in travel time from "morning" to "evening".

Intelligibility and comprehension of speech output

Despite advances in speech synthesis technology, the intelligibility and comprehension[6] of synthetic speech is still generally lower than that of

5 The iPhone's Siri voice is an example of concatenative synthesis voice.

6 Speech intelligibility measures whether users are able to recognize the spoken words; comprehension measures whether users are able to understand the meaning of the overall content.

natural speech. For simple tasks with high-quality synthetic speech, however, this may not be a serious problem: users may have similar levels of accuracy when identifying words spoken by natural and synthetic speech (Koul, 2003), although synthetic words take longer to identify. As task complexity increases, the listener requires more cognitive resources in order to process the message. This added burden may also negatively affect the comprehension of synthetic speech, as much of the listener's cognitive resources are used to identify words instead of attending to higher-level mental processes (Duffy & Pisoni, 1992). In a study by Papadopoulos and Koustriava (2015), participants with more intelligibility errors often had lower comprehension scores, although the problems interpreting individual words were not directly related to the errors in understanding the content. From these results, the researchers concluded that contextual cues could actually help participants identify words, and assist them in comprehending content.

Factors affecting content intelligibility

Though up to this point we have discussed the effects of natural and synthetic voices on speech intelligibility, there are several other factors that have the potential to reduce intelligibility while using an audio system. These factors may include: poor sound quality or interference caused by the speakers/headset, low volume levels, and loud ambient sounds (Georgia Tech's "Accessibility Assistant", n.d.). Ambient sound can be particularly problematic when using a voting system in a busy polling place, so it is important to provide users with audio options to compensate, such as volume controls and the ability to repeat audio.

Factors affecting human perception and processing

A person's age and cognitive ability can also affect how well they are able to identify and understand synthetic speech. Age-related hearing impairments cause older users to experience difficulty processing synthetic speech when it contains unfamiliar words or lacks contextual cues (Roring, Hines, & Charness, 2007; Wolters, Campbell, DePlacido, Liddell, & Owens, 2007). Individuals with intellectual and/or language impairments display even greater difficulty processing synthetic speech due to their pre-existing cognitive burdens.

Studies of synthetic speech intelligibility have shown that users with low vision or blindness perform better than sighted individuals, though this result is linked to the fact that synthetic speech intelligibility increases with experience and familiarity, regardless of disability. Thus, blind or visually impaired individuals who are regular users of screen readers and other synthetic speech technologies are likely to have less difficulty understanding synthetic speech. (Duffy & Pisoni, 1992; Papadopoulos, Katemidou, Koutsoklenis, & Mouratidou, 2010; Stent, Syrdal, Mishra, 2011).

Non-Speech Sounds

In addition to speech, interfaces may use a variety of non-speech sounds to convey information to the user, including alarms and sounds meant to provide contextual cues.

Alarms, alerts, and warning sounds

One of the most common types of sounds used in audio interfaces are alert or warning sounds. These sounds are generally short unsubtle sounds that are design to capture the user's attention and alert them to an action or issue (Nees & Walker, 2009). Since these sounds have no inherent meaning on their own, they should generally only be used when accompanied with visual representation of alert or warning, such as a on-screen image or message ("Accessibility Assistant", n.d.). Alert or warning sounds can lead to poor usability and accessibility in an interactive system, because users with visual or cognitive impairments may have trouble deciphering the reasons for the sound. Instead, simple tones or sounds should be used for correspondingly simple actions, such as providing feedback to indicate that a key or button has been pressed.

Auditory icons

Auditory icons are representations of naturally occurring sounds that relate to the action being performed or icon being selected (Nees & Walker, 2009). For example, when you take a picture on a digital camera or mobile device, the device often produces a sound similar to the shutter of a film camera. The sound indicates the successful action, using a metaphor of an older technology. Or if you empty the trash on your computer's desktop, you might hear the sound of crumpling paper; the metaphor of throwing a ball of paper into a wastebasket is not fully realized as the sound of the paper hitting the basket is missing. These sounds form a semantic relationship to the actions or objects they represent (Sanderson, 2006). Since auditory icons use everyday sounds and familiar metaphors, auditory icons are often easy to use and require little training (Walker & Nees, 2011). One drawback is that not all functions can be appropriately represented with a real world sound, and therefore auditory icons are few. Actions such as saving a document are difficult to represent with an auditory icon (Jeon and Walker, 2011).

Earcons

Earcons – a coinage on analogy with icons ('eye-cons') – are emblems of specific meaning, in which musical sounds of varying pitch, melody, timbre, and intensity indicate an object or action, provide a status, show a user's progress or location within a menu (Nees & Walker, 2009; Sanderson, 2006). For example, an earcon can indicate progress within a page by making a sound every time the user scrolls, with a rising pitch as they scroll up, and falling pitch as they scroll down. Earcons use musical notes to create motifs, each of which is later built upon or altered in order to convey the desired information.

Earcons are currently in use in many healthcare environments, as they allow healthcare practitioners to gather patient information without having to divert their attention to a visual interface. One such example is the set of blood pressure earcons developed by Marcus Watson at the University of Queensland (Watson, 2006). This blood pressure system gives the healthcare provider two initial tones that serve as reference points for the patient's systolic and diastolic pressure, respectively. Once the reading is taken, a third tone and fourth tone will sound, which represents the patient's actual systolic and diastolic pressure. If the systolic reading is lower than normal, the third tone will be lower and longer than its corresponding reference tone. The reverse is true for when the blood pressure is higher, and the diastolic tone functions the same as the systolic tone.

Though earcons have shown some promise in the areas of healthcare and auditory menu navigation (Brewster, 1998), earcons are less suited systems that are used infrequently or have numerous options. Since the sounds are arbitrary, it can be difficult for users to remember, and training is often required before a system that features earcons is usable (Nees & Walker, 2009).

Voting systems are unlikely to use earcons, as the voting happens infrequently, and often involves multiple contests with differing selection requirements (e.g., choose a single candidate for one office; choose up to 3 candidates for another). However, a more limited set of earcons and auditory icons to reinforce common actions, such as confirming a selection or indicating when the focus has moved to a new contest, might prove useful.

Spearcons

Though considered a non-speech sound, spearcons are actually synthetic speech sounds sped up to the point where they can become unintelligible (Walker, Nance, & Lindsay 2011). Spearcons have been shown to be effective in audio menus as each sound is unique, and retains an acoustic similarity to the word(s) it is meant to represent. Spearcons have several advantages over audio icons and earcons. The use of speech synthesis technology to create spearcons means that they are easy to produce, and can adjust to dynamic menus (Nees & Walker, 2011). Though spearcons are not always intelligible, they are distinct from one another, and also are not abstract or arbitrary, which can make them easier to use.

Speech output voice and tone

There is a body of research into the impact of the speech and tone used in an audio interface, though this review covers little of that work.

This consideration is especially important in application to voting systems that do not allow voters to customize the audio voice and, yet which need to avoid any bias caused by the interface. Noonan (2006), Nass & Gong (2000) and Los Angeles & IDEO (2015a) all discuss the need for designers to decide on the gender and tone. Los Angeles & IDEO state that a “firm, gentle, male” voice worked well.

Audience responses to other combinations of vocal qualities including gender identification, affective impressions of the voice deserve further review and research.

Speech output in election systems

Election systems today use either natural (recorded) voice or synthetic speech, depending on the capabilities of the voting system and the preference of the jurisdiction. Chapter 3 of the Voluntary Voting System Guidelines 1.1, which sets usability and accessibility standards for voting systems, lists very few specifications concerning the characteristics that make audio an effective component within an accessible system.

The VVSG guidelines that specifically concern speech output quality state that:

- 3.3.3.vii - The audio presentation of verbal information should be readily comprehensible by voters who have normal hearing and are proficient in the language. This includes such characteristics as proper enunciation, normal intonation, appropriate rate of speech, and low background noise. Candidate names should be pronounced as the candidate intends.

- 3.3.3.viii. - The audio system shall allow the voter to control the rate of speech throughout the voting session while preserving the current votes. The range of speeds supported shall include 75% to 200% of the nominal rate. Adjusting the rate of speech shall not affect the pitch of the voice.

Both of these guidelines can be met with either type of speech, natural or synthetic. However, election officials sometimes prefer recorded natural speech because it gives them greater control of the pronunciation of candidates' names. Recorded speech is also used to generate the audio for other languages. It is because of this need to offer ballots in languages other than English, that some election offices prefer to record the audio for English, as it can be easier and more consistent to present the same kind of audio for all languages offered.

When natural voice is recorded, it is critical that the quality of the recording is high. As mentioned previously, a low fidelity audio file can have a negative effect on intelligibility, especially when being listened to in a noisy polling place. The voting system that replays the audio must also have the ability to adjust the tempo of the recording rather than the speed of its playback. In some systems, changing the speed of playback drastically changes the pitch. Synthetic voice, on the other hand, allows for the rate of speech to be increased without raising the pitch of the voice. Instead, the space between the words is compressed.

Correct pronunciation of proper names using synthetic speech is possible, though more difficult to achieve as it requires the use of markup language. Markup language allows for greater control over synthetic speech, as the markup can inform the synthesis processor of what to do when it comes across a proper name, abbreviation, homograph, or a word/name in a foreign language. One such markup language is Speech Synthesis Markup Language (SSML), which was developed as a standard by the W3C's voice working group[7].

[7] The most recent recommendations of the W3C's Voice Browser Working Group web standards can be found at http://www.w3.org/Voice/

Audio technologies

Several types of technology to transform a visual display into audio output are in general use as assistive technology for people with disabilities. These technologies are important as a source of ideas for designing audio interfaces for election systems. For voters who use these technologies regularly, they are also important as their functions form expectations for how the voters will interact with elections systems.

Screen readers

Screen readers are one of the most prominent examples of speech technologies, and remains a widely used accessibly tool for people with low vision or blindness. There are several brands of screen readers, including those that are software-based such as JAWS by Freedom Scientific, or the free NVDA (NonVisual Desktop Access) reader; and those that are built into a platform, such as Android's "Talkback" or Apple's "VoiceOver" feature.

System-standard readers on mobile devices (such as Talkback and VoiceOver) have brought screen reading software and speech technology into the mainstream, allowing users with disabilities to access the web (and other mobile features) independently and without the use of separate (and sometimes expensive) assistive technology software. It is for this reason that platform-based, system-standard assistive technology may become increasingly important in elections, as consumer off-the-shelf (COTS) devices are being used in more voting systems.

Though both Android and Apple devices feature built-in screen readers, Apple's VoiceOver is currently the more mature product. VoiceOver also allows for greater control over what is said through a feature called *verbosity*, Verbosity affects the amount of information the user receives when the system reads text aloud. A system with high verbosity can alert users of punctuation, capitalization, misspellings, links, and changes in text attributes. Some other screen readers (such as JAWS) also have verbosity settings that allow users to customize the level of detail that is most helpful to them.

Though VoiceOver has increased the overall accessibility of products such as the iPhone and iPad, many websites and applications are not optimized for screen readers like VoiceOver, causing disconnects between the audio and visual interfaces. These issues, often caused by ignorance of the capabilities or requirements in website or application development. The result is sites and applications with low usability. The most common usability issues observed when VoiceOver is activated are (Leporini, Buzzi & Buzzi, 2012):

- Interactive elements (e.g. links and buttons) are not labeled in context, so users may not be aware when certain actions or functions are enabled.
- The user's focus does not move through the page's hierarchy in a logical way, when the page structure within the code of the document is does not match the visual or logical presentation.
- Difficulty filling out forms and checking what has been entered into a text field, as the field loses focus once the user begins entering text into the field.

Content highlighting

While screen readers often act as an alternative to a visual interface for users with visual impairments, some assistive technologies instead opt for bimodal presentation, combining screen magnification with audio output.

Some text-to-speech technologies, like WordTalk and Sayz Me, use a technique called content highlighting to draw visual attention to each individual word as it is read aloud. This type of technique is useful for both users with visual and cognitive impairments. Highlighting helps individuals with language/learning disabilities or low literacy learn the words being spoken aloud. Highlighting also helps low vision users follow along with the text.

Sitecues is an accessibility feature that acts as a screen magnifier and screen reader, filling a need for users who use both types of technology to navigate, then read a website. Sitecues works as a feature within the user's own web browser. Users can choose their preferred magnification level, and then navigate to the desired content. Once the target is found, the user is able to select the content and have it read aloud to them.

While strictly speaking, content highlighting by itself is not an "audio technology," it often works in coordination with auditory output.

Several projects, including The Trace EEU voting system, Raising the Floor, and AISquared SiteCues mention the value to low literacy and ESL readers in highlighting blocks of text as they are read aloud, providing both visual and audio support for reading tasks. The challenge for voters who may not be familiar with these kinds of features is discoverability. The challenge for the designers of systems is how to design in a touch-screen interface because of conflicts between interaction patterns for "touch to hear" and "touch to select."

An additional body of research beyond the scope of this review covers assistive technology support for complex reading, such as long-form reading by university students with dyslexia or other reading disabilities. Jackson, Swierenga & Hart-Davidson (2014), for example include a good description of how people use a multimodal approach to reading: "... these participants used [text-to-speech] strategically, much in the same way that an unassisted reader would skim a document to find relevant sections, and then read those sections more closely."

What more can we learn from close examination of tasks involving complex reading via assistive technology which might benefit voters with conditions that make print difficult to consume or interact with?

Audio interfaces and menus

When audio is used in an interactive system, it is often used to supplement the primary visual interface. GPS systems, mobile phones, ATMs and other service kiosks all use audio to support interaction with their graphical user interface (GUI). However there are some circumstances when a visual interface is impossible or not desirable, such as when user is driving a car and needs to perform a task without taking their eyes off the road.

One of the most common examples of audio interfaces are the Interactive Voice Response (IVR) systems that are used by large companies and call centers. Companies can use IVR systems to direct phone calls to the correct department, or to provide services to their customers, such as book tickets, make routine appointments, check flight status, or check account balance.

There is a body of work on speech output for automatic teller machines (ATM) (and transit fare machines) that might contribute useful guidelines for voting systems. As of this writing we have not digested those materials to determine what findings might be applicable to the interactions needed for kiosk-style voting systems.

Interactive Voice Response (IVR) systems

IVR systems use either recorded or synthetic speech to provide users with menu options and other information. Users are then able to interact with the system through speech or dual-tone multi-frequency (DTMF) signaling (accomplished by pressing numbers on the phone's keypad). Like many other types of audio interfaces, IVR systems often suffer from significant usability issues that are inherent to the technology (Holmes & Kortum, 2013). For one, systems that rely on synthetic speech may result in impaired intelligibility. Audio menus also tend to be much more difficult to use than visual menus. Without any visual cues or hierarchy, users can get lost in complex menus. Users must also listen to all options sequentially, as there is no way to skip or skim the options. Noonan (2006) provides a good example of guidelines to ensure better usability in an IVR system:

- Messages should use keywords and be brief, clear, and to the point. Highlight key words with verbal emphasis.
- Use terminology consistently throughout the entire IVR experience.
- Always announce the option first, then the button that must be pushed to activate it.
- Structure menus so that the most important or commonly used items are mentioned earlier (“at the top”). This way most callers will not need to listen to the whole menu in order to make their selection.
- Silence can help to convey a menu’s structure to the user. Use short pauses to distinguish menu items from one another, and somewhat longer pauses to distinguish different menus. However, avoid pauses or silences that are too long.
- Verbally confirm choices, and allow options to correct errors.
- Explain to users how to navigate through the IVR system and provide them with the ability to either go back to the previous option, to hear the current option again, or skip to the next option. These navigation options should be available at all times during the call.
- Tell users how to get help and ensure that the help button is always active.
- If an incorrect button is pressed, the user should be alerted to the issue and provided with information about what input the system requires at that point.
- Tell users if they are about to be transferred to a human attendant.

Auditory menus and navigation cues

In an attempt to improve on the IVR auditory menu model, Zhao et al. (2007) created an eyes-free menu technique called earPod. earPod uses a tactile interface with an inner disk and outer track called the dial. Every time the user enters a new layer of the navigation, the outer dial divides itself into functional segments, according to the number of menu options. When a user touches a finger to the dial, the earPod speaks the title of the option assigned to that segment. The user can explore all the menu options by sliding their finger around the dial. Each time the user enters a new segment, playback of the previous option is stopped, a click sound is made, and then the next menu option is announced. A menu option is selected by lifting the finger from the desired option, or aborted by sliding the finger over the inner disk and then lifting.

The earPod's design can accommodate users of multiple skill levels. Novice users can hear every speech option then make a selection, while intermediate users can quickly skim through the menu employing the interruptible audio feature. Additionally, expert users can rely on spatial memory to select options radially from the outer disk. Other research on visual access to menus (for example Venolia & Neiberg, 1994) has also showed that radial or pie menus are faster to use and easier to learn than linear menus. When comparing the earPod to a visual menu technique like that of an iPod, Zhao et al. showed that the two systems were comparable in accuracy. As for efficacy, the visual menu technique was initially faster to use, though within 30 minutes of training, participants (recruited from the university community, rather than visually impaired individuals) were able to use earPod more efficiently than the visual menu technique.

Spindex is an auditory index that uses non-speech sounds to help users navigate an audio menu (Jeon & Walker, 2011). Instead of voicing the full title of the menu option right away, the system instead produces an audio cue that is related to the menu option, such as the first letter (the sound or pronunciation). By truncating the sound, users are able to get to their target area (i.e. desired letter) at a much faster rate. Once at the target area, the user is then able to slow down and hear all the menu options, allowing them to select their desired choice. Results from the Spindex research found that while objective measures of performance showed increased speed while finding the target item, participants found some versions of the Spindex annoying. The "attenuated" Spindex, where the cue is spoken loudly when reaching a new letter in the alphabet, but is then spoken softly until reaching the next letter, performed best on tests of both preference and performance.

Spindex research may have some bearing on auditory ballot reviews, especially in states where the ballot contains many contests and may be tedious to review. In addition, research at the Trace Center on a universal voting system (Vanderheiden, 2004) included work to determine how long the pause in the system prompts should be so that users realize they can interrupt to make a choice or take action.

Audio instructions

The process of designing a system for accessible audio is twofold. Not only must the system's hardware and software support audio is that clear and understandable, but the content that is spoken by the audio must also be carefully designed to ensure comprehension. In this next section, we will discuss methods for writing content that is optimized for both the visual and aural pathways.

Multimodal instructions

Accessible systems require multimodal instructions in order to address the needs of individuals who use the visual interface, the audio interface, or a combination of both. For example, voting systems use audio instructions that duplicate on-screen instructions in order to provide equivalent access to information. Other systems use audio to supplement (rather than duplicate) visual text instructions. Supplementary audio instructions can be particularly advantageous for systems with small screens as the instructions provide additional information without taking up space or requiring a separate help window. Dedicated help screens can cause difficulties for users, especially new users, as the separate screen moves attention away from the task at hand (Kehoe & Pitt, 2006). Audio feedback and audio help accommodate users with low literacy skills (Medhi et al. 2011; Medhi, Sagar, & Toyama, 2007).

A study of dual-modal information presentation (Fang, Xu, Brzezinski, & Chan, 2006) suggested that auditory information that is presented while the user is reading text can actually improve user performance if the auditory information assists in accomplishing a single task rather than supporting a secondary one.

This result suggests that a voting system's audio can be designed to provide greater support to all voters, not just those with visual or cognitive disabilities.

Designing audio instructions

One of the challenges of creating audio instructions is that audio is transient. On a visual interface, users can quickly refer back to the persistent instructions or options listed on the screen, while users of an audio interface must remember what they have heard. Users also hear audio information sequentially, so they are unable to scan or skip through text as they could if it they were reading it. As such, audio instructions must be used thoughtfully and designed with awareness the limits of working memory.

To start, not all instructions are well suited for audio presentation. Users can benefit from audio information that is instructional or procedural (assuming the procedure is relatively short and task-oriented) (Kehoe & Pitt, 2006). Conceptual information can be more difficult to present via audio when it is long and complex, though still effective when the topic is short and simple. Reference information is often difficult to present using audio.

Audio instructions that use chunking, conversational language, and plain language reduce the user's cognitive burden by making the information easier to understand and remember.

Chunking

We know people are better able to remember phone numbers when the string of digits is broken down into small "chunks" of numbers. The same principle applies to the presentation of audio content and instructions. This can apply to numbers, letters, or keystroke sequences, but also to the content itself. Chunking content instead of presenting it all at once can make the information more conversational and easier to remember (Baddeley, Eysenck& Anderson, 2009; Kehoe & Pitt, 2006; Roter, 2011).

Conversational language

Conversational language involves the use of simple and short sentence structures, contractions, discourse markers (e.g. first… finally), and pointer words (e.g. there… here). This type of speech is more natural to users and can benefit comprehension (Kehoe & Pitt, 2006).

Plain language

The goals for plain language are somewhat similar to those of conversational language. Plain language emphasizes the use of reader-centric text written in the active voice, comprised of common everyday words, personal pronouns, and short sentences (www.plainlanguage.gov). These principles can be applied both to text or audio. The goal of plain language is to create content that is accessible to everyone, so that users can understand what their task is and how to complete it.

The plain language guideline to address the user directly (as "you") is especially applicable to audio instructions. Theofanos and Redish (2003, 2005) found that the phrasing of audio instructions and active elements such as links made a difference in users' satisfaction and how easily they can navigate through long lists such as menus.

Research on audio in the field of voting

Published research in the voting field is limited, as researchers rarely have access to electronic voting systems, and studies conducted by voting machine manufacturers are often kept confidential and never published. There are, however, several voting research projects that have touched upon audio to some degree. These projects are summarized below.

Vote-by-Phone

Vote-by-phone is an accessible voting method that is currently offered by some states[8]. These systems can utilize Interactive Voice Response (IVR) features in order to mark and cast a vote.

One such system is the Inspire Vote-by-Phone system. Pollworkers at the polling place set up the ballot by connecting a telephone line with a code, then handing the phone to the voter. The system only accepts DTMF tone input, so that a user must press one of the keypad buttons to vote[9]. After the voter makes all the choices, a final paper ballot is either printed at a central location, or it is printed and faxed from the polling place.

Research on vote-by-phone systems is fairly limited. One of the only usability studies to be conducted on an IVR voting systems was done in 2013 by Holmes and Kortum. The study consisted of three different IVR voting systems, which were tested using students with normal hearing and vision. One system used a recorded female voice, one used a synthetic male voice, and the third system used a synthetic male voice with the option to increase or decrease the audio speed. The systems displayed no significant differences in completion time or accuracy. The only significant difference observed was that the system that allowed users to increase the audio speed rated higher in satisfaction than the other synthetic speech IVR system.

When compared to other voting systems (i.e. mobile voting, bubble ballot, lever system, VoteBox electronic prototype system), all three systems were at least as accurate as other systems, and had similar levels of satisfaction. However, the IVR systems were shown to be less efficient than other voting methods as all three systems had significantly longer completion times. It is important to note that the participants used in this study do not reflect a representative sample of the voting population, and that further research is needed to establish how well users with disabilities are able to use an IVR

[8] As an example, Vermont currently uses a Vote-by-Phone system for voters with disabilities. https://www.sec.state.vt.us/elections/voters/accessibility.aspx#Vote-By-Phone

[9] Speaking the selection aloud would compromise the voter's privacy

voting system. Though IVR voting technology is not likely to be used by the general voting public due to its longer completion time, these results show some potential, at least from accessibility standpoint, for it to be used as a remote voting option for blind or low vision voters.

Improved write-Ins

For visually impaired voters, voting for a write-in candidate can be a lengthy and difficulty process, especially as systems without attached keyboards make voters scroll through an audio alphabet to input names. While several researchers are trying to improve upon the current method of interaction for write-ins (Dawkins, Eugene, Abegaz, & Gilbert, 2015; Xiong & Sanford, 2013), Gillette and Selker (2014) sought to improve voter speed and satisfaction by focusing on the system's audio during the write-in process. The proposed audio method would allow voters to quickly scroll through the alphabet by holding down either the back or forward navigation arrow. In order to prevent voters from unintentionally skipping past their target letter, the system uses waypoint letters (A, G, M, T, and Z), where it will briefly pause to announce the letter. These anchors help the voter maintain awareness of their approximate location in the alphabet, while helping to improve the speed of their task.

Voting systems designed for universal usability

Voters using audio systems typically use headphones for privacy. Although they are allowed to bring their own, they headphones provided at a polling place are usually simple over-the-ear type headsets with minimal filtering of outside noise.

Extended and Enhanced Usability (EEU) Voting Machine

The EEU voting machine (Vanderheiden, 2004) was designed with the intention of being a universally usable voting system. Though the machine's primary interface was a touchscreen, it also had an optional keypad interface that could be attached to the EEU machine.

Initially, the EEU machine only provided audio output when the keypad was used. This resulted in many voters moving away from the more intuitive touchscreen in favor of the keypad. Audio features were later added to the touchscreen as well, which allowed some voters, such as those with low vision or cognitive disabilities, to use the touchscreen instead. The two audio features that were added to the touchscreen interactions were: "touch to hear" and "voice confirm". The "touch to hear" feature allows those with low vision or cognitive disabilities to still use the touch screen, as it provides audio which supplements the visual content. The user is able to hear the instructions for a contest or listen to a candidate's name by touching the desired text area. To vote for a candidate, voters can touch the blank oval to the left of the candidate. After the desired candidate is selected, the system marks the oval with a check mark and the "voice confirm" feature speaks the name aloud as an added verbal confirmation.

EZ Ballot

After a formative user study with 12 participants with visual disabilities (low vision and blindness), several audio features were added to the EZ Ballot's user interface in order to improve its usability (Lee, 2014).

- First, the systems audio instructions were improved in order to provide greater clarity and enhance comprehension. The instructions were refined so that users with visual disabilities could better locate the "Yes" and "No" buttons that are critical to navigating the ballot.
- Instructions were broken down into smaller segments so that users weren't overwhelmed with large portions of text.
- A section where users could adjust audio speed was also added to the ballot.
- Finally, a feature was added to allow participants to directly select a candidate from a list on the screen.

Previously the user had each candidate's name read aloud to them, and the user had to chose either "Yes" or "No" for that candidate. This updated selection feature was designed for participants with some degree of vision and who would prefer to use the touchscreen. Somewhat similar to the "touch to hear" and "voice confirm" features on the EEU machine, the audio refinements to the EZ Ballot allowed users to hear the list of candidates by dragging their fingers across the touch screen and over the candidates names. Once the user heard the candidate of their choice, they simply lifted their finger to select the candidate and hear the confirmation of which candidate they had selected.

Prime III

The Prime III voting system (Dawkins & Gilbert, 2010) is unique because it not only produces speech output, but it can also accept oral input from the voter. Prime III receives commands from the voter through a combination of its touch interface and voice-activated headset. Voters who have difficulty either seeing or reaching the touch interface can hear the audio ballot using the headset, and can select their candidate by saying "vote" or by blowing into the microphone. These input commands allow the voter to select an option orally without compromising their privacy.

Los Angeles Voting System Assessment Project (VSAP)

Since 1999, the Los Angeles County Registrar-Recorder/County Clerk has been engaged in project to design a voting system for the large and complex electorate of Los Angeles County[10]. The Voting Systems Assessment Project (VSAP) is the largest election jurisdiction in the U.S., by both number of voters and number of languages supported. Working with the design firm IDEO, the County has conducted a number of research studies on the form and interaction of the voting system, including two studies on the audio ballot. These studies (Los Angeles & IDEO, 2015a, 2015b) were small qualitative sessions that tested the audio ballot in the context of exploring a new voting system including the discoverability of audio features. Participants included bilingual voters and people with a range of technical ability.

This project is still in process at the time of this report, so these summaries reflect investigations into an evolving prototype. However, they are among the few projects to test audio ballots extensively.

Their recommendations for designing the audio ballot:

- Voiceover narration, whether human-recorded or artificially synthesized, should be warm and authoritative. In this case, a middle-aged male voice worked well.
- Use headphone audio and other non-visual cues to immediately acknowledge the presence of a new voter and let them know that the system is ready to begin.
- Use a voter-driven approach to the audio+keypad interaction for users with visual impairments, rather than having the system drive the pace of the interaction.
- Incorporate simple ways for voters to manage the audio speed, audio volume, pausing, repeating, and advancing in the audio interaction.
- When designing the tactile interaction, ensure that buttons reliably provoke the same types of actions throughout the experience.
- Ensure that the system is sufficiently responsive that voters know that their actions are registered and understood.
- Do not automatically repeat audio in a “looping” function that can annoy voters.
- Provide clear cues about each step of the process, and when voting is complete.

An open question in this on-going research is whether the same audio script could work for voters using the audio with a tactile interface and for voters using audio as a companion to the visual interface.

[10] http://www.lavote.net/vsap/

Conclusion

This review of previous work has focused on two themes:

What work has been done with auditory outputs (and to a lesser extent, inputs) in interface design and standards development, irrespective of the application domain? And how much of that work is applicable to voting systems?

The review covers spoken output of content, the use of non-speech signals (as complementary alerts, auditory icons, and navigation for menus), and speech output in voting systems (with recognition of needs for privacy and security, while providing access).

Familiar technologies, such as screen readers, visual highlighting of content in coordination with spoken output, and experiments using audio as part of the interface including menus, are instructive as we consider translating them to the case of voting systems.

The review calls out instructional materials realized in audio form as a special area of focus, because voting systems are used infrequently, and the consequences of errors in reading, marking and submitting ballots are crucial.

What work has been done specifically in the realm of voting systems, including kiosks, electronic machines in polling places or other solutions, that includes audio interfaces especially to account for access by disabled individuals?

While a small area of research to date, the important contributions of studies in actual usage of accessible voting systems, such as remote voting system or auditory mechanisms for write-in votes, point the way to further work to enable universal usability and access.

The issue is not whether audio interfaces benefit voters. Audio does more than just provide basic accessibility; it enhances a user's experience of a system. Yet for audio to be usable, it must be designed for, and not added as an afterthought. Each modality must be capable of standing on its own, but should also enhance the overall quality of the system's experience when presented simultaneously.

If well designed, these features have the possibility to make voting more usable for a wider range of voters, including those who are blind, as well as those with a variety of challenges reading the screen because of low vision, low literacy, or other reading disabilities.

Guidelines for the design of the audio features of a voting system could be helpful in improving the usability of this key accessibility feature. It would also be important for any guidance to include both the use of the audio features with a tactile input device, as well as the coordinated use of the audio and visual systems together.

The work of designing audio systems that are usable by voters with a variety of visual impairments and with a variety of experience with audio interfaces is the current challenge. System designers need good research on how to integrate audio into voting systems in a way that is both usable and accessible in order to improve the voting experience for many voters. We have reviewed examples which will integrate well with existing equipment, if the voting could happen where the visually impaired individual's screen reader, or other technical accommodations are located. And we have pointed out how systems in assigned polling places might optimize the use of audio in sharing content, aiding navigation and preventing errors.

We also recognize the continuing issues for voters in using even a well-designed system:

- Discoverability: how does the disabled voter know where to find the controls to turn on or adjust the audio interface?
- Learnability: how does a print-disabled voter, who may be moderately visually impaired or dyslexic know that audio supports exist within the printed ballot or ballot represented on screen?

References

Anderson, L. (2013). Machine language: How Siri found its voice. **http://www.theverge.com/2013/9/17/4596374/machine-language-how-siri-found-its-voice**

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory* (1 edition). Psychology Press.

Brewster, S. A. (1998). Using nonspeech sounds to provide navigation cues. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 5(3), 224-259. **http://eprints.gla.ac.uk/3247/1/using22_non_speech.pdf**

Dawkins, S., Gilbert, J. (2010). Accessible, Private, and Independent Voting: The Prime III Voting System. *User Experience Magazine*, 9(2). UXPA **http://uxpamagazine.org/prime_iii_voting_system/**

Dawkins, S., Eugene, W., Abegaz, T., & Gilbert, J. E. (2015). Toward Private and Independent Accessible Write-In Voting: A Multimodal Prediction Approach. In *Universal Access in Human-Computer Interaction. Access to the Human Environment and Culture* (pp. 171-181). Springer International Publishing.

Duffy, S. A., & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35(4), 351-389. **http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3507427/**

Election Assistance Commission (2015) Voluntary Voting System Guidelines (VVSG) 1.1 **https://eac926.ae-admin.com/assets/1/Documents/VVSG.1.1.VOL.1.FINAL.pdf**

Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., & Pitrelli, J. (2004). A corpus-based approach to expressive speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis.* **http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.384.9018&rep=rep1&type=pdf**

Elkind, J., Black, M. S., & Murray, C. (1996). Computer-based compensation of adult reading disabilities. *Annals of Dyslexia*, 46(1), 159-186.

Fang, X., Xu, S., Brzezinski, J., & Chan, S. S. (2006). A study of the feasibility and effectiveness of dual-modal information presentations. *International Journal of Human-Computer Interaction*, 20(1), 3-17. **http://condor.depaul.edu/sxu/Publications/J4_AStudyOfTheFeasibilityAndEffectiveness.pdf**

Gillette, D., & Selker, T. (2014). Improving Write-In Candidate Text Entry for Audio-Only Voting Interfaces. **http://scholarworks.calstate.edu/bitstream/handle/10211.3/133385/JTPD201412-p176-193.pdf?sequence=1**

Holmes, D., & Kortum, P. (2013). Vote-By-Phone Usability Evaluation of an IVR Voting System. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 1308-1312). SAGE Publication. **http://www.danaeholmes.com/files/Holmes_HFES%20Submission.pdf, http://www.ivsllc.com/**

Jackson, J., Swierenga, S. J., & Hart-Davidson, W. (2014). Assistive technology support for complex reading. *Journal on Technology and Persons with Disabilities*, 1, 212-219. CSUN **http://scholarworks.csun.edu/handle/10211.3/121990**

Jeon, M., & Walker, B. N. (2011). Spindex (speech index) improves auditory menu acceptance and navigation performance. *ACM Transactions on Accessible Computing (TACCESS),* 3(3), 10. **http://sonify.psych.gatech.edu/publications/pdfs/2011TACCESS-JeonWalker.pdf**

Kehoe, A., & Pitt, I. (2006). Designing help topics for use with text-to-speech. *Proceedings of the 24th annual ACM international conference on Design of Communication* (pp. 157-163). ACM.

Koul, R. (2003). Synthetic speech perception in individuals with and without disabilities. *Augmentative and Alternative Communication*, *19*(1), 49-58.

Lee, S. (2014, April). Designing ballot interfaces for voters with vision disabilities. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (pp. 933-938). ACM.

Lepporini, Buzzi & Buzzi (2012). Interacting with Mobile Devices via VoiceOver: Usability and Accessibility Issues. *OZCHI'12*

Los Angeles, IDEO (2015a) VOX Research Report: Audio Experience, April 2015 **http://www.lavote.net/Documents/vsap/Audio%20Inspiration%20Report%20V1.5.pdf**

Los Angeles, IDEO (2015b) VOX Research Report: Audio Experience, April 2015. **http://lavote.net/Documents/vsap/Audio%20Test%20Report%20V1.3.pdf**

Medhi, I., Patnaik, S., Brunskill, E., Gautama, S. N., Thies, W., & Toyama, K. (2011). Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *18*(1), 2. **http://www.msr-waypoint.com/pubs/149090/ToCHI'11_Medhi.pdf**

Medhi, I., Sagar, A., & Toyama, K. (2007) Text-Free user interfaces for illiterate and semiliterate users. *Inf. Technol. Int. Devel.* 4, 1, 37–50. **http://msr-waypoint.com/pubs/143179/medhi_ictd2006.pdf**

Montali, J., & Lewandowski, L. (1996). Bimodal reading: Benefits of a talking computer for average and less skilled readers. *Journal of Learning Disabilities*, 29(3), 271-279.

Nass & Gong (2000). (2000) Speech interfaces from an evolutionary perspective. *Communications of the ACM*, Volume 43, Issue 9, Sept 2000, Pages 36-43

Nees, M.A., & Walker, B.N. (2009). Auditory interfaces and sonification. In Stephanidis, C. (Ed.), *The Universal Access Handbook*. New York: CRC Press, pp. 507- 522. **http://sonify.psych.gatech.edu/~ben/references/nees_auditory_interface_and_sonification.pdf**

Nees, M.A., & Walker, B.N. (2011). Auditory displays for in-vehicle technologies. In P. DeLucia (Ed.*), Reviews of Human Factors and Ergonomic*s, Vol. 7 (pp. 58-99). Sage Publications **http://sites.lafayette.edu/neesm/files/2011/10/NeesWalker2011-RHFEv7-.pdf**

Neerincx, M., Cremers, A., Kessens, J., van Leeuwen, D., & Truong, K. (2008). Attuning speech-enabled interfaces to user and context for inclusive design: technology, methodology and practice. *Universal Access in the Information Society*, 8(2), 109-122. **http://link.springer.com.proxy-ub.researchport.umd.edu/article/10.1007/s10209-008-0136-x/fulltext.html**

Noonan, T. (2006). Designing User Friendly IVR Telephone Voice Systems for All. **http://timnoonan.com.au/resources/information-accessibility-and-usability/designing-user-friendly-ivr-voice-systems/**

Oatsoft (nd) SayzMe **http://www.oatsoft.org/Software/sayz-me**

Papadopoulos, K., & Koustriava, E. (2015). Comprehension of Synthetic and Natural Speech: Differences among Sighted and Visually Impaired Young

Adults. *Enabling Access for Persons with Visual Impairment*, 147. **http://access.uoa.gr/ICEAPVI-2015/proceedingsFiles/149-153.pdf**

Papadopoulos, K., Katemidou, E., Koutsoklenis, A., & Mouratidou, E. (2010). Differences among sighted individuals and individuals with visual impairments in word intelligibility presented via synthetic and natural speech. *Augmentative and Alternative Communication*, 26(4), 278-288. **http://www.researchgate.net/profile/Konstantinos_Papadopoulos2/publication/271587023_Differences_amongst_sighted_individuals_and_individuals_with_visual_impairments_in_word_intelligibility_presented_via_synthetic_and_natural_speech/links/54cd109c0cf24601c08ca6df.pdf**

Papadopoulos, K., Koutsoklenis, A., Katemidou, E., & Okalidou, A. (2009). Perception of Synthetic and Natural Speech by Adults with Visual Impairments. *Journal of Visual Impairment and Blindness*, 103(7), 403-414. **http://www.researchgate.net/profile/Konstantinos_Papadopoulos2/publication/232221293_Perception_of_synthetic_and_natural_speech_by_adults_with_visual_impairments/links/0c96053a83f3137ae9000000.pdf**

Roring, R. W., Hines, F. G., & Charness, N. (2007). Age differences in identifying words in synthetic speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(1), 25-31.

Roter, D. L. (2011). Oral literacy demand of health care communication: challenges and solutions. *Nursing outlook*, *59*(2), 79-84.

Sanderson, P. (2006). Auditory Displays in Healthcare. *User Experience Magazine*, 5(3). UXPA **http://uxpamagazine.org/auditory_displays_healthcare/**

Stent, A., Syrdal, A., & Mishra, T. (2011, October). On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility* (pp. 211-218). ACM. **http://www.research.att.com/techdocs/TD_100505.pdf**

Theofanos, M. F. and Redish, J. C., 2005, Helping Low-vision and Other Users with Web Sites That Meet Their Needs: Is One Site for All Feasible? *Technical Communication,* 52 (1), February, 9-20. **http://www.redish.net/images/stories/PDF/Low_vision.pdf**

Theofanos, M. F. and Redish, J. C., 2003, Guidelines for accessible and usable web sites: Observing users who work with screen readers, *Interactions*, X (6), November-December, 38-51.

http://www.redish.net/images/stories/PDF/InteractionsPaperAuthorsVer.pdf

Vanderheiden, G.C. (2004). Using extended and enhanced usability (EEU) to provide access to mainstream electronic voting machines. *Information Technology and Disabilities, Vol. X, No. 2.* **http://www.rit.edu/~easi/itd/itdv10n2/vanderhe.htm**

Venolia, D., Neibert, F. (1994). T-Cube: A Fast, Self-Disclosing Pen-Based Alphabet Pen Input. *Proceedings of ACM CHI'94 Conference on Human Factors in Computing Systems* 1994-04-24 v.1 p.265-270

Vermont Secretary of State (n.d.) Accessibility: Vermont's Vote-By-Phone (At the Polls) Voting System. **https://www.sec.state.vt.us/elections/voters/accessibility.aspx#Vote-By-Phone**

Walker, B. N., Nance, A., & Lindsay, J. (2006). Spearcons: Speech-based earcons improve navigation performance in auditory menus. **https://smartech.gatech.edu/bitstream/handle/1853/50642/WalkerNance2006.pdf**

W3C (2010) Speech Synthesis Markup Language (SSML) Version 1.0 **http://www.w3.org/TR/speech-synthesis11/**

Walker, B. N., & Nees, M. A. (2011). 2 Theory of Sonification. **http://sonify.psych.gatech.edu/~ben/references/nees_theory_of_sonification.pdf**

Watson, M. (2006). Scalable earcons: bridging the gap between intermittent and continuous auditory displays. **https://smartech.gatech.edu/bitstream/handle/1853/50645/Watson2006.pdf**

Wolters, M., Campbell, P., DePlacido, C., Liddell, A., & Owens, D. (2007). Making Speech Synthesis More Accessible to Older People. *6th ISCA Workshops on Speech Synthesis (SSW-6).* **www.wordtalk.org.uk/**

Xiong, X. N., & Sanford, J. A. (2013, October). SlideType: universal design text-entry interface for touchscreen devices. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (p. 64). ACM. **http://elections.itif.org/reports/Xiong-Sanford-2013.pdf**

Zhao, S., Dragicevic, P., Chignell, M., Balakrishnan, R., & Baudisch, P. (2007, April). Earpod: eyes-free menu selection using touch input and reactive audio feedback. *Proceedings of the SIGCHI conference on Human factors in*

Computing Systems (pp. 1395-1404). ACM.
http://w.patrickbaudisch.com/publications/2007-Zhao-CHI07-EarPod.pdf

Technologies and Organizations

Accessibility Assistant from GTRI
http://accessibility.gtri.gatech.edu/assistant/roles.php

AI-Squared (nd) SiteCues
http://www.aisquared.com/products/zoomtext/

Apple (nd) Appendix B. VoiceOver Utility Options
http://help.apple.com/voiceover/info/guide/10.9/English.lproj/_1199.html

Freedom Scientific (nd) Job Access With Speech (JAWS)
http://www.freedomscientific.com/Products/Blindness/JAWS

Plain Language Action and Information Network (PLAIN)
http://www.plainlanguage.gov/

Prime III **http://www.primevotingsystem.org/**

Raising the Floor **http://raisingthefloor.org/projects/features/highlight**

VoiceOver for iOS
http://www.apple.com/accessibility/ios/voiceover/