Touch-Supported Voice Recording to Facilitate Forced Alignment of Text and Speech in an E-Reading Interface

Benett Axtell[1], Cosmin Munteanu[1, 2], Carrie DEMMANS EPP[1, 3], Yomna Aly[1], Frank Rudzicz[4]

[1] TAGlab, University of Toronto
Toronto, Canada
{benett, cosmin, yomna}@taglab.ca

[2] Institute of Communication, Culture, Information and Technology, University of Toronto Mississauga
Mississauga, Canada

[3] Learning Research and Development Center, University of Pittsburgh
Pittsburgh, United States
cdemmans@pitt.edu

[4] Toronto Rehabilitation Institute,
Toronto, Canada
frank@spoclab.ca

ABSTRACT

Reading a book together with a family member who has impaired vision or other difficulties reading is an important social bonding activity. However, for the person being read to, there is little support in making these experiences repeatable. While audio can easily be recorded, synchronizing it with the text for later playback requires the use of forced alignment algorithms, which do not perform well on amateur read-aloud speech. We propose a human-in-the-loop approach to augmenting such algorithms, in the form of touch metaphors during collocated read-aloud sessions using tablet e-readers. The metaphor is implemented as a finger-follows-text tracker. We explore how this could better handle the variability of amateur reading, which poses accuracy challenges for existing forced alignment techniques. Data collected from users reading aloud as assisted by touch metaphors show increases in the accuracy of forced alignment algorithms and reveal opportunities for how to better support reading aloud.

Author Keywords

Assistive technology; multi-modal interfaces; forced alignment; natural language and speech processing;

INTRODUCTION

Many assistive applications need to synchronize audio and text, such as closed captioning [27]. For many of these, the aim is purely to address an impairment, but there are many other applications where the ability to synchronize audio and text will enhance both the accessibility of the application and user experience. One such application is collocated family reading, which is an enjoyable social activity among family members, especially when one has impaired vision, limited literacy, or other difficulties reading. Reading aloud is a skill that is common in many professions, like lecturing, but for casual readers, such as within a family, it can be an unfamiliar or even uncomfortable activity.

Alignment of text with speech has been shown to benefit marginalized users (such as low-literacy adults) trying to read e-texts, especially when accompanied by a moving prompt that shows the synchronization of text and speech [30,31]. The audience for casual read-alouds, however, is also more accepting of errors such as mispronounced words or filled pauses (e.g., "um", in English). Additionally, recordings of these readings, while easy to create, are likely of a lower quality than those of professional readings (e.g., lectures or audio books) because of substandard recording conditions and equipment [24]. Due to these factors, forced alignment and available tools to align speech to text are unsuited for these settings.

To aid the alignment of text and speech, this research investigates how a classic forced alignment (FA) algorithm can be enhanced by using a human-in-the-loop approach with assistive touch metaphors. The motivation for this is to move towards options that better suit casually read-aloud speech. Human-in-the-loop approaches can aid in the development and use of adaptive systems [29] and other technologies [37]. This approach combines system and human efforts to provide the individualized support readers need. The system learns from the people who interact with it by deferring part of the decision process to those users [29,34]. Improving adaptive systems with similar systems has previously supported underrepresented users [29,36]. As such, we explore a human-in-the-loop system in this casual reading setting, using a family member to support those with various reading difficulties.

We explore and analyze the effects of these interaction metaphors and their text to speech alignment on the assessment of speech. We use measures of reading time and accuracy, and analyze how these measures change over the course of reading a text. We expand on these findings by building on a classic FA algorithm to create a new algorithm that accounts for the less precise nature of the expected

IUI'18, March 7–11, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-4945-1/18/03…$15.00
https://doi.org/10.1145/3172944.3172984

speech. Our Touch-Enabled Forced Alignment algorithm (TFA) proposes improvements to classic FA by accounting for expected human behaviours in casually read-aloud speech, which limits outlier data found in classic FA. These expected reading behaviours include interruptions, out of sync speech and touch, and variations in readers' timing of interaction, while also incorporating data from assistive touch interaction. We compare these to parallel results using a classic FA algorithm and human-aligned speech to text.

This paper contributes to the evaluation of intelligent user interfaces by demonstrating that touch interaction support for forced alignment more accurately aligns text to speech. We show that these alignment results better represent the accuracy of reading aloud for amateur readers by accounting for their normal variations is reading. We also explore opportunities to augment forced alignment in order to better support casually read-aloud data.

RELATED WORK

Previous research on forced alignment showed that, although such speech processing algorithms have greatly improved recently, there are significant shortcomings in the application of forced alignment to several application areas. Assisted reading research, especially when providing support to various marginalized groups of readers, commonly encounters some of these difficulties of aligning speech to text. The intersection of these areas reveals various possible approaches to address these shortcomings in order to better meet the needs of a wide range of readers.

Assisted Reading

Audio-books and e-readers have been proposed as alternatives to interacting with text for members of groups who have limited access to text-based information. This lack of access may be due to low literacy or impaired vision [31,32]. Professionally created audio books are expensive and time consuming to create [45]. Synthetic speech, which has been used to read to adults with vision impairments [39], is still found to be unnatural [21,38,42] and sometimes difficult to understand [12,18]. There are also few available texts that are properly formatted to be read aloud completely by commercial e-readers [1,3,32], and current accessibility features (e.g. screen readers) have limitations, such as navigation within a long text [5]. Other solutions allow people to record their own reading for children (e.g., http://explore.hallmark.com/recordable-storybooks) and adults [13], though these can take significant effort from the reader.

Alignment of text and recorded speech has been shown to benefit marginalized users [31] and to aid users in reading e-texts by using tools such as prompts that move along the text [13]. This alignment should enable access to those who cannot read on their own, but this is blocked by technical barriers [11]. Additionally, the variability of human reading behaviours, including self-correction and filled pauses, are a major factor in the high error rates found with standard options for alignment [11,15].

Speech to Text Alignment

There are several different methods to align long audio to text depending on the context [8,17,20,28]. Forced alignment is one such method of speech recognition that uses a transcription of what was expected to have been said and attempts to align the speech to the given transcript. It then produces time segments corresponding to particular words in the transcript. This information can be used to analyze speech data, including correctness through measures like Word Error Rate (WER). Since a transcript is provided, it is a logical approach when what is going to be said is known ahead of time, such as when a book is being read aloud. Classic FA is commonly used to align speech from a variety of settings such as lectures or other public speeches [16] or broadcast media subtitles [27]. It is also employed in the production of synchronized materials for students learning the pronunciation of a foreign language [4,40].

In some cases, the performance of FA is measured by the time differential of various anchors (e.g., words or sentences) between the gold standard manual alignment and some test data. This is useful for very long audio recordings [28] or when it is important to improve alignment of small phonetic units [19]. However, as found by Hazen, the role of FA is often to support "downstream" applications [16], including when the original text is known, such as for forced alignment of audio books with their corresponding texts [8], and even for medium-length audio recordings of broadcast news [27]. As our work is similar to these examples, we too use WER.

There are several examples of recent improvements to accuracy of forced alignment including those done in the areas of real time forced alignment [26] and precision requirements for different settings and ease of use [14]. Other work has looked at the shortcomings of forced alignment and how to improve upon them. Zhang et al [44] found that their system for scoring reading quality avoided known disadvantages of forced alignment by switching to large vocabulary continuous speech recognition (LVCSR). LVCSR was able to recognize more miscues (mistakes in reading) than forced alignment.

Anguera et al [1] argued that improving alignment accuracy and speech recognition will enable more effective synchronization of speech to texts. However, many challenges remain as speech recognition and alignment must account for human errors that make up 15% of speech [15]. These errors and human behaviours include mispronunciations [43] and added or off-script speech [6,23], including re-reading [29]. Other work further demonstrated the limitations of forced alignment [21,29,41], including inadequate miscue detection. All of this contributes to the at least 19% error rate of standard alignment methods [11], which interferes with the effectiveness of assistive tools and metaphors for reading on assessment.

Alignment of Assisted Reading

Much of the existing research into the assessment of read-aloud speech is focused on children or language learning [9,22]. One project not focusing on these groups investigated a more accurate alignment option for large amounts of coherent text, such as audio books [8], and supports that forced alignment, on its own, is not ideal for large read-aloud recordings. More recent work improves on long text speech recognition by improving the recognition of types of pauses including inhalations, which are more common in longer, casual readings [7]. Other research has explored the use of multimodal books for language learning [4] by developing a system that uses synchronized reading to aid language-learner spelling. This system plays the audio for a given text while a cursor follows, pointing at the current phoneme. Their preliminary findings show that spelling can be improved with the bimodal approach of audio and visuals. The bimodal interaction from this work is similar to the interaction that produced the data used in this paper.

THE ALLT PROJECT

For this work, our researchers were given access to the data from a study assessing the usability of assistive touch metaphors within the ALLT e-book reader (Figure 1) [13]. The ALLT (Accessible, Large-print Listening and Talking) e-book tablet reader uses a bimodal touch and speech interface and provides assistive interfaces for users to read aloud and record books that can be played back by other family members. ALLT is intended to support collocated reading between an older adult (likely with some vision impairment) and an adult grandchild (e.g. university-aged). A grandchild can record a read-aloud text with some level of tracking and a grandparent can later listen and follow (read) along while hearing the familiar voice of the grandchild. While recording, the user is presented with a karaoke-style reading prompt, based on the tracked position of the user's finger under the text they are reading. This prompt advances as the text is read aloud, as described below. As we have shown in [13], this assisted tracking can guide a user through audio content production. It is also expected to ease reading for low-vision older adults by highlighting each sentence as it is read [2,25]. This metaphor is grounded in prior work on finger-tracking of text to support reading [10,33].

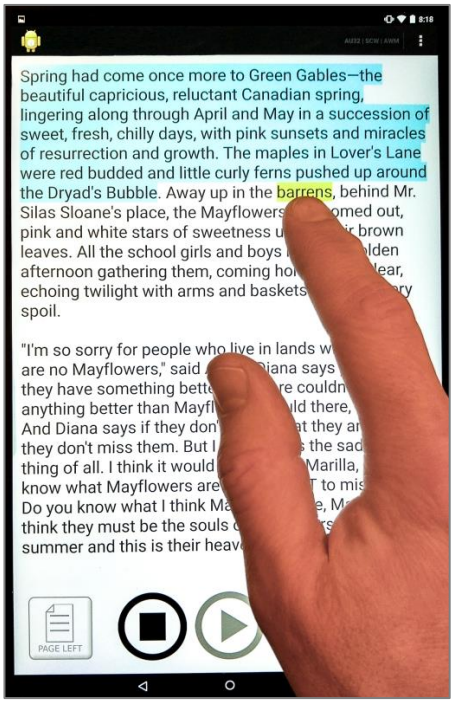

Figure 1. A version of the finger-tracking metaphor used to augment recording. The user is currently reading *barrens*.

Assistive Touch Metaphors

The ALLT e-book reader offers four assistive touch metaphors, which were tested empirically against a control condition (no touch interaction). The metaphors use touch interaction (F: finger) to highlight either the current word (W) or sentence (S) as the user reads through the text. Each level can either be advanced manually (M), or can use timer (T) to advance the highlighting automatically based on the user's initial reading speed. The four modes are:

- FSM – sentence level, manual highlight advance
- FST – sentence level, advance highlight with timer
- FWM – word level, manual highlight advance
- FWT – word level, advance highlight with timer

These touch metaphors help to align text to speech as the user interacts with the text and is expected to better support forced alignment for this data.

Initial ALLT Study and Observations of Reading

For the initial user study, users read sections of *Anne of Green Gables* aloud using an Android Nexus 7 tablet, with the different assistive modes that ALLT provides through touch interactions. This book was selected because it is publicly available. The assistive modes allowed the user to track what they were currently reading aloud using touch metaphors at the sentence or word level. Basic timing information from this input could be used to advance the tracking automatically after a certain amount of manual use.

The font size was approximately the size of a large-text print book to mimic the books most likely used by low-vision older adults. Users were not able to change the font size, and touch error was not measured in the original experiment. No participant's performance was noticably affected by touch error and none commented on font size as an issue.

The audio was recorded for each sentence separately based on a user's touch interaction. However, there were observed differences in how participants used this interaction. For example, at the sentence level, some users tapped exactly as they began to read, some paused their reading as they tapped and then began again, and others tapped after they began each sentence. As such, speech associated with a particular sentence may actually be captured in the audio recording for the next or previous sentence. Occurrences of disfluencies and mispronunciations are also not uncommon in this data due to the outdated nature of the language and the uncommon

nature of the activity. Additionally, during the study, users occasionally interrupted themselves to comment on their activity or to ask questions without first pausing the recording (as they had been instructed to do). This likely mimics the reality of casually recording reading, as readers may be interrupted while recording or may make comments themselves without pausing the recording. Forced alignment cannot currently account for these digressions from the expected text.

AUGMENTED FORCED ALIGNMENT FOR READING

Typically, forced alignment uses the Viterbi algorithm and a pre-trained language model with a pronunciation dictionary to extract speech features from the provided audio and force its alignment to the given script [16,35]. The Viterbi algorithm, commonly used to find the most likely path through hidden states, is used to provide the optimal alignment for the sounds in the audio file by using the transcription as a state machine (in this case, a sequence of words). Our proposed algorithm (TFA) augments the classic forced alignment algorithm, implemented here by the CMUSphinx tool (https://cmusphinx.github.io/), to measure the alignment and timings for a sentence of a read-aloud text with multimodal interaction.

To incorporate the different timings of users' touch inputs, this measure of alignment accounts for audio files that correspond to the sentences immediately preceding and following. It also disregards speech that is not part of the intended reading by skipping many concurrent insertions. This produces the following outputs: individual alignments for each word within a sentence, a measure of accuracy for each sentence (Word Error Rate - WER), and an adjusted total time for each sentence.

Modifications to Classic Forced Alignment

All use of forced alignment for both TFA and classic FA uses CMUSphinx's implementation of forced alignment with three minor changes: allowing for alignment of one-word sentences, allowing alignment to exclude a given number of milliseconds from the start of an audio file, and adding a standard count of substitutions (used for the WER measure described later). The original code only included deletions and insertions. Substitutions are defined as adjacent insertions and deletions, regardless of order.

Additionally, there were minor changes made to the pronunciation dictionary to include words in the read texts missing from the provided CMU Sphinx English language pronunciation dictionary (Table 1). It is not unexpected that some words would need to be added because of the source's use of regional language. These added words were either names (e.g., *Avonlea*) or already present in the dictionary as another part of speech (e.g., *bewitching* from *bewitch*; *marillas* from *marilla*; *scornfully* from *scornful*). The phoneme transcriptions were derived using a grapheme to phoneme tool within CMU Sphinx.

Word in text	Phoneme transcription
airily	EH R IH L IY
avonlea	AE V AH N L IY
bashfully	B AE SH F AH L IY
betaken	B IH T EY K AH N
bewitching	B IH W IH CH IH NG
birches	B ER CH IH Z
bootjack	B UW T JH AE K
budded	B AH D IH D
creakily	K R IY K IH L IY
dizzily	D IH Z IH L IY
dreamily	D R IY M IH L IY
dryads	D R AY AE D Z
heedlessness	HH IY D L AH S N AH S
marillas	M AA R IH L AH Z
mayflowers	M EY F L AW ER Z
mossy	M AA S IY
resignedly	R IH Z AY N IH D L IY
scornfully	S K AO R N F AH L IY
sloanes	S L OW N Z
sloped	S L OW P T
smilelessly	S M AY L L AH S L IY
starrier	S T AA R IY ER
unobservant	AH N AH B Z ER V AH N T
unvarying	AH N V EH R IY IH NG
upspringing	AH P S P R IH NG IH NG
woodbox	W UH D B AA K S
woodsy	W UH D Z IY

Table 1: Words added to the provided CMU Sphinx pronunciation dictionary and their phoneme transcriptions (using CMUBET phoneme codes)

Description of TFA

Because of how people interacted with the text being read, any audio file produced by reading with various touch-based assistive modes may not contain only and exactly the speech of the sentence that was expected to be read. The TFA algorithm produces an adjusted total time for each sentence representing the actual time spent reading the sentence, disregarding when users spoke off the script (e.g., to comment or ask a question of the researchers), and including the time spent reading the sentence before or after the associated audio file. If an audio file contains only exactly the speech corresponding to the expected sentence, this adjusted time would be the same as the total time of the original audio file.

Ideally, the chronologically-ordered speech files each correspond exactly to the text of one sentence in the same order. This is the case that classic FA algorithms expect; a provided text corresponds precisely to some provided speech. In this work, classic FA is modified to account for the case when the start or end of a sentence is in the preceding or following audio file, respectively.

When describing TFA, the terms "previous audio" and "previous sentence" refer respectively to the audio file from the previous round of alignment and the sentence of text that

the audio is expected to contain. Likewise, "current audio" and "current sentence" are the audio and sentence for the current round, and "next audio" and "next sentence" are those for the next round. This algorithm is applied to each audio file from a reading of a text in the recorded order and consists of three main steps:

1. If there was unaligned audio at the end of the previous sentence, align this sentence to that remaining audio.
2. Align the expected sentence to this audio. This excludes any text that was just aligned in step one and excludes any audio that was aligned to the end of the previous sentence in the last round's step three.
3. If there is unaligned text at the end of this sentence, align that remaining text to the next audio file. Compare this alignment with the next sentence's standard alignment. Use the alignment with the better (lower) resulting WER.

```
FOR EACH audio file for text by participant:
  //1: align missing audio from last round to
  this sentence
  IF previous audio had unaligned audio at end:
    Align current sentence to remaining
    unaligned previous audio
    Update adjusted time and WER for this
    sentence
  END IF

  //2: base alignment of current sentence to
  this audio
  // (exclude audio aligned in step 1 or the
  previous round's step 3)
  IF step 1 unaligned words left in this
  sentence:
    Align current unaligned sentence to this
    audio
    Update adjusted time and WER for this
    sentence
  END IF
  IF end of this alignment does not include the
  end of the audio:
    Set the time that unaligned starts for
    next round's step 1
    Remove this time from this sentence's
    adjusted time

  //3: align remaining text to next audio
  ELSE IF final alignment does not include end
  of sentence text:
    Align remaining expected sentence to next
    audio
    Align next sentence to next audio file
    IF remaining WER is lower (more accurate):
      Use this alignment
      Update adjusted time and WER for this
      sentence
    ELSE:
      Use next sentence alignment instead
    END IF
  END IF
END FOR EACH audio file
```

Figure 2: Pseudocode for the TFA algorithm, which accounts for amateur reading behaviours

The comparison in step 3 is to prevent the remaining alignment from using the entirety of the next audio that should truly be aligned to the next sentence. For example, if both the remaining text and the next sentence contain a common word, there is no way to know if the remaining alignment has correctly found the remaining audio or the next sentence's audio. To protect against this, both alignments are compared and the more accurate is used. This is not a perfect approach, but it should cover the majority of cases.

In order to handle off-script speech, insertions are excluded from WER calculations if there is a section at least two seconds long that contains only concurrent insertions. This is because it is unlikely that text would be incorrectly read in a manner that would produce a long series of insertions when the user is reading along with a provided text. In the case where many seconds worth of insertions are detected, it is most likely the user is no longer reading the text. The adjusted time for this sentence also excludes this time.

TFA is expected to improve upon classic FA to better suit the work of aligning amateur read-aloud text especially in a multimodal, assistive setting. See Figure 2 for a pseudocode explanation of TFA.

METHODS

Data

TFA was assessed using the data resulting from a previous study into touch-assisted casual reading [13]. This included speech data from 22 English-speaking participants. Each participant read the same five 2-page texts on the assistive ALLT tablet using a different touch interaction for each text or a control condition without touch interaction. This resulted in over 2,500 audio files each corresponding to either a sentence in a text (when using a touch metaphor) or the whole text (when in the control condition). The speech was recorded in separate audio files for each sentence based on the interaction from the reader. Recording for each sentence began when a user tapped on it or when a user's finger highlighted the first word, depending on the condition.

The data presented in this paper consist of Word Error Rate (WER) scores and durations of each sentence read by each participant as measured from three sources:

1) Classic FA algorithm as offered by CMU Sphinx
2) The augmented algorithm proposed in this paper (TFA)
3) Human alignment completed by a researcher not involved in this project

WER is a simple and standard measure of correctness in speech, which is regularly used to assess FA results, especially when the text is known ahead of time [8,16]. It assesses how correct speech is compared to a text by

	FSM		FST		FWM		FWT		Control	
	Mean	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**
Classic FA	30.26	(10.85)	44.69	(15.11)	41.32	(13.22)	42.42	(12.68)	47.97	(22.24)
TFA	32.14	(9.96)	45.22	(15.07)	41.16	(12.34)	41.11	(10.02)	44.93	(19.61)
Human	3.28	(2.54)	14.45	(14.78)	7.36	(6.37)	3.09	(2.58)	1.56	(1.09)

Table 2: WER scores per participant across assessment methods and touch metaphors

	FSM		FST		FWM		FWT		Control	
	Mean	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**	**Mean**	**(SD)**
Classic FA	358.39	(51.10)	283.05	(70.65)	353.42	(72.15)	372.73	(61.11)	340.49	(67.99)
TFA	340.73	(47.18)	310.87	(59.27)	322.96	(53.43)	362.38	(52.98)	336.20	(67.83)

Table 3: Reading time per participant across assessment methods and touch metaphors

measuring the number of deletions (D), insertions (I), and substitutions (S). A visual representation of how such errors are relevant to forced alignment is given by [28, Figure 2]. A deletion is a word that is expected to be said but was missing, an insertion is a word that was spoken that does not correspond to a word in the script, and a substitution is a replaced word. WER is defined as (D + I + S) / N where N is the number of words in the script. A WER of 0 represents a perfectly read sentence, and higher values are less correct. There is no upper bound on WER as there is no limit to how many insertions may be present. For example, a five-word sentence where nothing is read will have a score of 100 (N = 5, D = 5, I = 0, S = 0) as will that same sentence where each word is incorrect (S = 5), but the same sentence read correctly with a ten-word diversion in the middle will have a score of 200 (N = 5, D = 0, I = 10, S = 0).

Reading time was measured as (duration of the sentence reading in milliseconds / number of words in sentence). This was done to normalize times across the large variety in sentence lengths (1 to 53 words). Higher values of reading time indicate the user is reading at a slower pace (more time spent on each word) and vice versa. There is no human measure of reading time because of the nature of the audio files – the audio for individual sentences was split between different files, making it difficult to accurately measure manually.

We also analyze how WER and reading time changed over the course of reading a text. This is measured as the slope of the trendline for all the WERs or reading times for each sentence from a text as read by one participant. In order to compare trends in WER and reading times to each other, we scale all trends to be between -100 and 100. These trends can show how the effect of an assistive mode may change over the course of a session. For example, readers may become more comfortable with an interaction or may be rushed by the timer if it continues to get ahead of them.

RESULTS

Since WER measures insertions, deletions, and substitutions between a reference text and the automatic transcript of an audio recording, it is sensitive to the alignment between the reference text and the audio recording. As such, differences in alignment are expected to yield different WERs. We thus analyze the data between the three alignment methods tested in our investigation (classic FA, TFA, and human) as these are expected to measure WER in different ways. Beside these accuracy comparisons, we also analyze data that is relevant to the user's experience, such as reading time for the two automated alignments. We measure changes in WER and reading time over the course of a reading session for each of the assistive modes. This is not possible for the control sessions as there is no subdivision of the speech data. A summary of WER and reading time measures are provided in Tables 2 and 3, respectively.

Supporting Classic FA with Touch Metaphors

We discuss here the performance of the classic FA according to the measures stated above. It should be noted that, while we did not modify the classic FA, its performance has already benefitted from the assistive touch metaphors. This is in the form of intrinsic alignment of speech to text through the human-in-the-loop approach used during the touch-enabled assisted reading sessions from which data was collected. This further highlights the usefulness of a human-in-the-loop approach to intrinsically capture the start time of each audio segment from the users' touch interaction as it corresponds to each sentence. Additionally, we have labelled this baseline "classic" as this algorithm was not modified from the standard. This or similar algorithms have been previously employed both with segmented audio [8] and with unsegmented medium-length recordings [27,28].

WER and Reading Time

In classic FA, only the sentence-level metaphor (FSM) significantly improved WER scores over the control scores, as shown by a Wilcoxon Signed-rank test ($W = 237$, $Z = 3.59$, $p < 0.001$, $r = 0.77$). Reading time for the sentence-level with timer mode (FST) was significantly faster than for the control case ($W = 202$, $Z = 2.45$, $p < 0.014$, $r = 0.52$).

Using Spearman's correlation, the only significant correlation between WER and reading time was in the FST mode. WER was strongly negatively correlated with reading time ($r_s = -0.73$, $p < 0.001$). This could reflect two situations; either WER increases as reading time per word decreases (i.e., pace of reading increases), or WER increases as more insertions are introduced, both cases demonstrate that faster reading in this mode leads to more errors.

	FSM	FWT	FWM
r_s	0.75	0.73	0.505
p	< 0.001	< 0.001	0.019

Table 4: Correlation scores for reading times compared to the control case

There was, however, positive correlations for reading time between all other modes (FSM, FWT, FWM) and the control case (see Table 4). This shows that participants who read slower during the control case also read slower during other modes (with the exception of FST) and vise versa.

There was a significant positive correlation between WERs between the control case and FSM ($r_s = 0.52$, $p = 0.012$). This shows that if a participant's reading was more accurate in the control case, it would be better in FSM as well. This parallels the reading time correlation, but only in this mode.

Trends in WER and Reading Time

Trends in how WER and reading time changed over the course of reading show how the different metaphors affect reading. Using classic FA, three modes (FSM, FST, and FWT) had significantly higher trends in WER than in reading time (see Table 5). These modes also had negative correlations between the trend of the WER over the course of reading and the reading time (see Table 6). This demonstrates that, at the sentence level and when using a timed mode, readers either slowed their reading pace and decreased their error rate over the course of a reading, or sped up their reading and increased their error rate. When using a timer, this was likely caused by the timer getting ahead of the reader without the reader adjusting the timer. In FSM, it is likely users adjusted their speed to be slower as they were reading as accuracy was best in this mode and users had full control over their reading speed.

	FSM	FST	FWT
W	200	241	207
Z	2.39	3.72	2.61
p	0.017	< 0.001	0.009
r	0.51	0.79	0.56

Table 5: Wilcoxon test results for trends in reading time and WER over time

	FSM	FST	FWT
r_s	-0.62	-0.70	-0.50
p	0.002	< 0.001	0.019

Table 6: Correlation scores for trends in reading times and WERs

Touch-Enhanced FA

Results from TFA largely correspond to classic FA, which was already partially optimized for amateur reader speech through the assistive metaphors. FST is also the least accurate mode, and FSM is the most (see Table 1). TFA did not see significant improvements in WER across the assistive modes as compared to classic FA. The standard deviations for all modes are lower in TFA. This is because TFA detects human variations in reading and interaction, which are outliers in the data, and normalizes them to more realistically represent casually reading aloud. This shows that TFA handles the irregularities of casual reading aloud better than classic FA, even if not resulting in lower WERs.

Parallels and Differences Between Algorithms

Just as was found with classic FA, the WER for FSM shows significant improvement ($W = 210$, $Z = 2.71$, $p = 0.007$, $r = 0.58$) over the control case. There was also the same negative correlation between WER and reading time for FST ($r_s = -0.62$, $p = 0.002$), but TFA revealed a weak negative correlation in the FSM mode ($r_s = -0.43$, $p = 0.046$).

Similar changes are seen in the trends over WER and reading time. As in the results with classic FA, FST and FWT are found to have significantly higher trends in WER than reading time (FST: $W = 237$, $Z = 3.59$, $p < 0.001$, $r = 0.77$; FWT: $W = 189$, $Z = 2.03$, $p = 0.043$, $r = 0.43$), and a negative correlation between the WER trend and that of the reading time is found for FST ($r_s = -0.49$, $p = 0.024$), but this is no longer found for FWT. TFA also finds a similar negative correlation for FSM ($r_s = -0.43$, $p = 0.048$), but no longer finds a significant difference for this condition. So, both sentence-level conditions, but not the word-level ones, show correlations between faster reading and lower accuracy, and slower reading and higher accuracy. This also finds that FSM, in which readers have complete control over the interaction, has no correlation between reading speed and accuracy.

TFA also finds a positive correlation between the reading times for FSM and the control case ($r_s = 0.63$, $p = 002.$), and between WERs for the same conditions ($r_s = 0.56$, $p = 0.007$). However, there is no correlation in reading time with the control case for either FWM or FWT, which was found in classic FA.

Effects on WER and Reading Time

WER with TFA has a significant positive correlation with WER in classic FA for each mode. This is also true for reading time between the algorithms. The strengths of these correlations show how these measures have been changed by TFA. The two sets of reading times for the control case are nearly perfectly correlated, showing there was very little change in that data. This can be seen in Figure 3; the data for the control case are nearly all on the diagonal. Reading times for FWT are the least correlated. The spread of FWT data points in Figure 3 compared to that of the control demonstrate this. All other modes have similarly strong correlations (see Table 7 and Figure 3).

The control case also has a very strong correlation for WER. The strongest correlation for WER, though, is in FST, the least accurate mode, and the weakest correlation is found in the most accurate mode (FSM). The WERs for the word-level conditions are similarly correlated between algorithms, and are more correlated than the reading time correlations for

	FSM		FST		FWM		FWT		Control	
	r_s	p	r_s	p	r_s	p	r_s	p	r_s	p
WER	0.66	< 0.001	0.95	< 0.001	0.85	< 0.001	0.90	< 0.001	0.90	< 0.001
Reading time	0.74	< 0.001	0.76	< 0.001	0.72	< 0.001	0.53	< 0.001	0.99	< 0.001

Table 7: WER and reading time correlations between algorithms

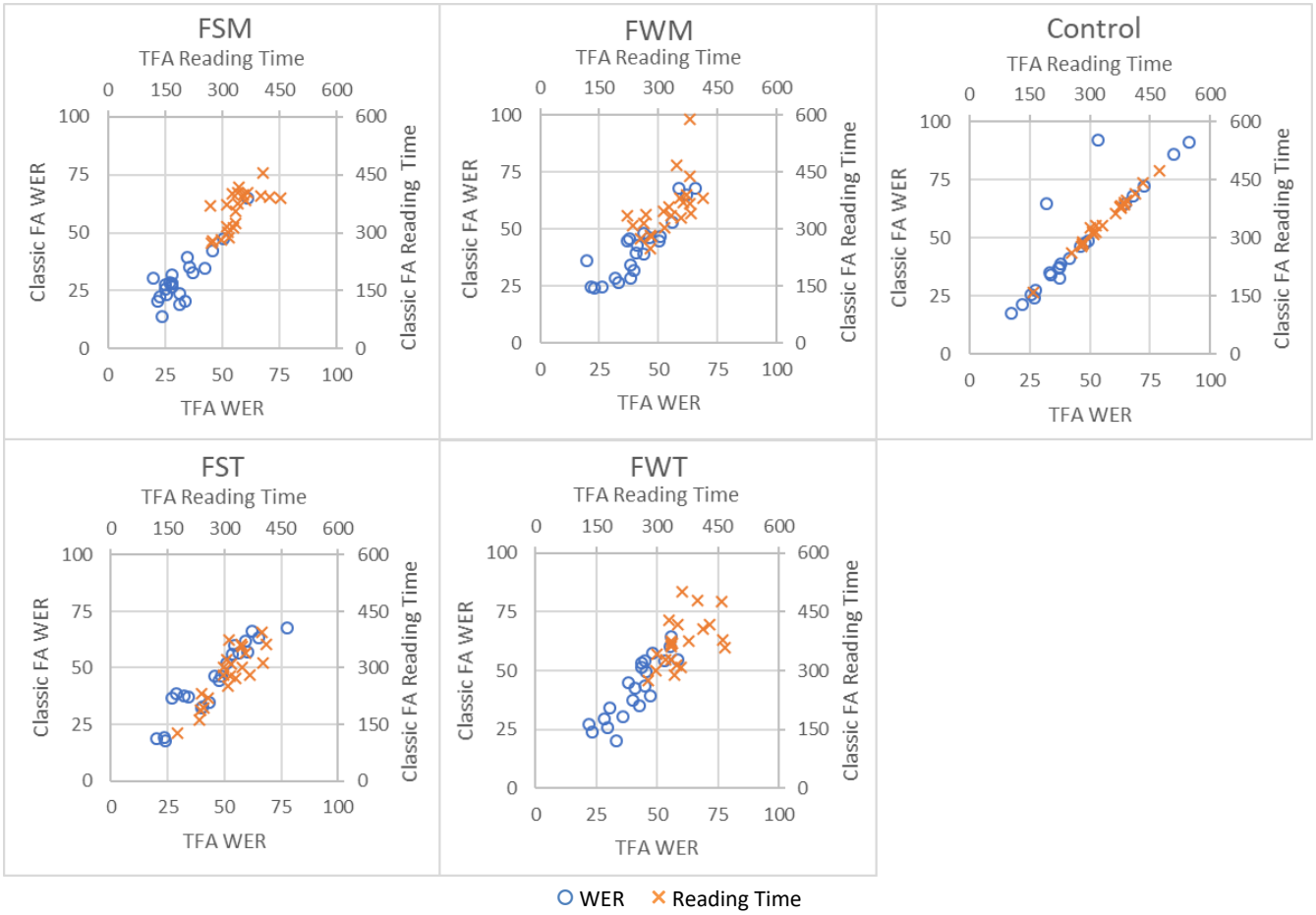

Figure 3: Mappings of WER and reading time between algorithms. Blue circles are WER, orange X's (diagonal crosshairs) are reading time.

the same modes. The more varied correlations across modes in reading time suggest that the changes to the algorithm had more of an effect on the time measures than the WERs.

Human Scoring

Unlike both algorithms, the human-scored WERs were significantly better in the control case than in any of the assistive modes (see Table 8). This is not unexpected since the assistive metaphors were new to the readers and the technical division of files caused some audio to be lost. Since the human scores showed so much improvement over classic FA and TFA, it is difficult to explore these differences meaningfully.

	FSM	FST	FWM	FWT
W	221	249	252	220
Z	3.07	3.98	4.07	3.04
p	0.002	< 0.001	< 0.001	0.002
r	0.65	0.85	0.87	0.65

Table 8: Wilcoxon test results for the human measured WER in each assistive mode compared to the control case

WER Between Algorithms

The human scoring confirms the general findings of both algorithms; FST was found to be least accurate, and FSM the most accurate (see Table 2). However, both algorithms were far more inaccurate than the human aligned data. This is expected as classic FA has been shown to have around a 19% WER in ideal settings [11]. For our study, the speech was recorded on commercially available tablets using the built-in microphone, so quality of audio is much lower than in professional recording settings. As this setting is conducive to even higher WERs for machine alignments as compared

to humans' (gold standard), the pairwise differences between the versions of machine alignment tend to not be significant.

DISCUSSION

The significantly lower WER scores for the FSM mode show that using a human-in-the-loop approach can improve accuracy in both classic FA and TFA. The manual tracking of the current sentence by an active reader aligned the speech to text at the sentence level, which simplifies the work of forced alignment after the fact. However, while some assistive reading modes clearly improve accuracy (FSM), others (like FST) seem to decrease it.

FST and FWT, metaphors that employed timers, caused users to speed up their reading over the course of a session and as reading time sped up, reading accuracy decreased. This suggests the auto-advancing function of the timed modes is going too fast for users and causing them to create more errors. Reading faster may also decrease the accuracy of alignment algorithms as readers will speak less clearly when rushed.

Though FSM has the same correlation between reading time and reading accuracy, this condition's higher rates of accuracy shows that this correlation can be interpreted in the other direction than the timed modes. That is, readers slowed their speed over the course of reading and increased accuracy. Readers were able to follow their own pace and minimize errors by not being rushed. As they were not rushed, their speech was likely clearer and therefore easier to align correctly.

TFA showed many of the same findings as classic FA. This is encouraging as TFA was designed to build on classic FA and the human-in-the-loop interactions to compensate for the differences in casual read-alouds. There were no clear and significant improvements in alignment, but the more normalized data, demonstrated by the lower standard deviations in each mode and the control case, show that some aspects of TFA work as intended. By checking neighbouring files for missing words and excluding off-text speech or sounds, TFA avoids large numbers of deletions or insertions that were not errors by the readers but natural variations in casual reading. This process seems to allow the data to more closely reflect the realities of casually reading aloud, and demonstrates how TFA handles differences in casually read-aloud speech better than classic FA. Overall, this suggests that leveraging touch interaction (where available) shows potential to improve on classic FA algorithms and better serve applications that are more natural and richer in an interaction context.

Where the classic FA results suggested that the timer negatively affected reading speed and WER over the course of a reading session, TFA only finds this to be true at the sentence level (FST). This shows that the word level conditions have little effect on reader behaviour and on the alignment of text to speech for forced alignment.

TFA also finds stronger evidence for FSM as the best metaphor for supporting reading as that is the only mode showing positive correlations with the control case in both WER and reading time. So, a reader using this mode can interact with the text and read similarly to how they would in the control case. Classic FA had shown similar correlations in reading time for word level modes as well. The changes implemented in TFA better interpreted outlier data in reading speed. These outliers may have implied a correlation with the control in classic FA that was not there.

The high correlations between reading time and WER in FST found for both algorithms demonstrate that readers using this condition made more errors and did not have as many acceptable variations in their reading, such as repeated insertions. Acceptable variations would have been recognized by TFA and weakened the WER and reading time correlation. The negative correlation between trends in WER and reading time for FST, also in both algorithms, showed that readers were rushed by the timer and accuracy decreased as they sped up. These findings together provide strong evidence that the poor accuracy in this mode is because of reader error and not because of the anticipated and acceptable reading differences that TFA accounts for.

The same high correlations between trends in reading time and WER are also present in both algorithms for FSM. However, this correlation can be interpreted in the opposite direction as FST since FSM was the most accurate mode and the only mode to show significant improvement. That is, the lower WERs support that readers slowed their speed over the course of reading and improved accuracy. The manual interaction of this condition allowed readers to follow their own reading pace, easily adjust is as they preferred, and aided them in reading the text correctly and clearly.

With the exception of FSM, TFA did not have a very strong effect on WER across the modes of interaction, as shown by the similarly high correlation between the algorithms. However, this same comparison for reading time had weaker correlations across modes than for the control, which had practically no change between algorithms. This suggests the changes in TFA successfully adjusted reading times to account for overlaps in the audio files before or after a given sentence. FWT has an even lower correlation than the other modes, implying that the overlaps in audio files were more extreme in this mode. Generally, these correlations between algorithms for all metaphors and their variation demonstrate how these interactions can affect alignment of speech data, positively or negatively.

While there are some improvements, TFA did not significantly improve WER accuracy in any of the assistive modes as compared to the base algorithm. We suspect this is due to the limitations of speech recognition when audio quality is lower. Even with the changes made specifically for this kind of speech, low audio quality and clipped words may have prevented TFA from reaping the benefits of the changes.

The human alignment, unsurprisingly, showed far more accurate WER scores than either algorithm. It does support the main findings of the effect of the assistive touch metaphors: FST is the least accurate across all three methods of WER scoring, and FSM has higher accuracy. Beyond this general support, as these results were significantly different than either of the forced alignment algorithms, the pairwise differences between the two machine algorithms were much smaller compared to the human scores. This is another example of the need for improvements in forced alignment and how there is still much work to be done in these spaces.

Limitations

The technology of the ALLT tablet has limitations in how the audio files are divided into sentence segments on the fly. This resulted in some missing or clipped audio data and may have generally affected the accuracy of any alignment and scoring. While this is not ideal, it is not uncommon for non-professionally recorded data to have inconsistencies.

Future Work

TFA is proposed as an initial exploration into the potential for adjusting existing tools to better suit amateur readers, and to raise awareness of the lack of alignment tools for this type of speech data. As such, it has not been designed with performance in mind. As TFA applies forced alignment multiple times for each file, there is clearly a performance cost. Future work should look to develop a stand-alone algorithm or variation on forced alignment building off the work done here, to create a better performing option.

In the future, the considerations for amateur readers presented here should be explored in a broader context than with the results of one study. Of particular interest will be whether the general changes of ignoring repeated insertions and accepting speech slightly off the given alignment (in this case, found in adjacent audio files) can generalize to other sources of amateur read-aloud speech.

CONCLUSION

The analysis presented here, shows that touch interaction can support the alignment of text to speech, especially for amateur readers in casual settings. This resulting alignment supports assessment using tools like forced alignment. A human-in-the-loop approach, where a human can track their current sentence (FSM), provides the best improvement to forced alignment algorithms. This is done by providing some initial alignment of text to speech from the reader's manual tracking interaction that also supported their reading accuracy.

We show the changes in TFA, that were designed to improve classic FA, account for the variability of amateur read-alouds, limit outlier data compared to classic FA, and build on assistive touch metaphors by demonstrating how these interactions affect reading in both quality and pace. As such, this work expands on the sources of speech that can be meaningfully assessed using existing methods of alignment, by incorporating the support of assistive touch metaphors to provide text and speech alignment.

ACKNOWLEDGEMENTS

This work was supported by AGE-WELL NCE Inc., a member of the Networks of Centres of Excellence (NCE), a Government of Canada program supporting research, networking, commercialization, knowledge mobilization and capacity building activities in technology and ageing to improve the quality of lives of Canadians. The authors also wish to acknowledge the sacred lands on which the University of Toronto operates. These lands are the traditional territories of the Huron-Wendat and Petun First Nations, the Seneca, the Haudenosaunee, and most recently, the Mississaugas of the Credit River. Today, the meeting place of Tkaronto is still the home to many Indigenous people from across Turtle Island, and we are grateful to have the opportunity to work in the community, on this territory.

REFERENCES

1. Xavier Anguera, Jordi Luque, and Ciro Gracia. 2014. Audio-to-text alignment for speech recognition with very limited resources. In *Fifteenth Annual Conference of the International Speech Communication Association*.
2. 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*, 440–443.
3. Jill Attewell and Carol Savill-Smith. 2004. Mobile learning and social inclusion: focusing on learners and learning. *Learning with mobile devices: research and development*: 3–11.
4. Gérard Bailly and William-Seamus Barbour. 2011. Synchronous reading: learning French orthography by audiovisual training. In *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, 1153–1156.
5. Valentina Bartalesi and Barbara Leporini. 2015. An Enriched ePub eBook for Screen Reader Users. In *International Conference on Universal Access in Human-Computer Interaction*, 375–386.
6. Joseph E. Beck and June Sison. 2006. Using knowledge tracing in a noisy environment to measure student reading proficiencies. *International Journal of Artificial Intelligence in Education* 16, 2: 129–143.
7. Norbert Braunschweiler and Langzhou Chen. 2013. Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS. In *Eighth ISCA Workshop on Speech Synthesis*.
8. Norbert Braunschweiler, Mark JF Gales, and Sabine Buchholz. 2010. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Eleventh Annual Conference of the International Speech Communication Association*.

9. Leen Cleuren, Jacques Duchateau, Alain Sips, Pol Ghesquière, and Hugo Van Hamme. 2006. Developing an automatic assessment tool for children's oral reading. In *Ninth International Conference on Spoken Language Processing*.
10. 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*, 272–275.
11. 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 *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 5155–5159.
12. Kathryn DR Drager and Joe E. Reichle. 2001. Effects of Discourse Context on the Intelligibility of Synthesized Speech for Young Adult and Older Adult ListenersApplications for AAC. *Journal of Speech, Language, and Hearing Research* 44, 5: 1052–1057.
13. Carrie Demmans Epp, Cosmin Munteanu, Benett Axtell, Keerthika Ravinthiran, Yomna Aly, and Elman Mansimov. 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*, 34.
14. Jean-Philippe Goldman. 2011. EasyAlign: an automatic phonetic alignment tool under Praat.
15. 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.
16. Timothy J. Hazen. 2006. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Ninth International Conference on Spoken Language Processing*.
17. Athanasios Katsamanis, Matthew Black, Panayiotis G. Georgiou, Louis Goldstein, and S. Narayanan. 2011. SailAlign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.
18. Simon King. 2014. Measuring a decade of progress in text-to-speech. *Loquens* 1, 1: 006.
19. Thea Knowles, Meghan Clayards, Morgan Sonderegger, Michael Wagner, Aparna Nadig, and Kristine H. Onishi. 2015. Automatic forced alignment on child speech: Directions for improvement. In *Proceedings of Meetings on Acoustics 170ASA*, 060001.
20. Benjamin Lecouteux, Georges Linares, Pascal Nocéra, and Jean-François Bonastre. 2006. Imperfect transcript driven speech recognition. In *InterSpeech*.
21. Nat Lertwongkhanakoola, Natthawut Kertkeidkachornb, Proadpran Punyabukkanac, and Atiwong Suchatod. 2014. An Automatic Real-time Synchronization of Live Speech with Its Transcription Approach. *ENGINEERING JOURNAL* 19, 5. Retrieved May 8, 2017 from http://engj.org/index.php/ej/article/view/703
22. Xiaolong Li, Li Deng, Yun-Cheng Ju, and Alex Acero. 2008. Automatic children's reading tutor on hand-held devices. In *Ninth Annual Conference of the International Speech Communication Association*.
23. Yan-Hua Long and Hong Ye. 2015. Filled Pause Refinement Based on the Pronunciation Probability for Lecture Speech. *PloS one* 10, 4: e0123466.
24. Yoshitaka Mamiya, Adriana Stan, Junichi Yamagishi, Peter Bell, Oliver Watts, Robert AJ Clark, and Simon King. 2013. Using Adaptation to Improve Speech Transcription Alignment in Noisy and Reverberant Environments. In *Eighth ISCA Workshop on Speech Synthesis*.
25. Michael Massimi, Rachelle Campigotto, Abbas Attarwala, and Ronald Baecker. 2013. Reading together as a Leisure Activity: Implications for E-reading. In *14th International Conference on Human-Computer Interaction (INTERACT)*, 19–36.
26. Petr Mizera, Petr Pollak, Alice Kolman, and Mirjam Ernestus. 2014. Impact of irregular pronunciation on phonetic segmentation of nijmegen corpus of casual czech. In *International Conference on Text, Speech, and Dialogue*, 499–506.
27. Pedro J. Moreno and Christopher Alberti. 2009. A factor automaton approach for the forced alignment of long speech recordings. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 4869–4872.
28. Pedro J. Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. 1998. A recursive algorithm for the forced alignment of very long audio segments. In *Fifth International Conference on Spoken Language Processing*.
29. Jack Mostow. 2012. Why and how our automated reading tutor listens. In *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, 43–52.
30. 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 *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*, 427–430.
31. Cosmin Munteanu, Heather Molyneaux, Julie Maitland, Daniel McDonald, Rock Leung, Hélène Fournier, and Joanna Lumsden. 2014. 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* 18, 6: 1455–1469.

32. 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.
33. Susan B. Neuman and David K. Dickinson. 2003. *Handbook of early literacy research*. Guilford Press.
34. Andrea Passerini and Michele Sebag. 2015. 4.2 Learning and Optimization with the Human in the Loop. *Constraints, Optimization and Data*: 21.
35. Kishore Prahallad and Alan W. Black. 2011. Segmentation of monologues in audio books for building synthetic voices. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 5: 1444–1449.
36. 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.
37. 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*, 2583–2592.
38. Kei Sawada, Shinji Takaki, Kei Hashimoto, Keiichiro Oura, and Keiichi Tokuda. 2014. Overview of NITECH HMM-based text-to-speech system for Blizzard Challenge 2014. In *Blizzard Challenge Workshop*.
39. 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. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2363–2372.
40. Ronanki Srikanth and Li Bo2 James Salsman. 2012. Automatic pronunciation evaluation and mispronunciation detection using CMUSphinx. In *24th International Conference on Computational Linguistics*, 61.
41. Richard K. Wagner, Andrea E. Muse, and Kendra R. Tannenbaum. 2007. *Vocabulary acquisition: Implications for reading comprehension*. Guilford Press.
42. Mirjam Wester, Matthew Aylett, Marcus Tomalin, and Rasmus Dall. 2015. Artificial personality and disfluency. In *Sixteenth Annual Conference of the International Speech Communication Association*.
43. 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 of Errors in Pronunciation Training, Stockholm, Sweden*.
44. Junbo Zhang, Fuping Pan, and Yonghong Yan. 2012. An LVCSR based automatic scoring method in English reading tests. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on*, 34–37.
45. LibriVox | free public domain audiobooks. Retrieved October 4, 2017 from https://librivox.org/