The Function of Nonspeech Audio in Speech Recognition Applications: A Review of the Literature

TR 29.3405
March 19, 2001

Melanie D. Polkosky
James R. Lewis

IBM Voice Systems

West Palm Beach, Florida

Abstract

We reviewed the literature on use of nonspeech audio in interface design, psycholinguistic aspects of interpersonal communication, and auditory cognition to determine the appropriate functions of auditory cues in telephony and speech recognition applications. In these interfaces, auditory cues may be an effective substitute for the cognitive and psychosocial communication cues typically conveyed through eye gaze, gesture, and other nonlinguistic, pragmatic communication forms. However, designers have rarely used auditory icons and earcons for these purposes. The review covers several perceptual and cognitive aspects of audition that are relevant to the design of effective auditory cues. Finally, we propose eight cognitive and psychosocial functions of auditory cues in a speech recognition interface.

ITIRC Keywords

Interpersonal Communication
Pragmatics
Man-Machine Interaction
Auditory Interface Design
Telephony Systems
Speech Recognition Systems

Contents

Introduction

Until relatively recently, nonspeech audio has received only minimal use in interface design, due largely to limitations in hardware and software, user irritation caused by annoying sounds, and an association with extravagance in design (Gaver, 1997). Auditory tones can serve as warning or alerting signals (Belz, Robinson, & Casali, 1999; Belz, Winters, Robins, Casali, 1997; Fusco & Katz, 1992; Graham, 1999; Hearst et al., 1997; Rauterberg, 1998), representations of complex data (Bly, 1982), auditory alternatives or complements to visual icons in desktop applications (Brewster, 1996; Brewster & Crease, 1999; Brewster, Wright, & Edwards, 1993; Mynatt, 1997), and as support for menu navigation (Brewster, 1997, 1998; Leplatre & Brewster, 2000). However, few studies have addressed the use of auditory tones in speech recognition applications (Williams & Cheepen, 1998), even though they can be an important component of these designs (Balentine & Morgan, 1999). This review of the literature summarizes the work on nonspeech audio, psycholinguistic aspects of human-computer communication, and auditory cognition to define some initial directions for using auditory cues to manage the unique cognitive and psychosocial aspects of human interaction with speech interface applications.

Choice of Sound: Auditory Icons vs. Earcons

A significant amount of previous work in using nonspeech audio in interface design has centered on the use of auditory icons or earcons (Blattner, Sumikawa, & Greenberg, 1998). Gaver (1986) initially described auditory icons and defined them as naturally occurring sounds used to represent specific actions or functions in an interface. A number of applications have used auditory icons (Albers, 1996; Albers & Bergman, 1995; Blez, Robinson, & Casali, 1999). For example, Gaver's (1989) SonicFinder application plays sounds during file manipulation. Table 1 summarizes the auditory icons used in this application.

Table 1. Auditory Icons Used in SonicFinder (adapted from Gaver, 1989, as cited in Gaver, 1997)

Events	**Auditory Icons**
Icons • Selection type (file, application folder, disk, trash) • Opening • Dragging • Drop-in • Copying	• Hitting sound source (wood, metal, etc.) • Whooshing sound • Scraping sound • Noise of object landing • Pouring sound (size of selection indicated by frequency)

Windows • Selection • Dragging • Growing • Scrolling	• Clink • Scraping • Clink on release • Ticking sound (size of window indicated by frequency)
Trashcan • Drop-in • Empty	• Crash • Crunch

By contrast, earcons are more abstract, musical sounds designed in "a rhythmicized sequence of pitches" (motives) and organized into families; these sounds represent actions and functions in an interface (Blattner, Sumikawa, & Greenberg, 1989, p. 23). The fixed characteristics of a motive are rhythm and pitch. Timbre, register, and pitch are variable characteristics of a motive. Earcons can consist of a single element or combinations that form compound earcons of several different tones in a distinct rhythm.

Brewster and his colleagues have used earcons in various interfaces (Brewster, 1996; Brewster & Crease, 1999; Brewster, Wright, & Edwards, 1993; Brewster, 1997, 1998; Leplatre & Brewster, 2000). An example of an earcon used with a graphical button is described in Brewster (1996): "A base sound was created for when the mouse was moved over a screen button. This was a continuous tone at C4 (130Hz). The volume of this was kept to just above the threshold level. This sound was played as long as the mouse was over a graphical button; it stopped when the mouse was moved off. When the mouse button was pressed down over a graphical button a continuous sound at pitch C3 (Middle C, 261 Hz) was played. This continued for as long as the mouse button was down and the mouse was over the graphical button. If the mouse was moved off the graphical button the sound stopped. If the mouse was released over the graphical button then a success sound was played. This consisted of two notes, played consecutively, at C1 (1046Hz)…" (p. 7). Brewster (1996, 1999) suggests earcons can improve the usability of graphical widgets and menus; however, his empirical evidence is not compelling due to problems with experimental design and measurement.

Most studies have used either auditory icons or earcons in interface design: a considerable controversy exists over which form of auditory tones is preferable (Edworthy, 1998; Gaver, 1997; Hearst et al., 1997). Few applications have integrated both types of sounds (Albers, 1995), although this may be a promising new direction for auditory interface design (Gaver, 1997). In a study that compared auditory icons and earcons, Sikora, Roberts, & Murray (1995) found that although users preferred earcons and rated them as more pleasant, auditory icons mapped more clearly to the intended function.

Beyond the controversy, several criticisms of both auditory icons and earcons are apparent: Gaver (1997) points to the lack of examples in working applications, significant learning demands, and long duration of earcons. He states that these "designs are often less subtle and more intrusive than might be possible" in an interface (Gaver, 1997, p.

1024). Similarly, mapping of an auditory icon may not be clear: even though the designer intends a specific meaning for a recognizable environmental sound does not ensure that the user will infer the same meaning, especially across cultures.

In addition to the problems identified by Gaver (1997), aspects of speech recognition and telephony applications may pose more difficulties for the choice of auditory cues. Much of the previous research is based on the desktop metaphor and uses sound to augment or substitute for visual icons. There is little need for scrolling, graphical buttons, files, and other visual elements of the desktop computer in telephony applications. Therefore, the mental model of the desktop may not be an appropriate design model for telephony and speech recognition systems. In these systems, there is a need to represent highly abstract, subtle, and automatic aspects of interaction, suggesting that cognitive mapping from the sound cue to its function is likely to be highly arbitrary. Next, the limited bandwidth of most telephony systems, as well as the likelihood of use in noisy environments (e.g., mobile phones), suggest that complex or overlapping sound patterns might not be audible or effective. Finally, because users will rarely want to extend their transaction by listening to sounds, any sounds used should be as short as possible.

Despite these problems, Balentine and Morgan (1999) assert that nonspeech audio does have a role in effective speech applications. They suggest that, with effective design, tones can relay information rapidly, provide feedback, and cue the user to application state. The literature on cognitive and psychosocial aspects of interpersonal communication point to several potential opportunities for the use of auditory cues.

Cognitive and Social Aspects of Interpersonal Communication

The use of nonspeech audio in a speech recognition application should support the primary function of human factors efforts in this area: "to develop human-computer communication modes that are both error tolerant and easily learned" (Ogden & Bernick, 1997). It is possible to interpret this goal as simply encouraging the design of a speech system that uses typical words and sentences but the cognitive and social aspects of human-computer communication are also vitally important to a successful interaction. The literature on the psychology of communication offers an understanding of how technology can adversely affect the cognitive and social aspects of communication. This literature reveals the expectations users have of a speech recognition system based on their knowledge of and skills for human-human communication.

In a comprehensive review of pragmatic cues adversely affected by telecommunication technology, Fussell and Benimoff (1995) identify several functions of eye gaze and gestural cues. These cues:

- Coordinate speaking turns;
- Reduce cognitive load when formulating messages;
- Provide feedback on listener attention, interest, and engagement;
- Indicate specific meanings (deitic gestures);

- Enhance the meaning of messages; and
- Aid message formulation.

In addition, Fussell and Benimoff (1995) cite 25 years of research, which "has shown that speakers attempt to take their addressee's background knowledge, beliefs, attitudes, and so on into account when they formulate messages" (p. 233-234). A speaker adjusts his or her message because of this information (known as perspective-taking), which also aids the addressee's comprehension (Fussell & Benimoff, 1995). Both the physical context (spatial location and visual prominence of communication partners) and social context (assumptions about partners' knowledge, attitudes, motives, etc.) significantly influence how individuals formulate messages and interact during interpersonal communication (Fussell & Benimoff, 1995).

The lack of eye gaze, gesture, and other reciprocal pragmatic cues in machine-human communication can negatively impact these interactions. Williams and Cheepen (1998), in their study of turntaking in a telephony application, found that novice users spoke significantly more quickly than expert users following a question prompt, often before a beep signaling that the recognition window was active. Similarly, Sadowski and Lewis (2000) found that 4 of 6 participants experienced turntaking errors in a half-duplex speech recognition system. The stuttering effect, spoke-too-soon, and spoke-way-too-soon are problems associated with ineffective turntaking in speech interfaces (Balentine & Morgan, 1999). Addressing a broader range of problems, Ogden and Bernick (1997) argue that user error with natural language interfaces generally arise from incorrect assumptions about the system's capabilities.

As shown by the literature on perspective taking in interpersonal communication (Fussell & Benimoff, 1995), the assumptions made by users are typical of human-human communication and are controlled by pragmatic cues. Errors occur because the cognitive and psychosocial aspects of human communication are automatic, highly integrated, and without conscious control by the user (Prabhu and Prabhu, 1997). Brems, Rabin, and Waggett (1995) demonstrate that use of these subtle cues (e.g., natural question prompts, brief pausing) in an interface resulted in a more rapid overall transaction, better user perception of the interface, and less user confusion than question-plus-option prompts. The researchers conclude that "the key to the successful development of computers that 'talk and listen' might well lie in discovery and implementation of such natural language conventions" (Brems, Rabin, Waggett, 1995, p. 282). Therefore, an overall goal of using auditory cues in a speech application may be to help users adapt to the pragmatic differences between human-machine communication and interpersonal communication (Kamm & Helander, 1997).

Cognitive and Perceptual Aspects of Sound

Sound works well for cueing the subtle cognitive and psychosocial aspects of communication for a number of reasons. Auditory tones may be:

- Rapid (Brems, Rabin, & Waggett, 1995; Balentine & Morgan, 1999);
- Associated with the source that makes them (Moore, 1989);
- Perceived as a group when they are similar in timbre, pitch, loudness, or spatial proximity (Bregman, 1993, 1990; Moore, 1989);
- Indicative of change in the sound source when frequency, intensity, or location changes are smooth and continuous (Moore, 1989);
- Indicative of a new sound source being activated when frequency, intensity, or location changes are abrupt (Moore, 1989);
- Perceived as coming from a single sound source (grouped) when they begin and end or change in intensity or frequency together (Moore, 1989);
- Perceived as "belonging" to only one sound source at a time (Moore, 1989).

In addition, complex sounds draw attention, especially when they are changing, and part of the sound stands out while the rest is not as prominent; this shifting of attention is known as the figure-ground phenomenon (Moore, 1989). Gaver (1997) notes that sound is effective at conveying information about spatial location, timing, dynamics or processes; further, sound can manipulate the hearer's mood and environmental awareness. Sound can be obscured temporarily (masked) by other sounds (Gaver, 1997; Moore, 1989); this characteristic is a possible drawback of sound for speech interfaces. Designers can exploit these characteristics of sound to create auditory tones that manage the cognitive and psychosocial aspects of speech recognition applications.

Recommended Functions of Auditory Cues in the Speech Interface

The previous literature suggests that auditory tones may effectively manage the complex cognitive and psychosocial aspects of human computer interaction (Balentine & Morgan, 1999). The following recommendations combine the information available on auditory tones, psycholinguisitics, and auditory cognition to help designers understand the purposes that auditory cues can serve in speech recognition applications. The basis for choosing a particular sound (auditory icon, earcon, or other tones) for these functions should be a consideration of the function of the sound itself, the overall design of the interface, and the results of empirical usability evaluation. However, auditory cues generally should be as pleasant, short, and unobtrusive as possible.

Cognitive Functions

1. **Auditory cues can focus or shift user attention.**

   When users need to focus their attention sharply (e.g., alert, alarm, or warning), provide abrupt changes in onset/offset, intensity, and/or frequency. This abrupt change plus a large acoustic difference from any background sound will also prevent masking. For more subtle attentional focus (e.g., preparation for the end of a user waiting period, attentional shift to new topic), gradual changes in onset/offset, intensity, and/or frequency are appropriate. Additionally, you can add or remove background audio to subtly shift user attention.

2. **Audio logos or familiar sounds make it easier to recall previous knowledge or associations.**

   The literature on auditory icons (Gaver, 1997) suggests this cognitive function; however, use only highly specific and familiar sounds that are consistent with the overall design. In speech interfaces, do not use familiar sounds unless you can articulate the specific knowledge or association you want to invoke.

3. **Accompany topic shifting by auditory cues to help the user refocus attention on the new topic.**

   Examples of topic shifting include mode changes (waiting, help, secure mode) and changes to a new content area, as in a menu structure. Signal a change of topic by playing an audio cue in a different timbre, frequency, and/or intensity than previous audio cues. You can also establish topic shifting by the onset of a continuous background tone or the offset of a background sound from another content area. For example, if a continuous, warbling tone has played during a processing time, a brief frequency change may cause the user to resume more active listening prior to the next prompt (Moore, 1989).

4. **Consider pairing auditory tones with natural conversational cues to help users learn their meaning.**

   Many of the cognitive and social aspects of interpersonal communication are abstract, automatic, and unconscious for users; therefore, users will need time to learn that these overlearned behaviors do not work with a speech recognition interface. When

appropriate, present auditory cues simultaneously with the natural conversational cue. For example, if using a turntaking tone, pair it with spoken prompts to acquire the same meaning as more natural cues to turntaking, including question form, rising or falling intonation, and pausing. After several instances of this paired, redundant prompting, the question and tone may cue experienced users to speak without the need for a second prompt listing response options (Brems, Rabin, & Waggett, 1995).

5. **Within a single topic area, all tones should use a similar timbre, pitch, and/or loudness to assist the user's cognitive organization.**

   In a menu structure, tones in one branch might be designed within a single timbre but vary in duration and pitch. For example, you might design a background tone with a violin timbre, with short turntaking and warning tones also designed from higher frequency violin notes. The violin timbre may help the user cognitively group various prompts within this single menu branch (Brewster, 1996; Brewster & Crease, 1999; Bregman, 1993, 1990). However, in complex menu structures, use of several timbres to differentiate menu branches may sacrifice overall design consistency. (This area of research needs more empirical work.)

Psychosocial Aspects

6. **Tones should signal relevant aspects of the computer's state, including communication breakdown, acceptance of input, and mode change.**

   As previously noted (recommendations 1 and 3), auditory cues can convey information about the system itself. Tones can provide quick and efficient feedback on communication breakdown (whether due to ineffective turn exchange, misrecognition, or out of grammar utterances). This immediate feedback on acceptance of user input and mode change cues can help the user develop a more accurate perception of the system's capabilities.

7. **Sounds manipulate the user's mood and attitude toward the interface.**

   It is possible to manipulate aspects of mood, emotion, and tension through sound, particularly music (Gaver, 1997). However, avoid random or meaningless use of sound because it may distract users from the primary purpose of an interface.

8. **When using spoken prompts that include natural cues to turn taking (pausing, rising/falling intonation, question form), tones can make the turn transfer more salient when presented concurrently with the end of the utterance.**

   In a half-duplex application, a brief tone presented simultaneously with the end of a computer prompt may help the user take a speaking turn by directing attention to the spoken prompt. If the tone occurs too long after the spoken prompt, users will automatically speak due to the natural cue instead of the tone. The literature does not provide any guidelines regarding the maximum permissible delay between the end of speech and the presentation of the turntaking tone. Research in human information processing (Massaro, 1975) suggests that people would probably not notice delays with durations of less than 50 msec.

References

Albers, M. (1996). Auditory cues for browsing, surfing, and navigating the www: The audible web. In *Proceedings of ICAD'96 The Third International Conference on Auditory Display* (pp. 85-90). Palo Alto, CA: ICAD.

Albers, M. (1995). The varese system, hybrid auditory interfaces, and satellite-ground control: Using auditory icons and sonification in a complex, supervisory control system. In *Proc. of ICAD'94 The Second International Conference on Auditory Display* (pp. 3-13). Santa Fe, NM: Santa Fe Institute.

Albers, M. & Bergman, E. (1995). The audible web: Auditory enhancements for Mosaic. In *Proceedings of CHI'95: The ACM Conference on Human Factors in Computing Systems* (pp. 318-319). Denver, CO: ACM.

Balentine, B. & Morgan, D. (1999). *How to build a speech recognition application: A style guide for telephony dialogues.* San Ramon, CA: Enterprise Integration Group.

Belz, S., Robinson, G., & Casali, J. (1999). A new class of auditory warning signals for complex systems: Auditory icons. *Human Factors, 41(4),* 608-618.

Belz, S., Winters, J., Robinson, G., & Casali, J. (1997). A methodology for selecting auditory icons for use in commercial motor vehicles. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, (pp. 939-943). Santa Monica, CA: Human Factors and Ergonomics Society.

Blattner, M., Sumikawa, D., & Greenberg, R. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction, 4*, 11-44.

Bly, S. (1982). Presenting information in sound. In *Proceedings of the CHI'82 Conference on Human Factors in Computer Systems*, (pp. 371-375). New York: ACM.

Bregman, A. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge: MIT Press.

Bregman, A. (1993). Auditory scene analysis: hearing in complex environments. In S. McAdams & E. Bigand, *Thinking in sound: The cognitive psychology of human audition.* (pp. 10-36). Oxford: Clarendon Press.

Brems, D., Rabin, M., & Waggett, J. (1995). Using natural language conventions in the user interface design of automatic speech recognition systems. *Human Factors, 37(2),* 265-282.

Brewster, S. (1996). The design of a sonically-enhanced interface toolkit. (Technical Report No. TR-1996-23). Glasgow, UK: Department of Computing Science, University of Glasgow.

Brewster, S. (1997). Navigating telephone-based interfaces with earcons. In Proceedings of BCS HCI'97 (pp. 39-56). Bristol, UK: Springer Verlag.

Brewster, S. (1998). Using nonspeech sounds to provide navigation cues. *ACM Transactions on Human Computer Interactions, 5(2),* 224-259.

Brewster, S. & Crease, M. (1999). Correcting menu usability problems with sound. *Behavior and Information Technology, 18(3),* 165-177.

Brewster, S., Wright, P., & Edwards, A. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In S. Aslund, K. Mullet, A. Henderson, E. Hollnagel, & T. While (Eds.), *Proceedings of InterCHI'93* (pp. 222-227). Amsterdam: ACM Press.

Edworthy, J. (1998). Does sound help us to work better with machines? A commentary on Rauterberg's paper 'About the importance of auditory alarms during the operation of a plant simulator.' *Interacting with Computers, 10,* 401-409.

Fusco, M. & Katz, R. (1992). Catch a rising tone: Selecting a tone for a new calling service. In Proceedings of the Human Factors Society Society 36th Annual Meeting, (pp. 227-231). Santa Monica, CA: Human Factors and Ergonomics Society.

Fussell, S. & Benimoff, I. (1995). Social and cognitive processes in interpersonal communication: Implications for advanced telecommunications technologies. *Human Factors, 37(2),* 228-250.

Gaver, W. (1986). Auditory icons: Using sound in computer interfaces. Human *Computer Interaction, 2,* 167-177.

Gaver, W. (1989). The SonicFinder, a prototype interface that uses auditory icons. *Human Computer Interaction, 4,* 67-94.

Gaver, W. (1997). Auditory interfaces. In M. Helander, T. Landauer, P. Prabhu (Eds.), *Handbook of Human-Computer Interaction,* 2nd ed. (pp. 1003-1041). Amsterdam: Elsevier.

Graham, R. (1999). Use of auditory icons as emergency warnings: Evaluation within a vehicle collision avoidance application. *Ergonomics, 42(9),* 1233-1248.

Hearst, M., Albers, M., Barrass, S., Brewster, S., & Mynatt, E. (1997). Dissonance on auditory interfaces. *IEEE Expert: Intelligent Systems and their Application, 12(5),* 10-16.

Kamm, C. & Helander, M. (1997). Design issues for interfaces using voice input. In M. Helander, T. Landauer, P. Prabhu (Eds.), *Handbook of Human-Computer Interaction*, 2nd ed. (pp. 1043-1060). Amsterdam: Elsevier.

Leplatre, G. & Brewster, S. (2000). Designing nonspeech sounds to support navigation in mobile phone menus. In *Proceedings of ICAD 2000* (pp. 190-199). Atlanta: ICAD.

Massaro, D. W. (1975). *Experimental psychology and information processing.* Chicago: Rand McNally.

Moore, B. (1989). *An introduction to the psychology of hearing.* 3rd ed. London: Academic Press.

Mynatt, E. (1997). Transforming graphical interfaces into auditory interfaces for blind users. *Human Computer Interaction, 12*, 7-45.

Ogden, W. & Bernick, P. (1997). Using natural language interfaces. In M. Helander, T. Landauer, P. Prabhu (Eds.), *Handbook of Human-Computer Interaction*, 2nd ed. (pp. 137-162). Amsterdam: Elsevier.

Prabhu, P. & Prabhu, G. (1997). Human error and user-interface design. In M. Helander, T. Landauer, P. Prabhu (Eds.), *Handbook of Human-Computer Interaction*, 2nd ed. (pp. 137-162). Amsterdam: Elsevier.

Rauterberg, M. (1998). About the importance of auditory alarms during the operation of a plant simulator. *Interacting with Computers, 10*, 31-44.

Sadowski, W. & Lewis, J. (2000). Usability evaluation of speech user interfaces for three currency conversion prototypes. (Technical Report No. TR 29.3308). Raleigh, NC: International Business Machines.

Sikora, C., Roberts, L., & Murray, L. (1995). Musical vs. real world feedback signals. In *Proceedings of CHI'95*. (pp. 220-221). New York: ACM.

Williams, D. & Cheepen, C. (1998). 'The sound of silence:' A preliminary experiment investigating non-verbal auditory representations in telephone-based automated spoken dialogues. In *Proceedings of ICAD'98 International Conference on Auditory Display*. Swindon, UK: British Computer Society.