A Paradigm Shift in Interactive Computing: Deriving Multimodal Design Principles from Behavioral and Neurological Foundations

Kay Stanney
Industrial Engineering & Management Systems Department
University of Central Florida

Shatha Samman
Psychology Department, University of Central Florida

Leah Reeves
Kelly Hale
Industrial Engineering & Management Systems Department
University of Central Florida

Wendi Buff
Psychology Department, University of Central Florida
NAVAIR Orlando, Training Systems Division

Clint Bowers
Psychology Department, University of Central Florida

Brian Goldiez
Industrial Engineering & Management Systems Department
University of Central Florida

Denise Nicholson
Stephanie Lackey
NAVAIR Orlando, Training Systems Division

This article is based on work supported in part by the Office of Naval Research (ONR) under Grants N00014-02-1-0927 and N000140310326. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or the endorsement of the ONR.

Requests for reprints should be sent to Kay Stanney, University of Central Florida, Industrial Engineering & Management Systems Department, 4000 Central Florida Blvd., Orlando, FL 32816-2993. E-mail: stanney@mail.ucf.edu
As technology advances, systems are increasingly able to provide more information than a human operator can process accurately. Thus, a challenge for designers is to create interfaces that allow operators to process the optimal amount of data. It is herein proposed that this may be accomplished by creating multimodal display systems that augment or switch modalities to maximize user information processing. Such a system would ultimately be informed by a user’s neurophysiological state. As a first step toward that goal, relevant literature is reviewed and a set of preliminary design guidelines for multimodal information systems is suggested.

1. **INTRODUCTION**

When computers first permeated society, thoughts of the Turing Test arose, but were soon abated as users struggled with perplexing interfaces and contrived interaction techniques. To allay much of this frustration, human interaction with computer systems has evolved over the past several decades into a relatively standard set of interaction paradigms. Experienced users readily adapt to a new application if it leverages common constructs, such as windows, icons, menus, and direct manipulation via pointing devices (WIMPs). These paradigms have transcended traditional computing platforms and found their way into information appliances, such as cell phones, pagers, and personal digital assistants. The pervasiveness of these paradigms provides some measure of their success; yet they are quite limited in their ability to adapt to the capabilities and limitations of an individual user and many users still find them difficult to master. In fact, these interaction paradigms rebuff Grandjean’s (1980) suggestion that designers should fit the task to the human.

Rather than constraining users to conform to contrived interaction conventions, adaptable systems would respond to the capabilities and limitations of the individual user. Information would be presented when needed, in a modality that was readily perceived, and in a form that was readily interpretable. A new class of interactive systems would ensue, which could learn from a user’s experiences, adapt to a specific user’s perceptual and cognitive needs, and respond to such needs by adapting system components to facilitate intuitive interactions with users. Through a deep understanding of the individual user, a genuine symbiosis could manifest between user and system.

Two converging decades of research can be leveraged to achieve such symbiosis. The Decade of the Brain (Levy-Reiner, 1999) has led to the development of powerful imaging techniques that enable mapping of distinct and detailed functions of the brain, such that a computer user’s cognitive state can be captured. Efforts in cognitive engineering (Rasmussen, Pejtersen, & Goodstein, 1994) can be leveraged to model the design problems associated with interactive systems. These two areas can be used to fundamentally change the nature of human communication with computers. Through advanced brain sensing technologies, electromechanical interaction devices (e.g., mouse, joystick) can be augmented or replaced with electrophysiological interaction (e.g., electroencephalography [EEG], positron emission tomography [PET], functional magnetic resonance imaging [fMRI], and functional optical imaging via near infrared [fNIR]), such that subtle human physi-
ological indicators can be used to direct brain-computer interaction (Allanson, 2002). Through neurophysiological interactive computer systems (NICS), the cognitive state of the computer user can be characterized, particularly in terms of current load on the limited capacity of working memory (WM). Once sensing technology has captured the current state of the brain (e.g., visual WM is overloaded), then cognitive engineering principles can be used to direct how best to present information to users given their cognitive state. More specifically, these principles can be used to specify such issues as how to couple or transpose multiple modalities (vision, audition, haptic) to leverage more WM capacity or how to support simultaneous use of multimodal user responses (e.g., voice and gesture) to optimize human performance. The result of coupling sensing with cognitive modeling would be a system that perceives, interprets, and appropriately responds to an individual user's "signatory physiology" (p. 64, Allanson, 2002, p. 64).

This article focuses on the cognitive engineering aspects of neurophysiological interactive computer systems. Once sensing equipment has detected a user's cognitive state, cognitive engineering principles can be used to direct how to coordinate between sensing and interface presentation, particularly how to design multimodal interaction to leverage more WM capacity.

2. LEVERAGING COGNITIVE ENGINEERING AS AN INFRASTRUCTURE FOR NICS

The infrastructure of NICS will consist of user models that holistically characterize human information processing (i.e., attention, perception, WM, and decision making), as well as models of the system (i.e., production rules to read, store, rehearse, and recall), and how its interface should best be presented. Cognitive engineering can be leveraged to develop such models. Cognitive engineering is a multidisciplinary approach to system design that considers the analysis, design, and evaluation of interactive systems (Rasmussen et al., 1994; Vicente, 1999). It involves developing systems through an understanding of human capabilities and limitations, and focuses on how humans process information, aiming at leveraging this knowledge to design interactive systems that are predictable (i.e., respond to the way users perceive, think, and act; Eberts, 1994). Human information processing models have provided a conceptual framework through which to theorize and predict how humans will interact with a NICS system. Such models have decomposed human information processing into a perceive-think-act cycle, that has been further parameterized in terms of capacity, decay, and processing time (Card, Moran, & Newell, 1983; Wickens, 1992).

Current understanding of human information processing suggests that information is perceived through multiple sensory processors. This information is then perceptually encoded (i.e., stimulus is identified and recognized), processed by a WM subsystem which may be supported by long-term memory (LTM), to arrive at a decision, which in turn triggers a human response (Baddeley, 1990, 2000; Wickens, 1992). Within human information processing, WM has been referred to as "the hub of cognition" (Haberlandt, 1997, p. 212), and hence is considered to play
an essential role in the design of a NICS system. In general, WM is described as a functional multiple-component of cognition "that allows humans to comprehend and mentally represent their immediate environment, to retain information about their immediate past experience, to support the acquisition of new knowledge, to solve problems, and to formulate, relate, and act on current goals" (Baddeley & Logie, 1999, p. 29). It is considered a temporary active storage area where information is manipulated and maintained for executing simple and complex tasks (e.g., serial recall, problem solving). WM is thought to be comprised of separate processes required for temporary storage, attention allocation, and coordination of maintained information (Baddeley & Hitch, 1974). Since the development of Baddeley’s original WM model, which included a central executive and two modality-specific slave systems—the visuospatial sketchpad and phonological loop—subsystems of WM have seen further fractionations. The phonological loop has been fractionated into a passive phonological store and an active articulatory rehearsal process (Baddeley & Logie, 1999). The phonological store represents material in verbal code, which, if not rehearsed to refresh information, may decay with time (about 2 sec). The visuospatial sketchpad has also been decomposed into a passive visual cache and an active spatially-based system called the inner scribe (Logie, 1995). The visual cache is considered to store visual information, whereas the inner scribe maintains rehearsal of body movement and retains information of paths between objects and locations coded through vision, sound, or touch (Baddeley & Logie, 1999). The central executive has also been divided into two subprocesses, where the central executive controls attention and coordinates the two slave systems, whereas the episodic buffer serves as a storage system to integrate information from LTM and the two subsidiary slave systems (Baddeley, 2001). Converging evidence in behavioral (Baddeley, 1999), neuroimaging research (Smith & Jonides, 1998; Smith, Jonides, & Koepppe, 1996), and neuropsychological studies of brain-damaged patients (Carlesimo, Perri, Turriziani, Tomaiuolo, & Caltagirone, 2001; Mendez, 2001; Pickering, 2001) concur with the notion of separate multiple WM systems, as originally proposed by Baddeley and colleagues; although there is little consensus on the number of subsystems.

Recently, there have been a number of frameworks proposed to conceptualize the modal nature of WM subsystems. Barnard (1999) developed the Interacting Cognitive Subsystems (ICS) framework, which specifies nine physically separable subsystems that are component resources of the “complete” mental mechanism. Within this framework there are three sensory subsystems (i.e., visual, auditory, body state sources), two effector subsystems (i.e., limb [control of skeletal action] and articulatory [speech output]), and four central subsystems (i.e., two to handle abstract structural descriptions of auditory–verbal information and visuospatial information; one propositional subsystem that handles items in terms of their semantic identity, properties, and interrelations; and one implicational subsystem, which is more abstract and schematic, that integrates products of processing propositional meaning with information obtained from all sensory inputs). An alternative model proposed by Schneider (1999) is the hierarchical connectionist control architecture of CAP2 (Controlled Automatic Processing, version 2). This hierarchically organized WM model suggests there are at least eight subsystems at the macrostructure level.
(i.e., visual, auditory, speech, lexical, semantic, motor, mood, context) and 500 to 1,000 cortical areas containing hundreds of thousands of cortical submodules at the microstructure level sharing similar input and output functions. These theories and numerous others (see review by Miyake & Shah, 1999) similarly argue that although they may be physically separable, the different WM subsystems share the same internal structure, are regulated by similar processing principles, and may share similar properties (e.g., capacity, decay rate). Other models similarly assume different domain-specific codes or partitions in WM (e.g., visual, tactile, motor, body state), even when they may not explicitly postulate physically separable subsystems (for a summary of alternative views, see Kintsch, Healy, Hegarty, Pennington, & Salthouse, 1999). It is important to note that few if any current WM models positively deny the existence of different codes or representations (e.g., separate storage buffers; Miyake & Shah, 1999). Thus, regardless of which theory one ascribes, from a processing perspective, multimodal interaction has promise because the WM subsystems are somewhat independent and tend to act cooperatively rather than competitively (i.e., do not entirely compete for the same processing resources). Further, improved performance (e.g., greater recall when information is organized by modality) is generally realized when tasks with different modalities are used as compared to a single modality, likely due in part to selective interference effects generally being less across modalities as compared to within modalities.

Beyond Barnard (1999) and Schneider (1999), another WM view (Sulzen, 2001) purports not only the existence of separate WM, but also separate processing capabilities for each sensory modality (i.e., visual, auditory, haptic, kinesthetic, olfactory, gustatory), as well as further higher-level subsystems for emotional and affective, linguistic, and spatial processes. This model suggests that if cognition is modally organized, each modality may have its own representational system, with its own resources. It further postulates that simple stimuli sets limited to a single modality will more likely primarily be encoded in that modality and not reencoded or cross-stored, with exceptions such as the tendency to recode linguistic terms among visual and auditory percepts. For example, an image of a horse (visual) or an auditory verbal sound of the word horse will both be cross-stored in a linguistic code. In terms of WM capacity, this model suggests and provides compelling empirical evidence to suggest that total capacity could be approximately equal to the sum of the capacities of the individual modal WM (ignoring duplicate cross-modal encoding effects). If this notion holds true, one could design multimodal displays to fill each modal WM independently of others. This should lead to increased WM capacity versus utilizing a more random design strategy.

Figure 1 provides a framework that can be used to direct multimodal interaction design based on a multiple WM subsystem model. If such a model were coupled with NICS, then neurophysiological sensors could be used to detect activation of various WM modalities, such that when one WM modality is overloaded the model would direct information to be transposed to another WM modality (e.g., present visual information via audition), thereby increasing human information processing capacity.

Given the framework depicted in Figure 1, a critical design question becomes, once neurophysiological sensors indicate that a WM subsystem is at capacity,
which design principles can be used to ensure information is redirected such that it is perceived and understood? To address this question it is important to consider both the particularities of each individual modality, as well as cross-modal effects among WM subsystems.

3. THE SENSORY MODALITIES

When designing multimodal interaction a central question is which information should be conveyed via which modality. The answer to this question lies in the determination of the types of information each modality is particularly suited to display. The sensory modalities can be differentiated by at least five criteria: (a) their receptive organs, (b) characteristic stimuli to which they respond, (c) nerves they innervate, (d) area of the brain they activate, and (e) sensations they convey (Kandel, Schwartz, & Jessell, 1995; Neff, 1960). Given these criteria, there are at least nine sensory modalities (i.e., vision, audition, tactile sense, kinesthetics, vestibular sense, temperature sense, olfaction, gustatory sense, and pain). Although any of the senses could be leveraged, given current technology only four (i.e., vision, audition, as well as tactile and kinesthetic [i.e., haptics]) are currently viable to incorporate into system design.

3.1. The Visual Modality

The visual modality (see Figure 2) is particularly effective in conveying spatial information, however, it can also be effective in presenting verbal–textual information, par-
particularly when long messages are being conveyed (Wickens, 1992). In general, visual presentation of information can be in the form of graphics, text, or animation. More specifically, visual displays are well-suited for conveying spatial relations (e.g., size, distance), two-dimensional localization, abstract or coded information, information representing change over time, information representing real-world physical objects, information requiring persistent attention, and information concerning absolute quantitative parameters (European Telecommunications Standards Institute, 2002; Watzman, 2003). Animation is good for drawing attention and can aid in identification, interpretation, transition schemes, orientation, choice decision making, demonstration, explanation, feedback, history, and guidance (Baecker, Small, & Mander, 1991). Considering the traditional speed-accuracy trade-off, if accuracy is essential or continuous feedback is required (e.g., tracking tasks); then again, visual presentation is preferred over other modalities (Mulgund, Stokes, Turieo, & Devine, 2002). When comparing graphics versus textual visual presentations, graphics (e.g., symbols) are preferred if speed is essential and text is preferred if accuracy is critical, whereas a combination may prove most effective (Watzman, 2003). To reduce attentional demand in multimodal interaction, one can leverage a number of basic Gestalt Principles (Koffka, 1935) that assist in determining how to group elements in visual displays (i.e., similarity, proximity, closure; see Palmer, 1992). Table 1 presents a summary of visual display guidelines that can guide multimodal interaction design.

3.2. The Auditory Modality

The auditory modality (see Figure 2) is highly effective at conveying instructions and other relevant information via speech (Hapeshi & Jones, 1992). Verbal instructions (rather than visual-textual) are particularly effective when a listener is performing a task or is in motion. Although the auditory modality is vastly adept for speech, it can also serve as a source of spatial localization (Blauert, 1996), although its spatial acuity is far worse than visual or proprioceptive localization (Shilling & Shinn-Cunningham,
<table>
<thead>
<tr>
<th>Presentation</th>
<th>Guideline</th>
</tr>
</thead>
</table>
| Graphics     | Graphics are better than text or auditory instructions for communicating spatial information:  
• Use graphics to illustrate components (e.g., equipment diagram) or spatial relationships (e.g., map, floor plan) and to clarify concepts/complex tasks (Williams, 1998).  
• Combine graphics with text to improve comprehension (European Telecommunications Standards Institute, 2002). |
| Text         | Text may be used to convey detailed long information, procedures, instructions, labels, annotations, and clarifications (Wetzel, Radtke, & Stern, 1994). |
| Animation    | • Animation may be effectively used as a redundant visual cue (Park & Hopkins, 1993).  
• Use motion to enhance detection of objects in periphery or overcome poor illumination.  
• Use animation to demonstrate sequential actions in procedural tasks. |
| General layout | Use “Gestalt Rules” to increase users’ understanding of relations between elements (Palmer, 1992):  
• Place related objects close together.  
• Enclose related objects by lines or boxes.  
• Move or change related objects together.  
• Related objects should look alike (e.g., shape, color, size or typography). |
| Color        | • Use color to aid visual search, indicate state, draw attention, communicate qualitative/quantitative differences (Post, 1997).  
• Design displays that require relative judgment via color (“differentiation tasks”); avoid absolute judgment (“recognition tasks”) via color (Sanders & McCormick, 1993). |
| Group relations | Use numbered lists to show groups of related items, with a specific order (Watzman, 2003).  
• Use flow charts to show relationships/steps involved in a process.  
• Use tables to show relationships between categories of ideas.  
• Use project plan tables to show relationships of tasks over time. |
| Evaluate and compare | Use rating tables to evaluate items against several criteria (Watzman, 2003).  
• Use comparison tables to evaluate items against one criteria.  
• Use matrix graphs to compare more than one item to more than one variable.  
• Use bar charts to compare several things in relation to one variable.  
• Use pie charts to compare relative parts that make up a whole. |
| Hierarchy concepts | Use organizational charts to show hierarchical structure (Watzman, 2003).  
• Use illustrations and/or text to show basic concepts.  
• Use illustrations with text and/or icons, other graphics, complex images, interactive components to show abstract concepts. |
| 2D/3D        | • To extract critical information use 2D graphs, as users often perform better with respect to accuracy and ease (Watzman, 2003).  
• Incorporate 3D into graphics to enhance aesthetics. |

Note. 2D = two-dimensional; 3D = three-dimensional.
The auditory sense can detect sounds ranging from 20 to 20,000 Hz, over a multitude of intensities, yet with varied precision, with localization within 1 m and along the transverse plane being superior to the median plane (Kandel et al., 1995). Such spatialized audio cues can be used to communicate direction, location, and movement, as well as in identifying auditory messages in noisy environments or guiding navigation tasks (Mulgund et al., 2002). One must be cautious when using sound localization for absolute judgments, as distance is generally overestimated and front-back directional reversals can occur (Caelli & Porter, 1980; Kramer, 1994). Supernormal auditory localization can be used to exaggerate normal auditory cues so listeners are better able to localize sounds (Shinn-Cunningham, Durlach, & Held, 1998a, 1998b).

Auditory cues are also well-suited for rapid cuing of critical information, such as for warning and alarms (Sanders & McCormick, 1993). When reaction time (RT) is essential, auditory warnings are generally superior to visual warnings, as they more efficiently draw attention to critical information (30 to 40 msec faster than vision; Welch & Warren, 1986; see also Bly, 1982). In fact, the acute temporal resolution of the auditory system is one of its greatest assets (Kramer, 1994). When it comes to the specific auditory cue to use, tones are good for communicating limited information sources (e.g., start vs. stop time); complex sounds are well-suited to alarms; speech is most effective for rapid communication of complex, multidimensional information sources (Blackwood et al., 1997); and timbre (i.e., sound quality) is an effective auditory grouping cue (Brewster, 2003). Auditory cues are not suitable for high resolution display of quantitative information, as it can be used to reliably convey only a very limited number of levels (e.g., distinct values of pitch, loudness, or timbre; Kramer, 1994). Table 2 summarizes auditory display guidelines that can guide multimodal interaction design.

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>General–thresholds</td>
<td>• Sounds should be approximately 500 msec in duration to be heard (Sanders &amp; McCormick, 1993).</td>
</tr>
<tr>
<td></td>
<td>• Sound frequencies generally should be between 20 and 20000 Hz to be heard: preferably between 500 and 3000 Hz.</td>
</tr>
<tr>
<td></td>
<td>• For sound traveling long distances (e.g., &gt; 1000 ft), use frequencies below 1000 Hz, as higher frequencies will not travel as far.</td>
</tr>
<tr>
<td></td>
<td>• Sound intensity should be between about 40 to 120 dB (SPL) to be heard, sounds above 120 dB can cause pain.</td>
</tr>
<tr>
<td>Speech</td>
<td>• Set speech output speed at 150 to 160 words per minutes; do not exceed 210 (European Telecommunications Standards Institute, 2002).</td>
</tr>
<tr>
<td></td>
<td>• Utilize intensity differences to aid in the discrimination of speech stimuli (Jancke, Shah, Posse, Grosse-Ryken &amp; Muller-Gartner, 1998).</td>
</tr>
<tr>
<td></td>
<td>• Never present two verbal tasks at the same time (e.g., two messages; ETSI, 2002).</td>
</tr>
<tr>
<td></td>
<td>• Avoid background speech (even a whisper at 50dB), which impairs reading and memory (Hapeshi &amp; Jones, 1992).</td>
</tr>
<tr>
<td></td>
<td>• Synthetic speech has been shown to hinder verbal learning and should be avoided if possible.</td>
</tr>
</tbody>
</table>

(Continued)
### Table 2 (continued)

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Guideline</th>
</tr>
</thead>
</table>
| Sound localization and pitch | Auditory cues can be spatialized to indicate direction, location, and movement:  
  - When using sound localization, use an ITD of at least 10 to 50 μsec in azimuth rather than elevation (Blauert, 1996).  
  - If using spatialized audio cues to communicate direction or location, it is important to note that localization performance decreases as the source approaches the median plane and distance to the source increases (especially beyond 1 m; Kandel, Schwartz, & Jessell, 1995).  
  - Use dynamic cues to aid sound localization in the median plane (i.e., IID/ITD of zero; Fisher & Freedman, 1968; Middlebrooks & Green, 1991).  
  - If using spatialized audio cues to communicate movement, position source in front of the listener (i.e., 0 deg azimuth; do not exceed ± 40 deg; Strybel, Manligas, & Perrott, 1992).  
  - Dynamic localization of sounds are poorest in extreme azimuth (> 40 deg) and elevation (> 80 deg of horizontal plane; Strybel et al., 1992).  
  - Use spatialized audio to aid identification of auditory messages in noisy environments (Mulgund, Stokes, Turiero, & Devine, 2002).  
  - Use spatialized audio to guide navigation tasks (e.g., convey waypoints or object/person locations; Mulgund et al., 2002).  
  - Use supernormal auditory localization to exaggerate normal auditory cues (Shinn-Cunningham, Durlach, & Held, 1998a, 1998b).  
  - Avoid using sound localization for absolute judgments (Caelli & Porter, 1980; Kraemer, 1994). |
| Alerts and warnings |  
  - Use auditory cues for rapid cueing of critical information, such as for warnings and alarms (Sanders & McCormick, 1993).  
  - Auditory warnings are effective for messages dealing with time-relevant events or continuously changing information and when requiring immediate action (Welch & Warren, 1986).  
  - Do not use auditory warning messages if information requires future referencing (Blackwood et al., 1997).  
  - Keep spectral components of a warning 15 dB above threshold imposed by background noise (Patterson, 1989).  
  - Incorporating apparent motion in the design of auditory warning sounds will make the warning more salient. |
| Auditory cues |  
  - Use tones for communicating limited information sources (e.g., start and stop time; Blackwood et al., 1997).  
  - Use complex sounds for alarms (e.g., deviations in rhythm, pitch, loudness, timbre).  
  - Speech is most effective for rapid communication of complex, multidimensional information sources.  
  - If auditory cues are used as a quantitative indicator, use speech as it involves minimum time or errors in attaining an exact value.  
  - Use timbres (i.e., sound quality) with multiple harmonics as an effective grouping cue (Brewster, 2003).  
  - Avoid using auditory cues for high resolution display of quantitative information (Kramer, 1994). |

**Note.** IID = interaural intensity differences; ITD = interaural time differences.
3.3. The Haptic Modality

Haptic displays are advantageous when a task is temporal in nature, requires the alarming feature of attention, and involves hand–eye coordination (e.g., object manipulation), where haptic sensing and feedback are key to performance (Biggs & Srinivasan, 2002; Mulgund et al., 2002; Popescu, Burdea, & Treffitz, 2002; Posner, 1976). In addition, when it comes to patterns and other “spatial and temporal” information, the potential exists to leverage the haptic modality. Although visuospatial information is thought to be best presented via visual imagery (Wickens, 1992), it could alternatively be conveyed via tactile or kinesthetic cues (e.g., touch, motion; see Figure 2). For example, Bach-y-Rita (1999) has demonstrated the ability to substitute spatial information presented visually via tactile “vision.” Rupert (1997) has used tactile cues to provide cues to resolve spatial disorientation in aviation environments. Tan, Lim, and Traylor (2000) have described how to design a haptic driving navigation guidance system by leveraging a spatiotemporal illusion of movement across the back known as “sensory saltation,” where three to six mechanical sensors are placed no greater than 10 cm apart along the back and emit vibratory pulses with an interstimulus duration of 50 msec.

Tactile cues, such as those conveyed via vibrations or varying pressures, can provide information concerning location, texture, softness, and surface viscosity, as well as serve as effective simple alerts (European Telecommunications Standards Institute, 2002; Mulgund et al., 2002). The vibrotactile sense is comparable in discriminatory ability to audition for frequencies up to about 50 Hz, after which point audition is far superior (Goff, 1967).

Kinesthesia is an awareness of the movements and relative position of body parts and is determined by the rate and direction of movement of the limbs and static position of the limbs when movement is absent through tension signals originating from sensory receptors in the joints, skin, muscles, and tendons (Biggs & Srinivasan, 2002). Kinesthetic cues can provide force feedback and convey gestures. Such cues can also be used to stimulate anticipation of a change, provide feedback confirming reception of a user input, provide an indication of current state, guide user interaction toward a desired position or location, and make clear distinctions between orthogonal directions (Miller & Zeleznik, 1999). They can also aid discrimination (e.g., length) and identification (e.g., shape), as well as spatial location memory (Tan, Stefanucci, Proffitt, & Pausch, 2002). Table 3 summarizes haptic display guidelines that can guide multimodal interaction design.

3.4. Other Sensory Modalities

Although the technology is not quite operational, it is informative to consider the design benefits that may be derived by incorporating additional modalities, such as olfaction, gustation, vestibular stimulation, pain, and temperature. Olfaction, which is best conveyed via different odors rather than variations of a given odor, holds particular promise as it can be used to manipulate mood, decrease stress, in-
Table 3: Haptic Design Guidelines

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Guideline</th>
</tr>
</thead>
</table>
| Tactile      | • Vibrotactile sense is comparable in discriminatory ability to audition for frequencies up to about 50 Hz (Miller & Zeleznik, 1999; Sherrick & Cholewiak, 1986).
• Detection of vibration for a single probe is about 28 dB (relative to 1 (u peak) for 0.4 to 3 Hz.
• To ensure user perceives individual signals, stimuli must be separated by at least 5.5 msec and preferably > 10 msec.
• To successfully activate a user’s pressure sensors, force exerted must be > 0.06 to 0.2 N/cm².
• Humans can detect pressure variations up to 0.002 in per 100 msec, with an optimal response at 400 Hz, and loss of sensitivity below 50 Hz and above 600 Hz.
• Tactile input must consider sensitivity to stimuli across various skin locations (i.e., 2-point threshold becomes smaller from palm to fingertips).
• Spatial resolution of a point stimulus on finger pad is about 0.15 mm and for a 2-point limen about 1 mm.
• Amplitudes above 0.6 mm to 0.8 mm are painful.
• Humans can detect presence of a 2 μm high single dot and a 0.075 μm high grating.
• Be aware that surface characteristics of a stimulus influence sensation of touch.
• Avoid use of tactual displays in low temperature environments because tactual sensitivity is degraded. |

Alerts and warnings
Tactile cues can provide effective alerts via vibrations or variations in pressure; they can be augmented by or substituted for auditory warning cues (e.g., automatic alerts, reception of coded messages such as Braille; Posner, 1976):
• If using tactile cues for warnings, it is important to note humans can identify about four haptic intensities, about five durations, and about nine different frequencies (20% difference needed between levels; Geldard, 1972).

Tactile localization
Tactile cues can be augmented by or substituted for visual tasks to aid localization (e.g., identification of controls, tactual maps as navigation aids, tracking-task displays):
• Humans can detect about seven haptic locations on the chest (Geldard, 1972).
• Use distal body parts if high spatial resolution is required (above 4 cm any body part can be used; Sherrick & Cholewiak, 1986).
• Tactile input can be incorporated into complex applications to provide orientation/direction (Rupert, 1997).
• Tactile location cues (e.g., up or down) can resolve spatial disorientation.
• To convey movement, one can leverage the spatiotemporal illusion of movement (i.e., sensory saltation) using 3 to 6 mechanical sensors placed no greater than 10 cm apart along the back, which emit vibratory pulses with an interstimulus duration of 50 msec (Sherrick & Cholewiak, 1986).

(continued)
<table>
<thead>
<tr>
<th>Presentation</th>
<th>Guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture, softness,</td>
<td>Tactile cues can be used to convey properties of simple objects (complex objects require multimodal presentation; Popescu, Burdea, &amp; Treffitz, 2002):</td>
</tr>
<tr>
<td>surface viscosity</td>
<td>• Sensation of textured surfaces requires some relative motion between surface and skin to be maintained.</td>
</tr>
<tr>
<td></td>
<td>• For a hard surface to be felt after initial contact active pressure must be maintained.</td>
</tr>
<tr>
<td></td>
<td>• Soft surfaces exert and maintain a slight positive reaction against the skin after the initial contact without active pressure or relative motion.</td>
</tr>
<tr>
<td>Kinesthetic</td>
<td>Kinesthetic cues can stimulate anticipation of a change, provide feedback confirming reception of a user input, provide an indication of current state, guide user interaction toward a desired position or location, make clear distinctions between orthogonal directions, and aid discrimination (e.g., length) and identification (e.g., shape; Miller &amp; Zeleznik, 1999):</td>
</tr>
<tr>
<td></td>
<td>• To maintain optimal hand–eye coordination, the delay (i.e., lag) between sensing and kinesthetic feedback should be less than 100 msec.</td>
</tr>
<tr>
<td></td>
<td>• When interacting with objects, allow adequate time for users to respond kinesthetically (minimum 250 msec for simple reaction time).</td>
</tr>
<tr>
<td></td>
<td>• Avoid static positions at or near end range of motion to reduce fatigue.</td>
</tr>
<tr>
<td>Gestures</td>
<td>Gestures can be used to communicate meaningful information in isolation (e.g., hand signals) or in combination with speech and/or visual information (e.g., ‘put that there’; Oviatt, 1997; Turk, 2002):</td>
</tr>
<tr>
<td></td>
<td>• Adding gesture to speech offers speed, high-bandwidth information, flexibility of input modes, enhanced error avoidance, and relative ease of use.</td>
</tr>
<tr>
<td></td>
<td>• Gestures should be intuitive and simple; avoid increasing user’s cognitive load with numerous or complex gestures.</td>
</tr>
<tr>
<td></td>
<td>Gestures can be used to manipulate the environment (Oviatt, 1997; Turk, 2002). Gestures are a natural, flexible input mode, and can effectively be used for control and navigation:</td>
</tr>
<tr>
<td></td>
<td>• Avoid temporal segmentation of gestures.</td>
</tr>
<tr>
<td></td>
<td>• Avoid frequent, awkward, or precise gestures to minimize user fatigue.</td>
</tr>
<tr>
<td></td>
<td>• Gestures can be effectively used for spatial tasks (e.g., resizing, moving objects; Popescu et al., 2002; Turk, 2002):</td>
</tr>
<tr>
<td></td>
<td>• Inform user of types of for which gestures allowed and what affect each will have on system interaction.</td>
</tr>
<tr>
<td></td>
<td>• Avoid precise motion gestures, as it is difficult to make highly accurate or repeatable gestures with no tactile feedback.</td>
</tr>
<tr>
<td>Kinesthetic localization</td>
<td>Kinesthetic cues (e.g., movement cues, impedance) can aid spatial location memory (Popescu et al., 2002):</td>
</tr>
<tr>
<td></td>
<td>• Kinesthetic is best coupled with other modalities to aid location memory of objects in space relative to one’s location.</td>
</tr>
</tbody>
</table>

(continued)
Table 3 (continued)

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haptics (i.e., tactile and kinesthetic—a.k.a force feedback)</td>
<td>Haptic feedback can direct manipulation of objects and improve performance on temporal tasks:</td>
</tr>
<tr>
<td></td>
<td>• Add haptic interfaces at distal body segments (e.g., hands) where more useful interaction can occur (Biggs &amp; Srinivasan, 2002).</td>
</tr>
<tr>
<td></td>
<td>• Provide two-handed haptic interfaces to maximize performance gains.</td>
</tr>
<tr>
<td></td>
<td>• Use caution when implementing finger tip haptic devices for tasks requiring sustained force because they can only exert 4.7 N over sustained periods (Shimoga, 1993).</td>
</tr>
<tr>
<td></td>
<td>• If response time is critical, do not use haptics alone because user will take longer to perceive and attend to information.</td>
</tr>
</tbody>
</table>

crease vigilance, and aid memory (European Telecommunications Standards Institute, 2002; Mulgund et al., 2002). Humans can detect approximately 10,000 different odors; however, if only intensity is varied, one can detect only about three to four different smells at a time (Goldstein, 1999). Olfactory cues are generally most appropriate for conveying affective or ambient information, as well as slow-moving, medium duration information, as odors linger and their persistence varies. Due to the nature in which the olfactory receptors are distributed, olfactory cues are not appropriate for conveying spatial information (Kandel et al., 1995).

Gustatory cues could convey categorical information along the different tastes (i.e., sweet, sour, bitter, salty) and may provide some benefits when coupled with olfaction or speech (European Telecommunications Standards Institute, 2002). If users need to judge differences in taste concentration, differences of 15% to 25% should be utilized to ensure users can differentiate them (Goldstein, 1999). Vestibular cues, which provide information associated with gravity and acceleration, can assist in spatial updating during virtual motion, particularly when visual turn information is degraded (Riecke, von der Heyde, & Bülthoff, 2002). Yet vestibular cues can provide confusing spatial orientation cues, particularly when there is a lack of equivalence between gravity and acceleration (e.g., when banking an airplane, feelings of acceleration overwhelm those of gravity and thus the sense of upright is distorted; Benson, 1999). Pain can signal information of a protective nature, as well as evoke strong emotional response, whereas thermal sensations can be used to affect motivational state (Kandel et al., 1995). Thus, all the senses have potential to augment multimodal interaction design.

A comparison of the sensory modalities and their theorized suitability for conveying various information sources has been summarized in Table 4. Research is needed to validate the suggested relations in this table.

4. CROSS-MODAL INTERACTION AND ILLUSIONS

Although each modality is well-suited to display certain information types, there is much to be garnered from a multimodal interaction design approach that leverages
Table 4: Theorized Suitability of Sensory Modalities for Conveying Various Information Sources

<table>
<thead>
<tr>
<th>Information Source</th>
<th>Visual</th>
<th>Auditory</th>
<th>Haptic</th>
<th>Olfactory</th>
<th>Gustatory</th>
<th>Vestibular</th>
<th>Pain</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal</td>
<td>o</td>
<td>++</td>
<td>+</td>
<td>--</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial</td>
<td>++</td>
<td>0</td>
<td>+</td>
<td>--</td>
<td>+/−</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2D localization</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+/−</td>
</tr>
<tr>
<td>3D localization</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alerts and warnings</td>
<td>−</td>
<td>−</td>
<td>++</td>
<td>+</td>
<td>−</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Fast reaction time</td>
<td>++</td>
<td>++</td>
<td>o</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Persistence</td>
<td>++</td>
<td>−</td>
<td>++</td>
<td>−</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memorability</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative</td>
<td>+</td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absolute</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private and confidential</td>
<td>o</td>
<td>−</td>
<td>++</td>
<td>−</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outside area of interest</td>
<td>−</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(periphery)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructions</td>
<td>o</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Object properties</td>
<td>++</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motion</td>
<td>+</td>
<td>0</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affective and emotive</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivational</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self (body) protective</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Adapted from European Telecommunications Standards Institute, 2002. ++ = best modality; + = next best; o = neutral; − = not well suited, but possible; −− = unsuitable; 2D = two-dimensional; 3D = three-dimensional.

Cross-modal effects. For example, in object recognition tasks, when various sensory modalities are combined, a process known as "perceptual integration" facilitates detection of elements of the object by amplifying sensory signals and combining them to form a new, multimodal representation of the object (O'Hare, 1991). Perceptual integration occurs even when sensory modalities are not present. For instance, event related potential (ERP) studies have shown that tactile stimuli can produce systematic modulations of visual ERPs, providing evidence of cross-modal links (Eimer & Driver, 2000). This was demonstrated in a tactile object recognition (TOR) task, where no visual feedback was provided, yet activation in cortical regions serving somatosensory, motor, and visual systems was found; thus suggesting that the tactile percept modulated visual ERPs (Diebert, Kraut, Kremen, & Hart, 1999). It was suggested that regardless of the external cues presented, TOR tasks may use the visual system to access an internal object representation. Hence, multimodal representation, whether sensorially or cognitively generated, enhances object recognition tasks.

The "redundant-signal" effect proposes that RT to stimuli containing redundant bimodal information is shorter than to unimodal stimuli (Miller, 1982). The
“coactivation model” suggests that the advantage of redundant signals is a result of parallel processing that occurs in various unimodal channels, which enhances human information processing at the sensory processing, response selection, or response execution stage. Several psychophysical studies have provided results showing that such parallel processing, and its related RT advantages, occurs at all three levels (Fournier & Eriksen, 1990; Hughes, Reuter-Lorenz, Nozawa, & Fendrich, 1994; Miller, 1982).

This RT advantage is also likely due to cross-modal neural activations that occur during multimodal interaction. Several neurophysiological studies suggest that the best-known region of multimodal convergence and integration occurs in the superior colliculus (SC) of the midbrain, which plays an essential role in orientation of behavior (Giard & Peronnet, 1999; Hillyard, Simpson, Woods, Van Voorhis, & Munte, 1984; Kohlrausch & Van de Par, 1999). Multisensory neurons in the SC receive and converge information from at least three sensory modalities (visual, auditory, and somatosensory), where each modality is represented by a sensory space map (King & Palmer, 1985). Not only does the SC map input stimuli from different modalities, it also integrates information from the modalities by increasing the number of impulses in a multiplicative ratio (Meredith & Stein, 1986). For example, a weak visual stimulus can be enhanced with a strong auditory stimulus, which has been demonstrated empirically by Neuman, Crigler, and Bove (1991). Multisensory interaction is further suggested to be mediated by the anterior part of the right temporal lobe or inferior lateral prefrontal cortex (or both; Giard & Peronnet, 1999). Both regions have been described as integration areas. Additionally, the right fronto-temporal region has been shown to activate when identifying a multimodal object but not a unimodal stimulus. Similarly, Macaluso, Frith, and Driver (2000), using PET studies, found that in addition to some modality-specific cortical regions, there was bimodal activation of at least one cortical area during a spatial attention task of two or more objects. This study indicated that sustained attention tasks are mediated by both unimodal and multimodal processing, thus supporting that unimodal processing may be slower than multimodal processing.

Neurophysiological evidence further characterizes the RT advantages of multimodal interaction. These studies suggest that multimodal integration begins very early in sensory processing. For example, activation of the right occipito-parietal region has been shown to occur as early as 40 msec after multimodal (visual and auditory) stimulus presentation; similar activation was not found with a unimodal stimulus (Giard & Peronnet, 1999). In addition, at 90 to 145 msec visual areas in the extrastriate cortex were activated, which was thought due to stimulus salience enhancement derived from the addition of an auditory cue. An interesting issue that emerged from the Giard and Peronnet study was that larger areas in the right hemisphere than the left hemisphere were activated. The strong predominance of the right hemisphere in multisensory perceptual interaction suggests that most multisensory integration may occur in that hemisphere.

Given the neurophysiological and behavioral benefits of multimodal interaction (i.e., enhanced perception, speeded RT, etc.), it is important to establish "integration rules" that direct how best to activate the multisensory neurons (Giard & Peronnet, 1999). Following is a set of proposed integration rules, which should be
further evaluated in terms of their effectiveness for directing multimodal interaction design.

The rule of “temporal and spatial coincidence” suggests that the largest integration of stimuli occur when information from different modalities are close temporally or spatially (Driver & Spence, 2000). For example, Bolla, D’Angelo, and McKinley (1999) investigated the use of spatialized audio in visual target detection and found simultaneous presentation of three-dimensional audio cues, emanating from the same spatial location as a visual target, decreased search times. Perrot, Sadralodabai, Saberi, and Strybel (1991) also found auditory cues to be useful in visual target detection especially when a shift in gaze was required. Similarly, a “frontal speech advantage” has been demonstrated, where participants’ driving performance increased when the focus of visual and auditory attention were from the same source (straight ahead) rather than when attention was divided between front (visual) and side (auditory; e.g., as with a cellular phone ear piece; Spence & Read, 2002). Frassinetti, Pavan, and Ladavas (2002) further defined the optimal spatial proximity coupling, demonstrating that acoustic stimuli produced a greater increase of visual response when cross-modal stimuli were spatially co-located at a distance of approximately 16 deg. Therefore, RT advantage may be contingent on how a cross-modal cue is presented. Similarly, when visual or auditory attention is endogenously directed via tactile stimuli to one side, RTs to both types of stimuli are found to be faster on the cued side, suggesting cross-modal links between spatial attention in vision, audition, and touch (Eimer, Cockburn, Smedley, & Driver, 2001; Eimer, van Velzen, & Driver, 2002). Such cross-modal cueing effects have been shown neurologically, for example through larger ERP signals for visual events following a tactile stimulus (Kennett, Eimer, Spence, & Driver, 2001). Several studies have reported that these cross-modal cueing effects follow an external spatial frame of reference (posture-independent) model rather than a hemispheric (anatomical) model (Eimer et al., 2001; Kennett et al., 2001). Based on the aforementioned findings, designers can take advantage of these cueing effects based on external location to decrease RT, and may be able to use a unimodal cueing strategy to direct multimodal interaction.

In terms of temporal coincidence, the “intersensory facilitation effect” specifies how temporal coupling should be integrated, suggesting that target latencies can be reduced by introducing a preceding sensory accessory (e.g., an auditory cue) to facilitate target acquisition in an alternative modality, say visual or haptic target detection (Heuermann & Colonius, 2001).

The concept of temporal coincidence (or immediately preceding) is altered, or shall we say clarified, when characterized neurophysiologically (Hillyard et al., 1984; Kohlrausch & Van de Par, 1999). More specifically, electrophysiological experiments using evoked potential of single cell recording of multisensory neurons have found that the delay for transduction from receptors to the cells in the SC vary by modality, where the range is 6 to 25 msec for auditory stimuli, 12 to 30 msec for somatosensory stimuli, and 40 to 120 msec for visual stimuli (Stein & Meredith, 1993; Stein, Wallace, & Meredith, 1995). Therefore, for both visual and auditory stimuli to occur simultaneously at the SC level, neurophysiological experiments suggest that auditory signals should be designed with a temporal lag to ensure they reach peripheral receptors 50 to 80 msec later than visual stimuli (Kohlrausch
& Van de Par, 1999). Further research is needed to examine the effects of various modality inputs and individual difference variability to ensure designers know how best to achieve desired effects, such as simultaneity.

Both spatial and temporal coincidence may be important for sensory integration. Grant and Seitz (2000) found that speech detection increased when visual cues (i.e., facial movements) were temporally and spatially concordant with auditory stimuli than when auditory stimuli were presented alone. Taken together, such studies provide a possible physiological basis for cross-modal integration and suggest that "cross-modal links ... can actually influence unimodal stages of sensory processing, presumably via feedback projections from multimodal areas" (Kennett et al., 2001, p. 474).

From Baddeley (1990, 2000), Barnard (1999), Schneider (1999), and Sulzen’s (2001) research, the role of “working memory capacity enhancement” can be derived; which suggests that to extend WM capacity, sensory stimuli should be directed to a multitude of sensory modalities, while avoiding extensive cross-encoding among visual and auditory percepts into linguistic terms.

The “congruency effectiveness rule” suggests that certain congruent combinations of cross-modal percepts will yield significantly faster RT than incongruent combinations (Dyson & Quinlan, 2002). For example, Melara and O’Brien (1987) found that RTs were significantly shorter for congruent pairings of high pitch-high position (object placed above fixation on visual display) and low pitch-low position (object placed below fixation on visual display) pairings relative to RTs of incongruent pairings. In a similar study, Melara (1989) used a combination of pitch and color and found shorter RTs for congruent stimuli of white color-high pitch or black color-low pitch, as opposed to incongruent pairings (e.g., black color-high pitch).

The “spatial orientation augmentation rule” suggests that if head movements are coupled with visual scene updating, then individuals are better able to maintain spatial orientation (Chance, Gauent, Beall, & Loomis, 1998). Similarly, head movements are known to enhance auditory localization (Fisher & Freedman, 1968). These findings are likely related to the belief that the concept of “locale” is independent of any single sensory system, and thus orientation is better maintained when the multimodal spatial senses yield congruent information (Freedman & Rekosh, 1968).

The “magnitude or inverse effectiveness rule” proposes that the less effective unimodal stimuli are, the larger the enhancement they can have when combined with other stimuli (due to the multiplicative factors). For example, Vroomen and de Gelder (2000) found that when participants were presented with a rapid sequence of visual displays and a series of tones, object detection was easier when visual detection was coupled with a unique sound.

Although this set of rules provides a foundation for directing multimodal interaction design, additional integration rules are likely required. In addition, when designing multimodal interaction, there are many instances in which it may be valuable to transform between two modalities, such as changing from visual-text to auditory-verbal instructions when an individual is visually involved in other activity. Thus, rules for how to transform between modalities (e.g., transpose versus augment) and what the timing of such transformations should be are also needed.
Notwithstanding the benefits of multimodal interaction, there are challenges to those seeking to establish multimodal design principles, particularly given intersensory illusions that can occur, such as visual dominance over audition and haptics, and other cross-modal perceptual phenomena, such as perceived duration effects (Storms, 2002). Perceived duration effects occur because information that is objectively the same in two modalities may not always be perceived as the same; there can be incongruency in the perceived duration of stimuli presented to different modalities (Behar & Bevan, 1961; European Telecommunications Standards Institute, 2002). In general, whereas temporal events are generally perceived congruent across audition and haptics, an auditory tone or haptic stimulus will seem to last longer than the same duration stimulus presented visually. Thus, rather than treating all modalities homogeneously, designers may have to deliberately manipulate the duration of a percept based on modality.

Visual dominance effects can be particularly challenging. If percepts of varying modalities are of the same relative intensity, then information gathered via vision tends to have a greater influence on perception, as compared to other modalities (Stein & Meredith, 1993). Although there are exceptions, such as during signal detection where an auditory percept tends to dominate visual cues (e.g., warning signal) or during size estimation where haptic approximation of angles has been found superior to visual estimates (Proffitt & Kaiser, 1995), typically vision dominates audition and haptics, as well as likely other modalities. Two interesting demonstrations of visual dominance involve loudness judgments associated with speech understanding and stiffness judgments (Storms, 2002). Rosenblum and Fowler (1991) found that visual cues of vocal effort (i.e., facial articulation) dominated the perception of speech loudness, as compared to veridical acoustic intensity cues. Thus, if one acts as if they are shouting but actually whispers, an observer will go with the visual articulation they saw rather than the actual voice inflection they heard. The McGurk Effect also demonstrates the bimodal nature of speech understanding (McGurk & MacDonald, 1976). This effect is elicited by presenting discrepant visual and auditory speech information (e.g., say “ba” but facially articulate “va”), with the percept being dominated by the facial articulation. Calvert (2001) found neurological evidence for the McGurk Effect by demonstrating that the lateral temporal auditory cortex was activated even during silent lip-reading. More specifically, their results revealed (a) support for previous research that the superior temporal gyrri (i.e., BA 41, 42, and 22) and the left hemisphere, in general, are involved with auditory speech perception; (b) valid-speech, mouth gestures activate auditory cortices as well as the extrastriate cortex and infero-posterior temporal lobe (which are involved in the detection of visual movement); (c) pseudospeech visual cues also activate auditory cortical regions; and (d) nonspeech, mouth gestures do not activate auditory cortices (Calvert et al., 1997). Similarly, but in a different cross-modal context, Srinivasan, Beauregard, and Brock (1996) demonstrated that visual cues representing the stiffness of a spring dominated veridical haptic cues. Consequently, if an observer saw a taut spring but felt a lax spring, they would believe the spring was firm.

Alternatively, research has shown auditory dominance over visual percepts in some context, particularly temporal tasks such as apparent motion. For example,
Shams, Kamitani, and Shimojo (2000) found that visual target recognition can be altered by auditory cues, even for unambiguous visual cues. More specifically, their results showed that when a single flash was paired with multiple beeps, the single flash was perceived as multiple flashes. Shams, Kamitani, Thompson, and Shimojo (2001) used visual evoked potentials to determine the neurological activity associated with this phenomenon. ERP data showed that both physical flashes and illusory flashes (induced by auditory cues) activated areas of the visual pathway. This phenomenon tended to occur more often in the periphery than in the fovea. Specifically, participants saw two flashes 81% of the time when the flash occurred in the periphery as compared to 21% of the time when the flash occurred in the fovea (Shams et al., 2001).

The dominance of one sensory modality over another is thought to be a result of differences in the suitability for perceptually coding certain features of a stimulus (Kohlrausch & Van de Par, 1999). Suitability of a modality is established by the accuracy of coding, which is based on the “modality appropriateness hypothesis” (Welch, DuttonHurt, & Warren, 1986). Based on this hypothesis, the visual modality is suggested to dominate audition in spatial acuity because its acuity is about 1 min of arc as opposed to 1 deg for hearing and even lesser acuity for other senses (Mills, 1958). Haptics also dominates audition during spatial tasks (European Telecommunications Standards Institute, 2002). During temporal tasks, it is suggested that the auditory modality dominates vision because of its acute temporal resolution; furthermore, haptics has also been proposed to dominate vision (European Telecommunications Standards Institute, 2002). In yet other circumstances, such as in the perception of surface roughness, the percept is clearly dominated by haptics (Ernst & Banks, 2002). Given the existence of perceptual illusion effects, efforts in multimodal design must determine if there are sensory thresholds that can overcome such perceptual phenomena when they are undesirable or detrimental.

Based on existing behavioral and neurophysiological cross-modal studies, many inferences can be made regarding implications for multimodal interaction design. Table 5 provides preliminary design considerations regarding cross-modal effects.

5. MULTIMODAL INTERACTION DESIGN

Given the limited capacity of WM, benefits of leveraging more than one WM subsystem and cross-modal activations, while maintaining prudent awareness of cross-modal perceptual illusions, how can one best determine how to couple multimodal stimuli? If, as postulated by Baddeley (1990, 2000), Barnard (1999), Schneider (1999), Sulzen (2001), and others, separable WM resources exist for each modality, then Wickens’s (1984) Multiple Resource Theory (MRT) could potentially be extended and leveraged with modality-based WM theories to develop a conceptual model of multimodal design. The multiple resource approach addresses the idea that given separable WM components, alternate resources can be strategically utilized at different points in user interaction to streamline a user’s cognitive load. In such case, the idea of central processing capacity is replaced with limits for spe-
Table 5: Preliminary Cross-Modal Integration Rules

- Temporal and spatial coincidence: When seeking a large amount of perceptual integration of multimodal stimuli and neural coactivation (e.g., for enhanced reaction time [RT], augmented perception, enhanced memory), ensure different modalities are close temporally and spatially (Driver & Spence, 2000).
- Working memory (WM) capacity enhancement: When seeking WM capacity enhancements, direct sensory stimuli to a multitude of sensory modalities while avoiding extensive cross-encoding among visual and auditory percepts into linguistic terms (Baddeley, 1990, 2000; Barnard, 1999; Schneider, 1999; Sulzen, 2001).
- Intersensory facilitation effect (IFE): When seeking enhanced target acquisition, introduce a preceding sensory accessory in an alternative modality than the primary percept (Heuermann & Colonius, 2001).
- Congruency effectiveness: When seeking enhanced RT via redundant-signals or neural coactivation, employ congruent combinations of cross-modal percepts as opposed to incongruent combinations (Dyson & Quinlan, 2002).
- Spatial orientation augmentation: When seeking enhanced spatial orientation, ensure multimodal spatial senses yield congruent information (Freedman & Rekosh, 1968), such as by coupling head movements with visual scene updating or auditory localization.
- Magnitude or inverse effectiveness: If a system has less than optimal displays (e.g., low resolution CRT), then couple with additional modality displays in order to garner the multiplicative effect of crossmodal integration (Meredith & Stein, 1986).

cific types of percepts, with total WM capacity depending on how dissimilar streams of information are in terms of modality. An expanded MRT, which considers additional WM subsystems and their associated resources, can be used to direct design of multimodal interaction. Such a model would address how to allocate WM resources in such a way as to allow attention to be time-shared amongst various modalities; thereby implementing modalities according to predicted available WM stores, as well as considering potential wanted or undesirable cross-modal sensory integration effects.

Figure 1 presents such a multimodal resource model. This model can be used to allocate processing resources among sensory modalities (visual, auditory, haptic, olfactory, gustatory), available WM codes (visual, spatial, verbal, nonverbal, tactile, kinesthetic), and their respective LTM modalities. These input modalities can leverage various output modalities such as manual, vocal, eye gaze, face and body gestures, and brain sensors. Within this model, the independence of multimodal resources can be leveraged in multimodal interaction design, while considering the effects of cross-modal integration. It is important to stress that WM codes are not exclusive from each other; thus any given task may employ more than one code, which may involve resultant activation of cross-modal neurons. This should result in enhanced perception and speeded RT, but may also result in unexpected perceptual illusions.

The multimodal resource model would include provisions to avoid the potential for thrashing. Thrashing occurs when there is a fixed boundary between two states; in this case between two modalities. As one state becomes saturated, there is a shift to the alternate state. Slight decay in the originally intended state causes a shift back to the original state and the cycle repeats itself. This type of thrashing characteristic could be common in the proposed model. Fortunately, thrashing character-
istic can be mitigated using a hysteresis loop. These types of loops are common in control system design. The specific implementation for the multimodal resource model needs to be empirically derived.

Beyond the multimodal resource model, the model proposed in Figure 2 couples some of the proposed multimodal presentation and multimodal user response by incorporating Wickens’s Stimulus-Central Processing-Response compatibility (S-C-R) schemes. In this model, the human information processing loop commences with a modal (e.g., visual, auditory, haptic) stimulus [S], which may in turn be perceived and processed through WM, that is, central processing [C] (verbal, spatial), and then responded [R] to by the human (e.g., vocally, manually, or via eye gaze), thereby providing an S-C-R processing loop (Wickens, 1992). Current understanding is that tasks demanding “verbal” WM, such as interpretation of system status, are thought to be best presented via audition (i.e., speech), but could alternatively be presented via text or Braille, particularly for those with visual impairments (see Information Format and Sensory Modalities in Figure 2). To optimize reaction time to such auditory information, a speech-based response is thought best (see vocal response [“R”] in Figure 2). For example, a user may use kinesthetic gestures, auditory speech, and printed text on a visual display to indicate the course of a ship while responding vocally to request the course change. “Spatial” information is thought to be best presented via visual imagery, but could alternatively be presented as sound localization or touch and motion (see Information Format and Sensory Modalities in Figure 2). The potential exists that those individuals who have low visual spatial ability may better interact and process auditory or haptic spatial information. If such substitutions can be successfully identified, it has tremendous potential to enhance interactive systems for those of varying ability. The optimal response mode to associate with such spatial information is thought to be manual (see manual response [“R”] in Figure 2). For example, a pilot may hear a warning sound, observe his or her visual display, and receive haptic cues on his left shoulder to indicate a left bank while responding manually by readjusting course via the yoke. If one generalizes this approach to the multisubsystem WM model in Figure 1 and couples it with the cross-modal design guidelines in Table 5, then multimodal interaction may be optimized by leveraging multiple WM subsystems through the presentation of information from different sensory types, expressly planning for optimal cross-modal integration, and when optimal user responses are engaged. Such an approach provides a framework through which multimodal interaction designers can determine how best to present multimodal couplings that are readily perceivable with minimal contradictory coupling and interference, as well as deal with intersensory illusions.

6. CONCLUSIONS AND FUTURE DIRECTIONS

Given multimodal design’s tremendous promise, there is a need to develop a much more rigorous understanding of how best to design information displays using multimodal information sources. In the earlier sections, a preliminary set of design
guidelines has been provided. It is important, however, to recognize that these suggestions are based on extrapolations from theories and studies that have emphasized one modality. It is unclear how valid these extrapolations may be. Further, there are some cases where the existing theoretical positions lead to contradictory guidance regarding multimodal system design. For example, MRT suggests that manual responses are most compatible with spatial tasks. However, data from neuropsychological studies suggest that right-hand motor activity interferes with spatial processing in women (Bowers & LaBarba, 1988). As such, there is a need for a more encompassing theoretical position. Consequently, there is a need for a program of research to serve as a foundation for recommendations in which we can have greater faith.

An obvious starting point for such a program is to explore the validity of existing models for multimodal design. As noted earlier, MRT has not been tested in many of the situations about which one might draw hypotheses. There is, therefore, a need for a series of studies that evaluate these hypotheses while keeping constant extraneous variables such as task demand. The challenge, here, is developing a test-bed that will allow combination of various modalities for such testing. This is a challenge that our lab is currently undertaking.

Future research should also evaluate the utility of the rapidly increasing body of research in neuropsychology. There is currently a substantial database regarding the cerebral areas that are associated with various types of task processing. Further, there seems to be general support for Kinsbourne and Hicks's "functional distance hypothesis" (LaBarba, Bowers, Kingsberg, & Freeman, 1987). This hypothesis suggests that one can calculate the functional distance between cerebral processing areas by observing the interference imposed by their concurrent processing. Those estimates of functional distance can then be used to build optimization models to allow maximum cortical processing by the operator. The advantage of this approach is that it more easily allows consideration of individual differences such as gender and ability level (cf. Bowers & LaBarba, 1988). Individual differences in sensory perception have been found, which consequently may create challenges for multimodal interaction design. Some individuals find it easier to identify an object using auditory cues, whereas others tend to be more visually dominant (Giard & Peronnet, 1999). Specifically, visually dominant individuals tend to show little early integration effects in the visual cortex, but show clear effects in the auditory cortex. Conversely, auditory dominant individuals tend to show little early integration effects in the auditory cortex, but clear effects in the visual cortex. Hence, multisensory integration occurs predominantly in the cortex of the nondominant sensory modality. Therefore, it is important to address such issues when designing a multimodal system, as this suggests that one design solution may not be suitable for all users.

The theories, from which the multimodal conceptual models proposed in Figures 1 and 2 have been derived, have mostly been tested with very simple tasks. Although early data suggest that hypotheses derived about more cognitively complex tasks have been supported (e.g., Bowers & LaBarba, 1991), there is clearly a need for a more thorough investigation. In fact, one could conceptualize all of the combinations in Figure 2 of this manuscript as a set of hypotheses about which the
different theoretical positions might make similar or different hypotheses, all of which need to be validated. Finally, there is a need for more applied research to discover how best to employ optimal design configurations once they are identified. For example, it might be that the resulting model would suggest that a certain type of information should be switched from one modality to another. How do we accomplish that switch? Should the switch be suggested, but approved by the user? Or should it happen seamlessly? Is training required to enable this switching? All of these questions must be addressed before the promise of multimodal display design is realized.

As is evident from the earlier discussion, the authors believe that multimodal displays offer tremendous potential to allow users to process more information. If true, this will address one of the most critical limiting factors confronting human performance professionals today. However, it is also clear that this is an incredibly complex problem that lacks a complete scientific foundation to guide design. As such, we hope that this article serves as a catalyst to begin this program of research so that the promise can be realized.

REFERENCES


