MODALITY AND ENCODING STRATEGY EFFECTS ON A VERIFICATION TASK WITH ACCELERATED SPEECH, VISUAL TEXT AND TONES

Michael A. Nees & Kathryn Best
Lafayette College, Department of Psychology
Oechsle Hall, Easton, PA, USA
neesm@lafayette.edu; bestk@lafayette.edu

ABSTRACT

An experiment examined performance on a speeded comparison verification task with accelerated speech (spearcons), visual text, and auditory tones (sonifications). Participants' task was to encode the state (increasing or decreasing) of a stock depicted in the first (study) stimulus for comparison with the state depicted in the second (verification) stimulus. We also instructed participants to remember the study stimulus according to a prescribed encoding strategy—either as words (a verbal working memory processing code) or tones (a tonal, auditory imagery working memory processing code). Results generally offered evidence that the accelerated speech stimuli assumed the same verbal working memory code as the visual text stimuli. Interestingly, however, both speech and tones also exhibited lingering, stimulus-specific perceptual effects during verification despite recoding in working memory. Results are discussed in terms of auditory imagery in working memory, and the relevance of results to auditory interface design is discussed.

1. INTRODUCTION

Advances over the last several decades have allowed for a multitude of modes of information display across a wide variety of everyday technologies. Dynamic auditory, visual, and multimodal information displays are ubiquitous in computing [1], mobile devices [2], and, increasingly, in vehicles [3], homes [4], and places of work [5]. In addition to traditional visual text and auditory speech displays, systems have begun to implement both nonspeech and altered or synthetic speech signals. A number of important questions for theory and practice remain unresolved for extant and emerging visual and auditory displays. The modality by which information is presented can impact how well a person is able to accomplish tasks in a human-machine system, but an additional and often overlooked consideration is the strategy used to encode and process the information. The current study examined how the effects of both modalities and encoding strategies affected reaction times on a simple speeded comparison task—often referred to as a *verification task* [6]—using text, nonspeech auditory tones, and accelerated speech stimuli. Across these stimulus categories, we examined similarities and differences in speeded processing of these stimulus categories in an attempt to identify the relative contributions of the modality and the working memory code.

1.1. Multimodal information display

Whereas visual displays have traditionally been the default mode of information presentation, many devices now have been equipped with the capability to make sounds. In general, auditory display of information may be appropriate in a variety of circumstances, especially those in which a person cannot use a visual display due to visual impairment, other concurrent visual tasks, or the constraints of a particular set of environmental circumstances [3], [7], [8].

Speech is a common auditory display in a variety of systems [9], and research [10], [11] has suggested that speech can be used to communicate information relatively effectively. One study even suggested that users perceive spoken messages with greater positive affect than written messages [12]. Speech displays can be administered as a pre-recorded spoken human voice, but synthetic speech is perhaps more common [13]. Text-to-speech (TTS) software, for example, can translate digital text to spoken auditory signals.

Speech displays, however, are susceptible to a number of potential problems. Information displays that use speech may interfere with concurrent speech communication with other people [7], [14]. Further, simply converting visual messages to speech can result in lengthy segments of audio that are inefficient for communication. One study [15], for example, suggested that brief speech alarms resulted in longer reaction times than visual text alarms, and another recent experiment showed that participants needed more time to complete a spoken audio version of a test as compared to a visual version of the same test [16].

Partly as a result of these limitations with speech auditory displays, researchers have described alternate display possibilities using nonspeech sounds. Nonspeech auditory displays have been broadly described as *sonifications* [17], [18]. *Earcons* are a type of sonification that use brief, abstract musical sounds to communicate information [19]. The abstract quality of earcons has consistently been identified as potentially problematic for their use in applications [20]–[22], because results have shown that people find earcons difficult to learn. Recently, however, research [23] has suggested that *spearcons*—accelerated speech earcons—may be preferable to earcons in auditory interfaces. Spearcons use brief speech messages that are time-compressed (and also frequency-shifted to avoid pitch increases) for faster delivery of the intended message, and initial studies have shown that spearcons result in

better performance than other nonspeech audio alternatives with respect to speed, accuracy, and learnability. Accelerated speech can overcome some of the temporal inefficiencies of regular spoken messages in systems, and one study [24] showed that synthetic accelerated speech was easier for listeners to process than natural speech delivered quickly (i.e., fast talking).

Accelerated and synthetic speech auditory displays—especially those like spearcons that can make the presentation of speech more time-efficient—offer perhaps the most viable path forward for auditory information display. Already screen readers that use TTS represent one of the most used and most successful instances of auditory information display**,** and some of the most promising lines of research in auditory display [23], [25] have sought to take advantage of the superior temporal processing of human audition [7] by maximizing the efficiency of serial presentations of audio in time (as opposed to complex layered or parallel audio signals). This approach can alleviate some, though perhaps not all, of the potential difficulties with audio information display.

1.2. Theoretical concerns with multimodal information display

The seminal Sonification Report [17] identified the interplay of visual and auditory stimuli as a critical component of a research agenda for auditory displays:

> Multimodal interactions (e.g., between visual and auditory displays) are poorly understood, yet critically affect most sonification applications…When does information presented in one modality interfere with the perception of information in another modality (i.e., cross-modal interference)? How can the total amount of information perceived across all modalities be maximized? Only by careful investigation of these issues can we optimize displays for the type of information conveyed. (pp. 21-22)

The potential for auditory and visual information to overwhelm the information processing capabilities of the human user is a primary concern when designing systems that use multimodal information displays. Models of human information processing have provided some insight into the circumstances in which interference between tasks and stimuli can occur. Wickens' Multiple Resource Theory (MRT)[26], [27] has been particularly influential, likely as a result of its coupling of cognitive theory with a concern for predicting performance and assisting in the design of practice applications.

With respect to the modality of information display, MRT [26], [27] predicts, for example, that two concurrent auditory or two concurrent visual displays have a greater potential to interfere with each than an auditory display paired with a visual display, because each modality represents an independent pool of information processing resources (see Figure 1). Another important aspect of MRT, however, involves the encoding of information in working memory—the active mental workspace that is typically characterized as being involved in the temporary maintenance and processing of active thought [28]. In most theoretical perspectives, including MRT [26], [27] and Baddeley's multi-component model of working memory [28], active thought can assume a verbal (e.g., words) or visuospatial (e.g., images) *processing code* in working memory. Like the modalities dimension of the model, two concurrent verbal or two concurrent visuospatial stimuli are more prone to interference than a paired verbal and visuospatial task.

Despite the usefulness of these heuristics for predicting performance, theoretical difficulties arise with multimodal display scenarios. Specifically, the extent to which visual text and speech, which are both presumably verbal in nature, are processed by the same cognitive mechanisms continues to be debated. Some evidence has suggested that verbal (i.e., linguistic, phonologically-based) processing assumes an amodal working memory code that is distinct from the modality [29] or acoustic features [30] of the external stimulus. From the amodal perspective, then, all linguistic information assumes a verbal code in working memory, regardless of the auditory or visual modality of input. Penney [31], on the other hand, argued in an extensive review of the literature that the mechanisms for processing auditory and visual verbal materials are overwhelmingly distinct and maintain separate perceptual information from the modality of input. Also, a recent series of experiments [32] suggested that appeals to abstract phonological encoding are unnecessary to explain the perceptual coding of verbal information. Further, some research [33], [34] has suggested that the acoustic properties of spoken language (such as the voice of the speaker) are maintained in memory and can influence performance on certain tasks. Still other research [35]; also see [36] has suggested that initially distinct, modal processing mechanisms for speech and text converge on a common, amodal, phonological code in working memory over time. Clearly, the amodal versus perceptual nature of verbal material in working memory remains debated.

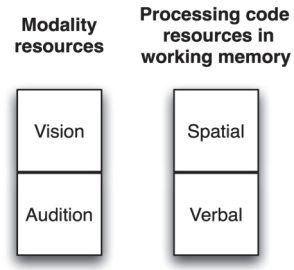

Figure 1: Simplified schematic of the modalities and processing codes dimensions of MRT.

Another theoretical complication is the indeterminate nature of how nonspeech sounds are processed in working memory. In most theory [26]–[28], verbal and visuospatial processing codes are the only representations available in working memory (see Figure 1). Jones and Macken [37] found that tones can interfere with verbal working memory and thus argued for shared mechanisms for speech and nonspeech sounds in working memory. Still other researchers found mutual interference by verbal and musical articulatory suppression with memory for tones and verbal materials [38]. People with congenital amusia show memory deficits that are specific to

pitch and timbre that do not extend to words [39]. Other research has suggested that nonspeech sounds may assume their own distinct working memory processing code--dubbed *auditory imagery*—whereby the acoustic properties of the sounds are remembered and rehearsed [40], [41]. Further, recent research [42], [43] has suggested that people may be able to flexibly encode nonspeech sounds in working memory as sounds, words, or images. With respect to accelerated speech (like spearcons), people could perceive and encode these stimuli as either speech or nonspeech (like tones) sounds, but to date it remains unclear whether accelerated speech is processed as speech in working memory [23]. A recent study showed that spearcons and speech interfered with recall of lists of words to the same extent [44], which suggests that in some instances spearcons may assume a verbal processing code in working memory.

Predicting task interference as a result of working memory processing codes has proven to be difficult [45], in part because the a priori designation of a particular stimulus as verbal or visuospatial (or any other format, e.g,. auditory imagery) in working memory has been difficult to make based on the observable attributes of the stimulus. The viability of multimodal information display in multitasking will hinge upon the ability of researchers to clarify a number of outstanding theoretical issues with respect to working memory processing codes. This issue is particularly relevant for both nonspeech and accelerated speech auditory displays.

1.3. Current experiment and hypotheses

The current experiment used a verification task to compare the encoding of visual text, accelerated speech sounds (spearcons), and nonspeech tones (earcons or brief sonifications). Verification tasks have been used to examine the encoding and encoding strategies. The general paradigm for verifications task involves: 1) the presentation of an initial stimulus that is to be remembered; and 2) the presentation of a second stimulus that either matches or does not match the initial stimulus on some important quality. The participant's task is to either confirm ("verify") or disconfirm the match between the first and second stimuli in a two-choice reaction time task. A classic version of the verification task is the sentence-picture verification task, whereby participants read a sentence (such as "plus is above star") and then compare the sentence to a picture that either matches or does not match the sentence. The picture, then could show either a "+" above a "*" or vice versa. A fundamental assumption of verification task paradigms is that people make faster comparisons (which are reflected in faster verification times) when the format of a stimulus matches their working memory processing code. For example, if the second stimulus is a picture, then people should be faster to make a verification response if they used a visuospatial working memory processing code to remember the first stimulus.

A recent experiment used a verification task for visual text, pictures, and nonspeech sounds [46] and offered some evidence that people seemed to be able to adjust their working memory processing code based on instructions. In general, people responded fastest to the second stimulus of the verification task (either words, a picture, or sounds), when the format of the second stimulus matched the strategy (verbal, visuospatial imagery, or auditory imagery) they had used to remember the first stimulus, regardless of whether the first stimulus was presented as words, a picture, or sounds.

In the current experiment, all stimuli depicted the simple state (either increasing or decreasing) of the price of a fictional stock. Visual text stimuli consisted of either the word "increase" or "decrease." Accelerated speech stimuli were the spoken word "increase" or "decrease," and nonspeech audio stimuli were two brief tones that indicated "increase" by an upward frequency change and "decrease" by a downward frequency change between tones. Participants saw or heard an initial stimulus—either text, accelerated speech, or tones—and were instructed to remember the stimulus according to a prescribed encoding strategy—either as words (a verbal working memory processing code) or tones (a tonal, auditory imagery working memory processing code). Participants were then presented with the second verification stimulus—again either text, accelerated speech, or tones, and responded to indicate whether the state of the second stimulus (stock increased or decreased) matched the first. Results were expected to show that verification times would be fastest for text when participants remembered the initial stimulus with a verbal strategy. For tones, verification times were expected to be fastest when participants remembered the initial stimulus using the tonal auditory imagery strategy. For accelerated speech, we were not certain whether the verbal or tonal auditory imagery working memory code would result in faster verification times, though some research has suggested that accelerated speech is encoded like a verbal stimulus in working memory.

2. METHODS

2.1. Participants

Participants (N = 51, 37 females, M age = 19.41 years, SD = 0.88) were recruited from undergraduate psychology courses and received course extra credit for their participation in the study.

2.2. Apparatus

All presentations of stimuli and collection of data used a program written with Macromedia Director 2004. Visual presentations were made on a 38.1 cm LCD monitor. Sounds were presented with Sony MDR-V6 headphones.

2.3. Stimuli

Verification tasks generally use a limited, simple stimulus set [6], [47], [48]. Thus, stimuli were designed to convey information about the stimulus state (increasing or decreasing) quickly and simply. Text stimuli described the state of the stock with one word—"increase" or "decrease"— presented in approximately 40 point font at the center of the screen. Tonal stimuli used two discrete notes—C4 (262 Hz) and C5 (523 Hz).

Each note was synthesized with the MIDI piano instrument and was 100 ms in length with 10 ms onset and offset ramps. An increase in stock price was represented with C4 followed by C5, and a decrease in stock price was represented with C5 followed by C4. Speech stimuli were the spoken words "increase" and "decrease" recorded to WAV files using the online AT&T Labs TTS demonstration[1] with a female voice ("Crystal, US English"). The resulting WAV files were approximately 600 ms in duration. The files were then imported into the Audacity sound editing program, and the sounds were compressed to 200 ms in duration with the program's "change tempo without changing pitch" function. The current study's use of accelerated speech stemmed from an interest in the applications of accelerated speech and also the desire to minimize any confounds by presenting auditory stimuli (tones and speech) that were equal in duration.

2.4. Procedure

Participants' task was to encode the state of the stock depicted in the first (study) stimulus for comparison with the state depicted in the second (verification) stimulus. Participants were introduced to the task with 36 practice trials. No encoding strategy was prescribed during practice trials. Participants completed blocks of 72 trials of the task with each of the two encoding strategies, with the order of the encoding strategies counterbalanced across participants. During the verbal encoding block, participants were instructed to encode the study stimulus as either the word "increase" or "decrease." During the auditory imagery encoding block, participants were instructed to encode the study stimulus in auditory memory as one of the two-note tonal stimuli by encoding an increase in pitch if the stock price increased or a decrease in pitch if the stock price decreased. Before they began using a prescribed encoding strategy, participants confirmed that they understood the strategy.

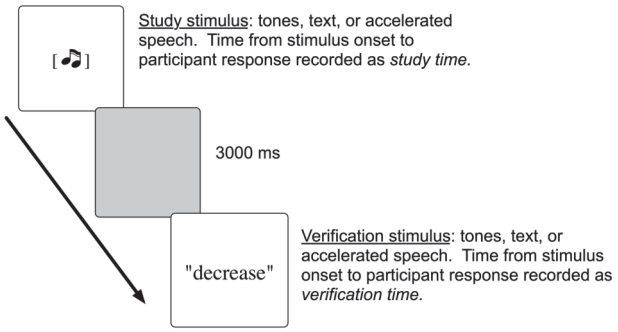

Figure 2: Structure of an experimental trial

The trial structure is represented in Figure 2. Participants positioned their left and right index fingers on the "Z" key and "?" keys, respectively, and pressed either key to initiate the study stimulus. Participants then encoded the study stimulus according to the assigned strategy. When the participants felt that they had satisfactorily encoded the study stimulus using the designated strategy, they pressed either key to continue. The time from the onset of the study stimulus until participants logged a response was recorded as the dependent variable *study time*. A blank grey screen was then displayed for 3000 ms, after which the verification stimulus was automatically displayed. Participants pressed the "Z" key for matches (e.g., the study stimulus and the verification stimulus both indicated an increase in stock value) or the "?" key for mismatches. The pairing of key presses to responses was counterbalanced across participants. Participants were instructed to respond to the verification stimulus as quickly as possible while following encoding instructions and avoiding errors**.** Participants saw feedback about their verification reaction times following every trial. All possible stimulus combinations (tones, text, and speech in increasing and decreasing states) appeared twice in the 72 trials, and the order of trials within a block was randomized. The primary dependent variable was *verification time*, the time from the onset of the verification stimulus until the participant pressed a response key to indicate a match or mismatch. Participants also completed the NASA-TLX [49] as a measure of workload for each type of encoding strategy.

3. RESULTS

One participant had many instances of reaction times under 200 ms and was excluded from all analyses for failing to follow experimental instructions. Greenhouse-Geisser corrections were used for all violations of sphericity assumptions, and corrected degrees of freedom were reported where appropriate.

3.1. Analyses for Verification Times

A 3 (study stimulus format: sound, speech, or text) by 2 (encoding strategy: verbal or auditory imagery) by 3 (verification stimulus format: sound, speech, or text) analysis of variance (ANOVA) was performed on the verification time dependent variable (see Figure 3). For this and all other analyses, Greenhouse-Geisser corrections were used in cases where sphericity assumptions were violated. Results showed significant main effects for strategy, $F(1,49) = 5.82$, $p = .02$, $\eta^2_p = .11$, and verification stimulus format, $F(1.66,81.34) = 12.32$, $p < .001$, $\eta^2_p = .20$. The interactions of strategy with verification stimulus format, $F(2,98) = 23.41$, $p < .001$, $\eta^2_p = .32$, and study stimulus with verification stimulus format, $F(2.93,143.36) = 9.80$, $p < .001$, $\eta^2_p = .17$, were both significant. The main effect of study stimulus, $F(1.48,72.41) = 0.70$, $p = .46$, the interaction of strategy with study stimulus, $F(1.39,68.26) = 0.72$, $p = .45$, and the three way interaction, $F(2.63,128.91) = 1.50$, $p = .22$, were not significant.

For the main effect of strategy, pairwise comparisons showed that participants were faster (all response times reported in milliseconds) to verify stimuli using the verbal strategy (M =961.93, SE = 54.62) as compared to the tonal auditory imagery strategy (M =1056.92, SE = 49.74). For the main effect of verification stimulus format, pairwise comparisons showed that participants were faster to verify text (M =954.93, SE = 54.02) than speech (M =999.16, SE = 43.23), $p = .036$, or sounds (M =1074.19, SE = 53.16), $p < .001$, and

[1] http://www2.research.att.com/~ttsweb/tts/demo.php

speech stimuli were also verified faster than sounds, p = .001. Main effects were qualified by the significant interactions.

For the interaction of strategy with verification stimulus format (collapsed across study stimulus format), simple effects analyses at each level of verification stimulus format showed that when the verification stimulus was a sound (left two bars in Figure 3), there was no difference between the tonal auditory imagery (M = 1037.68, SE = 49.78) and verbal strategies (M = 1110.69, SE = 64.65), p = .11. When the verification stimulus was text (middle two bars in Figure 3), participants were significantly faster to respond when they used the verbal strategy (M = 875.41, SE = 66.61) as compared to the tonal auditory imagery strategy (M = 1034.47, SE = 55.05), p = .008. When the verification stimulus was speech (right two bars in Figure 3), participants were significantly faster to respond when they used the verbal strategy (M = 899.71, SE = 40.08) as compared to the tonal auditory imagery strategy (M = 1098, SE = 52.36) p < .001.

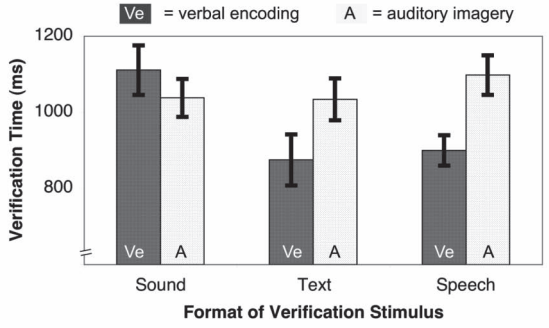

Figure 3:. Verification response times as a function of encoding strategy and the format of the verification stimulus.

For the interaction of study stimulus format with verification stimulus format (collapsed across encoding strategy, see Figure 4), simple effects analyses at each level of verification stimulus format showed that when the verification stimulus was a sound (left three bars in Figure 4), participants responded significantly faster if they had studied a sound (M = 992.76, SE = 47.74) as compared to text (M = 1101.55, SE = 54.65), p = .001, or speech (M = 1128.24, SE = 65.47), p = .001, and the difference between text and speech was not significant, p = .35. When the verification stimulus was text (middle three bars in Figure 4), there were no significant differences between participants who studied sounds (M = 955.82, SE = 69.01), speech (M = 961.45, SE = 56.29), p = .86, or text (M = 947.50, SE = 50.18), p = .88. The difference between having studied text and speech was also not significant, p = .70. When the verification stimulus was speech (right three bars in Figure 3), participants responded significantly faster if they studied speech (M= 921.19, SE = 35.91) as compared to sounds (M = 1053.45, SE = 51.14), p < .001, or text (M = 1022.75, SE = 47.35), p = .001. The difference between those who studied sounds and those who studied text was not significant, p = .17.

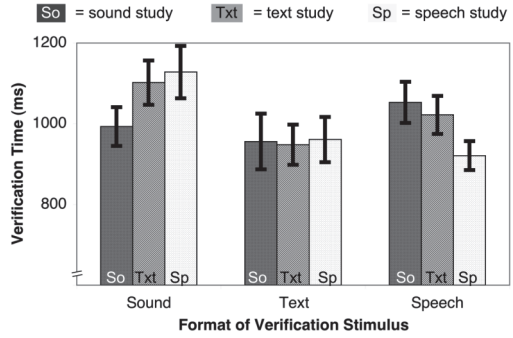

Figure 4: Verification response times as a function of the format of the study stimulus and format of the verification stimulus.

3.2. Analyses of study time

A 2 (encoding strategy: verbal or tonal auditory imagery) by 3 (study stimulus format: sound, speech, or text) analysis of variance (ANOVA) was performed on the dependent variable, study time. Results showed a significant main effect of encoding strategy, $F(1,49)$ = 14.29, p < .001, η^2_p = .23, a significant main effect of study stimulus format, $F(2,98)$ = 6.37, p = .002, η^2_p = .12, and a significant interaction of strategy with study stimulus format, $F(1.76,86.31)$ = 6.96, p = .002, η^2_p = .12.

The main effect of strategy showed that participants were significantly faster to encode the study stimulus using the verbal strategy (M =1442.51, SE = 107.20) as compared to the tonal auditory imagery strategy (M =1868.06, SE = 149.35). The main effect of study stimulus format showed that participants were significant faster to encode text (M = 1566.84, SE = 114.28), p = .003, and speech (M = 1626.30, SE = 122.65), p = .022, as compared to sounds (M = 1772.71, SE = 128.90). The difference between text and speech was not significant, p = .24. The main effects should be interpreted cautiously due to the significant interaction of strategy with study stimulus format (see Figure 5). Simple effects at each level of study stimulus format showed that when the study stimulus was a sound, there were no significant differences in study times using the tonal auditory imagery (M = 1893.67, SE = 163.63) versus verbal encoding strategy (M = 1651.75, SE = 121.43), p = .066. When the study stimulus was text, participants were significantly faster to encode the stimulus using the verbal encoding strategy (M = 1339.32, SE = 106.94) as compared to the auditory imagery strategy (M = 1794.37, SE = 150.43) p = .001. When the study stimulus was speech, participants were significantly faster to encode the stimulus using the verbal encoding strategy (M = 1336.46, SE = 108.36) as compared to the auditory imagery encoding strategy (M = 1916.15, SE = 159.05), p < .001.

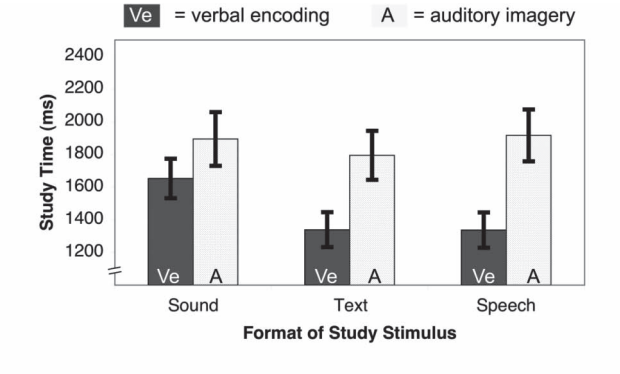

Figure 5: Study times as a function of encoding strategy and format of the study stimulus.

3.3. Analyses of Practice Trials

Both study times and verification times were analyzed for the first block of practice trials during which participants were given no encoding strategy. Results from this block, then, reflect spontaneously chosen or default strategies in working memory. For study times during the practice block, a 1 by 3 (study stimulus format: sound, speech, or text) ANOVA showed no significant differences, $F(2,98) = 1.24$, $p = .30$, which indicated that participants took the same amount of time to study and encode sounds (M =2537.43, SE = 192.95), text (M =2246.02, SE = 155.65), and speech (M =2524.04, SE = 220.83).

For verification times during the practice block, a 3 (study stimulus format: sound, speech, or text) by 3 (verification stimulus format: sound, speech, or text) ANOVA showed a significant main effect of verification stimulus format, $F(1.50,73.70) = 8.80$, $p = .001$, $\eta^2_p = .15$, and a significant interaction of study stimulus format with verification stimulus format, $F(3.31, 162.33) = 3.08$, $p = .025$, $\eta^2_p = .06$. The main effect of study stimulus format, $F(2,98) = 0.16$, $p = .85$, was not significant. Follow-up pairwise comparisons for the main effect of verification stimulus format showed that participants were faster to verify text (M = 1313.36, SE = 91.25), $p = .001$ and speech (M =1350.72, SE = 89.44), $p = .004$, as compared to sounds (M =1595.05, SE = 71.50), and the difference between speech and text was not significant, $p = .437$.

For the interaction of study stimulus format with verification stimulus format (collapsed across encoding strategy), simple effects analyses at each level of verification stimulus format showed that when the verification stimulus was a sound, participants responded significantly faster after having studied a sound (M = 1467.71, SE = 74.76) as compared to speech (M = 1689.70, SE = 94.86), $p = .02$, but they were not significantly faster for a sound as compared to text (M = 1627.75, SE = 94.37), $p = .07$. The difference between text and speech was not significant, $p = .49$. When the verification stimulus was text, participants were significantly faster to respond if they had studied speech (M = 1240.08, SE = 84.73) as compared to sounds (M = 1433.25, SE = 126.15), $p = .026$, and there was no significant difference for having studied sounds as compared to text (M = 1266.77, SE = 90.06), $p = .074$. The difference between having studied text and speech was also not significant, $p = .64$. When the verification stimulus was speech, there were not significant differences if participants had studied speech (M= 1285.25, SE = 94.82) as compared to sounds (M = 1410.53, SE = 115.38), $p = .14$, or text (M = 1356.42, SE = 121.26), $p = .56$, and the difference between having studied sounds and text also was not significant, $p = .69$.

3.4. Analysis of Workload

Analysis of workload as measured by the NASA-TLX [49] composite score showed no significant difference in perceived workload using auditory imagery (M = 11.48, SE = .38) as compared to verbal encoding (M = 10.98, SE = .35), $t(49) = 1.20$, $p = .24$.

4. DISCUSSION

Results suggested an interplay of both modality-based effects and the effects of working memory processing codes. Several findings suggested that accelerated speech and text were processed similarly in working memory, and, also that these stimuli were processed differently from sounds. The tonal auditory imagery strategy showed no advantage for sound stimuli (left bars in Figure 3), but for both text and speech, participants who used the verbal encoding strategy had an advantage and responded more quickly (middle and right bars in Figure 3). Similarly, when the initial stimulus was text or speech, participants were faster to encode the stimulus using the verbal strategy (middle and right bars of Figure 5). The strategy did not affect encoding times for sounds (left bars of Figure 5). This pattern of findings suggested that accelerated speech and text share a common working memory processing code. Spearcons, then, seemed to invoke verbal representations in working memory, though this result may not generalize to all auditory display scenarios, given the limited accelerated speech stimulus set used here.

Other results, however, showed ways in which accelerated speech and tones seemed to be processed differently from text and from one another. Particularly, a pattern emerged to suggest that both types of auditory stimuli (tones and accelerated speech) had lingering perceptual effects that persisted across working memory processing codes. For sound verification stimuli, participants were at an advantage during verification if they had studied sounds (left bars in Figure 4), and a parallel finding of initial and second stimulus congruity was found for speech verification stimuli (right bars in Figure 4). For text verification stimuli, the format of the initial stimulus that had been remembered did not matter (middle bars in Figure 4). This is interesting in that the advantage for congruities only occurred between study and verification stimuli in the auditory modality. Within that modality, however, advantages were format-specific (tones versus speech) and not simply a function of matching modality (vision versus audition). This is evidence in support of perceptual-based, auditory-specific effects in working memory that are also sensitive to whether the stimuli are verbal or tonal. A similar result for lingering auditory memory for tones was reported in [46]. This effect, then, appears to be robust across

experiments and different types of auditory stimuli and warrants further investigation.

Results generally did not support the predicted effects of auditory tonal imagery strategy in that there were no significant advantages for verifying sounds using this working memory processing code as compared to a verbal code. A previous study [46] did show a small advantage for using an auditory imagery strategy to verify tones. The TLX workload measure suggested that participants did not find the auditory imagery strategy to be more difficult to use, though the reaction time data suggested that the auditory imagery strategy was not particularly successfully implemented. The manipulation of the processing code *did* result in several notable differences for auditory imagery as compared to a verbal processing code, however, and these differences mostly manifested in slower processing of verbal stimuli during both encoding and verification when using auditory imagery. Whereas these results did not show the predicted facilitation for shorter study times and faster verification times when auditory imagery for tones was congruent with the perception of tonal stimuli during encoding or verification, the overall pattern of results did suggest that auditory imagery was distinct from verbal encoding. Previous researchers have suggested that verbal memory may simply be superior to tonal memory [38].

The current study's results have implications for auditory interface design. Most participants spontaneously defaulted to a verbal encoding strategy during the practice trials for which a specific encoding strategy was not specified. Even in conditions involving encoding and comparisons with tones, there were no significant differences between auditory imagery and verbal encoding. As a result, speech or text displays may be preferable to nonspeech displays for tasks that require speeded comparisons. Results did indicate, however, that speeded comparisons involving tonal stimuli can be facilitated when comparing tones to other tones (as opposed to comparing speech or text to tones).

5. REFERENCES

[1] S. Brewster, "Using non-speech sound to overcome information overload," *Displays*, vol. 17, pp. 179–189, 1997.

[2] S. Brewster, "Overcoming the lack of screen space on mobile computers," *Personal and Ubiquitous Computing*, vol. 6, no. 3, pp. 188–205, 2002.

[3] M. A. Nees and B. N. Walker, "Auditory displays for in-vehicle technologies," in *Reviews of Human Factors and Ergonomics*, P. Delucia, Ed. Thousand Oaks, CA: Sage Publishing/Human Factors and Ergonomics Society, 2011, pp. 58–99.

[4] M. R. McGee-Lennon, M. Wolters, and T. McBryan, "Audio reminders in the home environment," in *13th International Conference on Auditory Display*, 2007.

[5] P. M. Sanderson, "The multimodal world of medical monitoring displays," *Applied Ergonomics*, vol. 37, pp. 501–512, 2006.

[6] P. A. Carpenter and M. A. Just, "Sentence comprehension: A psycholinguistic processing model of verification," *Psychological Review*, vol. 82, no. 1, pp. 45–73, 1975.

[7] G. Kramer, "An introduction to auditory display," in *Auditory Display: Sonification, Audification, and Auditory Interfaces*, G. Kramer, Ed. Reading, MA: Addison Wesley, 1994, pp. 1–78.

[8] M. A. Nees and B. N. Walker, "Auditory interfaces and sonification," in *The Universal Access Handbook*, C. Stephanidis, Ed. New York: Lawrence Erlbaum Associates, 2009, pp. 507–521.

[9] D. S. Brungart, M. A. Ericson, and B. D. Simpson, "Design considerations for improving the effectiveness of multitalker speech displays," in *8th International Conference on Auditory Display*, 2002, pp. 424–430.

[10] T. Bonebright and M. A. Nees, "Most earcons do not interfere with spoken passage comprehension," *Applied Cognitive Psychology*, vol. 23, no. 3, pp. 431–445, 2009.

[11] S. E. Smith, K. L. Stephan, and S. P. A. Parker, "Auditory warnings in the military cockpit: A preliminary evaluation of potential sound types," 2004.

[12] H.-R. Pfister, S. Wollstadter, and C. Peter, "Affective responses to system messages in human-computer-interaction: Effects of modality and message type," *Interacting with Computers*, vol. 23, pp. 372–383, 2011.

[13] R. W. Massof, "Auditory assistive devices for the blind," in *International Conference on Auditory Display ICAD2003*, 2003, pp. 271–275.

[14] B. A. Schneider, L. Li, and M. Daneman, "How competing speech interferes with speech comprehension in everyday listening situations," *Journal of the American Academy of Audiology*, vol. 18, no. 7, pp. 559–572, 2007.

[15] N. A. Stanton and C. Baber, "Comparing speech versus text displays for alarm handling," *Ergonomics*, vol. 40, no. 11, pp. 1240–1254, 1997.

[16] M. A. Nees, "Correlations and scatterplots: A comparison of auditory and visual modes of learning and testing," in *Proceedings of the 18th International Conference on Auditory Display*, Atlanta, GA, USA, 2012, pp. 195–198.

[17] G. Kramer, B. N. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff, R. Bargar, S. Barrass, J. Berger, G. Evreinov, W. T. Fitch, M. Gröhn, S. Handel, H. Kaper, H. Levkowitz, S. Lodha, B. Shinn-Cunningham, M. Simoni, and S. Tipei, "The Sonification Report: Status of the Field and Research Agenda. Report prepared for the National Science Foundation by members of the International Community for Auditory Display," 1999.

[18] B. N. Walker and M. A. Nees, "Theory of sonification," in *Principles of Sonification: An Introduction to Auditory Display*, T. Hermann, A. Hunt, and J. Neuhoff, Eds. Berlin, Germany: Logos Publishing House, 2011, pp. 9–39.

[19] D. K. McGookin and S. Brewster, "Earcons," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. Neuhoff, Eds. Berlin, Germany: Logos Verlag, 2011, pp. 339–361.

[20] T. L. Bonebright and M. A. Nees, "Memory for auditory icons and earcons with localization cues," in *International Conference on Auditory Display (ICAD2007)*, 2007, pp. 419–422.

[21] D. Palladino and B. N. Walker, "Learning rates for auditory menus enhanced with spearcons versus earcons," in *International Conference on Auditory Display (ICAD2007)*, 2007, pp. 274–279.

[22] N. C. Perry, C. J. Stevens, M. W. Wiggins, and C. E. Howell, "Cough once for danger: Abstract warnings as informative alerts in civil aviation," *Human Factors*, vol. 49, no. 6, pp. 1061–1071, 2007.

[23] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon, "Spearcons (Speech-Based Earcons) Improve Navigation Performance in Advanced Auditory Menus," *Human Factors*, in press.

[24] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech," *Speech Communication*, vol. 42, pp. 155–173, 2004.

[25] D. Brock, B. McClimens, and S. C. Peres, "Evaluating listeners' attention to and comprehension of serialy interleaved, rate-accelerated speech," presented at the 18th International Conference on Auditory Display, Atlanta, GA, USA, 2012, pp. 172–179.

[26] C. D. Wickens, "Processing resources in attention," in *Varieties of Attention*, R. Parasuraman and D. R. Davies, Eds. New York: Academic Press, 1984, pp. 63–102.

[27] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, no. 2, pp. 159–177, 2002.

[28] A. D. Baddeley, "Is working memory still working?," *European Psychologist*, vol. 7, no. 2, pp. 85–97, 2002.

[29] E. H. Schumacher, E. Lauber, E. Awh, J. Jonides, E. E. Smith, and R. A. Koeppe, "PET Evidence for an amodal verbal working memory system," *Neuroimage*, vol. 3, no. 2, pp. 79–88, 1996.

[30] A. G. Samuel, "Central and peripheral representation of whispered and voiced speech," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 14, no. 3, pp. 379–388, 1988.

[31] D. M. Jones and W. J. Macken, "Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, no. 2, pp. 369–381, 1993.

[32] Z. A. Schendel and C. Palmer, "Suppression effects on musical and verbal memory," *Mem Cognit*, vol. 35, no. 4, pp. 640–650, Jun. 2007.

[33] C. G. Penney, "Modality effects and the structure of short-term verbal memory," *Memory & Cognition*, vol. 17, no. 4, pp. 398–422, 1989.

[34] D. W. Maidment and W. J. Macken, "The ineluctable modality of the audible: Perceptual determinants of auditory verbal short-term memory," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 38, no. 4, pp. 989–997, Aug. 2012.

[35] B. Tillmann, K. Schulze, and J. M. Foxton, "Congenital amusia: A short-term memory deficit for non-verbal, but not verbal sounds," *Brain and cognition*, vol. 71, no. 3, p. 259, 2009.

[36] S. D. Goldinger, "Words and voices: episodic traces in spoken word identification and recognition memory," *J Exp Psychol Learn Mem Cogn*, vol. 22, no. 5, pp. 1166–1183, Sep. 1996.

[37] T. J. Palmeri, S. D. Goldinger, and D. B. Pisoni, "Episodic encoding of voice attributes and recognition memory for spoken words," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, no. 2, pp. 309–328, 1993.

[38] D. S. Ruchkin, R. S. Berndt, J. Johnson, W. Ritter, J. Grafman, and H. L. Canoune, "Modality-specific processing streams in verbal working memory: Evidence from spatio-temporal patterns of brain activity," *Cognitive Brain Research*, vol. 6, pp. 95–113, 1997.

[39] S. Crottaz-Herbette, R. T. Anagnoson, and V. Menon, "Modality effects in verbal working memory: Differential prefrontal and parietal responses to auditory and visual stimuli," *Neuroimage*, vol. 21, pp. 340–351, 2003.

[40] A. R. Halpern and R. J. Zatorre, "When that tune runs through your head: A PET investigation of auditory imagery for familiar melodies," *Cerebral Cortex*, vol. 9, no. 7, pp. 697–704, 1999.

[41] E. G. Schellenberg and S. E. Trehub, "Good pitch memory is widespread," *Psychological Science*, vol. 14, no. 3, pp. 262–266, 2003.

[42] M. A. Nees and B. N. Walker, "Encoding and representation of information in auditory graphs: Descriptive reports of listener strategies for understanding data," in *International Conference on Auditory Display (ICAD 08)*, 2008.

[43] M. A. Nees and B. N. Walker, "Mental scanning of sonifcations reveals flexible encoding of nonspeech sounds and a universal per-item scanning cost," *Acta Psychologica*, vol. 137, pp. 309–317, 2011.

[44] M. Wolters, K. Isaac, and J. Doherty, "Hold that thought: are spearcons less disruptive than spoken reminders?," in *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2012, pp. 1745–1750.

[45] K. J. Sarno and C. D. Wickens, "Role of multiple resources in predicting time-sharing efficiency: Evaluation of three workload models in a multiple-task setting," *The International Journal of Aviation Psychology*, vol. 5, no. 1, pp. 107–130, 1995.

[46] M. A. Nees and B. N. Walker, "Flexibility of Working Memory Encoding in a Sentence-Picture-Sound Verification Task," submitted.

[47] C. M. MacLeod, E. B. Hunt, and N. N. Mathews, "Individual differences in the verification of sentence-picture relationships," *Journal of Verbal Learning and Verbal Behavior*, vol. 17, no. 5, pp. 493–507, 1978.

[48] N. N. Mathews, E. B. Hunt, and C. M. MacLeod, "Strategy choice and strategy training in sentence-picture verification," *Journal of Verbal Learning and Verbal Behavior*, vol. 19, no. 5, pp. 531–548, 1980.

[49] S. G. Hart and L. E. Staveland, "Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds. Amsterdam: North Holland Press, 1988, pp. 239–250.