Communicating Printed Headings to the ear

Robert F. Lorch, Hung-Tao Chen, Aqeel A Jawahir, Julie Lemarié

Communicating Printed Headings to the Ear

Robert F. Lorch, Jr., Hung-Tao Chen, Aqeel A. Jawahir

University of Kentucky

&

Julie Lemarié

Université de Toulouse – Jean Jaurès

Corresponding author:

Robert F. Lorch, Jr.

Department of Psychology

University of Kentucky

Lexington, KY 40506-0044

rlorch@email.uky.edu

Phone: 859-257-6826

Fax: 859-323-1979

Abstract

Two experiments compared three different methods of translating printed headings into an auditory format. In both experiments, college students listened to a text with instructions to stop the recording whenever they heard a heading and type the hierarchical level and exact wording of the heading. In both experiments, listeners were relatively poor at identifying headings and their levels if the headings were not signaled in any way. Listeners were very good at identifying headings if headings were preceded by a tone whose frequency indicated the hierarchical level (Experiment 1) or if the headings were accompanied by a change of speaking voice where a different voice was associated with each hierarchical level (Experiment 2). In both experiments, listeners were very good at identifying headings if the headings were preceded by a label that explicitly signaled the heading and communicated its level. The labeling method was superior to tones or voice changes in communicating information about hierarchical level. Thus, the study identifies a simple method of effectively communicating headings in spoken text.

Keywords: text-to-speech software, headings, text signaling devices, text processing

Communicating Printed Headings to the Ear

Text-to-speech (TTS) applications communicate continuous prose very intelligibly, but they are less successful at communicating other representational formats. For example, tables, graphs and signaling devices (e.g., typographical contrast, headings) are rarely well-communicated by TTS applications. Often the text content is not coded in a form that would allow the TTS application to interpret important formatting information. Even when the text content is appropriately coded, not all TTS applications are programmed to process and translate formatting information. And even when the text content is appropriately coded and the application attempts to process it, conversions of graphical and tabular formats to an auditory format often leave much room for improvement (Spiliotopoulos, Xydas & Kouroupetroglou, 2007; Tsonos, Xydas & Kouroupetroglou, 2007). The challenge of converting a visual, analog representation (e.g., graphs) to an auditory, serial representation (e.g., speech) is in constructing an auditory representation that communicates the relevant information efficiently enough to avoid overburdening the memory limitations of the human user (Richardson, 2010; Spiliotopoulos, Xydas, Kouroupetroglou, Argyropoulos & Ikospentaki, 2010). This is a formidable challenge for TTS applications. However, a second challenge seems more tractable. Namely, it should be relatively easy to find an auditory means of effectively communicating relevant formatting information (e.g., change of typeface or font) associated with signaling devices (Lorch, 1989). In this study, we evaluate options for communicating one common type of signaling device, headings, via TTS (Lemarié, Eyrolle & Cellier, 2006; Pascual, 1996).

Printed headings are often not effectively communicated when translated to speech because formatting (e.g., contrasting typeface or font, white space) information is lost in the translation. Formatting information is critical because it carries information about the structure

of the text and the function of the heading. Formatting information is often not communicated when printed text is translated to speech either because the text code does not retain formatting information or because the TTS application is not programmed to interpret the formatting (e.g., the Kindle *e*-reader). Whatever the reason, the loss of the information carried by the formatting of headings disrupts listeners' attempts to understand the spoken text (Lorch, Chen & Lemarié, 2012). Therefore, it is important that the text code retain formatting information and that TTS applications communicate the information in a manner that is readily interpretable and usable by the listener.

The first step in determining how best to communicate the information carried by the formatting of headings is to identify what information needs to be communicated. Headings can vary a great deal in their form and function (Lemarié, Lorch, Eyrolle & Virbel, 2008). However, most headings to consist of a verbal label that states a topic or function of the ensuing section of text, as well as being formatted in a distinctive way. For example, the title at the top of this page designates the topic of the article. The boldfaced type, centering and white space separating the title from the subsequent content conveys additional, important information that must be retained when the printed text is rendered in an auditory form. In principle, it is possible to analyze the meaning of the formatting and find a way to render the information in some auditory form. In fact, Lorch et al. (2012) were successful in doing just that. They used a theoretical analysis of signaling devices (Lemarié et al., 2008) to determine that the printed formatting of the headings in their texts communicated three distinguishable types of information: (1) the formatting clearly demarcated the sections of the texts; (2) the formatting also provided emphasis to the labels comprising the headings which, in turn, marked their functions as headings and emphasized the content of the labels; and (3) the systematic variation in the formatting of different headings

communicated which of three hierarchical levels were associated with any given heading. Based on this analysis, the researchers attempted to restore the missing information by inserting a pause after each heading to more clearly separate it from the main content of the text. In addition, they preceded each heading by a phrase that identified it as a heading and explicitly stated its hierarchical level (e.g., "Level 1 heading: Energy problems"). This approach is very similar to how some common screen-readers (e.g., *JAWS*) communicate headings in speech. Indeed, these simple modifications resulted in greatly improved performance on an outlining task relative to control conditions of no headings, or headings rendered by the TTS application without the above modifications. In fact, outlining improved to the point of being indistinguishable from outlining based on the original printed text.

The Lorch et al. (2012) study was successful in two goals: It demonstrated that the failure to translate the printed formatting of headings into an auditory format had negative consequences for text comprehension. And it demonstrated that one approach to remedying that problem was quite successful under the conditions investigated in the study. The current study follows up Lorch et al. (2012) to address two questions. First, we wanted to determine whether the incomplete renderings of the headings by TTS resulted in listeners having trouble correctly perceiving the headings at the time they were spoken. The Lorch et al. procedure did not allow separation of the contribution of perceptual factors from memory factors to poor outlining performance. Participants in the study were required to write an outline as they listened to the text being spoken. Because the participants were asked to simultaneously listen and write, they may well have adopted a buffering strategy in which they did not immediately update their outlines upon hearing a new heading. Rather, they may have tentatively identified the heading but listened to additional content to gather more evidence that they were, indeed, dealing with a

new topic. If the participants did use a type of buffering strategy, memory lapses may have been an important source of errors in outlining. The procedure is also problematic in that the act of writing may interfere with listening to the presentation of the text.

To minimize the delay between when a heading was spoken and when participants responded to the heading, we developed a new task for the current investigation. Participants were told that the purpose of the study was to compare different ways of presenting headings in a TTS environment. Their task was to identify each heading as soon as it was spoken. When they heard a heading, they were to immediately stop the recording and type the hierarchal level and the exact wording of the heading. This procedure should reveal when listeners misperceive a heading or completely fail to recognize that a heading has been presented.

The second issue we addressed concerns the relative effectiveness of alternative methods of rendering heading information via TTS. The Lorch et al. study demonstrated that outlining performance greatly improves if each heading is preceded by a simple verbal label that warns of the upcoming heading and specifies its hierarchical level (e.g., "Level 3 heading: Environmentally-sensitive areas"). However, there are alternative ways of communicating this information. For example, tones might be used to warn of an upcoming heading and the frequency of the tones might be used to code the hierarchical level of a heading. Alternatively, a change of voice might be used to indicate that a phrase is a heading and different voices might be used to code different hierarchical levels. In Experiment 1, we compared the relative effectiveness of using verbal labels vs. tones to signal headings and convey their hierarchical levels. In Experiment 2, we compared verbal labels with the use of different voices to communicate headings and their hierarchical levels.

Experiment 1

Verbal labels that precede a heading and announce the hierarchical level of the heading (e.g., “Level 1 heading: Energy problems” or “Level 3 heading: Storage of radwaste“) have already been shown to be very effective at supporting outlining of a text presented by a TTS application (Lorch et al., 2012). Verbal labels have the advantage of communicating level information in a very straightforward way; that is, there is no need to interpret level information because the label provides the number of the level directly. Further, preceding each heading with a label that follows the same simple syntax for every heading probably aids listeners’ identification of section boundaries. However, labels have the potential disadvantage of not being auditorily distinctive relative to the spoken text content. An inattentive listener might therefore miss a heading if distracted at the time a label is spoken. The parallel situation in a printed text would be if a heading was presented without any distinguishing typography and no white space to set it off from the body of the text. In that situation, a reader might easily overlook a heading.

An alternative to preceding each heading with a label is to precede each heading with a tone whose frequency indicates the hierarchical level of the upcoming heading. The use of non-speech sounds to convey visual information is a common strategy “in eyes-free context applications” (Ferati, Pfaff, Mannheimer, & Bolchini, 2012, p. 936). Tones might be more effective than verbal labels at helping listeners to identify headings because the simple but distinctive auditory signal should alert readers whose attention has wandered. However, the hierarchical information communicated by tones may not be easily processed because – in contrast to the correspondence between a verbal level and a heading’s level – the correspondence between a tone’s frequency and the level it indicates must be learned. Further, to the extent that

listeners have difficulty identifying the hierarchical level associated with a particular tone, processing of the heading itself may be disrupted.

We compared verbal labels and tones in a task we developed to study online processing of headings. Participants listened to a text containing headings and were instructed to stop the recording as soon as they identified a new heading. Whenever they stopped the recording, a textbox was presented and they were to type the level of the heading and the exact wording of the heading. We were interested in several measures of a participant's performance: (1) How many of the text's headings did the participant detect? (2) How quickly did the participant respond after a heading was presented? (3) How often did the participant wrongly indicate the presence of a heading (i.e., false alarm)? (4) How accurate was the participant in recording the exact wording of a heading? (5) How accurate was the participant in processing the hierarchical level of a heading?

The four conditions in the experiment were formed by combining two independent variables: (1) the presence vs. absence of a verbal label preceding each heading, and (2) the presence vs. absence of a tone preceding each heading. We anticipated that participants would perform better if the heading was preceded by a label or tone or both than if there was no signal preceding the headings. That is, the presence of either or both types of signals should produce better and faster detection of headings, fewer false alarms, more accurate memory for the wording and better identification of the hierarchical level. If any differences were to be observed in comparing labels and tones, we anticipated that tones might produce better and faster detection of headings and fewer false alarms because tones should be more effective at alerting an inattentive listener. However, we anticipated that labels would produce better recognition of

the hierarchical level of a heading because labels communicate this information in a more accessible form than tones do.

Method

Participants. A total of 64 volunteers from the Psychology subject pool participated in the experiment to satisfy a research participation requirement. Participants' ages ranged between 18 and 24. Data from two participants were dropped because they failed to follow instructions. A third participant's data was dropped because of experimenter error (i.e., the wrong audio file was played).

Materials. Two audio text files were created for the experiment. One audio text was on the topic of energy problems and their solutions; the second audio text was on the topic of firefighting and prevention. The texts were adopted from previous experiments (Lorch et al., 2012; Lorch, Lemarié & Grant, 2011a, 2011b). Each text had a total of 20 headings organized into three hierarchical levels. The 20 headings included two top-level headings, four mid-level headings, and 14 bottom-level headings. The energy text had a total of 1596 words and the fire-fighting text had a total of 1584 words in the control condition. The Flesch reading ease scores for the energy and fire texts were 38.8 and 50.3, respectively. The Flesch-Kincaid grade level scores for the energy and fire texts were 11.3 and 10.0, respectively.

The audio texts were created by converting Word document files into TXT format files that were readable by Amazon Kindle DX. The printed texts were converted to audio using Amazon Kindle DX's Text-to-Speech function. The default setting was chosen for the Text-to-Speech processing, which was represented on Kindle' function menu as medium speech rate with female voice. The medium speech rate corresponded to an average of 2.92 words per second and the female voice had an average pitch frequency of 165.17 Hz. The audio output from Kindle

was recorded to a Windows computer using Windows' built-in audio recording software. The energy text was 10 minutes and 51 seconds long and the fire text was 10 minutes and 12 seconds long in their control versions.

The original recording from Kindle served as the control condition for both texts. The audio headings in the control condition had a 650 millisecond (ms) pause before and after a heading. No modifications were made to the recording for the control condition.

Three more versions of audio texts were created for each text. These versions differed in how they rendered the headings. Each version was designed to clearly distinguish each heading from the text content and explicitly communicate the hierarchical level of each heading. The "label" versions of the audio texts added a phrase before each heading that identified it as a heading and stated the hierarchical level of the heading. The form of each preceding label was the same: A 650 ms pause preceded the label; the label was of the form "Level # Heading"; a 650 ms pause followed the label; the heading was then verbalized and followed by a 650 ms pause. For example, the audio heading for the first, major heading of the energy text was: "[pause] Level 1 Heading [pause] Energy Problems [pause]." This version of the audio texts was an exact replication of the corresponding condition in Experiment 2 of Lorch et al. (2012).

The "tones" versions of the audio texts were so-named because a tone preceded each heading. A 1000 Hz tone preceded the two major headings in a text; a 690 Hz tone preceded the four mid-level headings; and a 408 Hz tone preceded each of the 14 low-level headings. We assumed that having higher frequencies associated with higher levels in the text structure would constitute a correspondence between tone frequency and hierarchical level that would be easy for participants to learn. The tones were created using Audacity, version 2.0.4 and they each had a duration of 900 ms. The duration time was chosen to be equal to the average playtime duration of

the labels in the Label condition. There were also 650 ms pauses before and after each tone. The tones had an amplitude of 1 while the rest of the text had an amplitude of 0.5.

The "labels + tones" versions of the audio texts preceded each heading with a tone and a label. The tone always occurred first and its frequency was varied just as in the tones condition; the label followed the tone and had the same format as the label condition; the heading was verbalized last. A 650 ms interval intervened between the tone and label, and between the label and heading. In addition, a 650 ms interval preceded the tone and followed the heading.

The audio texts were presented on Dell desktop computers. Participants listened to the recordings on JVC HA-V570 headphones with individual volume dials that participants could use to adjust to their preference.

Procedure. The presentation of the audio texts and the collection of responses were controlled by Dell desktop computers. Participants were given an instruction sheet that contained basic keyboard control instructions and a sample outline. The sample outline consisted of three levels of headings with bulleting used to format the three levels. Participants were told that they would be listening to two audio texts and their task was to detect headings in each text. They were told to listen for headings in the text and pause as quickly as possible once a heading was detected. In order to pause the recording, participants used the left arrow key on the keyboard. When the left arrow key was pressed, the audio recording paused and a textbox appeared on the video display. Participants were instructed to type a number to indicate the hierarchical level of the heading and then type the exact wording of the heading into the textbox. Once this information was entered, participants pressed the right arrow key to continue the recording.

After reading the instructions and having any questions answered, each participant listened to a sample text to make sure they understood the procedure. The sample text explained

what a heading was and reiterated that the participant's task was to listen for headings and to pause the recording as quickly as possible once a heading was detected. The sample text referred to the sample outline to explain how different hierarchical levels of headings were represented in the audio text. The sample text provided one example for each hierarchical level of heading and asked participants to practice responding by pausing, entering the correct information, and then resuming the audio recording.

Two different sample text recordings were created. Half of the participants were assigned to receive one experimental text in the Tones condition and one experimental text in the Tones + Labels condition. Because these two conditions both involved the use of tones, the sample text they were presented also used tones to communicate the hierarchical level of the headings. The other half of the participants were assigned to receive one text in the Control condition and one text in the Label condition. Because neither of these two conditions involved the use of tones, the control version of the sample text presented to these participants. After completing the sample text and having any final questions answered, participants were then presented the two experimental texts.

Design. The experiment had a mixed-factors design. There were two within-subject factors: (1) whether or not the headings of a text were preceded by a label, and (2) text versions (energy or fire). There were three between-subjects factors: (1) whether tones were used to signal each heading; (2) the order of presentation of the two text versions was counterbalanced; (3) the assignment of the two text versions to the label vs. no label condition was counterbalanced.

Results & Discussion

Preliminary analyses showed no effects of the counterbalancing variables or the text topic so these variables were excluded from the final analyses. In addition, a MANOVA demonstrated

that the performance across the set of dependent variables varied as a function of the Tones manipulation and as a joint function of the Tones and Label manipulations; $F(10, 50) = 4.59$ and $F(5, 55) = 5.64$, respectively. Therefore, separate ANOVAs were conducted on the five dependent variables. The design of each ANOVA included the within-subjects factor of whether the text included labels before each heading, and the between-subjects factor of whether the text included tones before each heading. For efficiency of presentation, the results of Experiment 1 are combined with the results of Experiment 2 in all figures (the numbers associated with the condition labels on the X-axis indicate the relevant experiment) because both experiments included a replication of the Control (no labels, no tones) and Label (label, no tones) conditions. The bars in the figures represent the upper-limits of the 95% confidence intervals on the condition means. In considering the results of Experiment 1, focus on the left-hand and middle pairs of bars in the graphs; ignore the right-hand pairs of bars. For all reported statistical tests, the adopted level of significance is .05.

First, we tabulated the number of headings that participants identified based on their written recordings of the headings. A heading was credited as correctly identified if most of the wording was identical to the text. The maximum possible score was 20. The key result from the ANOVA is the interaction of the presence/absence of Labels with the presence/absence of Tones; $F(1, 59) = 19.89$, $MSe = 3.67$, *partial* $\eta^2 = .252$. It can be seen in Figure 1 that performance in the Control condition (no labels, no tones) was lower than in the other three conditions; performance in the three conditions containing signaling did not differ and, in fact, was near ceiling.

Insert Figure 1 about here

Second, we calculated the latency between the end of a heading and how quickly a participant pressed the spacebar to indicate detection of the heading. Average latency to respond is presented as a function of condition in Figure 2. (Note that latencies less than zero are theoretically possible because participants could indicate detection before completion of a heading. However, most negative latencies probably reflect variance in the actual completion times of the recorded headings across replications of the procedure.) Again, the key result is the interaction of the presence/absence of Labels with the presence/absence of Tones: $F(1, 59) = 20.74$, MSe = 1.13, *partial* $\eta^2 = .260$. Responses in the Control condition were slower than in the other conditions; responses were virtually immediate in the other three conditions and times did not differ across those conditions.

Insert Figure 2 about here

Third, we tabulated the number of times a participant stopped the recording to indicate a heading when, in fact, no heading had been presented. These results are presented in Figure 3. We again found an interaction; $F(1, 59) = 9.12$, MSe = 8.07, *partial* $\eta^2 = .134$. The nature of the interaction is a bit different from the interactions reported above. It is again the case that performance in the Control condition was considerably poorer than in the other three conditions. However, there were differences among the other three conditions. Specifically, there were more false alarms in the condition where only labels preceded headings than in the condition where only tones preceded headings; $t(59) = 2.18$, or the condition in which both tones and labels preceded headings; $t(59) = 2.44$. In other words, the use of tones to signal headings is more

effective than the use of labels with respect to preventing listeners from making false alarms. Note, however, that participants who were presented only labels still made very few false alarms (i.e., *M* = .79, *SE* = .31 for a text with 20 headings).

Insert Figure 3 about here

Fourth, we looked at the accuracy of participants' wording of the headings. If there was a label preceding the headings, participants made fewer errors in reproducing the heading verbatim (*M* = .04 errors, *SE* = .010) than if there was no label (*M* = .08, *SE* = .013); $F(1, 59) = 6.86$, *MSe* = .006, *partial* $\eta^2 = .104$. There were no other reliable effects for this measure.

Finally, we looked at the accuracy of participants' reports of the hierarchical levels of the headings. Figure 4 illustrates the interaction between the presence/absence of Labels and the presence/absence of Tones; $F(1, 59) = 5.02$, *MSe* = .016, *partial* $\eta^2 = .078$. The interaction reflects the fact that performance was equally good when headings were preceded by labels (*M* = .91, *SE* = .019) or by a combination of labels and tones (*M* = .94, *SE* = .018), and both conditions produced better performance than tones alone (*M* = .79, *SE* = .032), which produced better performance than having neither labels nor tones (*M* = .66, *SE* = .033). That is, listeners gain some information about hierarchical level from the tones but they do much better when a label explicitly states the level.

Insert Figure 4 about here

Integrating the results across all measures of performance, listeners perform relatively poorly on all measures if neither labels nor tones signal an upcoming heading and communicate its hierarchical level. However, if labels or tones or both precede the headings in a text, listeners do quite well on all measures of performance. Nevertheless, there are some differences between the two types of signaling devices: First, listeners are somewhat less likely to falsely recognize a heading if tones are used to signal headings than if only labels are used to signal headings. We attribute this slight advantage of tones to their attention-alerting properties. Second, listeners do much better at processing hierarchical level information from labels than from tones. Evidently, participants had some difficulty identifying tones and interpreting them with respect to the hierarchical level they denoted. The combination of labels and tones to signal headings resulted in performance at least as good – and often better – than the other three conditions on all five measures of performance.

Experiment 2

Experiment 2 compared the relative effectiveness of labels to the relative effectiveness of changes in voice to signal headings and their hierarchical levels. Using different voices to distinguish headings from content and to communicate headings at different hierarchical levels has the attractive feature of not lengthening the text in any way. Rather than preceding headings with a label or tone, the heading label is simply spoken in a distinctive voice. Thus, this approach to rendering headings is relatively seamless.

Our manipulation of voice changes was analogous to the manipulation of tones in Experiment 1. Just as three different frequencies were used to communicate the hierarchical levels of the headings in Experiment 1, three different voices were used to communicate the hierarchical levels of the headings in Experiment 2. Further, both approaches distinguished

headings from text content in a way that should be perceptually salient (i.e., did not require processing of the meaning of the headings and content). Given the parallels between the two approaches to rendering the headings, we anticipated that the manipulation of voice changes would have advantages and disadvantages parallel to those observed for tones. Specifically, the attention-alerting properties of voice changes to indicate headings were expected to result in very good and fast identification of headings with very few false alarms compared to the use of labels to signal headings. However, we expected that labels would produce better performance with respect to identifying hierarchical levels because the correspondence between a label and the hierarchical level it signified was explicit whereas the correspondence between a voice change and the hierarchical level it signified was less familiar. Recognizing this, we made one change in the procedure of Experiment 2 relative to the procedure of Experiment 1. Namely, we provided participants who received texts containing voice changes with practice trials to learn the correspondence between a particular voice and the hierarchical level it signified.

Method

Participants. A total of 72 volunteers from the Psychology subject pool participated in the experiment in order to satisfy a research participation requirement. Participants' ages ranged between 18 and 24. Data from eight participants were dropped because the participants failed to follow instructions. The types of participant errors included consistently recording only heading information without hierarchical information, or consistently including only hierarchical level information.

Materials. The texts used in Experiment 2 were identical to those in Experiment 1. New base recordings were made using a Kindle Fire Tablet (2nd generation). The new base recordings

had an average pitch frequency of 180 Hz. Other than the pitch frequency, the new recordings sounded similar to the recordings in Experiment 1.

There were four text versions in Experiment 2. The Control and Label versions of the texts were the same as the corresponding versions of Experiment 1. Two new text versions were created to implement the new factor in Experiment 2; namely, the use of a change of voice instead of tones to communicate headings and their hierarchical levels.

In the "Voice" version of an audio text, three different combinations of pitch, tempo and amplification of the audio headings were used to create three voices that were distinct from each other and from the voice that spoke the main content of the text. The three voices thus distinguished headings from content and communicated the hierarchical levels of the headings. The variations in pitch, tempo and amplification were created using Audacity. For the highest level heading (Level 1), relative to the voice for the main content the pitch of the voice was decreased by 6 semi-tones, the tempo was decreased by 10%, and amplification was set to +10db. This voice was lower, slower and louder than any of the other voices. For the mid-level heading (Level 2), relative to the voice for the main content the pitch was decreased by 4 semitones, tempo was decreased by 5%, and the amplification was set to +5db. This voice was lower, slower and louder than all of the voices except the voice associated with the highest level heading. For the lowest level heading (Level 3), relative to the voice for the main content the pitch was decreased by 2 semi-tones, the tempo was decreased by 2%, and the amplification was set to +2db. This voice was lower, slower and louder than only the voice used to speak the text content. Subjectively, the voice communicating the highest level heading was a deep masculine voice; the voice communicating mid-level headings was more androgynous but clearly a masculine voice; the voice communicating low-level headings was distinctly feminine.

The "Voice + Labels" condition of Experiment 2 combined the Voice and Labels conditions. This version of the audio texts simply applied the voice manipulations of the Voice condition to the Label versions of the text so that each heading was preceded by a label, and both the label and heading were spoken in the voice appropriate to the hierarchical level of the heading.

Procedure. The procedure for Experiment 2 was the same as Experiment 1 with one important change. As in Experiment 1, participants were presented either the Control and Label conditions or the Voice and Both conditions. All participants received instructions in the task and were presented the same sample text as Experiment 1 for practice. Participants in the Control and Label conditions were then presented the two experimental texts. Participants in the Voice and Both conditions, however, did not proceed immediately to the experimental texts. Instead, they received practice in associating the voice changes with the hierarchical levels they were intended to communicate. Their training consisted of being presented the phrase "What is the level of this heading?" in one of the three voices used to communicate hierarchical level in the Voice condition. Presentation of the question was followed by a 90 second pause during which the participants were to stop the recording and type 1, 2 or 3 to indicate their response. The computer then revealed the correct response as feedback and the program continued on to repeat the question in a new voice. A total of 20 training questions was included in the training phase. The choice of 20 training trials was based on pilot data indicating that participants' accuracy in correctly detecting heading levels peaked between 16 and 25 training questions at 93.4% accuracy. After this training phase of the procedure was completed, participants in the Voice and Both conditions were presented the two experimental texts.

Results & Discussion

The design of Experiment 2 was analogous to that of Experiment 1 so the analyses of the data were analogous to the analyses in Experiment 1. As in Experiment 1, a MANOVA demonstrated that the performance across the set of dependent variables varied as a function of the Tones manipulation and as a joint function of the Tones and Label manipulation; $F(10, 53) = 9.22$ and $F(5, 58) = 9.01$, respectively. Therefore, separate ANOVAs were conducted on the five dependent variables. The results for Experiment 2 are summarized in the second and fourth pairs of bars in each of the figures presented earlier.

We first tabulated the number of headings that participants identified based on their written recordings of the headings. Overall, headings were more consistently identified if they were accompanied by a change of voice than if they were not; $F(1, 62) = 46.44$, $MSe = 9.38$, *partial* $\eta^2 = .428$. The effect of a voice change was much less if a label preceded a heading than if there was no label; $F(1, 62) = 39.29$, $MSe = 4.49$, *partial* $\eta^2 = .388$. This interaction is illustrated in Figure 1, which shows that performance in the control condition (no labels, no voice) was considerably lower than in the other three conditions. However, performance in the label condition was somewhat lower in Experiment 2 ($M = 18.30$, $SE = .34$) than in Experiment 1 ($M = 19.41$, $SE = .21$) with the consequence that the difference between the label condition and the label, voice condition was reliable in Experiment 2; $t(62) = -2.79$, $SE = 0.48$. Nevertheless, performance in the label condition was very good and much better than performance in the no label, no voice control condition; $t(32) = 7.97$, $SE = .65$. In short, signaling of any sort greatly facilitated identification of headings but a change of voice was somewhat more effective than a label in aiding identification performance.

As shown in Figure 2, average latency to respond to headings was much faster if headings were accompanied by a change of voice than if there was no change of voice; $F(1,62) =$

23.77, MSe = 1.61, *partial* η^2 = .277. However, the benefit of a change of voice was smaller if headings were preceded by labels than if there were no labels: F(1, 62) = 13.43, MSe = 1.61, *partial* η^2 = .178 for the interaction.

Figure 3 shows that listeners were less likely to incorrectly indicate the presence of a heading (i.e., false alarm) when there was a change of voice than when there was no change of voice; F(1, 62 = 14.07, MSe = 19.34, *partial* η^2 = .185. There was a tendency for this effect to be smaller if labels preceded headings than if there were no labels; F(1, 62) = 3.74, MSe = 21.26, *partial* η^2 = .057, *p* = .058.

Unlike Experiment 1, there were no reliable effects with respect to errors in reproducing headings verbatim.

Figure 4 illustrates that listeners were much better at reproducing the hierarchical levels of the headings if labels preceded the headings than if there were no labels; F(1, 62) = 95.67, MSe = .026, *partial* η^2 = .607. The presence of a significant interaction indicates that there was some benefit of a change of voice in the absence of a label; F(1, 62) = 5.55, MSe = .026, *partial* η^2 = .082. Performance was better in the no label, voice change condition (M = .78, SE = .04) than in the no label, no voice change condition (M = .65, SE = .04).

Finally, we note the similarity of the results for the conditions involving a voice change to the results for the conditions involving tones in Experiment 1 (i.e., compare the last two pairs of bars in Figures 1-4). We conducted ANOVAs on each of the five dependent measures to compare the results for the two experiments. There were differences in the mean levels of performance on most dependent measures across the two experiments. However, the levels of performance in conditions involving tones or voice changes were very similar relative to the other experimental conditions (i.e., control and labels conditions). This is supported by the

observation that the Label x Tones/Voice interaction did not vary across the two experiments for any of the five dependent measures; $F(1, 121) = 2.44$, $p = .121$ for the 3-way interaction test on the number of correctly-identified headings and the other four F-tests were all < 1.

The summary of the results of Experiment 2 is very similar to the summary of the results of Experiment 1: Listeners perform relatively poorly on all measures if neither labels nor a change of voice are used to signal headings and their hierarchical levels. If either a label or a voice change signals a heading, listeners are very good at detecting the heading, they are quick to do so, and they make few false alarms. A change of voice is somewhat more effective than a label at supporting identification of headings, eliciting a quick response and preventing false alarms. These effects are reliable but small. A label is considerably more effective than a change of voice in communicating information about the hierarchical level of a heading. Analogous to Experiment 1, the combination of labels and voice changes to signal headings results in performance as good as – or better than – the other three conditions on all measures of performance.

General Discussion

The results of our experiments are quite clear. First, headings are often poorly communicated when translated from print to speech. Sometimes this is because the TTS application does not attempt to translate formatting information and sometimes it is because the coding of the printed text fails to retain such information. In the task developed for this study, listeners were able to focus all of their attention on the headings of the texts. Nevertheless, participants performed relatively poorly on the control texts: They failed to identify several headings, they were slow to respond to headings, they made many false alarms, and they were particularly poor at identifying the hierarchical levels of headings. This outcome is consistent

with previous findings in the literature (Lorch et al., 2012) and it reinforces the point that it is important that the information carried by the formatting of headings be communicated in the auditory text. Given that listeners in the current study were to stop the recording as soon as they detected a heading, the slow responses of the participants and the relatively high rate of false alarms are attributable to difficulties in discriminating headings from other text content; that is, the poor performance is due to problems perceiving the headings rather than to memory problems.

Second and fortunately, it is easy to correct the problems listeners have in processing headings as rendered by common TTS applications. All of the signaling conditions investigated in our experiments resulted in substantial improvements in the processing of headings. Any form of signaling of headings resulted in rapid and near-perfect identification of headings with good discrimination of headings from other text content. Simply inserting a tone or label before each heading or using different voices for headings vs. text content was sufficient to produce great improvements in the processing of headings.

Finally, there are a couple of important differences in how the alternative signaling methods influence the processing of headings. Tones and voice changes share the characteristic that they allow discrimination of headings from text content without requiring the processing of meaning. This characteristic gives tones and voice changes a slight advantage over labels in helping listeners distinguish headings from other text content: There are fewer false alarms if headings are signaled by tones or voice changes than if headings are signaled only by labels. On the other hand, labels have a clear advantage over tones and voice changes when it comes to communicating a heading's hierarchical level. This is not surprising because labels explicitly state the level of a heading. In contrast, when presented tones or voice changes, listeners must

identify the particular tone or voice and remember the association between that information and the hierarchical level. In other contexts, it is possible to use a more ecological approach where the selection of the specific tones is not arbitrary but relies on familiarity and everyday usage of sounds (Tuuri, Eerola, & Pirhonen, 2011; see also the distinction between auditory icons and earcons made by Sodnik, Jakus, & Tomažiĉ, 2011). However, it is not immediately apparent whether such an approach can be adapted for communicating headings or hierarchical information.

No single method of signaling headings resulted in perfect performance, but a combination of devices did. Combining labels with either tones or voices produced near perfect performance on all measures. The combination of labels with tones is a particularly attractive method of augmenting current TTS applications because it should be very straightforward to modify current software to introduce labels and tones. Programming languages (e.g., HTML, XML) already exist that code headings and distinctions in their hierarchical levels (Fourli-Kartsouni, Slavakis, Kouroupetroglou & Theodoridis, 2007). Adding a tone before each heading and a label with a regular, simple syntax would be easy to automate. However, there is still a question to be resolved with respect to this approach. The label should explicitly state the hierarchical level of the heading given our finding that this is the most effective way to communicate such information. However, should the tone's frequency be varied to redundantly code hierarchical level? Or should a single tone frequency be used before every heading? We suspect that a single frequency might be most effective at alerting listeners to an upcoming heading whereas variation in tone frequency is unnecessary. However, this is an empirical question at this point.

As clear as the findings are for Experiments 1 and 2, there are at least two important limitations of the studies. One limitation has implications for the conclusion that labels are a more effective way to communicate hierarchical level than tones or voice changes. Although that was the finding under the conditions of Experiments 1 and 2, there may be other circumstances in which tones or voice changes are much more effective. Although we gave participants practice with voice changes to familiarize them with the voices and their associations to specific hierarchical levels, extensive practice over time would surely have resulted in listeners becoming faster and more accurate in associating voices with hierarchical levels. Also, although we devoted a good deal of attention to creating tones and voice changes that were maximally discriminable, perhaps another set of tones and voice changes could be created that would be more discriminable. This consideration is important because difficulty in discriminating signals would translate to lower accuracy in identifying hierarchical level. However, even if more discriminable stimuli could be constructed for three-level hierarchies, the challenge is greater for texts with four- or five-level hierarchies. Thus, the use of tones or voice changes to communicate hierarchical level has an inherent limitation that labels do not.

From a broader perspective, our research so far has only examined tasks that emphasize the processing of headings to the virtual exclusion of processing of text content. Most common text-processing tasks place much greater emphasis on the processing of text content and often no explicit emphasis on the processing of headings. The literature on processing of printed texts suggests that headings influence text-processing in a variety of situations, including searching text (Klusewitz & Lorch, 2000), recalling text (Krug, George, Hannon & Glover, 1989; Lorch & Lorch, 1995, 1996a, 1996b; Lorch, Lorch & Inman, 1993), summarizing text (Hyona, Lorch & Kaakinen, 2002; Lorch & Lorch, 1996a; Lorch, Lorch, Ritchey, McGovern & Coleman, 200),

and learning from text (Mayer, Dyck & Cook, 1984). We might therefore expect that clearer communication of headings via TTS would have broad benefits for audio text-processing, as well. Nevertheless, headings are not as commonly used in auditory contexts (e.g., lectures) as other means of communicating topic changes (e.g., preview statements). Thus, an important question for future research is whether more effective communication of headings in a TTS environment has benefits for text-processing under conditions that do not specifically emphasize processing of headings.

References

Ferati, M., Pfaff, M., Mannheimer, S., & Bolchini, D. (2012). Audemes at work: Investigating features of non-speech sounds to maximize content recognition. *International Journal of Human-Computer Studies*, 70 (12), 936–966.

Fourli-Kartsouni, F., Slavakis, K., Kouroupetroglou, G., & Theodoridis, S. (2007). A Bayesian network approach to semantic labeling of text formatting in XML corpora of documents. *Lecture Notes in Computer Science, 4556*, 299-308.

Hyönä, J., Lorch, R.F., Jr., & Kaakinen, J. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology, 94*, 44-55.

Klusewitz, M.A., & Lorch, R.F., Jr. (2000). Effects of headings and familiarity with a text on strategies for searching a text. *Memory & Cognition, 28,* 667-676.

Krug, D., George, B., Hannon, S. A., & Glover., J. A. (1989). The effect of outlines and headings on readers' recall of text. *Contemporary Educational Psychology, 14*, 111-123.

Lemarié, J., Lorch, R.F., Jr, Eyrolle, H., & Virbel, J. (2008). SARA: A text-based and reader-based theory of signaling. *Educational Psychologist, 43*, 27-48.

Lemarié, J., Eyrolle, H., & Cellier, J-M. (2006). Visual signals in text comprehension: How to restore them when oralizing a text via speech synthesis. *Computers in Human Behavior, 22*, 1096-1115.

Lemarié, J., Lorch, R. F., Jr., & Péry-Woodley, M.-P. (2012). Understanding how headings influence text processing. *Discours, 10*.

Lorch, R.F., Jr. (1989). Text signaling devices and their effects on reading and memory processes. *Educational Psychology Review, 1*, 209-234.

Lorch, R. F., Jr., Chen, H-T., & Lemarié, J. (2012). Communicating headings and preview sentences in text and speech. *Journal of Experimental Psychology: Applied, 18,* 265-276.

Lorch, R. F., Jr., Lemarié, J., & Grant, R.A. (2011a). Signaling hierarchical and sequential organization in expository text. *Scientific Studies of Reading, 15,* 267-284.

Lorch, R. F., Jr., Lemarié, J., & Grant, R.A. (2011b). Three information functions of headings: A test of the SARA theory of signaling. *Discourse Processes, 48,* 139-160.

Lorch, R.F., Jr., & Lorch, E.P. (1995). Effects of organizational signals on text processing strategies. *Journal of Educational Psychology, 87*, 537-544.

Lorch, R.F., Jr., & Lorch, E.P. (1996a). Effects of headings on text recall and summarization. *Contemporary Educational Psychology, 21*, 261-278.

Lorch, R.F., Jr., & Lorch, E.P. (1996b). Effects of organizational signals on free recall of expository text. *Journal of Educational Psychology, 88*, 38-48.

Lorch, R. F., Jr., Lorch, E. P., & Inman, W. E. (1993). Effects of signaling topic structure on text recall. *Journal of Educational Psychology, 85*, 281-290.

Lorch, R.F., Jr., Lorch, E.P., Ritchey, K., McGovern, L., & Coleman, D. (2001). Effects of headings on text summarization. *Contemporary Educational Psychology, 26,* 171-191.

Mayer, R. E., Dyck, J. L., & Cook, L. K. (1984). Techniques that help readers build mental models from scientific text: Definitions pretraining and signaling. *Journal of Educational Psychology, 76*, 1089-1105.

Pascual, E. (1996). Integrating text formatting and text generation. In M. Adorni & M. Zock (Eds.), Trends in natural language generation: An artificial intelligence perspective (pp. 205-221). Berlin: Springer-Verlag.

Richardson, M. L. (2010). A text-to-speech converter for radiology journal articles. *Academic Radiology, 17,* 1570-1579.

Sodnik, J., Jakus, G., & Tomažič, S. (2011). Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users. *International Journal of Human-Computer Studies, 69(1-2), 100-112.*

Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G., & Argyropoulos, V. (2005). Experimentation on spoken format of tables in auditory user interfaces. *Proceedings of the 11th International Conference on Human-Computer Interaction (HCII '05)*, 361–370.

Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G., Argyropoulos, V., & Ikospentaki, K. (2010). Auditory universal accessibility of data tables using naturally derived prosody specification. *Universal Access in the Information Society, 9,* 169-183.

Tsonos, D., Xydas, G., & Kouroupetroglou, G. (2007). Auditory accessibility of metadata in books: A design for all approach. *Proceedings of the 4th International Conference on Universal Access in Human-Computer Interaction (HCII '07)*, vol. 4556 of *Lecture Notes in Computer Science (LNCS)*, 436-455.

Tuuri, K., Eerola, T. & Pirhonen, A. (2011). Design and Evaluation of Prosody Based Non-Speech Audio Feedback for Physical Training Application. *International Journal of Human-Computer Studies, 69*(11), 741–757.

List of Figures

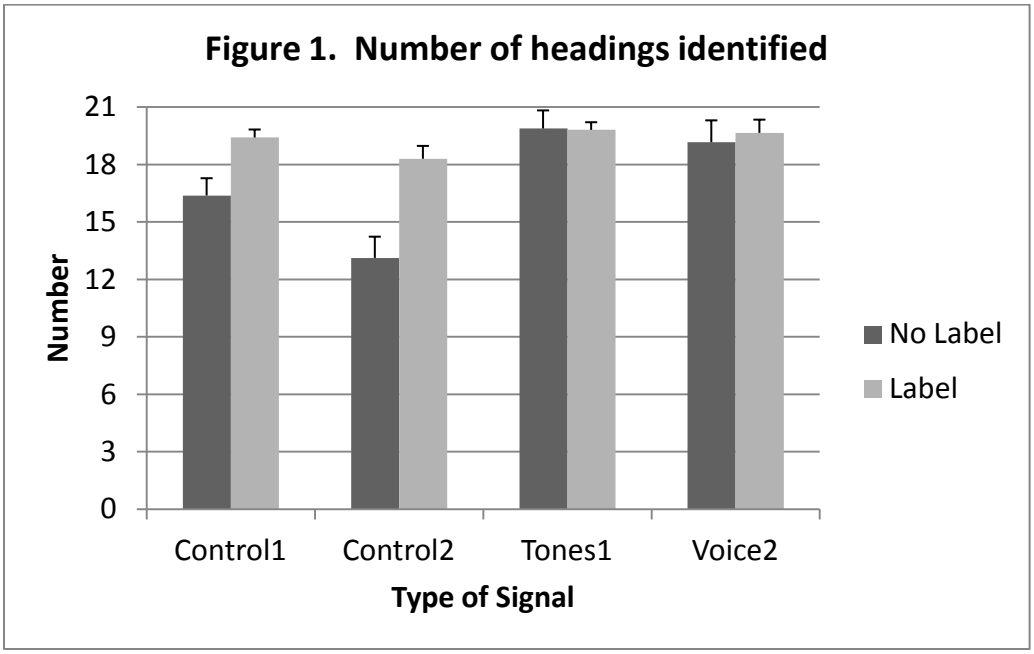
Figure 1. Number of headings identified
21
18
15
12
9
6
3
0
Number
Control1
Control2
Tones1
Voice2
Type of Signal
No Label
Label

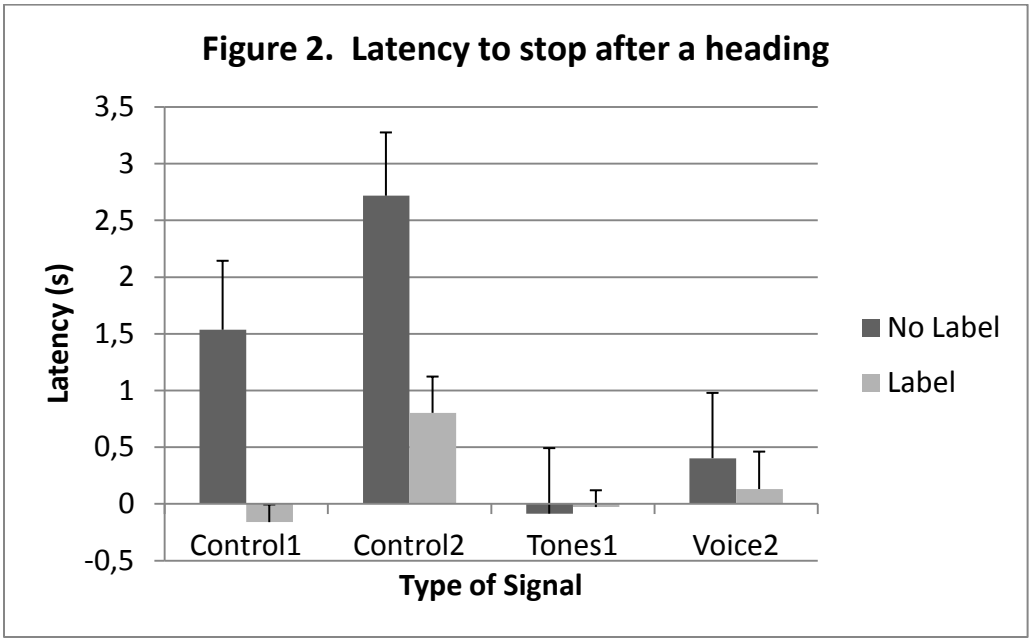
Figure 2. Latency to stop after a heading
3,5
3
2,5
2
1,5
1
0,5
0
-0,5
Latency (s)
Control1
Control2
Tones1
Voice2
Type of Signal
No Label
Label

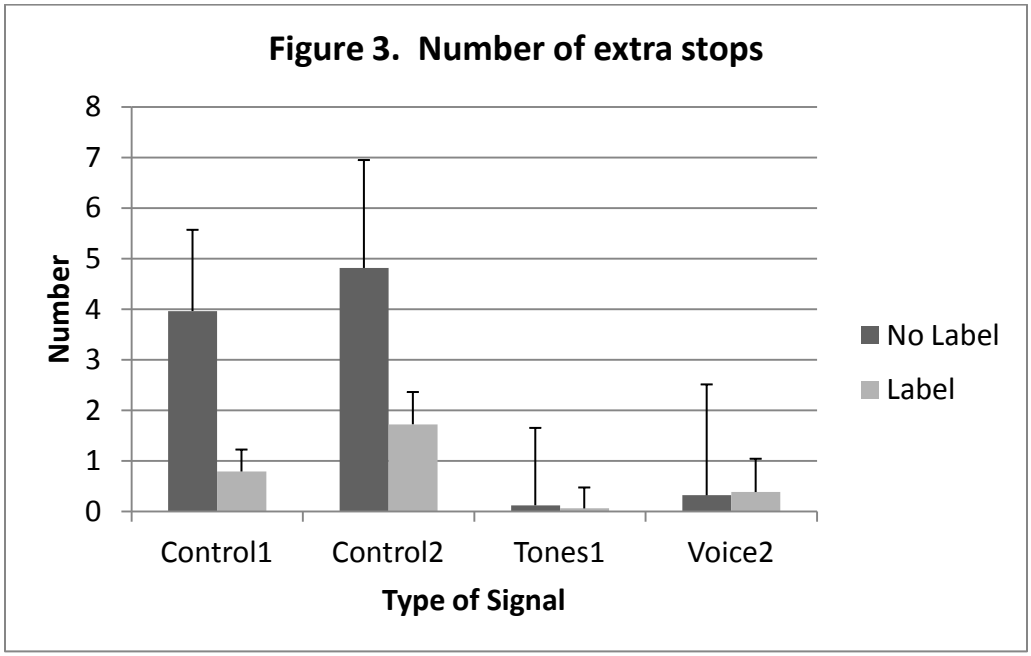
Figure 3. Number of extra stops
Number
8
7
6
5
4
3
2
1
0
Control1
Control2
Tones1
Voice2
Type of Signal
No Label
Label

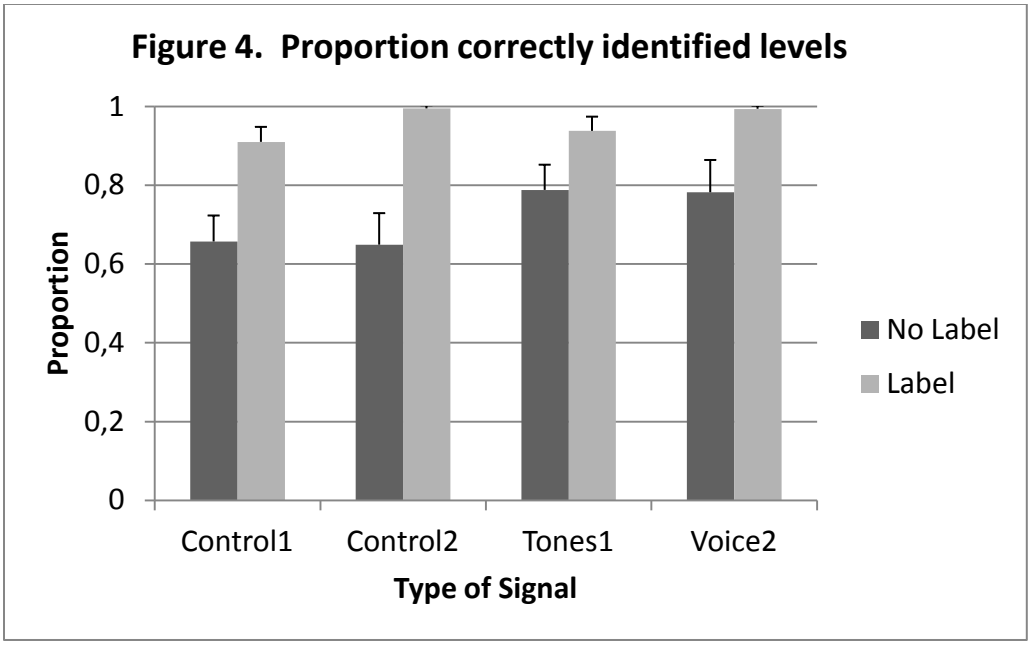
Figure 4. Proportion correctly identified levels
1
0,8
0,6
0,4
0,2
0
Proportion
Control1
Control2
Tones1
Voice2
Type of Signal
No Label
Label