Demmans Epp, C., Munteanu, C., Axtell, B., Ravinthiran, K., Aly, Y., & Mansimov, E. (2017). Finger Tracking: Facilitating Non-commercial Content Production for Mobile e-Reading Applications. In Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 34:1–34:15). Vienna: ACM. https://doi.org/10.1145/3098279.3098556

Finger Tracking: Facilitating Non-Commercial Content Production for Mobile E-Reading Applications

Carrie DEMMANS EPP
Learning Research and Development Center,
University of Pittsburgh
Pittsburgh, United States
cdemmans@pitt.edu

Cosmin Munteanu
Institute of Communication, Culture, Information and Technology, University of Toronto Mississauga
Mississauga, Canada
cosmin@taglab.ca

Benett Axtell
TAGlab, University of Toronto
Toronto, Canada
benett@taglab.ca

Keerthika Ravinthiran
Institute of Communication, Culture, Information and Technology, University of Toronto Mississauga
Mississauga, Canada
keerthika.ravinthiran@mail.utoronto.ca

Yomna Aly
TAGlab, University of Toronto
Toronto, Canada
yomna@taglab.ca

Elman Mansimov
Dept. of Computer Science, University of Toronto
Toronto, Canada
elman.mansimov@mail.utoronto.ca

ABSTRACT
Limited literacy and visual impairment reduce the ability of many to read on their own. Current e-reader solutions rely on either unnatural synthetic voices or professionally produced audio e-books. Neither provide the same enjoyment as having a family member read to a user, especially when the user requires assistive reading (following printed text while listening to it being read). Unfortunately, the support for non-commercial production of such e-books is limited and requires significant effort. We evaluate a novel, assistive mobile interaction technique that facilitates the recording of audio e-books and their synchronization with the read text. We show that a technique based on a finger tracking metaphor provides optimal support with respect to reading speed. These human-in-the-loop, adaptive techniques can now be used to reduce the content-creation burden that is associated with supporting those who cannot read on their own.

MobileHCI '17, September 04-07, 2017, Vienna, Austria

ACM ISBN 978-1-4503-5075-4/17/09 … $15.00
http://dx.doi.org/10.1145/3098279.3098556

Author Keywords
Reading; education; visual impairment; literacy; mobile e-readers; assistive technology.

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces; I.7.4 [Document and Text Processing]: Electronic Publishing

MOTIVATION AND BACKGROUND
The increased digitization of media within our information-centric society leads to some segments of our population being left on the wrong side of the digital divide. The move towards online-only access to information about topics, such as health or finances, introduces barriers that reduce the ease with which many users can access information [11,43,52]. The vast amount of text-based media also reduces the ability of certain population segments to access the cultural knowledge and education that are paramount to their economic success [15]. The ones who are being left behind include those speaking highly-local and under-represented languages or dialect [55], those with limited literacy, or those with vision impairments [15,34,42,43]. In the United States, this accounts for approximately 15 percent of the population [15,27]. Technology could allow these people to gain better access to the programs and information that result in improved literacy [61] or social and economic benefits [45].

Identifying appropriate techniques for enabling members of these marginalized groups to interact with text will support improvements in the accessibility of text-based media for a variety of users across different contexts. Some might argue

that audio books, e-readers, screen-reading software, or e-book reading applications meet the needs of these people. This is partially true, in that these tools can support a number of accessibility and literacy needs through their use of speech synthesis to verbalize text or recordings of someone reading text. However, as detailed later, the experiences provided by existing technologies and their basic accessibility features fail to fully address users' needs [30,38,43]. Furthermore, current solutions are rather expensive, often due to the substantial effort and technical requirements needed to produce accessible content [49] or the lack of infrastructure and tools on the part of volunteers who are willing to help produce accessible content [65]. As a result, audio-book and e-reader alternatives have been proposed. Some use customized hardware to read the text to low-vision adults [59]. Others allow people to record themselves reading a text for children (e.g., http://explore.hallmark.com/recordable-storybooks/) or adults [4,5] who cannot access the text for a variety of reasons, such as the many languages that do not have computational resources available (as in recent work on audio books in local Indian dialects [55]). These technologies demand considerable effort on the part of a loved one, or they rely on often-inadequate synthetic speech as their output [30,68]. Few resources are formatted so that they can be read in their entirety by current e-readers [2,6,43] and few tools provide support for those with low-literacy [15], with the computational underrepresentation of highly-local languages or dialects [55] further hindering support for many in developing regions, such as India. Moreover, users find listening to speech synthesis tiring [13] because of the increased cognitive effort [43,47] that is required to process the impersonal [68], unnatural [35,68] (even for recent state-of-the-art systems [58]), and sometimes unintelligible speech [26,35]. While automatic text-to-speech (TTS) systems that produce natural voices are available [7], these are carefully fine-tuned and thus expensive.

Recording someone reading a text could alleviate the cognitive demand that speech synthesis places on a listener. This has the benefit (for the listener) of experiencing a sense of "togetherness" that is afforded by the voice of a family member [54], which is not yet possible even with the most natural TTS systems. Unfortunately, human-produced recordings and support materials take considerable effort, time, and money to develop [72]. Furthermore, technical barriers prevent the full alignment between the text and the recordings that are meant to enable information access for those who cannot read on their own [21]. This alignment of text with speech has been shown to benefit marginalized users (such as low-literacy adults) when trying to read e-texts, especially when accompanied by a moving prompt that shows the synchronization of text and speech [41,42]. While considerable progress has been made with respect to aligning a transcript from a known text to its accompanying audio, several challenges remain. Technological limitations and widespread human behaviours [29] (e.g., the use of fillers, self-correction, or filled pauses [29,37]), contribute to the 19% or higher error rates that accompany the use of standard alignment procedures [21]. Such error rates are at best distracting. At worst, they are mis-educative because they muddy the mapping between sounds and characters [67].

The Socio-Technical Gap Motivating this Work

The ability of audio books and e-readers to improve access to texts for many categories of marginalized users creates a need for appropriate applications and materials that can support these users. Targeted marginalized users include older adults who enjoy reading together with loved ones and those who have low-literacy, are visually-impaired, or speak under-represented local dialects. However, the cost and technical requirements of producing audio-synchronized e-book content for such categories of users, in an accessible format, represent a significant barrier. Fully automated solutions do not always produce adequate or adoptable results [57]. Yet, producing such content manually is a costly proposition. Thus, a hybrid solution is needed to facilitate the non-commercial production of audio-synchronized e-texts.

Our Solution

We propose a hybrid approach that uses human interactions to augment natural language processing algorithms so that we can better enable the creation of technologies that support those who cannot read on their own. Facilitating these content-creation processes should make e-books more accessible and affordable. This is particularly relevant for groups such as low-vision readers, native speakers of highly-local dialects without audio books or synthetic speech in their language, or older adults with limited literacy (either native or due to immigration) to enjoy following the text of a book while listening to the familiar voice of a loved one.

Building these technologies using appropriate interactions will enable those who support such marginalized groups to more easily create quality resources that can be reused across contexts [19,57], thus, saving time, effort, and money. This paper introduces several such interaction techniques that facilitate the production of accessible e-books. Our goal is to support the ability of users to produce audio content in non-commercial settings, with the purpose of enhancing the accessibility of e-books. These techniques were evaluated for their ability to enable users to support the reading activities of others by preparing appropriate audio scaffolding.

To support this content creation, a human-in-the-loop approach is used because it can aid in the development and use of adaptive systems [40] and other technologies [57]. We evaluate the inclusion of humans in the reading loop from a novel perspective: that of content creator. Since much of the work in reading support has focused on individual reader experiences, our work instead focuses on the experience of the person who is part of the adaptive support system. This first attempt at identifying the interaction techniques that are appropriate to supporting the human contribution to this socio-technical process provides a foundation for further explorations of how to effectively combine human-in-the-loop approaches with automated techniques, such as speech

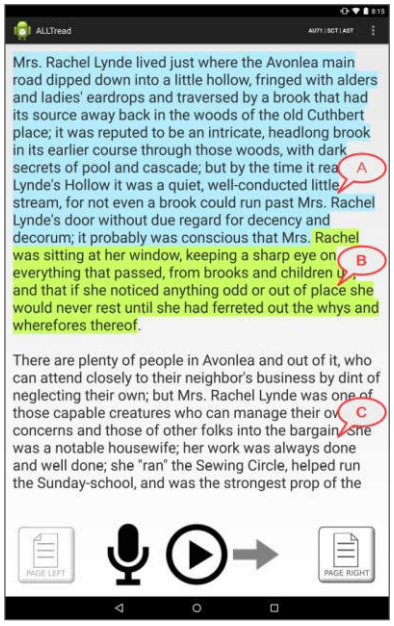

Figure 1. The assistive reader's recording interface. Blue highlighting indicates text that has been recorded (A). Green highlighting (B) indicates the sentence that is being read, and text that is not highlighted (C) has yet to be recorded.

processing. Moreover, combining these modalities could increase the accuracy of this class of technologies [40,48].

Building on the design proposed by Attarwala et al. [5], we implemented metaphor-based interaction techniques (one is illustrated in Figure 1) that support the simultaneous reading, recording, and synchronization of audio with text. These techniques are both adaptable and adaptive. They exploit how readers follow text with their fingers to improve synchronization. This interaction technique exploration helped us overcome the technical and design challenge of recording and synchronizing the audio with the read text.

We present the results of a controlled experiment (n = 22) demonstrating that this class of assistive techniques effectively support the combined reading and recording activities of users. This study showed that the adaptive versions of the finger-tracking metaphor enable faster recording. It further demonstrated that moderately-long text segments (i.e., sentences rather than words or pages) support the needs of the content creator: they are short enough to provide additional alignment information to the speech processing algorithms, result in a better user experience, and minimize the amount of time users invest in content creation.

This study demonstrates how user interactions can be designed to accommodate natural language processing techniques that facilitate the content creation efforts of those who support marginalized populations. It provides early work and design guidance with respect to how human-computer interaction techniques can be adjusted to account for the imperfect nature of algorithms from complementary areas of computer science. This augmentation is achieved while maintaining a natural and comfortable interaction experience for the user. Moreover, mobile interfaces have yet to support this much-needed synchronization while enabling people to aid those who cannot read on their own.

Statement of Contribution

The above technical and interface challenges are addressed through the development of a mobile interaction technique that will facilitate the production of audio books for marginalized users. This hybrid solution gathers interaction data from human users as they are reading aloud a text to someone else. Our solution is implemented as a reading support interface that tracks the readers' finger position as they follow the text. This does not require additional effort on the reader's part, and it has the added benefit of collecting reading pace and tracking data to improve the synchronization of audio with text.

PRIOR APPROACHES TO SUPPORTING READING

Many people struggle with accessing text-based media. Among them are 1) those with low-literacy levels (young [67] and old [15,18,42]) who lack the knowledge of how sounds map to characters [17,50,67] and who may have learning disabilities; 2) those who have yet to learn the language of the text [15,20,50]; and 3) those with low-vision who know the mappings between characters and sounds but who have trouble seeing the text [5]. Regardless of the reason behind their inability to access text on their own, these users need assistance. This assistance is received through the efforts of loved ones [5], teachers, or peers [20]; technologies [5,15,23,25]; and education or community programs [15,53]. However these supports are rarely enough [15,16] and their failures are leading to socio-technical approaches that hold the potential to support readers' multiple needs. In this section, we are surveying such recent approaches that have informed our own work of developing more suitable interaction metaphors to support accessible reading.

Peer Support for Individual Reading

Most solutions that support those struggling with reading focus on second language learning or bringing literary materials to those without access instead of addressing the user's ability to process text [6,12,33]. Notable exceptions considered approaches to supporting reading comprehension through word substitution [66], annotation for fluent readers [32], and the reading fluency of children [10].

Even with the potential for some of these technologies to support reading, they often do not leverage other sources of support (such as family members, peers, or teachers) that are available in the reader's environment [4,5]. Instead, assistive technologies focus on providing direct support to users (e.g., [31,59]) and ignore the larger context of use or the situations they cannot support, which may explain why these technologies are often not widely adopted [60]. Another reason for their lack of adoption may be the limited use of socio-collaborative approaches where people support one another's access to and comprehension of a text directly [5,8,20] or indirectly by developing and sharing support

materials. We argue for this type of socio-collaborative approach to developing reading support resources.

Adaptive Support for Reading

Some approaches to enabling reading through mobile devices have included adaptive support, which can be implemented as a human-in-the-loop approach [5,39]. This approach combines human efforts with those of the system to provide the individualized support readers need. This approach can help because the system learns from the people who interact with it by deferring part of the decision process to those users [40,51]. This approach to improving adaptive systems has previously supported underrepresented users [40,56]. We, therefore, argue that introducing a human can help overcome barriers to the types of support that are needed by those who cannot access text. In particular, we propose that the human user takes the role of content producer.

Natural Language Processing and Reading Support

Advances in text-to-speech may enable people to access text because speech synthesis verbalizes the text both for those who live with impaired vision [59] and those who struggle with achieving literacy [17,22,24,25]. However, natural language processing techniques are limited by regional pronunciation variations [21], the intelligibility of produced speech [35], the opaque orthography of English [50], and pronunciation errors [69]. In cases such as highly-local dialects for which no computational models or resources exist, language processing is not possible at all. We, therefore, propose the use of recorded audio where possible.

Like Anguera et al [2], we argue that improving alignment accuracy and speech recognition will enable the effective synchronization of audio to texts, enhancing user access. However, attaining this goal remains an open challenge because speech recognition must account for much slower speech rates that aim to support learning [3], phonemic and prosodic errors [69], mispronunciations [69], low-frequency words that increase error rates [29], and off-script utterances [9,37], such as re-reading [40]. These challenges and work demonstrating the limitations of forced alignment [36,40,71], such as the inability to detect miscues, indicate the integration of complementary approaches or techniques would benefit both the algorithms and users.

A MOBILE READING SUPPORT APPLICATION

For many users, such as those affected by vision loss, having someone read to them is the only way they can access the content of books or other text-based materials. Recent applications, such as ALLT [5], allow the recipient to enjoy listening to an audio book in a familiar voice and liberate users from the constraints of sharing a common location and schedule when reading jointly. Some interactive books and ALLT [5] allow readers to record their voice and synchronize their reading to the text by using record and stop buttons. These buttons allow the user to play the recording back at a later time so that she or he can enjoy a reading experience that visually links the audio recording to the text using highlighting, much like that used in karaoke. However, the burden of recording a novel or lengthy informative text using the current method is prohibitive to widespread adoption. Better mechanisms for enabling a reader to record the audio of a text that is synchronized to the words on the page are needed. Accurate synchronization is useful not only for a human reader but also for algorithms that perform various tasks that enhance user experience. For example, topic detection [64] could be used to separate the recording of the targeted text from the side discussions, explanations, or reminiscences that occur during joint reading activities.

Users should be provided with a seamless audio-recording experience where the text is automatically synchronized to the recording. Initial attempts at fully-automated approaches to synchronizing or aligning recordings with the text, using speech recognition, fell short of user expectations; this is a known problem due to the computational complexity and error prone nature of forced alignment [1]. These challenges can produce inconsistent artifacts when users have accents, background noise is present, or when mobile devices that are not connected to the Internet (for server-side speech recognition) do not have the computing power to handle forced alignment on-device. These limitations have prevented the inclusion of speech recognition techniques within reading-support applications.

To overcome such limitations, we propose a human-in-the-loop approach because others have shown that using complementary modalities, such as gestures, can lead to significant increases in the accuracy of the underlying natural language processing techniques [40,48]. The use of human input to enable lightly supervised approaches has been shown to improve alignment [14]. As such, the metaphor we evaluate serves as an example of an interaction that can integrate complementary automatic processing techniques while maintaining a natural experience for users. Our proposed metaphor can help improve automatic alignment techniques by incorporating the reading pace and reading position data collected from user interactions.

Proposed Metaphor Implementation

The reading support metaphor we implemented within the ALLT e-book reader relies on an implicit finger gesture to collect audio/text alignment data (Figure 2). This finger-tracking metaphor exploits the mobile device's touch screen. Finger tracking was selected because it is an established technique for teaching and supporting reading [20,44]. In finger tracking, the reader traces the words or sentences that are being read with a finger. This technique is known to help with several aspects of reading, including the reader's ability to turn characters into sounds and mark his or her place. Similarly to how readers trace their finger across the page of a book, users trace the text that is displayed in the mobile ALLT e-book reading app as they read. The position of the finger is marked by highlighting the word under which the finger is positioned. Already-read words are highlighted with a different colour. Different levels of granularity are possible for the finger tracking and highlighting (e.g. word, sentence),

which we explore in the next section. Our interaction metaphor was built as an extension of the ALLT mobile e-reader, which fully implements the DAISY standard for accessible reading [63]). We have not used DAISY or other synchronization tools in this study (neither have we explored TTS voices), as our focus was on how well the finger tracking metaphor supports the creation of content in a user's familiar voice. Our implementation of this novel finger-tracking metaphor allows users to continue following the text as they read even without using the finger. This is possible by learning a user's reading pace and automatically "following" the text, allowing users to rest their finger. Several variations of this automatic tracking are discussed and evaluated in the following sections.

Early implementation efforts showed that finger slipping or wandering could pose problems for readers who were using this approach with a touch-screen device rather than a paper book. We, therefore, implemented a heuristic that prevents accidental slips, by restricting touches so they are only registered in the immediate vicinity of the current sentence or word. The vicinity includes the same or adjacent rows and it centres just below the row that is currently in focus, unless the finger has been lifted from the screen. This definition was empirically determined during pre-piloting and allows users to adjust their recording speed. It also best avoids text occlusion. In other words, the area covered by the finger is below the row that is being read. This heuristic also considers the vicinity as a continuous space that follows the text. When the virtual cursor reaches the right margin of the page, the vicinity moves to the next row and repositions itself on the left margin. The vicinity does not extend across pages. In our early pre-piloting, we tested this but found it required the pages to change or turn automatically, which users found distracting and error prone. Furthermore, some users continued to touch the same spot on the screen after a page had turned, which provided false information to the system. As such, pages need to be turned manually through the explicit gesture of touching the "page left/right" buttons. Additional standard interface widgets support other explicit gestures, such as initiating audio recording and playback or accessing customization features, such as font size.

While relying on this metaphor supports the an interaction that has the potential to seem more natural to users, there are many possible variations that could benefit users of mobile e-readers, and none of them has been evaluated until now.

Use of the Metaphor during Audio Book Production

The ALLT e-book reader has been specifically designed for collocated reading, such as a grandchild reading to their grandparent, or a family member reading to a low-literacy relative. By introducing the finger-tracking metaphor, such "reading together" activities can be captured as audio recordings that are synchronized with the text. This results in a seamless production of audio books. These books can be "played back" at a later time by the target user (e.g., an older adult or low-literacy user). The synchronization of audio and text and the use of highlighting to visually indicate this can assist both during the recording phase but also during playback. The assistive recording afforded by our finger-tracking metaphor can also be useful in recording an audio book (still synchronized with the text) by a single user.

METHODS

A laboratory study was conducted following pilot testing. This study explored the user experience and performance implications of different text-recording methods.

Pilot Studies

Initial pilot studies were used to refine the study protocol, train experimenters, and refine the interaction techniques that relied on a combination of adaptive and adaptable controls. The interface controls that were selected as independent variables implement the finger tracking metaphor and either auto-advance the recording prompt using a timer (FST, FWT) or require the user to manually advance the recording (FSM, FWM), with another distinction being the granularity of the recording prompt (Sentence or Word – S/W). See Table 1 for a complete description of these variables.

Twelve computer science undergraduate and graduate students participated in the pilot studies, sharing similar demographic characteristics and reading abilities as those in the main study (described later).

After piloting, text lengths were reduced to prevent user fatigue. The amount of text a user had to read before the application automatically advanced recording was also adjusted. In FST, the user had to record 3 sentences because averaging the times for this many sentences seemed to produce reasonable behavior from the adaptive auto-advance feature. In FWT, a minimum of 5 words were needed to allow the application to auto-advance at a reliable pace.

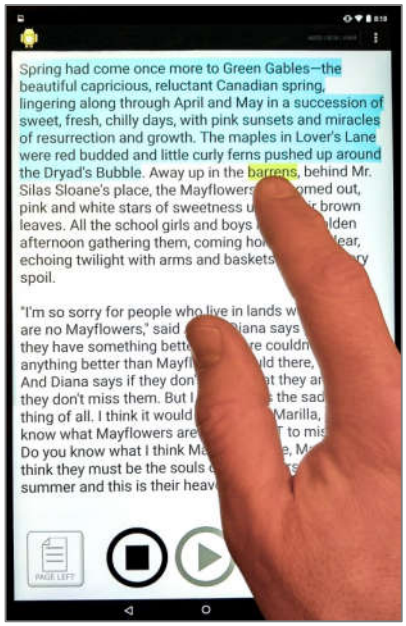

Figure 2. A version of the finger-tracking metaphor used to augment recording. The user is currently reading *barrens*.

Condition	Description
Baseline	This mode is used to record an entire page from start to finish at once. It is consistent with the current state of affairs, which provides no support to users
FSM	This mode employs the finger tracking metaphor (F) to advance recording. The user touches the sentence (S) that he or she wants to record and manually (M) progresses through sentences one at a time. Sentences are considered separated by the following types of punctuation: periods, exclamation marks, question marks, and semi-colons. Users can tap anywhere on a sentence.
FST	This mode employs the finger tracking metaphor (F) to advance recording. The user touches the sentence (S) that he or she wants to record and progresses through sentences one at a time. Users can tap anywhere on a sentence and the application determines the user's reading speed after he or she has recorded 3 sentences. Once users lift their finger or stop advancing the sentence, the application uses a timer (T) and the learned reading speed to advance the recording to the next sentence until users stop the recording or until they start touching the screen again, which results in the recalibration of the timer.
FWM	This mode is similar to FSM. The only difference is the recording unit: users trace individual words (W) with their finger (F) to manually (M) advance the recording from one word to the next. While not all users may require word-level synchronization, some with low-vision want it, and it does not harm other user populations, making it an appropriate target given its potential for increasing word-alignment accuracy.
FWT	This mode is similar to FST except that each word is recorded individually. Users follow the text with their finger (F) and the app shows them which word to read. Their finger needs to point right below the word that is being read. As in FST, they can lift their finger after tracing 3 sentences and the app will advance automatically. Users can touch the screen again if they feel the timer (T) is advancing either too quickly or too slowly. It takes about one sentence for this action to recalibrate the timer so that the user can once again lift his or her finger.

Table 1. The interaction techniques that were evaluated.

However, we set the minimum recording threshold for FWT to the maximum between 5 words and the total length (in words) of 3 sentences. This adjustment was made because piloting revealed that users were occasionally confused by the differences between these two threshold conditions, and it was important to ensure consistency from the user's perspective to avoid confounds.

Reading Materials

All readings were taken from L.M. Montgomery's Anne of Green Gables. Text segments with similar lengths (M = 403 words, SD = 20.64; M = 30.4 sentences, SD = 7.02) were selected from five chapters (Table 2). According to the Flesch-Kincaid measure of reading ease (where higher numbers are easier to read), the texts were fairly easy (M = 76.01, SD = 6.78). As with most e-readers, sentences wrapped lines – we did not modify or control the text display.

Hardware and Software

We have implemented several variations of the finger-tracking metaphor within the ALLT e-book "reading together" app [4,5]. The app runs on the Android operating system. For this study, we used Android Nexus 7 tablets (screen size 7 inches), running Android OS version 5.1.

Chapter Title	No. Words	No. Sentences	Reading Ease
Morning at Green Gables	384	24	74.81
Matthew Insists on Puffed Sleeves	430	22	68.21
Marilla Makes Up Her Mind	393	38	76.20
Anne's History	388	35	86.84
A Good Imagination Gone Wrong	420	33	74.00

Table 2. The texts that were read and their characteristics

Study Design

A laboratory-based usability study was conducted to explore which of the five conditions best supported the reader's experience of recording audio for consumption by another. These conditions are described in Table 1 and can be seen in the video that supplements this paper. They relied on a combination of explicit and implicit controls, some of which were adaptive. The presentation order of the conditions and texts were counter-balanced using a Graeco-Latin square [28]. This method superimposes Latin squares for each independent variable (i.e., text and condition). It helped ensure order effects for either independent variable did not influence the dependent variable measures (i.e., reading time and user experience).

This study tested four hypotheses, which are directional when the literature supports one:

H1: Auto-advancing the recording prompt enables users to record their reading of a text in less time.

H2: Auto-advancing the recording prompt influences user perception of effort.

H3: The length of the auto-advancing recording unit (i.e., word or sentence) influences user perception of effort.

H4: The length of the auto-advancing recording unit (i.e., word or sentence) influences the amount of time needed to record a text.

Testing H3 (user effort) and H4 (recording time) allows us to determine the appropriate recording granularity. That is, it answers the question of how much text should be recorded at once when using the proposed assistive features. Testing H1 (recording time) and H2 (user effort) indicates whether assistive features that are both adaptive and adaptable better

support the end goal, which is to enable the recording of texts using as little human time and effort as possible.

Instruments and Measures

A **demographics** form was used to collect information about participants' reading habits, language proficiency, and use of mobile devices. This data was collected at the end of the study. Two members of the research team also assessed participants' English fluency on a scale from 1 (native-like proficiency) to 5 (difficult to understand).

After each experimental task (i.e., reading), information about participant **experiences** of that condition was collected using a questionnaire. Participants rated various attributes on 5-point, semantic-differential scales, where 5 is the best or most positive response (e.g., very easy or much faster). These attributes represent dependent variables and include user perceptions of task difficulty, how tiring the task was, the speed with which they completed the task, and the ease of maintaining their recording pace. Participants were asked to compare the reading task to their regular leisure reading activities and to estimate the number of pages that they could read, without taking long breaks, for each condition.

The application logged information about participants' actions. This included how long, in microseconds, it took participants to read a text. This **recording time** information or dependent variable is represented through 3 measures: the *total time* spent reading a text, the average reading time per *sentence*, and the average per *word* reading time.

Both experimenters (authors 3 and 4) observed participant behaviors. Experimenters noted when participants adjusted the text highlighting because the auto-advance feature was moving too quickly or too slowly, whether they kept their finger on the screen when in an auto-advancing condition, and when their finger skipped a word while reading. This observation also included monitoring participant reading habits, how well the app suited their reading style, their physical posture and postural changes, their reading pace, any aspects of a condition that seemed to confuse them, and anything that participants may have found difficult.

Data Analysis

Descriptive statistics are provided for quantitative data. Inferential statistics are used to compare conditions for the relevant measures: time and perceived user experience. Non-parametric tests (i.e., Friedman ANOVA by Ranks and Wilcoxon Signed Rank) were used when one or more of the variables were not normally distributed. Parametric tests (i.e., two-tailed paired t-tests) were used to compare participant perceptions and performance across conditions when both variables were normally distributed. Bonferroni correction was applied to pair-wise tests to control for multiple comparisons. Due to the number of pair-wise comparisons performed, the significance threshold for the p-value is .005 rather than .05.

Ideally, the best condition (i.e., recording technique) will enable users to record the most text in the least amount of time while providing for the best possible experience.

Participants

Following institutional review board approval, participants were recruited through posters at a research-intensive North American university because students' backgrounds are consistent with those who would be creating content (e.g. young adults reading to their low-vision grandparents). A total of 31 people consented to participate and were compensated 30 dollars. Data from 9 people were discarded because participants failed to follow experimental protocols (e.g., started a conversation with the experimenter while in a condition), the application crashed, or an experimenter erred. To ensure high-quality data was available to support future research on better text alignment, we used strict inclusion criteria. Only the data from the remaining 22 participants is reported. They were 25.75 years old (SD = 6.79) on average.

Participants had varied language backgrounds and all of them spoke some English at home. On average, participants reported using English at home 80.46% of the time (SD = 18.38). English was used exclusively in 6 homes.

Participant home language was not tied to their reading fluency, which was rated as highly proficient (M = 1.48, SD = 0.63). The lowest fluency level observed was a 3 (good or acceptable) and only 2 participants received this score. Note that all participants either spoke English as a first language or had achieved sufficiently high English-language test scores to be admitted to an English-language institution.

Half (n = 11) of the participants reported reading at least one novel per month, and all but one reported reading at least one article per week (Mdn = 7, IQR = 13). Like the articles they read, the novels typically read by participants tended to be short (Mdn = 225 pages, IQR = 100). However, participants also read reasonably long books (Mdn = 550 pages, IQR = 325), and 14 reported that they had read aloud to others. Of those 14, 8 read books or stories to family members and young children; the others read articles or textbook segments to friends or religious texts to a group.

Participants reported a high level of comfort with using mobile technologies (M = 1.32, SD = 0.48), which is reflected in the number of hours (M = 3.91, SD = 2.67) that they spend using mobile devices in a typical day. This comfort is confirmed by their mobile reading habits: all participants had read at least one novel on a mobile device and they typically read multiple articles (Mdn = 4, IQR = 3) on their mobile devices each week.

We did not purposefully control for participants' reading habits or abilities as our app's target users are representative of a broad segment – younger adults who want to help family members access audio-enhanced e-books by recording the audio of such books.

Condition	Word Time			Sentence Time			Total Time		
	M	*SD*	*95% CI*	*M*	*SD*	*95% CI*	*M*	*SD*	*95% CI*
Baseline	0.34	0.07	[0.31, 0.37]	4.53	1.47	[3.88, 5.18]	136.44	28.68	[123.72, 149.16]
FSM	0.35	0.06	[0.33, 0.38]	4.89	1.51	[4.21, 5.56]	141.77	22.75	[131.69, 151.86]
FST	0.28	0.07	[0.25, 0.31]	3.78	0.94	[3.36, 4.20]	112.26	29.06	[99.37, 125.14]
FWM	0.32	0.09	[0.28, 0.36]	4.49	1.74	[3.72, 5.27]	128.18	33.69	[113.25, 143.12]
FWT	0.35	0.05	[0.32, 0.37]	5.03	1.53	[4.35, 5.71]	139.17	23.17	[128.90, 143.12]

Table 3. Participant reading times per condition (in seconds) for each unit of text: words, sentences, and the text as a whole (total).

RESULTS

Participants adjusted the experimental environment to meet their varied preferences for handling the tablet from which they were reading. Some chose to hold the tablet in their hands, others placed the tablet flat on the table, and some alternated between these options. In addition to participants demonstrating their comfort by modifying the experimental environment to suit their preferences, participants reported enjoying their overall reading experience. They also commented on how the highlighting of text helped them to keep track of their location within a reading.

Reading Time

Total reading time (Table 3) differed by condition, $\chi^2(4) = 15.13$, $p = .004$. Post-hoc comparisons showed that recording the text using finger tracking at the sentence level was faster ($t(21) = 4.10$, $p = .001$, $r = .67$) when the application automatically advanced the recording (H1) than when the participant had to manually advance recording using finger tracking (FST was faster than FSM). Finger tracking with adaptive support was also faster ($t(21) = -3.94$, $p = .001$, $r = .65$) when the recording unit was the sentence than when it was a word (H4). That is, FST was faster than FWT.

No significant effect of condition was found for the average sentence reading times, $\chi^2(4) = 8.69$, $p = .069$. This is likely due to the high variability of sentence lengths within each text influencing average reading time: some sentences were only a word long while others exceeded 19 words in length. We observed that some participants were distracted by single word sentences when the recording condition used sentence-level highlighting, especially when the text highlighted was a title, such as Mr. or Mrs. Participants also commented on this artifact of how sentences were defined.

Unlike the reading times observed at the sentence level, participant per word reading speeds (Table 3) differed by condition ($\chi^2(4) = 15.66$, $p = .004$). Post-hoc comparisons showed using the application to maintain the pace that was set by the user (FST) resulted in faster recording times ($t(21) = 4.02$, $p = .001$, $r = .66$) than having the participant manually advance the recording (FSM) when reading one sentence at a time (H1). This finding is consistent with the text-wide results. Similarly, FST was faster than FWT ($t(21) = -4.09$, $p = .001$, $r = .67$), which indicates that using sentences as the recording unit is the most expedient choice as is the choice to have the user set his or her recording pace and then have the application maintain that pace (H4). This finding may partly result from an observed user behavior: they tended to quicken their reading pace, without noticeable changes in their intonation, when the automatic highlighting advanced faster than their manually controlled pace.

As expected, participants felt their normal reading speed was similar to their recording speed when they were in the baseline condition (Table 4). Differences, $\chi^2(4) = 10.14$, $p = .038$, were found when participants were asked to compare their reading speed to their normal reading activities for each of the treatment conditions. Post-hoc analysis shows that participants only perceived a difference ($Z = -2.81$, $p = .005$, $r = .60$) between their baseline or normal reading speed and the FWM condition, with participants indicating their regular reading process was faster. This finding is consistent with experimenter observations that participants were pressing the screen hard and taking their time to ensure that each word was highlighted, thus reducing their reading speed.

User-Perceived Effort

Here we report on user perceptions of the amount of fatigue that they experienced in each condition, the ease with which

Condition	Reading Speed			Recording Pace			Difficulty			Tiredness		
	M	*SD*	*95% CI*	*M*	*SD*	*95% CI*	*M*	*SD*	*95% CI*	*M*	*SD*	*95% CI*
Baseline	2.77	0.43	[2.58, 2.96]	3.55	0.80	[3.19, 3.90]	3.00	0.69	[2.69, 3.31]	3.86	1.04	[3.40, 4.32]
FSM	2.50	0.51	[2.27, 2.73]	3.18	0.50	[2.96, 3.40]	2.73	0.77	[2.39, 3.07]	3.41	0.80	[3.06, 3.76]
FST	2.55	0.86	[2.17, 2.93]	2.32	0.78	[1.97, 2.66]	2.23	0.81	[1.87, 2.59]	2.77	0.97	[2.34, 3.20]
FWM	2.14	0.89	[1.74, 2.53]	3.00	1.07	[2.53, 3.47]	2.41	0.85	[2.03, 2.61]	2.64	1.00	[2.19, 3.08]
FWT	2.41	1.05	[1.94, 2.88]	2.09	0.68	[1.79, 2.39]	2.18	0.96	[1.76, 2.61]	2.32	0.89	[1.92, 2.71]

Table 4. Participant questionnaire responses to semantic differential items about their experience within each condition.

they could maintain the recording pace that was employed within each condition, and their general perceptions of the difficulty associated with each condition (Table 4). These measures are proxies for the perceived effort that is associated with each interaction technique. As such, they provide insight into users' experiences of those interactions.

When asked how the difficulty of an experimental task compared to their leisure reading activities, participant reports differed ($\chi^2(4) = 19.52$, $p = .001$) based on the recording condition. Post-hoc analysis revealed the FWT condition was more difficult than the baseline ($Z = -3.35$, $p = .001$, $r = .71$) as was the FST condition ($Z = -2.86$, $p = .004$, $r = .61$), indicating participants found the adaptive versions of the support more difficult than their normal reading activities (H2). This result is consistent with 16 of them tapping on a word or sentence to highlight what they wanted to read and adjust the recording pace: 2 adjusted the recording pace to go faster and 14 reported the auto-advance pace was too fast. Some participants were visibly confused when the automatic highlighting moved to the next sentence before they had completed recording the current one. Participant reports that the baseline was similar to their usual reading activities indicate the auto-advancing feature influenced the perceived difficulty level of recording (H2).

Like difficulty, a difference in participants' rating of how easy it was to maintain the recording pace for each condition ($\chi^2(4) = 45.63$, $p < .001$) was found. Participants found it more difficult to maintain their recording pace when using the auto-advancing conditions for recording by sentence (FST: $Z = -3.71$, $p < .001$, $r = .79$) or by word (FWT: $Z = -3.89$, $p < .001$, $r = .83$) than when they were using the non-adaptive baseline condition (H3). Differences were also noticed between the auto-advancing and manual versions (H2) when recording by word (FWT vs. FWM: $Z = -3.26$, $p = .001$, $r = .70$) and by sentence (FST vs. FSM: $Z = -3.38$, $p = .001$, $r = .72$), with the manual advancing option being the easier one to control. Additional differences ($Z = -3.87$, $p = .001$, $r = .83$) were seen in the ease with which participants could maintain the recording pace when manually advancing the recording sentence by sentence (FSM) and when the application advanced the recording word by word (FWT).

These differences between conditions, including the manual and adaptive methods for advancing to the next recording unit, indicate users feel they can more easily maintain a recording pace over which they exercise complete control (H2). It is worth noting that while many participants were observed adjusting their reading pace to match that of the timer, most touched the sentence or word they wanted to read. If the timer was advancing faster than they were reading, touching the text they wanted to read resulted in it being re-highlighted so they could record that text and have it synchronized with their recording. If they were reading faster than the system predicted, this action moved the highlighting forward so the system's pace would match that desired by participants. Participants then continued reading from the selected text segment.

The above differences in perceived difficulty and recording pace were further reflected in how tiring (H3) participants found each of the interaction techniques ($\chi^2(4) = 41.80$, $p < 0.001$). The baseline or regular reading condition was seen as less tiring than 3 of the 4 finger tracking conditions (FWT: $Z = -3.99$, $p < .001$, $r = .85$; FWM: $Z = -3.25$, $p = .001$, $r = .69$; and FST: $Z = -3.24$, $p = .001$, $r = .69$). FWT was also perceived to be more tiring than FSM ($Z = -3.75$, $p < .001$, $r = .80$), and there was no observed difference between manually advancing the recording one sentence at a time and recording an entire 2-page text in one sitting (i.e., baseline condition). The results of these tests indicate manual methods were less tiring than automated methods (H1) and recording a few large units of text is less tiring than recording multiple small text units (H3).

Differences in the amount of energy and cognitive effort participants invested when using each technique are supported by significant differences ($\chi^2(4) = 47.95$, $p < 0.001$) in the number of pages participants would be willing to record (Table 5). These users said they would be willing to record more pages using their regular reading technique (baseline) than in 3 of the proposed approaches (FST: $Z = -3.73$, $p < .001$, $r = .80$; FWT: $Z = -3.84$, $p < .001$, $r = .82$; and FWM: $Z = -3.70$, $p < .001$, $r = .79$). The sentence-level recording where users manually controlled the pace (FSM) was the only new technique that was not measurably different from the baseline ($Z = -3.35$, $p = .026$, $r = .71$): note the corrected p-value is more than 5 times the significance threshold for what would be a large effect. These users appear to prefer exercising control while recording at the sentence level (H3) as is evidenced by differences across the FST, FSM, and FWT conditions. These users said they were willing to read fewer pages when using FST than when using FSM ($Z = -3.23$, $p = .001$, $r = .69$), and they would read more pages ($Z = -3.35$, $p = .001$, $r = .71$) using the manually advancing sentence interface (FSM) over the automatically advancing word-level version (FWT).

INTERACTION POTENTIAL: IMPROVED ALIGNMENT

To demonstrate the secondary study goal, the forced audio-text alignment accuracy of each interaction technique was calculated. The word error rate (WER) is reported as a

	No. Pages		
Condition	***M***	***SD***	***95% CI***
Baseline	9.45	6.74	[6.47, 12.44]
FSM	7.73	5.87	[5.12, 10.33]
FST	5.09	4.67	[3.02, 7.16]
FWM	6.32	5.57	[3.85, 8.79]
FWT	5.41	4.68	[3.34, 7.48]

Table 5. The amount of text participants would feel comfortable recording using the evaluated approaches.

measure of this accuracy. WER factors in substitutions, deletions, and insertions between the original text and the text produced by passing the recorded audio through a speech recognizer and forced alignment tool. We used CMU Sphinx and its forced alignment tool (http://cmusphinx.org/) with the finger tracking timestamps included as input. The baseline WER was 100% (when recording the entire page without assistance), which is unsurprising as per prior work [1]. The largest improvement, of 34%, was observed for the FSM condition (WER = 66%). The other conditions showed a 12 – 17% improvement, with the WER for FST being 88%, FWM being 83%, and FWT being 85%. This suggests sentences are the proper granularity for extracting timestamp data from finger tracking, which is consistent with the user-test results. The largest WER reduction was exhibited in a manual mode. This provides evidence for the usefulness of this interaction metaphor and suggests opportunities for additional research on how to best combine intelligent supports, such as the auto-advancing timer, with natural metaphors to further improve forced alignment algorithms.

DISCUSSION AND LIMITATIONS

While participants read larger segments of text faster (H4) and felt that these were easier to read (H3), the largest segment of text recorded (a page) is too large to enable improvements between the alignment of a recording and the text that is associated with that recording. Based on the evaluated text segments, the data for reading speed and user effort indicate the appropriate text segment size is the sentence. This recording unit size results in a savings of approximately 20 seconds for every two pages of text read.

The text was also read more quickly when the personalized timer automatically advanced recording (H1). The adaptive auto-advance feature resulted in a savings of approximately 30 seconds per two-page recording. However, participants felt that it was harder to maintain their reading pace when the system was in an adaptive mode (H2). This perception is interesting given that the adaptive timer had been set to match the user's initial reading pace.

These differences could be partly due to interface design: the highlighting of individual words and the latency introduced through this process appears to have resulted in users slowing their reading pace and attending to individual words rather than larger, more fluid text segments. This is supported by experimenter observations and user reports: participants felt the word-based recording conditions were slower than the others. Additionally, participants indicated that maintaining their initial pace was more difficult in the auto-advancing conditions, likely as a result of their setting an unrealistic pace in the opening sentences (H2). This pace may have been unintentionally fast because the first sentences were simpler in nature or because participant working memory had been cleared when they changed conditions. As participants read the text, their processing of the text would have increased cognitive load [62] because more information would need to be held in participants' short-term memory. The automated highlighting that indicates when the timer has advanced to the next text segment may have further increased cognitive load or added a sense of urgency for users, which helps explain why they felt that it was less tiring to manually advance recordings.

That said, participants were able to complete the tasks in less time, indicating the presence of a trade-off between user task enjoyment and recording pace. The adaptive techniques allowed users to correct the pace (by tapping on previously-read sentences or words), and we speculate the higher reading speeds were within users' capabilities given that many participants adjusted their reading speed to match that of the application. This suggests users may have a wide comfort zone with respect to reading pace that does not fully overlap with their perception of difficulty. Like other assistive technologies, our app is not equipped to determine the user's most comfortable reading speed so we cannot compare the observed reading speeds to that ideal. However, we compared them to the regular reading habits that were observable (e.g., baseline). We expect using on-device sensors (e.g., the camera) to measure aspects of cognition [70] or advances in lightweight brain-computer interfaces to facilitate incorporating these aspects into future interfaces.

Additional work with respect to how to indicate the recording is advancing may lessen the difference in user experience that exists between the adaptive auto-advancing and the manual advancing interfaces. For example, a focus window that highlights the current sentence and lowlights the upcoming sentence may help reduce the sense of urgency that is sometimes communicated through the current highlighting scheme. This reduction in perceived urgency may improve the user's experience and reduce or eliminate the trade-off that appears to exist between enabling the fastest possible recording speed while allowing for sentence alignment without harming user experience. This approach of highlighting a collection of sentences using a sliding window, where the highlight saturation indicates the current focus, may also reduce the confusion that users reported when short phrases, such as Mr. or Mrs., were highlighted.

The design of new interaction techniques and their evaluation through a controlled experiment shed light on how mobile interfaces can assist users in their reading and content-creation tasks. Going forward, we will explore whether other text unit lengths (e.g., paragraphs) influence the trade-off between users' cognitive load, perception of effort, and speed of recording, while supporting the need to automatically align the text and audio so as to enable the later consumption of that text. Due to practical considerations, the present study was limited in the range of unit lengths that we could evaluate within a 2-hour participant session.

We plan to conduct follow-up investigations that measure the effect assistive metaphors, such as finger tracking, have on reading quality as perceived by those the content is meant to support (e.g., older adults or low-literacy adults). Additionally, we plan to measure the accuracy of users'

recordings in the different assistive modes, particularly focusing on how these techniques can be used to augment existing speech processing and alignment algorithms.

While the results indicate an implicit bias toward users wanting to exercise full control, it is possible this bias would be reduced or eliminated over time as the user interacts with the system, better understands how it works, and begins to trust it [46]. This lack of trust in adaptive systems is commonly observed in initial evaluations. It indicates longer-term evaluations are needed now that we know these adaptive recording approaches result in the same activity being completed in less time.

CONCLUSIONS

We have introduced and evaluated a natural metaphor-based interaction technique for mobile e-readers. This technique enables loved ones and professionals, such as teachers, to create assistive materials for those who cannot read on their own. Enabling the non-commercial production of accessible reading materials can benefit marginalized users such as low-vision older adults by providing them with synchronized audio recordings in the familiar voice of a loved one.

Several versions of the finger-tracking metaphor (i.e., following the text with a finger while reading) were implemented and evaluated through a controlled experiment (n = 22). These implementations varied the recording unit (word, sentence, or page) and the level of control (manual or adaptive) that users could exercise over the recording pace. The study evaluated the influence these variations of the finger-tracking metaphor had on user recording time and perceptions. This evaluation revealed that users read at their fastest pace, while maintaining reading quality, when they recorded a text one sentence at a time with the system adaptively controlling the recording pace through the use of pre-attentive visual cues. This method of interaction helps ensure accuracy and saves the user or content creator between 20 and 30 seconds for every two pages of text read. This is a substantial gain when considering longer texts.

Our future work will further investigate the appropriate presentation and granularity of text to best support assisted reading. This work will also study how cognitive factors (e.g., load or fatigue), measured through on-device sensors, can be incorporated to personalize and improve the content-creation experience. Additional natural language processing techniques will be investigated for automatically extracting when the reader engages in discussion that is related to the topic or makes tangential comments. The collected data will be used to improve natural language processing techniques for aligning audio with text when users make errors, self-correct, or otherwise vary their speech from the text that they are reading. Furthermore, we plan to evaluate how the supported readers (e.g., older adults with low vision) perceive the quality of the recordings made under different assistive conditions.

Overall, our analysis of objective and subjective data shows that interaction metaphors, such as using the finger to track text on a mobile e-reader, are suitable for use in assistive tools. We demonstrated their use within a specific domain where these metaphors were shown to facilitate the development of support materials for those who cannot read on their own. Our evaluation, which demonstrates the feasibility of this new metaphor, serves as a starting point for designing more natural interactions for human-in-the-loop approaches to assistive technologies. As such, this study provides an example of how interaction techniques can be used to collect data that can later inform algorithm development. Furthermore, it demonstrates how adaptive systems, whether simple or complex, can be augmented through implicit user interaction techniques.

ACKNOWLEDGEMENTS

This work was supported by AGE-WELL, a national research network supporting research, networking, commercialization, knowledge mobilization and capacity building activities in technology and aging to improve the quality of lives of Canadians. AGE-WELL is a member of the Government of Canada's Networks of Centres of Excellence program. We are also thankful for the contribution of undergraduate student developers Jaisie Sin and Hubert Hu.

The authors would also like to acknowledge the sacred lands on which the University of Toronto Mississauga operates. These are the traditional territories of the Haudenosaunee and of the Mississaugas of the New Credit First Nation. The territories were the subject of the Dish With One Spoon Wampum Belt Covenant, an agreement between the Iroquois Confederacy and the Ojibwe and allied nations to peaceably share and care for the resources around the Great Lakes. We are grateful to have the opportunity to work in the community, on this territory.

REFERENCES

1. Aitor Álvarez, Haritz Arzelus, and Pablo Ruiz. 2014. Long audio alignment for automatic subtitling using different phone-relatedness measures. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6280–6284. https://doi.org/10.1109/ICASSP.2014.6854812
2. Xavier Anguera, Jordi Luque, and Ciro Gracia. 2014. Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*, 1405–1409. Retrieved from http://www.isca-speech.org/archive/interspeech_2014/i14_1405.html
3. J. Archibald and W. O'Grady. 2008. *Contemporary Linguistic Analysis*. Pearson.
4. Abbas Attarwala, Ronald M. Baecker, and Cosmin Munteanu. 2012. Accessible, Large-print, Listening & Talking e-Book (ALIT). In *Proceedings of the Fifth ACM Workshop on Research Advances in Large*

Digital Book Repositories and Complementary Media (BooksOnline '12), 19–20. https://doi.org/10.1145/2390116.2390129
5. Abbas Attarwala, Cosmin Munteanu, and Ronald Baecker. 2013. An Accessible, Large-print, Listening and Talking e-Book to Support Families Reading Together. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services* (MobileHCI '13), 440–443. https://doi.org/10.1145/2493190.2494658
6. Jill Attewell and Carol Savill-Smith. 2004. Mobile learning and social inclusion: focusing on learners and learning. In *Learning with mobile devices: research and development*. 3–12.
7. Matthew P. Aylett, Graham Pullin, David A. Braude, Blaise Potard, Shannon Hennig, and Marilia Antunes Ferreira. 2016. Don't Say Yes, Say Yes: Interacting with Synthetic Speech Using Tonetable. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '16), 3643–3646. https://doi.org/10.1145/2851581.2890245
8. Scott Bateman, Rosta Farzan, Peter Brusilovsky, and Gord McCalla. 2006. OATS: The Open Annotation and Tagging System. In *3rd Annual International Scientific Conference of the Learning Object Repository Research Network*, 10. Retrieved April 1, 2012 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.9222
9. J.E. Beck and J. Sison. 2006. Using Knowledge Tracing in a Noisy Environment to Measure Student Reading Proficiencies. *International Journal of Artificial Intelligence in Education (IJAIED)* 16, 2: 129–143.
10. Joseph Beck, Peng Jia, and Jack Mostow. 2003. Assessing Student Proficiency in a Reading Tutor that Listens. In *International Conference on User Modeling (UM)* (Lecture Notes in Computer Science), 323–327.
11. Shirley Ann Becker. 2004. A Study of Web Usability for Older Adults Seeking Online Health Resources. *ACM Trans. Comput.-Hum. Interact.* 11, 4: 387–406. https://doi.org/10.1145/1035575.1035578
12. Nicola J. Bidwell, Thomas Reitmaier, Gary Marsden, and Susan Hansen. 2010. Designing with mobile digital storytelling in rural Africa. 1593. https://doi.org/10.1145/1753326.1753564
13. Christian Boitet. 1990. Towards Personal MT: General Design, Dialogue Structure, Potential Role of Speech. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2* (COLING '90), 30–35. https://doi.org/10.3115/997939.997945
14. N. Braunschweiler, M. J. F. Gales, and S. Buchholz. 2010. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, 2222–2225. Retrieved August 29, 2016 from http://publications.eng.cam.ac.uk/323723/
15. Jennifer Bravo, Eileen Bartholomew, Christopher Frangione, Jil Gravender, and Matt Keller. 2014. *XPRIZE Adult Literacy Landscape Analysis*. XPRIZE Foundation, Culver City California. Retrieved from http://www.xprize.org/sites/default/files/adult_literacy_landscape_analysis_2014.pdf
16. Nadia Caidi and Danielle Allard. 2005. Social inclusion of newcomers to Canada: An information problem? *Library & Information Science Research* 27, 3: 302–324. https://doi.org/10.1016/j.lisr.2005.04.003
17. Joanne F. Carlisle. 1988. Knowledge of Derivational Morphology and Spelling Ability in Fourth, Sixth, and Eighth Graders. *Applied Psycholinguistics* 9, 03: 247–266. https://doi.org/10.1017/S0142716400007839
18. Joanne F Carlisle, C Addison Stone, and Lauren A Katz. 2001. The effects of phonological transparency on reading derived words. *Annals of Dyslexia* 51, 07369387: 249–274. https://doi.org/10.1007/s11881-001-0013-2
19. Daniel Churchill. 2006. Towards a useful classification of learning objects. *Educational Technology Research and Development* 55, 5: 479–497. https://doi.org/10.1007/s11423-006-9000-y
20. Luca Colombo, Monica Landoni, and Elisa Rubegni. 2012. Understanding Reading Experience to Inform the Design of Ebooks for Children. In *Proceedings of the 11th International Conference on Interaction Design and Children* (IDC '12), 272–275. https://doi.org/10.1145/2307096.2307143
21. Rasmus Dali, Sandrine Brognaux, Korin Richmond, Cassia Valentini-Botinhao, Gustav Eje Henter, Julia Hirschberg, Junichi Yamagishi, and Simon King. 2016. Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5155–5159.
22. S. Hélène Deacon, Nicole Conrad, and Sébastien Pacton. 2008. A statistical learning perspective on children's learning about graphotactic and morphological regularities in spelling. *Canadian Psychology/Psychologie canadienne* 49, 2: 118–124. https://doi.org/10.1037/0708-5591.49.2.118
23. Carrie Demmans Epp. 2016. Noticing: ELL use of MALL for filling the gap. In *CALICO Conference*.
24. Carrie Demmans Epp. 2016. Supporting English Language Learners with an Adaptive Mobile Application. University of Toronto, Toronto, ON, Canada. Retrieved from http://hdl.handle.net/1807/71720
25. Carrie Demmans Epp. 2017. Migrants and Mobile Technology Use: Gaps in the Support Provided by Current Tools. *Journal of Interactive Media in Education, Special Collection on migrants, education*

and technologies 2017, 1: 1–13. https://doi.org/10.5334/jime.432
26. Kathryn D. R. Drager and Joe E. Reichle. 2001. Effects of Discourse Context on the Intelligibility of Synthesized Speech for Young Adult and Older Adult Listeners: Applications for AAC. *Journal of Speech, Language, and Hearing Research* 44, 5: 1052–1057. https://doi.org/10.1044/1092-4388(2001/083)
27. W. Erickson, C. Lee, and S. von Schrader. 2015. *Disability Statistics from the 2013 American Community Survey (ACS)*. Cornell University Employment and Disability Institute (EDI), Ithaca, NY. Retrieved from http://www.disabilitystatistics.org/
28. Ronald A. Fisher. 1966. *The design of experiments*. Hafner Publishing Company, Inc., New York, NY, USA.
29. Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication* 52, 3: 181–200. https://doi.org/10.1016/j.specom.2009.10.001
30. João Guerreiro and Daniel Gonçalves. 2014. Text-to-speeches: evaluating the perception of concurrent speech by blind people. 169–176. https://doi.org/10.1145/2661334.2661367
31. Chandra M. Harrison. 2004. Low-vision reading aids: reading as a pleasurable experience. *Personal and Ubiquitous Computing* 8, 3–4: 213–220. https://doi.org/10.1007/s00779-004-0280-0
32. Ken Hinckley, Xiaojun Bi, Michel Pahud, and Bill Buxton. 2012. Informal Information Gathering Techniques for Active Reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 1893–1896. https://doi.org/10.1145/2207676.2208327
33. Matt Jones, Emma Thom, David Bainbridge, and David Frohlich. 2009. Mobility, digital libraries and a rural Indian village. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 309–312. https://doi.org/10.1145/1555400.1555451
34. M. J. Kieffer, N. K. Lesaux, M. Rivera, and D. J. Francis. 2009. Accommodations for English Language Learners Taking Large-Scale Assessments: A Meta-Analysis on Effectiveness and Validity. *Review of Educational Research* 79, 3: 1168–1201. https://doi.org/10.3102/0034654309332490
35. Simon King. 2014. Measuring a decade of progress in Text-to-Speech. *Loquens* 1, 1: e006. https://doi.org/10.3989/loquens.2014.006
36. Nat Lertwongkhanakool, Natthawut Kertkeidkachorn, Proadpran Punyabukkana, and Atiwong Suchato. 2015. An Automatic Real-time Synchronization of Live speech with Its Transcription Approach. *Engineering Journal* 19, 5: 81–99. https://doi.org/10.4186/ej.2015.19.5.81
37. Yan-Hua Long and Hong Ye. 2015. Filled pause refinement based on the pronunciation probability for lecture speech. *PloS One* 10, 4: e0123466. https://doi.org/10.1371/journal.pone.0123466
38. Catherine C. Marshall and Sara Bly. 2005. Turning the Page on Navigation. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '05), 225–234. https://doi.org/10.1145/1065385.1065438
39. Gord McCalla. 2004. The Ecological Approach to the Design of E-Learning Environments: Purpose-based Capture and Use of Information About Learners. *Journal of Interactive Media in Education* 2004, 7: 1–23.
40. Jack Mostow. 2012. Why and How Our Automated Reading Tutor Listens. In *International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, 43–52.
41. Cosmin Munteanu, Joanna Lumsden, Hélène Fournier, Rock Leung, Danny D'Amours, Daniel McDonald, and Julie Maitland. 2010. ALEX: Mobile Language Assistant for Low-Literacy Adults. In *Proc. International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)* (MobileHCI '10), 427–430. https://doi.org/10.1145/1851600.1851697
42. Cosmin Munteanu, Heather Molyneaux, Julie Maitland, Daniel McDonald, Rock Leung, Hélène Fournier, and Joanna Lumsden. 2013. Hidden in plain sight: low-literacy adults in a developed country overcoming social and educational challenges through mobile learning support tools. *Personal and Ubiquitous Computing*: 1–15. https://doi.org/10.1007/s00779-013-0748-x
43. Emma Murphy, Ravi Kuber, Graham McAllister, Philip Strain, and Wai Yu. 2008. An empirical investigation into the difficulties experienced by visually impaired Internet users. *Universal Access in the Information Society* 7, 1–2: 79–91. https://doi.org/10.1007/s10209-007-0098-4
44. Susan Neuman and David K Dickinson. 2011. *Handbook of Early Literacy Research, Volume 3*. Guilford Publications, New York.
45. Nicholas R. Nicholson. 2012. A Review of Social Isolation: An Important but Underassessed Condition in Older Adults. *The Journal of Primary Prevention* 33, 2–3: 137–152. https://doi.org/10.1007/s10935-012-0271-2
46. John O'Donovan and Barry Smyth. 2005. Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces* (IUI '05), 167–174. https://doi.org/10.1145/1040830.1040870
47. Nobuko Osada. 2004. Listening Comprehension Research: A Brief Review of the Past Thirty Years. *Dialogue* 3: 53–66.

48. S. Oviatt. 2003. Advances in robust multimodal interface design. *IEEE Computer Graphics and Applications* 23, 5: 62–68. https://doi.org/10.1109/MCG.2003.1231179
49. Joyojeet Pal, Manas Pradhan, Mihir Shah, and Rakesh Babu. 2011. Assistive Technology for Vision-impairments: An Agenda for the ICTD Community. In *Proceedings of the 20th international conference companion on World wide web*, 513–522.
50. Adrian Pasquarella, Alexandra Gottardo, and Amy Grant. 2012. Comparing Factors Related to Reading Comprehension in Adolescents Who Speak English as a First (L1) or Second (L2) Language. *Scientific Studies of Reading* 16, 6: 475–503. https://doi.org/10.1080/10888438.2011.593066
51. Andrea Passerini and Michele Sebag. 2015. Learning and Optimization with the Human in the Loop. In *Constraints, Optimization and Data*, 21–24. Retrieved from http://drops.dagstuhl.de/opus/volltexte/2015/4890/pdf/dagrep_v004_i010_p001_s14411.pdf
52. Jennifer Pearson, George Buchanan, and Harold Thimbleby. 2011. The Reading Desk: Applying Physical Interactions to Digital Documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 3199–3202. https://doi.org/10.1145/1978942.1979416
53. Jennifer Pearson, Tom Owen, Harold Thimbleby, and George Buchanan. 2012. Co-reading: investigating collaborative group reading. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 325–334.
54. Hayes Raffle, Glenda Revelle, Koichi Mori, Rafael Ballagas, Kyle Buza, Hiroshi Horii, Joseph Kaye, Kristin Cook, Natalie Freed, Janet Go, and Mirjana Spasojevic. 2011. Hello, is Grandma There? Let's Read! StoryVisit: Family Video Chat and Connected e-Books. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 1195–1204. https://doi.org/10.1145/1978942.1979121
55. Preeti Rao, Prakhar Swarup, Ankita Pasad, Hitesh Tulsiani, and Gargi Ghosh Das. 2016. Automatic Assessment of Reading with Speech Recognition Technology. In *24th International Conference on Computers in Education (ICCE)*, 1–3.
56. Frank Rudzicz, Rosalie Wang, Momotaz Begum, and Alex Mihailidis. 2014. Speech recognition in Alzheimer's disease with personal assistive robots. In *Proceedings of the 5th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 20–28. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=41D9A49475DE3BD5D8636885D56FB4B7?doi=10.1.1.477.4901
57. Nithya Sambasivan, Ed Cutrell, Kentaro Toyama, and Bonnie Nardi. 2010. Intermediated Technology Use in Developing Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 2583–2592. https://doi.org/10.1145/1753326.1753718
58. Kei Sawada, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, and Keiichi Tokuda. Overview of NITECH HMM-based text-to-speech system for Blizzard Challenge 2014. In *Blizzard Challenge Workshop*.
59. Roy Shilkrot, Jochen Huber, Wong Meng Ee, Pattie Maes, and Suranga Chandima Nanayakkara. 2015. FingerReader: A Wearable Device to Explore Printed Text on the Go. 2363–2372. https://doi.org/10.1145/2702123.2702421
60. Eva Siegenthaler, Pascal Wurtz, and Rudolf Groner. 2010. Improving the Usability of E-Book Readers. *J. Usability Studies* 6, 1: 3:25–3:38.
61. Statistics Canada. 2004. *International Adult Literacy Survey (IALS)*. Human Resources Development Canada.
62. John Sweller, Paul L Ayres, and Slava Kalyuga. 2011. *Cognitive Load Theory*. Springer, New York.
63. Elsebeth Tank and Carsten Frederiksen. 2007. The DAISY Standard: Entering the Global Virtual Library. Retrieved February 6, 2017 from https://www.ideals.illinois.edu/handle/2142/3763
64. Gökhan Tür, Dilek Hakkani-Tür, Andreas Stolcke, and Elizabeth Shriberg. 2001. Integrating Prosodic and Lexical Cues for Automatic Topic SegmentatioN. *Computational Linguistics* 27, 1: 31–57. https://doi.org/10.1162/089120101300346796
65. Ashwini Venkatesh, M. V. Lalitha, Jyothi Narayana, and Kavi Mahesh. 2015. Wikiaudia: Crowd-sourcing the Production of Audio and Digital Books. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*.
66. Rafael Veras, Erik Paluka, Meng-Wei Chang, Vivian Tsang, Fraser Shein, and Christopher Collins. 2014. Interaction for Reading Comprehension on Mobile Devices. In *Proc. of the 16th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '14)* (MobileHCI '14), 157–161. https://doi.org/10.1145/2628363.2628387
67. Richard K Wagner, Andrea E Muse, and Kendra R Tannenbaum (eds.). 2007. *Vocabulary Acquisition: Implications for Reading Comprehension*. Guilford Press, New York.
68. Mirjam Wester, Matthew Aylett, Marcus Tomalin, and Rasmus Dall. 2015. Artificial personality and disfluency. In *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*.
69. Silke M. Witt. 2012. Automatic Error Detection in Pronunciation Training: Where we are and where we need to go. In *International Symposium on automatic detection on errors in pronunciation training (IS ADEPT 6)*.
70. Xiang Xiao and Jingtao Wang. 2015. Towards Attentive, Bi-directional MOOC Learning on Mobile

Devices. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (ICMI '15), 163–170. https://doi.org/10.1145/2818346.2820754

71. Junbo Zhang, Fuping Pan, and Yonghong Yan. 2012. An LVCSR Based Automatic Scoring Method in English Reading Tests. 34–37. https://doi.org/10.1109/IHMSC.2012.14
72. 2005. *LibriVox*. Librivox. Retrieved from http://librivox.org/