14 AUDITORY-VISUAL INTERACTIONS

Thomas G. Ghirardelli
Angélique A. Scharine

Multisensory Perception

Most displays or other devices designed to communicate information focus on a single sensory modality. This is perhaps a consequence of the fact that most research examining human perceptual capabilities has focused on a single sensory system at a time. However, most events in the natural environment generate physical information affecting multiple sensory modalities. This information is typically co-located in both space and time, and our perceptual systems have evolved to create a single coherent representation of our environment. Recent research has increasingly acknowledged the importance of these regularities, and the topic of multisensory integration became critical for understanding global awareness of the environment. This chapter presents an overview of the most recent research into the integration of auditory and visual information and presents considerations for the design of multisensory (i.e., auditory-visual) displays. The initial section of the chapter presents comparison of the basic features of the auditory and visual modalities and examines the effects of interacting visual and auditory stimuli. Included in this section is a discussion of the relative strengths and weaknesses of each modality and the situations in which one modality dominates. This section is followed by a discussion of perceptual effects when information from both modalities is complementary or conflicting and the role of multisensory attention on the perception of auditory-visual information. The final section of the chapter presents a discussion of the applications of auditory-visual integration principles to audio-visual display (e.g., helmet-mounted displays [HMDs]) and signal designs. Research on the interaction of a third modality, tactile or haptic displays, is just beginning and will not be discussed extensively here, although many of the same considerations can be expected to apply (see Chapter 18, *Exploring the Tactile Modality for HMDs*).

Dominant Characteristics of Audition and Vision

Many of the interactive effects that will be discussed are driven by the unique characteristics of each modality. These affect which modality is preferable for a specific type of information and when a sensory mode will dominate. First, auditory information is characterized by temporal changes in the sound pressure wave arriving at the ear, and a complete sphere of receptivity around the head, albeit with differential sensitivity. In many cases, sounds are transient, meaning that they have terminated prior to the observer's response. The observer must either remember the features of the sound or supplement the sound information with visual information in order to make the response. On the other hand, visual information is characterized largely by changes in the intensity and/or spatial frequency of light waves across a limited spatial region the field-of-view, or field-of-regard.[1] Although some objects can move or change with time, the majority of the scene elements will remain constant over time.

Perhaps, because auditory information is primarily temporal, the temporal resolution of the auditory system is more precise. We can discriminate between single and pairs of clicks when the gap is only a few tens of microseconds (Krumbholz et al., 2003; Leshowitz, 1971). Perception of temporal changes in visual modality is much poorer, and the fastest visible flicker rate in normal conditions is about 40-50 Hertz (Hz) (Bruce, Green and Georgeson, 1996).

[1] Field-of-regard includes head movements, but not torso movements; field-of-view refers to the eye only as limited by whatever stops are in the field, e.g. glasses, NVG, etc.

In contrast, the maximum spatial resolution (contrast sensitivity) of the human eye is approximately 1/30°, a much finer resolution than that of the ear, which is approximately 1°. Furthermore, the relative temporal stability of visual information means that the observer has time to visually locate a visual object in his or her environment before resolving its details. An auditory object must be localized while it is still sounding, or remembered after the auditory event. Consequently, the visual modality tends to dominate spatial perception. This has consequences when visual and auditory information conflict; and will be discussed in the section on the capture effect.

Conversely, as noted previously, humans are sensitive to sounds arriving from anywhere within the environment; whereas, the visual field is limited to the frontal hemisphere, and "good" resolution is limited to the foveal region. Therefore, while the spatial resolution of the auditory modality is cruder, it can serve as a cue to events occurring outside the visual field-of-view.

Information presented by a display system must be remembered for at least as long as it takes the user to respond to it. This "short-term memory" (Klatzky, 1975) or "working memory" (Baddeley, 1982) refers to the limited storage capacity where we first process the stimuli originating from the environment. Its capacity is very limited and varies with modality. One common technique used to test the capacity of short-term memory is to present a list of words and then test for recall. This usually results in a pattern of results called the *serial position effect*, where the items at the beginning and end of the list are more likely to be recalled than those in between. When the mode of presentation is varied so that one can compare the effect of visual or auditory presentation, there is no difference in recall of items at the beginning of the list, but there is a slight improvement in memory for auditory items at the end of the list. However, since sound is transient and vision can be static, an auditory message is best accompanied by a visual message that can remain on the display until dismissed.

The modality effect appears to be eliminated for long term memory. Visual and auditory events are equally likely to be recalled. There does seem to be an effect of level of processing; so redundancy is advantageous. As the number of modes that information is presented in increases, the amount of processing of that information and the probability that it will be attended to also increases.

The perceived intensity of sound is referred to as *loudness* and the perceived intensity of light is referred to as *brightness* (Stevens and Marks, 1965), and each of these depends on the characteristics of the specific stimulus (sound or light) and the context in which the stimulus appears. Since intensity can be used to convey the importance or urgency of a signal, it is important to consider how the two modalities compare perceptually. When Stevens compared perceived intensity of a 75-4800 Hz band of noise and of a white light, he found close functional similarity between both sensory functions. The levels for which the two stimuli were perceived as equal depended somewhat on the test constraints (experimenter-paced or self-paced) and are shown in Figures 14-1 and 14-2.

Interaction of Audition and Vision

An interesting question is: What will happen to the perception of a visual stimulus when presented simultaneously with an auditory stimulus? Does the neural stimulation combine, improving detection? Or, does sensory input from one modality inhibit that of the other modality? The answers depend on several factors. For example, both Kravkov (1934) and Hartmann (1933) found facilitative effects of auditory tonal stimulation on visual thresholds and visual acuity. Others have found similar effects for broadband signals (Watkins, 1964; Watkins and Feehrer, 1964). Maruyama (1957, 1959) found that this effect is dependent on the frequency and intensity of the auditory stimulus. Kravkov and subsequent researchers have found that sensitivity to green light increases as sound intensity increased, but that this effect is reversed for orange-red light (Allen and Schwartz, 1940; Kravkov, 1936, 1939; Letourneau, 1972; Letourneau and Zeidel, 1971). Other studies have found inhibitory effects (Davis, 1966; Maloney and Welch, 1972) and that the effect is dependent on the temporal relationship between the stimuli in each modality (Broussard, Walker, and Roberts, 1952; Coleman and Krauskopf, 1956; Ince, 1968) or with no effect whatsoever (Symons, 1963).

Even if a stimulus in one modality is known to be irrelevant to the observer's response, any signal in the irrelevant modality may serve to enhance processing in the relevant modality. For example, Stein et al. (1996) found that observers' judgments of the intensity of a light-emitting diode (LED) were increased by the co-occurrence of an irrelevant noise burst, regardless of whether the noise originated from the same location as the LED or not.

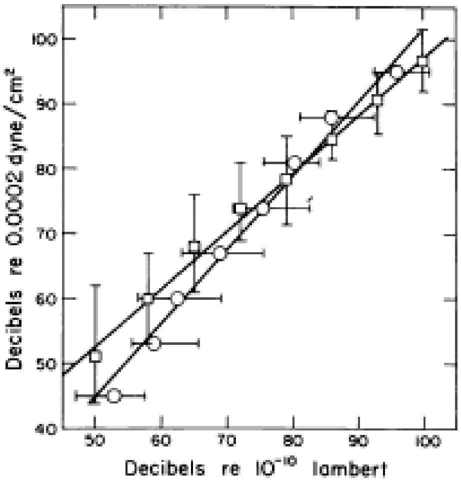

Figure 14-1. Equal-sensation functions for loudness and brightness, showing the levels of luminance and sound pressure that appeared equal in subjective intensity (Expt. 1). Squares: sound adjusted to match light; (Expt. 2). circles: light adjusted to match sound. The vertical and horizontal line segments show the interquartile ranges of the adjustments. Duplicate sentence deleted These ranges become much smaller when the intercept variability is removed.[2]

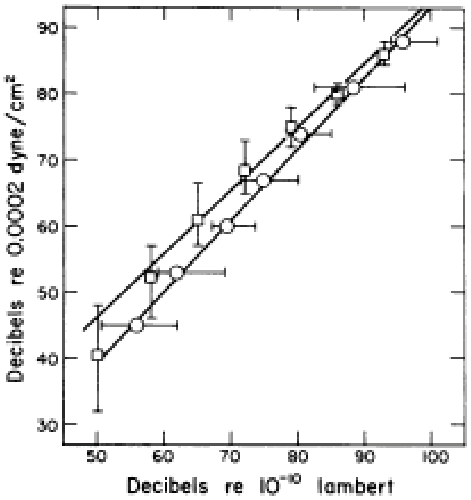

Figure 14-2. Equal-sensation functions for loudness and brightness, showing the levels of luminance and sound pressure that appeared equal in subjective intensity. Squares: sound adjusted to match light; circles: light adjusted to match sound.

[2] Note: Figures taken from "Cross-modality matching of brightness and loudness" by J. C. Stevens, and L. E. Marks 1965, *Proceedings of the National Academy of Sciences of the United States of America*, 54, 407-411. (Reprinted with permission.)

Because the findings are so inconsistent, it is tempting to dismiss them altogether. Often the reported findings were obtained under greatly restricted experimental conditions. For example, dark adapted observers seated in a dark room with no extraneous distractions showed small improvements in their ability to detect simple Gabor patches,[3] colors, pure tones and narrowband noise bursts. When one considers the contrast provided by the normally rich environment, these minute differences in threshold may not be significant. However, there do seem to be two consistent effects observed in auditory-visual studies. The first is that the combined neural activation of the two sensory modalities increases the probability that at least one sensory event will be detected. The second is that there does seem to be a limit to the amount of sensory input that can be attended to at one time. Therefore, if a light is flashed at the same time that a tone is presented, the observer may or may not detect the tone. If they are offset in time, so that the flash precedes the tone, this is less likely to happen.

Colavita (1974) seems to be the first one who described the above inhibitory effect and the effect when observers failed to respond to an auditory stimulus occurring simultaneously with a visual stimulus, known as the Colavita effect. The Colavita effect is more likely to occur if the auditory and visual events have a lower probability of co-occurrence, and if the events are spatially co-located (Koppen and Spence, 2007; Sinnett, Spence, and Soto-Faraco, 2007). In a natural environment, when an auditory and a visual stimulus occur simultaneously and in the same location, there is a good probability that they were caused by the same event and less obvious characteristic of the event is more easily missed.

The Colavita effect is reduced if the sound occurs prior to the visual event (Koppen and Spence, 2007). In a natural environment, sound often serves as a cue to visual events occurring outside our visual focus. Although a sound may sometimes distract an observer from the task of detecting a visual target (Turatto, Benso, Galfano, and Umilta, 2002), in general it signals a possible visual event – drawing the observer's attention to the visual target.

Auditory-Visual Synergy and Redundancy

Given that both the auditory and visual information from a single event will inform the observer about that event, it is not surprising that research has focused on the effects of information redundancy. The primary finding from such studies is that observers are faster when responding to redundant bimodal stimuli (e.g., a light and a sound) than they are to either of the component unimodal stimuli alone. Such signals are redundant because observers are instructed to make the same response to the presence of either the light or the tone. This *redundant signals effect* (RSE) (Miller, 1982, 1986) is greater for spatially congruent than for spatially incongruent stimuli (Gondan et al., 2005) and might be the result of the separate processing of the two different signals, with the response triggered by whichever processing finishes first. Based solely on the theory of probability, the resulting *race model* predicts the improvement because the two signals would produce a faster response than either signal alone.

If the response time is faster than that predicted by the race model, then the evidence supports the *coactivation model* which proposes that the two separate signals are integrated and processed together. (For a detailed discussion of the race and coactivation models, see Miller [1982]). Recent behavioral and electrophysiological studies have provided strong evidence for the coactivation by redundant bimodal stimuli of separate brain areas responsible for processing unimodal sensory information, and thus support the coactivation model (Giard and Peronnet, 1999; Molholm et al., 2002). For example, Giard and Peronnet devised two objects. Each object could be defined by a visual feature alone, an auditory feature alone, or by a combination. Object A consisted of a circle that morphed into a horizontal ellipse and/or a 540 Hz tone. Object B consisted of a circle that morphed into a vertical ellipse and/or a 560 Hz tone. Each of the six possible stimuli was presented equally often and observers made a speeded discrimination. Observers identified the objects more rapidly and more accurately when both features were presented than when presented with either visual or auditory features alone. Neurophysiologically, they found that event-related potential (ERPs) to multimodal objects were temporally, spatially, and functionally

[3] A *Gabor patch* is a luminance profile where the intensity at the center is the maximum grayscale value and the intensity at the edge of the diameter is one grayscale step above the background.

distinct from those to unimodal objects and these differences appeared very early in the processing of the objects (e.g., within 200 ms poststimulus).

Auditory-visual search

In addition to altering perceptual judgments and facilitating processing of redundant targets, information from a different modality may facilitate processing in other ways. Bolia, D'Angelo, and McKinley (1999) found that auditory cues speeded responses to targets in a visual search task. The targets were configurations of 2 or 4 LEDs amongst distractors consisting of 1 or 3 LEDs. The total number of targets and distractors (i.e., the set size) was 1, 5, 10, 25, or 50 items. Auditory cues were pink noise that was presented either from a loudspeaker at the same location as the target or at the same virtual location via spatialized headphone presentation. Auditory cues that were co-located with targets resulted in search times that did not increase significantly with increasing set size. Virtual auditory cues produced response-times (RTs) that increased with set size but only by 40 ms per item compared to increases in search time of more than 240 ms per item for trials in which no auditory cues were presented. This study showed the benefit of adding redundancy via auditory information by speeding localization of a visual target, even in the presence of non-target distractors. Although Bolia, et al., do not explicitly identify attention as the source of this facilitation, this is consistent with findings from studies that specifically address the role of multisensory attention as we shall see later in this chapter.

Auditory-visual synchrony

When designing or purchasing a HMD device that has both auditory and visual displays, it is important to consider the degree to which the auditory output is synchronized with the visual output. Obviously, perfect synchrony, though optimal, may not be possible due to technical constraints. The human brain is accustomed to a certain amount of asynchrony between auditory and visual information due to the fact that sound travels more slowly than light and as such, our tolerance for asynchrony is asymmetric (Stone et al. 2001). A number of studies have been conducted to determine the limits of our ability to detect auditory-visual asynchrony. To some extent, these limits depend on the type of auditory-visual information being transmitted. Vatakis and Spence (2006a) found that the stimulus onset asynchrony (SOA) required for detecting asynchrony between video and audio clips was lowest for simple non-speech sounds, higher for speech, and highest for piano and guitar music. They suggest that tolerance increases as the source familiarity decreases and the complexity increases (Vatakis and Spence, 2006b). The visual portion of speech can lead audition by more than 240 ms before asynchrony becomes noticeable (Dixon and Spitz, 1980; Grant and Greenberg, 2001; Grant, van Wassenhove, and Poeppel, 2003; Munhall et al., 1996). This limit is supported by neurophysiological research that shows that the temporal interval during which multisensory enhancement can occur in animals is about 200 ms (King and Palmer, 1985; Meredith, 2002; Meredith, Nemitz, and Stein, 1987; Stein and Meredith, 1993). The window is smaller for nonspeech items; auditory lags of 112 to 188 ms can be detected (Dixon and Spitz, 1980; Lewkowicz, 1996). These same studies show that there is less tolerance for lagging vision; the limits found ranged between 40 to 80 ms, with the exception of Dixon and Spitz, who found tolerance for lags up to 131 ms for speech passages. Therefore, a conservative guideline might be that visual output should lead sound by no more than 100 ms and lag by no more than 40 ms. Any asynchrony larger than this may be noticeable, depending on the source.

Capture effect

Another important consideration when designing or choosing an audio system for an HMD is to realize that information received in one sensory channel can be affected by information received through another channel. This phenomenon is called the *capture effect*. One of the most familiar examples of this phenomenon is the *ventriloquism effect* (VE) (Howard and Templeton, 1966). The VE refers to our tendency to perceive sounds as coming from the same location as a visual event, as would be the case of perceiving the sound as coming from the ventriloquist's dummy. In this case, the location of a visual object, the dummy, captures the perceived location of the sound source, the ventriloquist. Thomas (1941) describes the tendency for listener judgments of sound source location to be biased in the direction of a flickering light, especially if the light is in sync with the sound. The perceived location can either be fused with the visual source, or shifted towards the source.

The effect is strong and compelling for smaller angles of 20° to 30°. Thurlow and Jack (1973) report that it is greatly decreased at 60° (but still occurred at least some of the time for 6 out of 10 participants). Although Thurlow and Rosenthal (1976) observed some capture at 170°, it is probable that this is due to the human tendency to confuse the auditory location of sounds near 0° and 180° (see Chapter 12, *Visual Perceptual Conflicts and Illusions*).

In the case of the ventriloquist, the percept of the sound source location is *fused* with the apparent visual source of the sound (Figure 14-3). Cognitive factors affect the strength of the VE by increasing the likelihood that the visual and auditory sources will be fused (Radeau and Bertelson, 1977). For example, researchers have varied the apparent probability that the visual object is the source of the sound, using video monitors, puppets and stationary objects as the visual targets (Thurlow and Jack, 1973; Warren, Welch, and McCarthy, 1981) (Figure 14-4). As might be expected, the sound is more likely to be fused with the visual object if the visual object appears to be a probable source of the sound.

At times a visual object will capture the location of the auditory object and bias it towards the visual object even without them actually being perceived as a fused object. For example, Bertelson and Radeau (1981) reported that the attraction of auditory localization towards visual objects may occur even when fusion is not present, that is, the stimuli are not correlated. Thus the visual capture may depend strongly on the synchrony of auditory and visual stimulations and not necessarily on the realism of the auditory-visual pair (Radeau and Bertelson, 1977). The extent of the visual capture depends on the distance between the locations of the visual and auditory stimuli and is the strongest around the midline (Hairston, Wallace, Vaughan, Stein, Norris, and Schirillo, 2003).

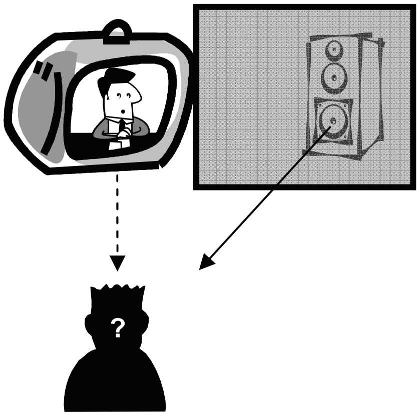

Figure 14-3. Schematic of typical demonstration of the ventriloquism effect. Listener is presented with visual stimulus along with an auditory stimulus for which the source is unseen. The location of the sound source is perceived to be collocated with the visual stimulus.

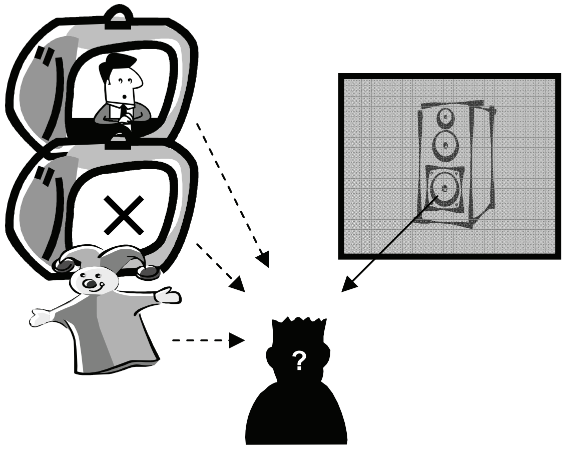

Figure 14-4. Schematic of experiment testing the effect of realism on the strength of the ventriloquism effect. In this example, the image of the human speaker is more likely to be fused with the perceived location of the sound than the video monitor with an X taped on it or the puppet. However, all three visual objects can bias the perceived location of the sound towards their location.

Vision can affect recognition of speech as well. Lip reading, called also speech reading, is used unconsciously to clarify ambiguous speech information. The posterior lateral surface of the superior temporal gyrus (located in the auditory cortex) has been found to be involved in the processing of audiovisual speech (Reale, et al., 2007). The fact that the auditory cortex has a region that processes visual information highlights the interdependence of the senses. The signal to noise ratio required for intelligibility is less for speech accompanied with a visual display of the speaker (Binnie, Montgomery and Jackson, 1974). However, in some cases, adding lip reading to auditory perception lip reading may also result in the change of the perceived sound. McGurk and MacDonald (1976) describe a phenomenon where phonemic categorization is biased depending on the visual display accompanying the auditory track. In their experiment listeners were asked to report the heard syllable (e.g., /aba/, /aga/, or /ada/). When the visual display and the auditory display were inconsistent, such as when a visual /aga/ was combined with a heard /aba/ it was often reported heard as /ada/.

If vision is biased by an auditory object, the condition is called *auditory capture*. Auditory capture occurs less frequently than visual capture. The instances where it does occur suggest that it only occurs when participants are given a reason to distrust the visual information. For example, 17 percent of the participants in Warren et al's (1981) experiment perceived the visual stimulus as shifted towards the sound source in the highly compelling condition. However, this effect is probably due to the instructions that suggested that the visual image was unreliable; they were given goggles and told that the image presented through the goggles might be "distorted." Radeau and Bertelson (1976) found auditory capture when the visual stimulus was a single light occurring in an otherwise dark environment. However, the effect disappeared as soon as the visual environment was enriched by a textured background.

Generally, vision dominates audition with respect to localization. Auditory capture of source location is most likely to occur only when visual information is ambiguous. Since vision is a more reliable source of location information, it usually results in a more compelling percept of location. Therefore, if it is necessary to convey spatial location information in a HMD display, auditory cues are best used to signal the general region of interest with a visual cue giving the precise location. However, when auditory- and visual-temporal information conflict, audition may capture vision. An elegant demonstration of this phenomenon can be found in an experiment conducted by Shams, Kamitani, and Shimojo (2000). They found that if participants were presented with stimuli consisting of a single flash and multiple auditory beeps (1-4), their perception of the number of flashes was captured by the auditory stimuli, and they consistently saw multiple flashes.

The term "capture" can be applied also to the perception of the direction of motion. In most cases, just as for stationary objects, visual motion will capture auditory objects and the sound will be heard as moving with the visual object even when the sound is stationary or moving opposite the visual motion (Kitajima and Yamashita, 1999; Mateeff, Hohnsbein and Noack, 1985). However, there is some evidence that auditory capture of motion can occur when the visual motion is ambiguous (Addie, 2003; Alais and Burr, 2004; Shimojo, Miyauchi and Hikosaka, 1997).

Outside the visual focal range, visual information is relied upon less. Audition has the advantage of being able to alert the listener to events occurring anywhere in the 360° range horizontally, as well as below and above the listener. Therefore, when visual information occurs outside one's focal range in the periphery, auditory information is relied upon more than visual motion information (Strybel and Vatakis, 2004). Further, less synchrony is needed for fusion to occur when the objects are outside of one's focal region (Noesselt et al., 2005). This may increase the risk that unrelated visual and auditory events will be perceived as a fused object. More likely, it can increase the overall uncertainty about the information presented because if the auditory information has terminated, it may be difficult to locate its source in space. This can be alleviated by presenting visual and auditory information jointly. This redundancy will allow the user to be alerted to the event occurring outside his visual range, and allow him to locate it and see it after its onset.

Whether capture is a factor for HMDs depends on the display and the situation. A 3-D or stereo audio display will provide fairly accurate location information for auditory information presented through the display and the visual and auditory information should match. If the system is monaural or bi-aural (same signal presented to both channels), the user will probably attribute the auditory information to one of the visual events in the display, but no spatial information is being given. However, capture can easily occur in events outside the display. A sound event triggered by an unknown source can be attributed to any plausible visual object in the vicinity. This can be the source of a tragic error. Capture can also be used to protect oneself. For example, if one has hidden a howitzer, noises made by it and the Soldiers operating it can be redirected to a decoy target placed in full view a few feet away.

Multisensory Attention

In addition to assessing processing of multisensory information, designers of HMDs also must be concerned with information that may be lost or missed, or that may cause other information to be lost or missed, particularly given the concerns about information overload. In order to address these concerns, we need to examine the role of attention within and across sensory modalities. With information available in multiple sensory modalities occasionally providing inconsistent information, it is sometimes necessary to select the information that is to be given additional processing, to the exclusion of the information not selected. While at other times, it is necessary to process both streams of information simultaneously. These two situations are commonly referred to as requiring *selective attention*, and *divided attention*, respectively.

We typically encounter more information from the environment than we can process at any one time (Johnston and Dark, 1986). This is usually true in any one sensory modality and is certainly true under normal circumstances where all manner of multimodal stimuli are present. As noted at the beginning of this chapter, most perception research has focused on a single sensory modality, and the same is true for research investigating attention. However, in the same way that perception in the natural world is multimodal in nature, attention is multimodal as well. Recognizing the multimodality of attention, recent work (within the last 10 years) has begun to investigate attention within and between individual sensory modalities.

Spence and Driver (e.g., 1994; 1996; 1997) have demonstrated extensive spatial links between touch, audition, and vision. Most of this work involves a variation of the spatial cuing task first used by Posner and colleagues (e.g., Posner, 1980). In the spatial cuing task, participants respond to a target. Prior to the appearance of the target, pre-cues appear that either correctly indicate the location of the subsequent target (a valid cue), indicate an incorrect location (an invalid cue), or provide no location information (a neutral cue). Posner and colleagues used visual targets and visual cues, but Spence and Driver have demonstrated the same cuing effects with all possible combinations of auditory, visual, and tactile targets and cues. These findings suggest that there exists a supramodal spatial attention system; such that spatial attention can be directed to an area of extrapersonal space (e.g., a portion of a visual display) by non-visual cues, and these cues will still facilitate processing of the visual information presented there. For example, Spence and Driver (1996) showed that observers more accurately localized an auditory or visual target as being above or below the midline of a display when a cue (either auditory or visual) directed them to attend to the side of the display on which the targets appeared. (For a more complete review, see Driver and Spence [1998] and Spence and McDonald [2004].)

Driver (1996) used a unique task to demonstrate an advantage for speech shadowing by the introduction of the ventriloquism effect (VE). The task was to shadow (i.e., repeat) target words presented from a loudspeaker. Distractor words were presented along with the target words from the same loudspeaker, and the task was to report only the target words while ignoring the distractor words. Above the loudspeaker was a television monitor, and an identical secondary pair of monitor and loudspeaker was positioned next to the first one. A video that showed a full frontal face view of a person speaking the target words was presented on one of the monitors. The video could appear in the monitor above the speaker that presented the words (same-side condition), or in the monitor opposite (different-side condition). Shadowing performance was significantly better in the different side condition than in the same side condition, suggesting that the presence of the visual information aided in the spatial separation of (and thus the selection of) the relevant auditory signal, the target words, from the irrelevant distractor words. The implication of this finding is that the integration of auditory and visual information can be used to functionally increase the signal-to-noise-ratio (SNR).

Consistent with Driver (1996), Santangelo and Spence (2007) found that auditory-visual cues captured attention under conditions of high and no perceptual load, however equivalent unimodal (e.g., auditory or visual) cues only captured attention in the no-load condition. They used the same spatial cuing task as Spence and Driver (1996) described above but in this study the cues were not informative as to which side of the display the subsequent target would appear. As a result, any effect of the cues on RT performance indexes the involuntary capture of attention. In addition they presented the cues under two conditions. In the high perceptual load condition, observers also had to monitor a rapidly presented stream of letters presented at fixation for occasionally presented target digits and in the no-load condition there was no centrally presented stream. The fact that such capture was not found for unimodal cues in the high load condition suggests that bimodal auditory-visual cues may be important for disengaging attention from a concurrent perceptually demanding stimulus. This may have important implications for warning signals.

Human Factors Issues in Auditory-Visual Displays

A commonly expressed argument for the inclusion of audio in a display system is that the visual system is overloaded and that additional information can be presented to the user via the auditory system. Unfortunately, this statement fails to take into account the costs of switching attention between modalities, the fit of the information to the modality, the resources shared by modalities, and the overall limits on attentional capacity. The Colavita effect illustrates this problem. When an auditory and a visual signal occur simultaneously, often the auditory signal will not be detected. However, the fact that they co-occurred increases the probability that at least one signal will be detected. Therefore, the advantage of adding an auditory component to a display system is by providing redundancy (Selcon, Taylor and McKenna, 1995) and facilitation to information presented visually. Otherwise, the auditory information is competing with the visual information for attention and may cause loss of information transfer.

The case for auditory warning signals

Redundancy is one technique to prevent or decrease information loss. Another way is to ensure that the modality used to convey information is a good match for the type of information to be conveyed. The goal is to present information in such a way that it requires very little attention and memory to recognize and respond to. For example, the fact that auditory information is dominant temporally makes it an ideal cue for observing visual changes in the environment. Morein-Zamir, Soto-Faraco, and Kingstone (2003) presented flashes from two lights placed vertically and asked participants to report which light (top or bottom) was the lagging light. For stimulus onset asynchronies shorter than the participants' visual temporal acuity, they found that an auditory cue presented just before and after the visual flashes captured the visual percept and allowed them to answer correctly. Humans are usually much quicker to detect changes in the auditory scene than in the visual scene – making sound cues ideal for alerting the user to situations requiring a fast response. Further, the visual range is limited to the frontal region of the surrounding environment, from -90° to 90° in azimuth. One's ability to focus is limited to only a small portion of that range. The auditory system, however, is able to hear sounds from the full 360° range. However, because sound is transient and spatial resolution is better for vision, visual display can serve as a redundant system, allowing an early auditory warning to be followed up by attention to the visual display.

Auditory signals by themselves should convey meaning and ideally, their meanings should be intuitive, rather than assigned (Patterson, 1990; Perry et al., 2007). For example, we have learned to expect that approaching objects will get louder. Therefore, a signal announcing an approaching aircraft should get louder as it approaches. High frequencies can indicate physical height or urgency. Urgency or severity can also be conveyed by increasing the repetition rate of a sound. Sounds in the real world are rarely tonal, and tones used in an auditory display need not be either. Timbre, the quality given to a sound by its overtones, is a natural way to convey meaning as well as to add a dimension to a signal. For example, each signal can be created from the sound inherent to the equipment it is informing the user about, and then the urgency for all can be conveyed using repetition rate or another dimension. If three-dimensional (3-D) auditory information is available, signals can be made even more meaningful by being co-located with the object of interest, drawing attention directly to the location requiring a response. It is important to consider the other auditory signals in a display and to be cautious about the meanings assigned to a dimension. If signals vary on an unimportant dimension, it will be more difficult to attend to the relevant ones (Pollack, 1970).

Earcons, auditory icons, auditory tactical signals and auditory warnings are all names given to auditory signals commonly included in a display system to represent specific events or objects. Earcons refer to arbitrary tones or tonal sequences used to convey a message in a user-computer interface (Blattner, Sumikawa and Greenberg, 1989; Gaver, 1994). An auditory icon is the mapping of a computer event to a sound, usually one with an intuitive mapping (Lucas, 1994). These are most easily understood if they are easily detected, understood and attended to. There are a number of things that should be considered when designing auditory signals for use in tactical displays.

First, the SNR should be sufficient to allow detection. Although detectability depends on a number of factors (Handel, 1989; Yost, 1994), a few basic guidelines are presented here. Ideally, the sound should be 15 dB higher than the ambient noise at all possible listening locations. However, it should not be so loud so as to cause hearing damage (Patterson, 1990). In cases where a 15 dB SNR is not possible, one must consider other factors that cause masking, such as the frequency content of the target and the background and factors that aid in sound segregation, e.g., grouping or spatial separation.

Knowledge of the way frequency components cause masking can help in the design of signals with a higher probability of detection. For example, if the noise in the environment consists primarily of speech, masking can be avoided by choosing frequency components or profiles outside the range of speech. Further, remember that low frequencies mask higher ones due to the upward spread of masking (Egan and Hake, 1950); therefore, very high frequencies should be avoided. It is also necessary to be conscious of the range of sensitivity of hearing (Fletcher and Munson, 1933). Humans are not very sensitive to sounds below 100 Hz. In addition, noise-induced and age-related hearing loss first occurs at approximately 4000 Hz and above. Thus, these frequency ranges should be avoided for allocation of signal energy. Finally, the probability that the noise will contain precisely the same frequency as the warning signal can be reduced by using a signal comprised of multiple frequencies (a complex signal). A pure tone is not a good choice for a warning tone. Instead, using a tone or complex signal that alternates between two fundamental frequencies improves the probability of detection by decreasing the probability that the auditory signal will share the same frequency content of the environmental noise.[4] The ideal choice of frequency for an auditory signal is a complex signal with a varying fundamental frequency that is lower than most of the ambient noise in the environment but with most of its spectral energy occurring outside the frequency range of the dominant ambient noise.

One can capitalize on the randomness of the noise in the environment by choosing a signal that repeats rhythmically (Patterson, 1990). This technique has several advantages. First, the regularity of the rhythm will draw attention, if the sounds in the background have irregular tempos. Second, humans are more sensitive to changing sounds than to steady state sounds. Therefore, a long continuous signal could be missed, while a repeating one will have multiple onsets to draw attention. Finally, the use of different repetition rates can add meaning and make the sound more memorable; this will be discussed later.

There should be a balance between the urgency of the sound and the annoyance and this balance should take into account the importance of the message conveyed by the sound. Further, the sound should not be so intrusive that the user is unable to respond to its message or perform other tasks. This may require a task analysis of the types of alarms that may co-occur and the kinds of tasks that will be required to respond to those alarms. Several features can make a sound seem more urgent: intensity, speed of repetitions, frequency content and envelope (onset, decay) (Edworthy, Loxley, and Dennis, 1991).

More intense sounds will seem more urgent. However, as stated before, intensity must be limited by safe presentation levels in order to avoid hearing damage. Further, a very loud sound may make communication difficult, making it difficult to respond to the emergency that triggered the sound. However, there are a couple of strategies that utilize intensity while attempting to avoid intrusiveness. For example, an auditory signal can start out at a normal level and increase in intensity if the problem is not resolved or if the problem severity increases. For example, a particular hotel clock alarm started soft, paused, and then got louder on the next repetition. Thus, if it didn't wake an individual the first time, it was more likely to be heard later. However, the individual had the option of shutting it off as soon as soon as it was heard, before it got louder. Another tactic is to present the signal initially at a very loud level, and then to drop it to a lower level in order to allow the listener time to respond. If no response is made, the signal can return to the loud level again, cycling between levels as needed. This strategy can be combined with increasing the speed of repetitions.

High priority signals can be made to sound more urgent by increasing the energy in the higher frequency components and by adding dissonance (Patterson, 1982). Unpleasant, dissonant sounds will stand out from the auditory environment and convey urgency. Dissonance refers to the lack of harmonicity of the spectral components in a sound. If the frequencies that make up a sound occur in multiples of the fundamental frequency, they will be harmonic and pleasant to the ear. If they don't, they will be dissonant and unpleasant. Three psychoacoustical properties describe how the amplitude envelope can make sounds unpleasant, sharpness, roughness and fluctuation strength. Sharpness refers to the proportion of higher frequency content in a sound. Increasing the level of frequency components above about 2700 Hz will increase the sharpness. Modulating the amplitude of a sound creates either roughness or fluctuation strength. Roughness refers to amplitude modulation between 15 and 300 Hz. Roughness is greatest at a modulation rate of 70 Hz. Roughness can be created by modulating the whole signal, but spectral variation can have a similar effect. Fluctuation strength refers to modulation below about 20 Hz. This effect is similar to that of a siren. Finally, abrupt onsets will also make the sound seem more urgent.

[4] For example, German police sirens alternate between two tones, in contrast to U.S. fire vehicles with a sinusoidal wailing sound.

In order for an individual auditory signal to be useful, the user must be able to remember quickly and accurately what the signal means. Pollack and his colleagues (Pollack, 1952, 1953, 1956, 1973, 1976; Pollack and Ficks, 1954; Sumby, Chambliss and Pollack, 1958) investigated the use of auditory signals for information transmission. Their findings are quite relevant to the design of memorable auditory signals. Despite the fact that listeners are able to discriminate different loudness levels and frequencies quite well, they aren't able to remember them well enough to identify the specific signals (Pollack, 1952). Therefore, designing an auditory display that uses different frequencies to distinguish between types of warning is a poor design. The listener may confuse a particular frequency with a neighboring one and misidentify the frequency. This is true even if the frequencies are spaced across a large frequency range (Pollack, 1953). At most, listeners were able to identify four or five levels of frequency; but it is recommended to limit the selection to two or three frequencies. This is true of other dimensions such as, loudness levels, repetition rate, and duration as long as they are discriminable (Pollack and Ficks, 1954). Memory for frequencies can be improved slightly if the cue frequency is combined with a reference frequency especially if the cue frequency is near to the reference frequency. It is likely that the reference is forming a salient interval that is recognizable, just as one recognizes the first few notes of a tune even if they cannot accurately identify the first note.

Despite the fact that one can identify five levels of a dimension, it is probably better to limit the set to two or three levels, especially since the user will need to be performing multiple tasks concurrently. In order to increase the number of recognizable signals, multiple dimensions should be combined. Pollack and Ficks (1954) tested listener ability to identify sounds based on levels of frequency, loudness, rate, continuity (percentage of the time "on"), duration, and spatial location. They found little improvement in the number of signals identified for sets divided into more than three levels per dimension but they could learn to identify signals distinguished by a large number of dimensions having binary values. Therefore, rather than having five different alarms that are assigned to different frequencies, they can be assigned to one or two frequencies but also vary in loudness, repetition rate, duration or location.

In an environment that has more than one signal present, care should be taken to avoid the requirement that the user have to memorize an extensive list of auditory signals. Signals should be designed to be inherently informative. One way to achieve this is to locate the sound source near the object requiring a response or the information it is cueing. If the "low battery" signal comes from the telephone, it is clear what the meaning is. Whenever possible, the sound should convey its own meaning. One way to make an auditory signal meaningful is to use a speech signal. Obviously, if the sound is, "the washer fluid is low", there's no need to memorize its meaning (Simpson, 1987). However, there are three potential problems with this approach. First, if there's already a lot of speech present in the environment, the auditory signal may be easily masked by informational masking. Further, speech can be easily susceptible to noise, especially if the spectral content is similar to or higher than the environmental noise. Finally, not all users may be as familiar with the language used and therefore may have trouble understanding the alarm.

If speech is to be used, consider the noise in the environment, the voice of the speaker, the vocabulary set and attention. Intelligibility of speech depends on the perception of its high frequency components, the consonants and these components are easily masked by noise. Speech can be preprocessed with a 3 dB/octave boost or peak clipped in order to reduce masking effects. Synthetic speech allows control of parameters such as pitch, speech rate, sex and accent, allowing it to be more easily perceived over noise. However, it is more difficult to understand and may require more attention for processing (Pisoni, 1982). Polysyllabic words are generally more intelligible than monosyllabic words. Similarly, sentences are more intelligible than single words, as they give a context that allows a listener to fill in masked information. However, since deciphering a long sentence is not recommended for time critical information it is recommended that sentences are limited to 4-8 syllables (Simpson et al., 1987). Depending on the context, a tonal alert signal, or a distinctive voice can serve to draw attention to the speech signal.

A warning should be given about the problem of excessive false alarms. Usually, a warning signal presented by a display system is a mechanical and automatic way of informing the human user of a problem or event. However, this may lead to a warning being triggered erroneously (a false alarm) or not at all (a miss). To some extent, it may be preferable for the system to err on the side of caution. However, if false alarms occur often, this may lead to a tendency by the user to ignore (Hancock, Parasuraman, and Byrne, 1996; Parasuraman, Hancock, and Olofinboba, 1997) or attempt to permanently shut off the signal (Sorkin, 1989) deeming it as an annoyance. Ideally, the system should be made as accurate as possible, with as few false alarms and misses as possible. Given that any system will have a certain amount of error, the number of false alarms can be controlled by setting the response criterion of the machinery that produces the alarm to a higher value. In some cases this will not raise the "miss rate" significantly. If this is not the case, the choice of a response criterion should be dependent on the potential danger incurred if the problem is not detected. Using an alarm that is incremental, that is one that varies in response to the changing probability that a problem exists, can reduce annoyance and increase compliance with the signal (Sorkin, Kantowitz, and Kantowitz, 1988). Finally, the user should be trained to understand the tradeoff between misses and false alarms. These considerations obviously apply not only to auditory signals but also to other types of warning signals including visual, tactile, and the signals of mixed modality.

Visual warning signals

Wickens, Gordon and Liu (1998) identify four features that are analogous to considerations for auditory warnings and should be considered when designing visual signals: visibility, discriminability, meaningfulness and location. Visibility is a concern for HMDS because not only do warnings need to be detected, but the user needs to interact with the environment while wearing it. If the display device is see-through (transparent) or monocular, care must be taken so that the display doesn't not carry so much information that it distracts the user from the real world around him.

The permanence of vision allows warnings that don't require immediate action to be postponed until the user is able to respond. In order to reduce visual clutter, information should be located in a window that can be minimized until desired. An icon can be used to remind the user of a message that is awaiting attention. If possible, the message can reside in a peripheral region of the display until it is retrieved.

Although omnidirectional auditory signals carry an advantage when it comes to quickly drawing the user's attention to information not necessarily in line of sight with the new information, they are transient and temporary in nature. Verbal messages that are longer or more complex should be transmitted visually so that the user can refer back to them and ensure that the full message is understood. If the message requires immediate action, an auditory cue can be used to call attention to the message and the message can be presented in both modalities, however, the primary mode should be visual. Care should be given that the visual message doesn't interfere with other tasks currently underway.

Some operational environments are simply too noisy for reliance on auditory displays. In others, a task analysis may reveal that the user is overburdened with auditory information. For example, a commander may be required to monitor multiple radio channels simultaneously. In these cases, visual indicators of information are preferable. Short messages that occur frequently and are part of the standard "vocabulary" can be represented by icons and other symbols. Just as with auditory warnings, care should be given to make signals as discriminable and meaningful as possible. Minimize the number of signals requiring memorization and maximize the meaningfulness of each icon.

A careful task analysis can highlight which messages are likely to be most important, and as with auditory warnings, visual warnings should be designed to reflect the urgency of the message. For example, there can be three levels of alerts: warnings, cautions, and advisories. Cautions and advisories can be presented visually because action can be postponed. When conditions are too noisy for urgent warnings to be heard, visual cues such as flashing lights can be use. Other non-essential display information can be dimmed and minimized. Estimates of likelihood can be used in order to avoid excessive false alarms (Sorkin, Kantowitz, and Kantowitz, 1988)

Indicators of locations, whether it is in form of overlays on the real world or on map displays, are best presented visually. An auditory signal can signal the general location, but a visual display has the benefit of occupying a location and remaining there as long as the information remains true or until the user is able to respond to it.

Table 14-1 summarizes some basic guidelines of when warnings should be visual and when they should be auditory. In many instances, as will be discussed in the next section, both modalities can be used effectively.

Table 14-1.
When to Use the Auditory Versus Visual Form of Presentation.

Use auditory presentation if:	Use visual presentation if:
1. The message is simple.	1. The message is complex.
2. The message is short.	2. The message is long.
3. The message will not be referred to later.	3. The message will be referred to later.
4. The message deals with events in time.	4. The message deals with location in space.
5. The message calls for immediate action.	5. The message does not call for immediate action.
6. The visual system of the person is overburdened.	6. The auditory system of the person is overburdened.
7. The receiving location is too bright or dark adaptation integrity is necessary.	7. The receiving location is too noisy.
8. The person's job requires him or her to move about continually.	8. The person's job allows him or her to remain in one position.

Source: Deatherage (1972: Table 4-1).

Auditory-visual warning signals

One way to increase meaning is to use redundant features. Sound can be combined with speech (Simpson and Williams, 1980) or visual icons to increase the probability of comprehension. We do not have to guess why our car is beeping at us in the morning because the seatbelt light is also on, and often is flashing with the same pattern as the tone. A visual cue can alert a listener to an impending auditory message. Auditory cues can signal a viewer to updates on a tactical display. An auditory cue can signal the arrival of a new message that is sent via both modalities so that if the user is busy, the message can be reviewed at a later time.

When considering the design or purchase of HMDs, one should consider the ways in which the visual and auditory displays interact with each other and with the environment in which they are used. Visual and auditory information should be consilient and thus redundant if at all possible. Rather than trying to increase information conveyed by presenting some information visually and other information auditorally, cognitive load should be decreased by coherent multimodal presentations that facilitate quick reactions. However, it is important to conduct a task analysis in order to determine when job tasks are likely to interfere with each other and incoming information from the display. The *multiple resource theory* framework consists of the following four dichotomies: stages (cognitive vs. response), sensory modalities (auditory vs. visual), codes (visual vs. spatial), and channels (focal vs. ambient) (Wickens, 2002). If two tasks are to be performed simultaneously, one task will usually suffer; however, the secondary task will usually be less difficult if they share fewer resources (Wickens, Dixon and Seppelt, 2005). Helleberg and Wickens (2001) demonstrated that verbal information presented auditorally interfered most with a visual scanning task, due to the need to writes notes – interference cause by competition for response resources. Performance was not necessarily improved for the redundant condition, perhaps because the auditory instructions disrupted focal attention and participants still relied on the visual instructions. In this case, performance was best when the information was presented visually. Multiple resource theory will be discussed in greater detail in Chapter 19, *The Potential of an Interactive HMD*. When redundancy is not feasible, care should be taken to present information via the most appropriate modality as suggested by Table 14-1.

It should also be stressed that transmission of information through auditory and visual channels must be synchronized because such synchrony facilitates the realism of the display, accurate attribution of percept to object and faster reaction times. The window during which asynchrony is undetectable depends partly on the mode and partly on the information presented, but can be conservatively defined as an visual lag of no more than 40 ms and a visual lead of no more than 100 ms.

It is desirable to have 3-D or at least, stereo sound presentation if feasible. The spatial separation of different events allows the user to attend to them better and to filter out irrelevant noise. If vision and sound are co-located in space, they are intuitively understood to be a single event and detection and response is quicker. Although capture allows us to tolerate some spatial dislocation between auditory and visual information, spatial dislocation reduces display fidelity. Further, dislocated auditory signals can, through capture, be attributed to the wrong visual events. However, a visual "master" signal located in the front of the system operator may be effectively used as a cue signal before an auditory warning signal presented in a 3-D space attracts operator's attention to the specific location in space.

In summary, the inclusion of well-designed auditory displays in a multi-sensory HMD system can greatly reduce information loss and cognitive load. Careful considerations of the limitations of each modality allow the design of supplemental signals in the other modality that provide redundancy and prevent errors. By capitalizing on the temporal and spatial advantages of each modality, information can be easily understood and the correct responses quickly performed. This makes the auditory system an important consideration in the design or purchase of an HMD system.

References

Addie, J.D. (2003). *Spatial multi-sensory interaction: Auditory capture of visual apparent motion.* Unpublished master's thesis, Arizona State University.

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 15*,* 257-262.

Allen, F., and Schwartz, M. (1940). The effect of stimulation of the senses of vision, hearing, taste, and smell upon the sensibility of the organs of vision. *Journal of General Physiology,* 24, 105-121.

Baddeley, A.D. (1982). *Your Memory: A User's Guide*. New York: Macmillan.

Bertelson, P., and Radeau M. (1981). Crossmodal bias and perceptual fusion with auditory – visual discordance. *Perception and Psychophysics*, 29, 578-584.

Binnie, C.A., Montgomery, A.A., and Jackson, P.L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech and Hearing Research,* 17(4), 619-630.

Blattner, M.M., Sumikawa, D.A., and Greenberg, R.M. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction,* 4(1), 11-44.

Bolia, R.S., D'Angelo, W.R., and McKinley, R.L. (1999). Aurally aided visual search in three-dimensional space. *Human Factors,* 41(4), 664-669.

Broussard, I.G., Walker, R.Y., and Roberts, E.E. (1952). The influence of noise on the visual contrast threshold (No. 101). Fort Knox, KY: U. S. Army Medical Research Laboratory.

Bruce, V., Green, P.R., and Georgeson, M.A. (1996). *Visual Perception: Physiology, Psychology, and Ecology* (3rd Ed.). East Sussex, UK: Psychology Press.

Colavita, F.B. (1974). Human sensory dominance. *Perception and Psychophysics,* 16(2), 409-412.

Coleman, P.D., and Krauskopf, J. (1956). The influence of high intensity noise on visual thresholds (No. 222). Fort Knox, KY: U. S. Army Medical Research Laboratory.

Davis, E.T. (1966). Heteromodal effects upon visual thresholds. *Psychological Monographs,* 80, 24.

Deatherage, B.H. (1972). Auditory and other sensory forms of information presentation. In: van Cott, H.P., and Kinkade, R.G. (Eds.), *Human Engineering Guide to Equipment Design*. Washington, DC: Government Printing Office.

Dixon, N., and Spitz, L. (1980). The detection of audiovisual desynchrony, *Perception*, 9*,* 719-721.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature,* 381*,* 66-68.

Driver, J., and Spence, C. (1998). Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society, Section B,* 353*,* 1319-1331.

Edworthy, J., Loxley, S., and Dennis, I. (1991). Improved auditory warning design: Relations between warning sound parameters and perceived urgency. *Human Factors,* 33, 205-231.

Egan, J.P., and Hake, H.W. (1950). On the masking pattern of a simple auditory stimulus. *Journal of the Acoustical Society of America,* 22, 622-630.

Fletcher, H., and Munson, W.A. (1933). Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America,* 5, 82-108.

Gaver, W. (1994). Using and creating auditory icons. In: G. Kramer (Ed.), *Auditory Display: Sonification, Audification and Auditory Interfaces, SFI Studies in the Sciences of Complexity*, Vol. 18. Reading, MA: Addison Wesley.

Giard, M.N., and Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience,* 11*,* 473-490.

Gondan, M., Niederhaus, B., Rösler, F., and Röder, B. (2005). Multisensory processing in the redundant target effect: A behavioral and event-related potential study *Perception and Psychophysics,* 67*,* 713-726.

Grant, K.W., and Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. Presented at the ISCA 2001 International Conference on Auditory-Visual Speech Processing, Aalborg, Denmark.

Grant, K.W., van Wassenhove, V., and Poeppel, D. (2003). Discrimination of auditory-visual synchrony. Presented at the ISCA 2003 International Conference on Auditory-Visual Speech Processing, St. Jorioz, France.

Hairston, W.D., Wallace, M.T., Vaughan, J.W., Stein, B.E., Norris, J.L., and Schirillo, J.A. (2003). Visual localization ability influences cross-modal bias. *Journal of Cognitive Neuroscience,* 15, 20-29.

Hancock, P.A., Parasuraman, R., and Byrne, E.A. (1996). Driver-centered issues in advanced automation for motor vehicles. In: Parasuraman, R., and Mouloua, M., (Eds.), *Automation and human performance: Theory and applications*. Mahwah, NJ: Lawrence Erlbaum.

Handel, S. (1989). *Listening: An Introduction to the Perception of Auditory Events*. Cambridge, MA: MIT Press.

Hartmann, G.W. (1933). Changes in visual acuity through simultaneous stimulation of other sense organs. *Journal of Experimental Psychology,* 16, 393-407.

Helleberg, J., and Wickens, C.D. (2001). *Auditory vs. visual data link: Relative effectiveness.* Paper presented at the *Human Factors and Ergonomics Society,* 45th Annual Meeting, Minneapolis, MN.

Howard, I.P., and Templeton, W.B. (1966). *Human Spatial Orientation.* New York: Wiley.

Ince, L.P. (1968). Effects of low-intensity acoustical stimulation on visual thresholds. *Perceptual and Motor Skills,* 26, 115-121.

Johnston, W.A., and Dark, V.J. (1986). Selective attention. *Annual Review of Psychology,* 37*,* 43-75.

Kitajima N, Yamashita Y. (1999). Dynamic capture of sound motion by light stimuli moving in three-dimensional space. *Perceptual and Motor Skills,* 89, 1139-1158.

Klatzky, R.L. (1975). *Human Memory: Structures and Processes*. San Francisco: Freeman.

King, A.J., and Palmer, A.R. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus, *Experimental Brain Research,* 60*,* 492-500.

Koppen, C., and Spence, C. (2007). Seeing the light: exploring the Colavita visual dominance effect. *Experimental Brain Research,* 180(4), 737-754.

Kravkov, S.V. (1934). Changes of visual acuity in one eye under the influence of the illumination of the other or of acoustic stimuli. *Journal of Experimental Psychology,* 17, 805-812.

Kravkov, S.V. (1936). The influence of sound upon the light and colour sensitivity of the eye. *Acta Ophthalmologica,* 14, 348-360.

Kravkov, S.V. (1939). The influence of the loudness of the indirect sound stimulus on the colour sensitivity of the eye. *Acta Ophthalmologica,* 17, 324-331.

Krumbholz, K., Patterson, R., Nobbe A, and Fastl, H. (2003). *Journal of the Acoustical Society of America,* 113, 2790-2800.

Leshowitz, B. (1971). Measurement of the two-click threshold. *Journal of the Acoustical Society of America*, 49, 462-466.

Letourneau, J., and Zeidel, N.S. (1971). The effect of sound on the perception of color. *American Journal of Optometry and Archives of American Academy of Optometry,* 48, 133-137.

Letourneau, J.E. (1972). The effect of noise on vision. *The Eye, Ear, Nose and Throat Monthly,* 51, 441-444.

Lewkowicz, D.J. (1996). Perception of auditory-visual temporal synchrony in human infants, *Journal of Experimental Psychology: Human Perception and Performance,* 22*,* 1094-1106.

Lucas, P.A. (1994, November 7-9, 1994). *An evaluation of the communicative ability of auditory icons and earcons* Paper presented at the Second International Conference on Auditory Display, Santa Fe, New Mexico.

Maloney, D.M., and Welch, R.B. (1972). The effect of accessory auditory stimulation upon detection of visual signals. *Psychonomic Science,* 29, 345-347.

Maruyama, K. (1957). The effect of tone on the successive comparison of brightness. *Tohoku Psychologica Folia,* 15, 56-69.

Maruyama, K. (1959). Effect of intersensory tone stimulation on absolute light threshold. *Tohoku Psychologica Folia,* 17, 51-81.

Mateeff, S., Hohnsbein, J., and Noack T (1985). Dynamic visual capture: apparent auditory motion induced by a moving visual target. *Perception,* 14(6) 721-727.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature,* 264, 746-748.

Meredith, M.A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research,* 14, 31-40.

Meredith, M.A., Nemitz, J.W., and Stein, B.E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors, *Journal of Neuroscience,* 7*,* 3215-3229.

Miller, J.O. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology,* 14*,* 247-279.

Miller, J.O. (1986). Timecourse of coactivation in bimodal divided attention. *Perception and Psychophysics,* 40*,* 331-343.

Molholm, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E., and Foxe, J.J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research,* 14*,* 115-128.

Morein-Zamir, S., Soto-Faraco, S., and Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research,* 17*,* 154-163.

Munhall, K.G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect, *Perception and Psychophysics,* 58*,* 351-362.

Noesselt, T., Fendrich, R., Bonath, B., Tyll, S., and Heinze, H. (2005). Closer in time when farther in space: Spatial factors in audiovisual temporal integration. *Cognitive Brain Research,* 25*,* 443-458

Parasuraman, R., Hancock, P.A., and Olofinboba, O. (1997). Alarm effectiveness in driver-centered collision warning systems. *Ergonomics,* 40, 390-399.

Patterson, R.D. (1982). *Guidelines for Auditory Warning Signals on Civil Aircraft* (CAA Paper No. 82017). London: Civil Aviation Authority.

Patterson, R.D. (1990). Auditory warning sounds in the work environment. *The Philosophical Transactions of the Royal Society: Series B,* 327, 485-492.

Perry, N.C., Stevens, C.J., Wiggins, M.W., and Howell, C.E. (2007). Cough once for danger: Icons versus abstract warnings as informative alerts in civil aviation. *Human Factors,* 49, 1061-1071.

Pisoni, D.B. (1982). Perception of speech: The human listener as a cognitive interface. *Speech Technology,* 1, 10-23.

Pollack, I. (1952). The information of elementary auditory displays. *Journal of the Acoustical Society of America,* 24(6), 745-749.

Pollack, I. (1953). The information of elementary auditory displays. II. *Journal of the Acoustical Society of America,* 25(4), 765-769.

Pollack, I. (1956). Identification and discrimination of components of elementary auditory displays. *Journal of the Acoustical Society of America*, 28(5), 906-909.

Pollack, I. (1970). Depth of sequential auditory information processing. III. *Journal of the Acoustical Society of America,* 50(2 , Part 2), 549-554.

Pollack, I. (1973). Multidimensional encoding within the temporal microstructure of auditory displays. II. *Journal of the Acoustical Society of America*, 54(1), 22-28.

Pollack, I. (1976). Multidimensional encoding within the temporal micro-structure of auditory displays. III. Multistate displays. *Journal of the Acoustical Society of America,* 59(1), 148-152.

Pollack, I., and Ficks, L. (1954). Information of elementary multidi-mensional auditory displays. *Journal of the Acoustical Society of America,* 26(2), 155-158.

Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology,* 32, 3-25.

Radeau, M., and Bertelson, P. (1976). The effect of a textured visual field on modality dominance in a ventriloquism situation. *Perception and Psychophysics*, 20*,* 227-235.

Radeau, M., and Bertelson, P. (1977). Adaptation to auditory–visual discordance and ventriloquism in semirealistic situations. *Perception and Psychophysics,* 22*,* 137-146.

Reale, R.A., Calvert, G.A., Thesen, T., Jenison, R. L., Kawasaki, H., and Oya, H. (2007). Auditory-visual processing represented in the human superior temporal gyrus. *Neuroscience,* 145, 162-184.

Selcon, S.J., Taylor, R.M., and McKenna, F.P. (1995). Integrating multiple information sources: Using redundancy in the design of warnings. *Ergonomics,* 38, 2362-2370.

Shams, L., Kamitani, Y., and Shimojo, S. (2000). What you see is what you hear. *Nature,* 408*,* 788.

Shimojo, S., Miyauchi, S., and Hikosaka, O. (1997). Visual motion sensation yielded by non-visually driven attention. *Vision Research,* 37*,* 1575-1580.

Simpson, C. (1987). Speech controls and displays. In: Salvendy, G., (Ed.), *Handbook of human factors*. New York: Wiley.

Simpson, C., and Williams, D.H. (1980). Response time effects of alerting tone and semantic context for synthesized voice cockpit warnings. *Human Factors,* 22, 319-330.

Sinnett, S., Spence, C., and Soto-Faraco, S. (2007). Visual dominance and attention: The Colavita effect revisited. *Perception and Psychophysics,* 69, 673-686.

Sorkin, R.D. (1989). Why are people turning off our alarms? *Human Factors Bulletin,* 32, 3-4.

Sorkin, R.D., Kantowitz, B.H., and Kantowitz, S.C. (1988). Likelihood alarm displays. *Human Factors,* 30, 445-460.

Spence, C.J., and Driver, J. (1994). Covert spatial orienting in audition: Exogenous and endogenous mechanisms facilitate sound localization. *Journal of Experimental Psychology: Human Perception and Performance,* 20, 555-574.

Spence, C., and Driver, J. (1996). Audiovisual links in endogenous covert spatial attention. *Journal of Experimental Psychology: Human Perception and Performance,* 22, 1005-1030.

Spence, C., and Driver, J. (1997). Audiovisual links in exogenous covert spatial attention. *Perception and Psychophysics,* 59, 1-22.

Spence, C., and McDonald, J. (2004). The cross-modal consequences of the exogenous spatial orienting of attention. In: Calvert, G.A., Spence, C., and Stein, B.E. (Eds.), *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.

Stein, B.E., London, N., Wilkinson, L.K., and Price, D.D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A Psychophysical analysis. *Journal of Cognitive Neuroscience*, 8, 497-506.

Stein, B., and Meredith, M.A. (1993). *The Merging of the Senses.* Cambridge, MA: MIT Press.

Stevens, J.C., and Marks, L.E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences of the United States of America,* 54(2), 407-411.

Stone, J.V., Hunkin, N.M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M., and Porter, N.R. (2001). When is now? Perception of simultaneity, *Proceedings of the Royal Society of London* 268B*,* 31-38.

Strybel, T.Z., and Vatakis, A. (2004). Effect of crossmodal distractors on auditory and visual apparent motion in the periphery. *Abstracts of 45th Annual Meeting of the Psychonomic Society, 9,* 97.

Sumby, W.H., Chambliss, D., and Pollack, I. (1958). Information transmission with elementary auditory displays. *Journal of the Acoustical Society of America,* 30(5), 425-429.

Symons, J.R. (1963). The effect of various heteromodal stimuli on visual sensitivity. *Quarterly Journal of Experimental Psychology,* 15, 243-251.

Thomas, G.J. (1941). Experimental study of the influence of vision on sound localization. *Journal of Experimental Psychology,* 28*,* 163-177.

Thurlow, W.R., and Jack, C.E. (1973). Certain determinants of "Ventriloquism Effect." *Perceptual and Motor Skills,* 36*,* 1171-1184.

Thurlow, W.R., and Rosenthal, T.M. (1976). Further study of existence regions for the "Ventriloquism Effect." *Journal of the American Audiology Society,* 1*,* 280-286.

Turatto, M., Benso, F., Galfano, G., and Umilta, C. (2002). Nonspatial attentional shifts between audition and vision. *Journal of Experimental Psychology-Human Perception and Performance,* 28(3), 628-639.

Vatakis, A., and Spence, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research,* 1111(1), 134-142.

Vatakis, A., and Spence, C. (2006b). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience Letters,* 393(1), 40-44.

Warren, D., Welch, R., and McCarthy, T. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception and Psychophysics,* 30*,* 557-564.

Watkins, W.H. (1964). Effect of certain noises upon detection of visual signals. *Journal of Experimental Psychology,* 67, 72-75.

Watkins, W.H., and Feehrer, C.E. (1964). *Investigations of acoustic effects upon visual signal detection.* (No. TR-64-577). Bedford, MA: United States Air Force Electronic Systems Division.

Wickens, C.D. (2002). Multiple resources and performance prediction*. Theoretical Issues in Ergonomic Science,* 3, 159-177

Wickens, C.D., Dixon, S.R., and Seppelt, B. (2005). *Auditory preemption versus multiple resources: Who wins in interruption management?* Paper presented at the *Human Factors and Ergonomics Society* 49th Annual Meeting, Orlando, FL.

Wickens, C.D., Gordon, S.E., and Liu, Y. (1998). *An Introduction to Human Factors Engineering*. New York: Longman.

Yost, W.A. (1994). *Fundamentals of Hearing* (3rd Ed.). San Diego, CA: Academic Press.