A Prototype Reading Coach that Listens

Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, and Matthew Kane

Project LISTEN, 215 Cyert Hall, Carnegie Mellon University Robotics Institute
4910 Forbes Avenue, Pittsburgh, PA 15213-3890
mostow@cs.cmu.edu

COPYRIGHT NOTICE

newpage

The following notice appears on page xiv of the proceedings:

AAAI-94
Outstanding Paper Award

A Prototype Reading Coach that Listens

Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, and Matthew Kane

This year, AAAI's National Conference on Artifical Intelligence honors a paper that exemplifies high standards in technical contribution and exposition. Papers were nominated for the Outstanding Paper Award by members of the program committee during the NCAI review process. These nominations were then reviewed once again by a smaller subset of the program committee to select the winning paper. Care was taken during the review process to ensure that our final decisions were based on the opinions of impartial readers who are free from personal biases and conflicts of interest.

A Prototype Reading Coach that Listens

Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, and Matthew Kane

Project LISTEN, 215 Cyert Hall, Carnegie Mellon University Robotics Institute
4910 Forbes Avenue, Pittsburgh, PA 15213-3890
mostow@cs.cmu.edu

Abstract[1]

We report progress on a new approach to combatting illiteracy -- getting computers to listen to children read aloud. We describe a fully automated prototype coach for oral reading. It displays a story on the screen, listens as a child reads it, and decides whether and how to intervene. We report on pilot experiments with low-reading second graders to test whether these interventions are technically feasible to automate and pedagogically effective to perform. By adapting a continuous speech recognizer, we detected 49% of the misread words, with a false alarm rate under 4%. By incorporating the interventions in a simulated coach, we enabled the children to read and comprehend material at a reading level 0.6 years higher than what they could read on their own. We show how the prototype uses the recognizer to trigger these interventions automatically.

1. Introduction

This paper is about a problem where even a partial solution would quickly pay back every dollar this nation has ever invested in artificial intelligence research. The problem is illiteracy. Its scope is widespread (NCES, 1993a, OTA, 1993). Its economic costs exceed $225 billion per year (Herrick, 1990). Its human and social costs are incalculable. Individuals with low reading proficiency are much likelier to be unemployed, poor, or incarcerated (NCES, 1993b).

Although a large body of software exists to teach reading, it is limited in its ability to listen and/or intervene. Most systems do not listen at all. Some systems try to help children anyway by providing speech output on demand (Wise et al, 1989, Roth & Beck, 1987, McConkie & Zola, 1987, Reitsma, 1988). This capability is now available in some commercial educational software, e.g., (Beck et al, 1987, Discis, 1991). However, young readers often fail to realize when they need such help (McConkie, 1990). Moreover, these systems cannot tap the unique motivation that listening to a reader can engender (Kantrov, 1991).

Other systems do listen, but use isolated word recognizers that cannot monitor the oral reading of connected text. Such systems have been used for reading (Kantrov, 1991, Cowan & Jones, 1991), speech training (Watson et al, 1989, Umezaki, 1993), and foreign language learning (Molholt, 1990).

More recently, some systems have used continuous speech recognition to detect errors in reading (Phillips et al, 1992, Mostow et al, 1993a) or pronunciation (Bernstein et al, 1990, Bernstein & Rtischev, 1991). However, the pedagogical interventions performed by published systems were either rudimentary or missing altogether.

Project LISTEN is addressing these various limitations by **adapting continuous speech recognition to listen to children read connected text, automatically triggering pedagogically appropriate interventions.** We present evidence for the claim that these interventions are both **pedagogically effective** to perform, and **technically feasible** to automate.

The rest of this paper is organized as follows. Section 2 describes the interventions performed by our prototype oral reading coach, which we have named after Emily Latella (a character on *Saturday Night Live* created by the late Gilda Radner and known for her difficulties in distinguishing among words that sound alike). Section 3 describes the speech analysis required to make Emily work. Section 4 concludes.

2. Emily's interventions

Emily is designed to help a child read and comprehend a given story. (One can imagine alternative goals, such as correcting pronunciation or giving explicit instruction in phonics.) Emily is intended to maintain a fluent, pleasant reading experience that gives the child practice in reading connected text, plus enough assistance to be able to comprehend it. It therefore uses a combination of reading and listening which we have named "shared reading," in which the child reads wherever possible, and the coach helps wherever necessary.

Emily intervenes when the reader misreads one or more words in the current sentence, gets stuck, or clicks on a word to get help. We do <u>not</u> treat hesitations, sounding out, false starts, self-corrections, or other insertions as misreading; by "misread," we mean "fail to speak the correct word" (though see Section 3.1). Emily's current set of interventions targets two obstacles that interfere with children's reading comprehension (Curtis, 1980).

First, young readers often have trouble identifying printed words. Some of Emily's interventions are therefore primarily intended to assist **word identification**:

- Retry a misread word by highlighting it and asking the child to reread it. This intervention

[1]This research was supported primarily by the National Science Foundation under Grant Number MDR-9154059, and the Defense Advanced Research Projects Agency, DoD, through DARPA Order 5167, monitored by the Air Force Avionics Laboratory under contract N00039-85-C-0163, with additional support from the Microelectronics and Computer Technology Corporation (MCC). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the sponsors or of the United States Government.

For a short summary of Project LISTEN, see (Hauptmann et al, 1994).

prompts the child to attend more carefully to the word, and signals that the first attempt may have been incorrect.

- Recue or "jumpstart" the last misread word by speaking the text that leads up to it, and then flashing the word to prompt the child to reread it. The jumpstart serves to put the child back in the context where the word occurred, which may help in identifying it. However, this intervention does not apply if the word occurs near the beginning of the sentence.
- Speak a word if the child gets stuck on that word.
- Speak a word if the child clicks the mouse on it.
- Speak a word after a retry or recue. This feedback is confirmatory if the child's second try was correct, and corrective if it was not.

Second, struggling readers spend so much of their attention figuring out the words that even when they get the words right, they may still not comprehend the overall meaning. Emily's other interventions address this **attentional bottleneck**:

- To avoid disrupting the flow of reading, ignore a misread word if it is on a list of 52 common function words unlikely to affect comprehension.
- Speak the entire sentence if the child misreads three or more words in it, or misreads a word after a retry or recue. Either condition means the child is unlikely to have comprehended the sentence. Hearing the sentence frees the child to focus on comprehension (Curtis, 1980).

For both pedagogical and technical reasons, Emily waits to intervene until the end of the sentence, unless the reader gets stuck or clicks for help. The pedagogical reasons are to give the reader a chance to self-correct, and to avoid disrupting the flow of reading. The technical reasons are that Emily cannot gracefully interrupt the reader, both because its speech recognizer lags too far behind to respond instantaneously, and because it lacks the subtle nonverbal cues that humans use to interrupt each other.

To finesse the interruption problem, Emily displays the text incrementally, adding one sentence at a time. When the reader reaches the end of the sentence, Emily has an opportunity to intervene without having to interrupt. It does not display the next sentence until it has completed any such interventions.

For natural speech quality, Emily normally outputs predigitized human speech. However, two synthesized voices (ORATOR™ (Spiegel, 1992) and DecTalk™ (DEC, 1985)) are available as alternatives.

To evaluate and refine these interventions while we were still working on the speech analysis, and independently of recognition accuracy, we developed a simulated coach that appeared automatic to the subjects, but was controlled behind the scenes by a human experimenter, as shown in (Mostow et al, 1993b). To design the interventions, we used the following development process:

1. Observe individual reading assistance provided by human experts.
2. Select the most frequent interventions that seem feasible to automate.
3. Codify interventions as written instructions for the human experimenter.
4. Implement interventions as actions the experimenter selects from a menu.
5. Automate the triggers for the interventions.

At this point, the experimenter's role consisted of listening to the reader, following along in the text, and marking each word as correct or misread. The rest of the simulated coach was automatic, and used the marking information to trigger its interventions.

2.1. Pedagogical evaluation

We performed a pilot study to test the overall effectiveness of our interventions. Another purpose of this experiment was to refine our interventions and experimental protocols before performing larger scale studies with more subjects and subtler effects.

Hypothesis: Our hypothesis was that these interventions would enable struggling readers to read and comprehend material significantly more advanced than what they could read on their own. Therefore we selected as our subjects 12 second graders at an urban public school in Pittsburgh who had been identified by their reading teachers as having problems with reading.

Dependent variables: To minimize the effect of inter-subject variability, we compared three conditions for each subject. The control condition measured their <u>independent reading level</u>, that is, the level of material they could read and comprehend without assistance. The experimental condition measured their <u>coach-assisted reading level</u>, that is, the level of material they could read and comprehend by using the coach. A third condition measured their <u>"potential" reading level</u>, that is, the level of material they could comprehend when it was read aloud to them.

Method: To measure these three levels, we adapted materials and procedures from a widely used test of oral reading (Spache, 1981). This test includes one-page passages at carefully calibrated grade levels ranging from early first grade to mid-seventh-grade. Each passage has an accompanying list of comprehension questions. For obvious reasons, once a subject read a passage, it was "contaminated" and could not be reused for that subject in the other conditions. Fortunately (Spache, 1981) has two complete series of passages. Therefore we used one series to determine each subject's independent reading level, and the other to determine his or her assisted reading level. To measure reading level in a given condition, we presented successively higher passages until the subject exceeded a limit on the number of oral reading errors or failed over 40% of the comprehension questions. The subject's reading level for that condition was then defined as the grade level of the previous passage. The subject then listened to the subsequent passages until his or her comprehension score dropped below 60%. We defined potential reading level as the level of the highest passage successfully comprehended.

To avoid confounding effects, we randomized the order of the subjects and counterbalanced both the order of the control and experimental conditions, and the choice of passage series for each condition. We recorded the children at their school in November 1993 (month 3 of the school year), taking them one at a time out of their regular class to a separate room. Whenever subjects exceeded their attention span or got restless, we excused them and continued the session on the next day of school.

Apparatus: The apparatus for the experiment consisted of a NeXT workstation with two monitors, one for the subject and the other for the human experimenter. A color monitor was used to display the text and interventions to the subject. To avoid unnecessary variability, all the passages, spoken interventions, and comprehension questions were digitally prerecorded in a pleasant female voice. The experimenter used the keyboard, mouse, and second monitor to select the passage to display, mark each word as correct or misread, and administer the comprehension questions. The subject was given a button to push for help on the current word. The button simply operated a flashlight that signalled the human experimenter, who then selected the appropriate menu item. This configuration avoided the need to train the subjects to operate a mouse, and allowed us to run the experiments on a single workstation.

Data: We digitally recorded the children's oral reading, using a Sennheiser noise-cancelling headset microphone to keep the speaker's mouth an appropriate distance from the microphone, and to filter out some of the noise typical of a school environment. "Event files" captured every action performed by the experimenter or the system in response to the subjects' oral reading. (See Figure 2-1.) We also recorded the results of the comprehension tests, including which specific questions were answered correctly.

Key Results: The outcome of this experiment supported our hypothesis. The subjects' assisted reading level was higher than their independent reading by an average of 0.6 years (2.7 vs. 2.1). This effect was statistically significant at the 99% level.

The interventions also dramatically reduced the frustration experienced by the children in their effort to read. When they used the coach, our subjects misread only 2.6% of the words. Without assistance, they misread 12.3% of the words on passages of matched difficulty; anything over 10% indicates that the reading material is too difficult (Betts, 1946, Vacca et al., 1991).

Based on a study (Curtis, 1980) of similar students reading the same materials, we expected that listening comprehension would be about two years higher than independent reading level; instead, we found that it was slightly (though not significantly) lower than the coach-assisted level. We observed that when we asked the subjects to listen to an entire story, their attention wandered, perhaps because they lacked a natural visual focus such as a talking face.

```
#:   TIME:    EVENT:    TEXT WORD:
```
At time 1179648, measured in samples (16,000 per second) of the digitized oral reading, the coach displays "Spotty thought he had caught a black and white kitten":
```
78>  1179648  NEXTSEN   Spotty#49
79>  1239040  OK        Spotty#49
```
After hesitating 4 seconds on "thought", the child pushes the help button.
```
80>  1306624  SAYWORD 4 thought#50
81>  1314816  OK        thought#50
82>  1325056  OK        he#51
83>  1337344  OK        had#52
```
The child misreads "caught":
```
84>  1384448  MARK      caught#53
85>  1400832  OK        a#54
86>  1417216  OK        black#55
87>  1429504  OK        and#56
88>  1439744  OK        white#57
89>  1458176  OK        kitten#58
90>  1458176  START_EOS .#58
91>  1458176  NUM_ERRS 1
```
The coach recues "caught":
```
92>  1458176  GOMARK    caught#53
93>  1458176  JUMPSTART caught#53
94>  1458176  JUMPEND   caught#53
```
The child misreads it again...
```
95>  1458176  MARK      caught#53
```
... so the coach speaks it:
```
96>  1458176  THISWORD  caught#53
97>  1458176  END_EOS
```

Figure 2-1: Annotated excerpt from an event file

Our analysis of the data suggested how our interventions might be made more effective. We found that the coach read fewer sentences to the subjects than it should, because thanks to the help button they hardly ever misread three or more words in one sentence. We plan to make the trigger for this intervention sensitive to reader hesitations that may indicate comprehension difficulties.

3. Speech analysis

Unlike conventional speech recognition, whose goal is to guess what the speaker says, Emily has a discrimination task, whose goal is to find where the speaker deviates from the text. It can also be viewed as a classification task, whose goal is to classify each word of text as correctly read or not. This task is easier than recognition in that it does not require identifying what the speaker said instead, but it is harder in that the speaker's deviations from the text may include arbitrary words and non-words.

Thus the interventions in Section 2 require the following speech analysis capabilities:

1. Given a starting point in the text and a

possibly disfluent reading of it, detect which words of text were misread. The starting point may be the beginning of a sentence, a word the reader selected for help, or a word the reader is asked to reread.

2. Detect when the reader reaches the end of a given fragment of text. This fragment may be the current sentence or a word to reread.

3. Detect when the reader gets stuck.

We now describe how Emily implements these capabilities.

Emily consists of two basic components -- an intervenor that runs on a color NeXT workstation and interacts with the reader, and a speech recognizer that runs on a DEC 3000 or HP 735. The intervenor tells the recognizer where in the text to start listening -- either at the beginning of a new sentence, after a word spoken by the coach, or at a word the coach has just prompted the reader to reread. Four times a second, the recognizer reports the sequence of words it thinks it has heard so far. Capability 1 is implemented by aligning the output of the recognizer against the text. Capability 2 is implemented by checking if the recognizer has output the last word of the fragment. Capability 3 is implemented by a time limit for progressing to the next word in the text; the intervenor assumes that the reader is stuck on this word if the time limit is exceeded without a previously unread word appearing in the recognizer output. The intervenor is invoked whenever the reader reaches the end of a sentence, gets stuck, or clicks the mouse on a word for help.

The speech recognizer, named Sphinx-II (Huang et al, 1993), requires three types of knowledge -- **phonetic**, **lexical**, and **linguistic** -- as well as several parameters that control its Viterbi beam search for the likeliest transcription of the input speech signal. The recognizer evaluates competing sequences of lexical symbols based on the degree of acoustic match specified by its **phonetic models**, the pronunciations specified by its **lexicon**, and the *a priori* probability specified by its **language model**. Thus Sphinx-II's recognition accuracy is limited by how well these three representations model the speech input. These representations must approximate the broad range of speech phenomena contained in disfluent reading, which include omission, repetition, and hesitation, as well as substitution and insertion of words, non-words, and non-speech sounds. These phenomena (especially words and non-words outside the vocabulary used in the text) compound the variability that makes connected speech recognition so difficult even for fluent speech.

(Mostow et al, 1993a) assumed that phonetic models trained only on female speakers would work better for children's speech because of its high pitch. However, we found that models trained on combined male and female speech seemed to work just about as well. Therefore Emily uses 7000 phonetic Hidden Markov Models trained on 7200 sentences read by 84 adult speakers (42 male and 42 female), though we can also run it on the male-only and female-only models. To retrain these models from scratch, we need to collect and transcribe a much larger corpus of children's oral reading. In the meantime, we plan to adapt the adult phonetic models to work better on children's speech by using an interpolative training method.

The recognizer's accuracy at detecting misread words depends on its ability to model deviations from correct reading. We model several different phenomena of oral reading in Emily's lexicon (illustrated in Figure 3-1) and language model, which are automatically generated from a given text, such as "Once upon a time a...."

Subscripts denote word numbers:

$\texttt{Once}_1$	W AH N S
Alternate pronunciations are parenthesized:	
$\texttt{TRUNCATION}_1\texttt{(W)}$	W
$\texttt{TRUNCATION}_1\texttt{(W AH)}$	W AH
$\texttt{upon}_2$	AX P AO N
$\texttt{TRUNCATION}_2\texttt{(AX)}$	AX
$\texttt{TRUNCATION}_2\texttt{(AX P)}$	AX P
$\texttt{a}_3$	AX
$\texttt{time}_4$	T AY M
$\texttt{TRUNCATION}_4\texttt{(T)}$	T
$\texttt{a}_5$	AX

Figure 3-1: Lexicon for "Once upon a time a..."

To model **correct reading**, we include the text words themselves in the lexicon, numbering them to distinguish among multiple occurrences of the same word (e.g., "a_3" vs. "a_5"). Each word's pronunciation, represented as a sequence of *k* phonemes, is taken from a general English dictionary. If not found there, it is computed by the pronunciation component of a speech synthesizer, such as MITalk (Allen et al, 1987) or ORATOR[TM] (Spiegel, 1992). In our language model, each word w_{i-1} (e.g., "Once_1") is followed with probability .97 by the correct next word w_i ("upon_2").

To model **repetitions** and **omissions**, word w_{i-1} (e.g., "Once_1") is followed with probability $.01/(n-1)$ by any word w_j of the other $n-1$ words in the same sentence. A non-uniform probability would be more realistic, but can cause problems, as discussed later. Repetitions and omissions correspond respectively to jumps backward ($j < i$, e.g., back to "Once_1") and forward ($j > i$, e.g., to "a_3").

To model **false starts** and **near misses**, we include a truncation symbol TRUNCATION_i for each text word w_i. Besides modelling actual truncations of the word, these pronunciations approximate many phonetically similar substitution errors. The truncation symbol TRUNCATION_i follows the word w_{i-1} with probability .02. For the example text in Figure 3-1, this model assigns a probability of 2% to the prediction that after reading the word "Once_1", the reader will next truncate the word "upon_2." We give this symbol $k-2$ alternate pronunciations, consisting of proper prefixes of the complete pronunciation, as illustrated in Figure 3-1. We found that including truncations where only the last phone is omitted, e.g., "AX P AO", seemed to cause recognition errors, especially for speakers of dialects that tend to drop the last phone.

To model **repeated attempts**, **self-corrections**, and **substitution errors**, respectively, each truncation symbol TRUNCATION$_i$ is followed with equal probability by itself, by the complete word w_i, or by the following word w_{i+1}. That is, after truncating the word "upon$_2$", the reader is considered equally likely to truncate it again, read it correctly, or go on to the word "a$_3$."

This language model reflects some lessons from previous experience. First, although the words in the lexicon are intended to model correct reading, in practice **words are often used to model deviations.** For example, if the word "elephant" is not in the lexicon, it is liable to be recognized as the sequence "and of that." Anyone who designs a language model for this task without anticipating this phenomenon is liable to be surprised by the results.

Second, the ability to detect deviations depends on having a **phonetically rich repertoire** of symbols for matching them. The word-only lexicon used in (Hauptmann et al, 1993) was surprisingly successful despite its limitations because it included all the words in an entire passage, which was enough to provide considerable phonetic variety.

Third, **over-constrained search can impede error recovery.** One of our earlier language models tried to exploit the characteristic structure of disfluent oral reading. It assigned low or zero probabilities to transitions that children seldom take, such as long jumps. These probabilities were estimated from the transcribed oral reading corpus described in (Mostow et al, 1993a). We expected that this model would produce more accurate recognition than simpler ones, but we have not (yet) succeeded in making it do so. We suspect the reason is that when the recognizer follows a garden path, this more constrained model makes it difficult to recover. For example, suppose the reader says "Oncet upon a time," and the recognizer recognizes "Oncet" as "Once$_1$ a$_3$ TRUNCATION$_4$(T)." To recover from this garden path, the recognizer must be able to jump to "upon$_2$" without incurring an excessive penalty (low probability) from the language model; otherwise it may misrecognize "upon a time" as "a$_5$...." Since Emily's phonetic models and lexicon can only crudely approximate the virtually infinite range of speech sounds produced by disfluent young readers, it appears impossible to keep the recognizer from starting down such garden paths. Therefore the language model must be designed to recover from them as quickly as possible. That is, since we cannot prevent recognition errors from occurring at all in these cases, we must instead try to minimize their extent.

3.1. Accuracy

We evaluated Emily's accuracy off-line on 514 sentences read by 15 second graders as they used the simulated reading coach described in Section 2. (To avoid testing on our training data, we were careful in developing our language model and tuning parameter values to use a separate set of 457 sentences by 30 second graders from a different school.) Our test utterances averaged 10 text words in length and 15 seconds in duration, including 5 seconds of silence due to struggling readers' frequent hesitations. The readers misread only 1.6% of the words in this corpus; we attribute this low rate partly to the help button, which they used on 6% of the words, and partly to how we operationalized "misread."

We relied on the human experimenter to flag misread words in real-time, causing the simulated coach to record "MARK" in the event file. (An UNMARK command allowed self-corrections.) This scheme was faster and cheaper than conventional detailed transcriptions, especially since disfluent reading is difficult to transcribe. Moreover, it solved the sticky problem of when to consider a word misread -- we simply told the experimenter to follow the instructions in (Spache, 1981), which caution against treating dialect substitutions and minor mispronunciations (e.g. "axe" for "ask") as reading errors.

Our purpose in evaluation was to measure the ability of our recognizer to trigger the coach's interventions. Therefore in computing the list of words Emily treated as misread, we filtered out the same function words that the coach ignored. This step substantially reduced the incidence of false alarms (correct words treated as misread), since these function words were rarely misread by the reader but were often misrecognized by the recognizer. Similarly, we ignored misreadings of words where readers used the help button, both because the coach does not require them to echo these words (though they often do), and because our digital recording apparatus often failed to record the beginnings of these words, since it stops recording during the coach's spoken interventions, and there is a slight delay before it resumes.

The evaluation results according to this methodology are shown in Table 3-1. Emily's sensitivity in classifying words as misread or correct is demonstrated by the fact that its detection rate is significantly (over 10 times) greater than its false alarm rate.

For comparison, we reanalyzed the results for Emily's predecessor, named Evelyn (Mostow et al, 1993a, Hauptmann et al, 1993). These results were obtained for a corpus of children's oral reading that was similar except that each utterance was an entire Spache passage, read without assistance. They were computed by averaging the individual accuracies on each passage, which reduced the effect of the passages where most of the recognition errors occurred. Without such averaging, Evelyn's detection rate was lower than Emily's. The difference is not significant with respect to the $\pm 2\sigma$ confidence intervals, which are wide because so few words were misread. However, Evelyn's false alarm rate was significantly (over three times) worse than Emily's.

We attribute Emily's higher accuracy to several factors. First, Evelyn was evaluated based on a different, more literal criterion, which treated any word not spoken exactly correctly as "missed." In contrast, Emily was evaluated

Table 3-1: Comparative Accuracy in Detecting Misread Words

System:	Corpus:	Definition of Misread Words:	Detection Rate:	False Alarm Rate:
Evelyn	99 passages	all substitutions and omissions	$.370 \pm .093$ (40 of 108)	$.126 \pm .010$ (567 of 4516)
Emily	514 sentences	only pedagogically relevant errors	$.488 \pm .110$ (40 of 82)	$.0366 \pm .0053$ (187 of 5106)

Detection rate = (misread words detected) / (words misread); false alarm rate = (false alarms) / (words read correctly)

Confidence intervals shown are $\pm 2\sigma$, where $\sigma = \sqrt{\frac{p \cdot (1-p)}{n}}$ is the standard error for rate *p*, sample size *n*

based on the more pedagogically relevant criterion applied by the experimenter who flagged words as "misread." We have now transcribed enough of our corpus to compare these two schemes. Almost no correctly read words were erroneously flagged by the experimenter, but for every word flagged as misread, several minor substitutions (such as adding or dropping a plural ending) were not flagged. Treating such near-miss substitutions as reading errors would erode Emily's detection rate. However, this difference in criteria does not account for Emily's much lower false alarm rate.

Second, Evelyn was evaluated on a corpus of page-long passages, and its language model had equiprobable transitions to any word on the page other than the next word in the text. In contrast, Emily recognizes one sentence at a time rather than a complete passage, and its language model preserves state by avoiding transitions out of the current sentence.

Third, filtering out function words cut Emily's false alarm rate by roughly half.

Fourth, Emily's richer lexicon enables it to model non-text-words using truncations, not just sequences of other words. We plan to further enrich the lexicon based on analysis of the transcribed oral reading and of Emily's recognition errors. We also need to optimize both the heuristic probabilities used in our language model, and the various input parameters to Sphinx-II.

Emily embodies a somewhat "lenient" tradeoff between detection and false alarms. The 97% transition probability between successive words of the text represents a strong expectation of correct reading, which can be overcome only by compelling acoustic evidence. A weaker bias would improve detection but increase the false alarm rate. But Emily's false alarms already outnumber misread words (187 to 82 on our test corpus). The reason is that even poor readers misread fewer than 10% of the words if the material is appropriate to their reading level (Betts, 1946, Vacca et al., 1991). To avoid swamping the reader with unnecessary interventions, we must reduce false alarms.

To help diagnose Emily's recognition errors, we split up the utterances into four subsets based on whether the reader pushed the help button (which tended to corrupt the recording) and/or misread a word (which indicated disfluency). We found that each of these factors multiplied the false alarm rate by about 1.5, reaching 6.1% on utterances with both SAYWORD and MARK events, compared to 2.8% on utterances with neither. The detection rate was insignificantly better on the utterances with no SAYWORDs (53% vs. 46%).

It is important to point out that speech recognition errors in this domain are not devastating. Some errors are masked by the interventions. For example, if Emily correctly detects that three or more words were misread, it will reread the sentence to help the reader comprehend it -- even if it is wrong about which words were misread. At worst, failure to detect a misread word merely loses one opportunity for corrective feedback. Conversely, false alarms merely slow down the flow of reading by asking the student to reread text unnecessarily. In practice, they encourage clearer enunciation. At worst, they may irritate the student if they become too frequent.

We have not yet used recorded speech to measure Emily's ability to detect when the reader reaches the end of the sentence or gets stuck. Such a test would need to determine how often, given the recorded reading, Emily would have intervened within an acceptable delay.

3.2. Other Improvements

Speed: Emily's speech processing consists of some signal processing performed on a NeXT in close to real time, plus a beam search performed on a more powerful machine (DEC 3000 or HP 735). In our off-line evaluation, this search was consistently faster than real time, averaging roughly 50% times real time. In contrast, Evelyn's search took 1-2 times real time on the same machine (Mostow et al, 1993a). We attribute this two- to four-fold speedup to Emily's sentence-based language model.

Flexibility: The Sphinx-II recognizer used in (Mostow et al, 1993a) and (Hauptmann et al, 1993) required a separate language model for each passage. This requirement precluded interrupting the reading before the end of the passage, because there was no way to tell the recognizer where to resume listening other than at the beginning of the passage.

We overcame this limitation by modifying Sphinx-II to accept a starting point as a parameter. Simply by changing its starting point, Emily can listen to one sentence at a time, resume listening in mid-sentence after the reader clicks on a word, jump back to listen to the reader retry a misread word after an intervention, or switch to another story.

We plan to further improve Emily's flexibility by reimplementing its language model to take constant space, instead of space proportional to (or even quadratic in) the total amount of text. Eliminating the need to reload language models for different text could enable Emily to monitor oral reading of text generated on the fly.

4. Conclusion

Emily **improves in measurable ways** on previously published attempts to use connected speech recognition to monitor and assist oral reading. First and foremost, it provides meaningful assistance, using interventions that (when implemented in our simulated coach) enabled struggling second graders to read material 0.6 years more advanced than they could on their own, and with much less frustration. Second, its detection rate for misread words is higher, its false alarm rate three times lower, and its search phase two to four times faster, than the system in (Mostow et al, 1993a).

These results are based on a number of **conceptual contributions**. First, Emily's interventions were derived from a combination of theory, expertise, and experiment. They embody an interesting new type of human-machine interaction -- shared reading -- and express in machine-applicable form some basic rules for helping children read. Second, Emily's language model, sentence-based processing, and mechanism for multiple starting points have improved the automated analysis of oral reading. Third, the development process by which Emily was designed, with its parallel interacting tracks for the interventions and the speech analysis required to support them, may serve as a useful model for other multidisciplinary applications.

Finally, the lessons we have learned from this work reflect the scientific value of the reading assistance task as a **carrier problem for research on two-way continuous speech communication with machines**. Because the computer knows the text the reader is trying to read, the speech analysis is tractable enough to study some limited but natural forms of such interaction now, without waiting for real time recognition of unconstrained spontaneous speech to become feasible.

One lesson is the **identification of novel criteria for evaluating speech recognition accuracy.** The conventional off-line evaluation criteria take the endpoint of each utterance as a given. These criteria do not test the ability to detect when the reader reaches the end of a sentence or gets stuck. In general, two-way speech communication will require a way to decide when the speaker is done speaking.

Another lesson is that **perplexity can be less important than recovery.** Perplexity measures the language model's average uncertainty about what the speaker will say next at any given point. Lower perplexity normally leads to higher accuracy. But when we tried to reduce perplexity by modelling patterns of disfluent oral reading, detection accuracy actually fell, apparently because the very constraints that reduced the perplexity of the language model impeded its ability to recover from garden paths. Further work is needed to test this hypothesis, define a formal measure of a language model's error recovery ability, and analyze how it affects recognition accuracy.

We are now trying out Emily on children and modifying its interventions to tolerate errors by the speech recognizer, as illustrated in our video of Emily in action (Mostow et al, 1994a, Mostow et al, 1994b). We hope to test soon how well the fully automated coach helps children read. But our longer-term goal is to scale up the coach to help children learn to read on their own.

Acknowledgements

We thank our principal reading consultant Leslie Thyberg; Raj Reddy and the rest of the CMU Speech Group (Filleno Alleva, Bob Brennan, Lin Chase, Xuedong Huang, Mei-Yuh Hwang, Sunil Issar, Fu-hua Liu, Chenxiang Lu, Pedro Moreno, Ravi Mosur, Yoshiaki Ohshima, Paul Placeway, Roni Rosenfeld, Alex Rudnicky, Matt Siegler, Rich Stern, Eric Thayer, Wayne Ward, and Bob Weide) for Sphinx-II; Adam Swift for programming; Paige Angstadt, Morgan Hankins, and Cindy Neelan for transcription; Maxine Eskenazi for transcript analysis; Lee Ann Kane for her voice; Murray Spiegel and Bellcore for ORATOR™; Dave Pisoni and Digital Equipment Corporation for DecTalk™; CTB Macmillan/McGraw-Hill for permission to use copyrighted reading materials from George Spache's *Diagnostic Reading Scales*; the students and educators at Colfax Elementary School, East Hills Elementary School, Turner School, and Winchester Thurston School for participating in our experiments; and many friends for advice, encouragement, and assistance.

References

Allen, J., Hunnicutt, S. and Klatt, D.H. (1987). *From Text to Speech: The MITalk system*. Cambridge, UK: Cambridge University Press.

I. L. Beck, S. F. Roth, and M. G. McKeown. (1987). *Word-Wise*. Allen, Texas: Developmental Learning Materials, Educational software for reading.

J. Bernstein and D. Rtischev. (1991). A voice interactive language instruction system. *Proceedings of the Second European Conference on Speech Communication and Technology (EUROSPEECH91)*. Genova, Italy.

J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. (1990). Automatic evaluation and training in English pronunciation. *International Conference on Speech and Language Processing (ICSLP-90)*. Kobe, Japan.

E. A. Betts. (1946). *Foundations of Reading Instruction*. New York: American Book Company.

H. Cowan and B. Jones. (September 1991). Reaching students with reading problems; Optimum Research Reading Program, The Sentence Master, Autoskill CRS evaluation. *Electronic Learning*, Vol. *11*(1).

M. E. Curtis. (1980). Development of components of reading skill. *Journal of Educational Psychology, 72*(5), 656-669.

Digital Equipment Corporation. (1985). *DecTalk: A Guide to Voice*. Maynard, MA: Digital Equipment Corporation.

Discis Knowledge Research Inc. *DISCIS Books*. 45

Sheppard Ave. E, Suite 802, Toronto, Canada M2N 5W9. Commercial implementation of Computer Aided Reading for the MacIntosh computer.

A. G. Hauptmann, L. L. Chase, and J. Mostow. (September 1993). Speech Recognition Applied to Reading Assistance for Children: A Baseline Language Model. *Proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH93)*. Berlin.

A. G. Hauptmann, J. Mostow, S. F. Roth, M. Kane, and A. Swift. (March 1994). A Prototype Reading Coach that Listens: Summary of Project LISTEN. C. Weinstein (Ed.), *Proceedings of the ARPA Workshop on Human Language Technology*. Plainsboro, NJ: ARPA SISTO, Morgan Kaufmann Publishers, Inc..

E. Herrick. (1990). *Literacy Questions and Answers*. Pamphlet. P. O. 81826, Lincoln, NE 68501: Contact Center, Inc. A private non-profit organization that operates the National Literacy Hotline ((800)228-8813).

X. D. Huang, F. Alleva, H. W. Hon, M. Y. Hwang, K. F. Lee, and R. Rosenfeld. (April 1993). The SPHINX-II speech recognition system: An overview. *Computer Speech and Language*, *7*(2), 137-148.

I. Kantrov. (1991). *Talking to the Computer: A Prototype Speech Recognition System for Early Reading Instruction* (Tech. Rep.) 91-3. Education Development Center, 55 Chapel Street, Newton, MA 02160: Center for Learning, Teaching, and Technology.

G. W. McConkie. (November 1990). Electronic Vocabulary Assistance Facilitates Reading Comprehension: Computer Aided Reading. Unpublished manuscript.

G. W. McConkie and D. Zola. (1987). Two examples of computer-based research on reading: Eye movement tracking and computer aided reading. In D. Reinking (Eds.), *Computers and Reading: Issues for Theory and Practice*. New York: Teachers College Press.

G. Molholt. (February-April 1990). Spectographic analysis and patterns in pronunciation. *Computers and the Humanities*, *24*(1-2), 81-92.

J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth. (July 1993). Towards a Reading Coach that Listens: Automated Detection of Oral Reading Errors. *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93)*. Washington, DC, American Association for Artificial Intelligence.

J. Mostow, S. Roth, A. Hauptmann, M. Kane, A. Swift, L. Chase, and B. Weide. (August 1993). Getting Computers to Listen to Children Read: A New Way to Combat Illiteracy (7-minute video).

J. Mostow, S. Roth, A. Hauptmann, M. Kane, A. Swift, L. Chase, and B. Weide. (August 1994). A Reading Coach that Listens (6-minute video). *Video Track of the Twelfth National Conference on Artificial Intelligence (AAAI94)*. Seattle, WA, American Association for Artificial Intelligence.

J. Mostow, S. Roth, A. Hauptmann, M. Kane, A. Swift, L. Chase, and B. Weide. (August 1994). A reading coach that listens: (edited) video transcript. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI94)*. Seattle, WA.

National Center for Education Statistics. (September 1993). *NAEP 1992 Reading Report Card for the Nation and the States: Data from the National and Trial State Assessments* (Tech. Rep.) Report No. 23-ST06. Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (September 1993). *Adult Literacy in America* (Tech. Rep.) GPO 065-000-00588-3. Washington, DC: U.S. Department of Education.

Office of Technology Assessment. (July 1993). *Adult Literacy and New Technologies: Tools for a Lifetime* (Tech. Rep.) OTA-SET-550. Washington, DC: U.S. Congress.

M. Phillips, M. McCandless, and V. Zue. (September 1992). *Literacy Tutor: An Interactive Reading Aid* (Tech. Rep.). Spoken Language Systems Group, 545 Technology Square, NE43-601, Cambridge, MA 02139: MIT Laboratory for Computer Science.

P. Reitsma. (1988). Reading practice for beginners: Effects of guided reading, reading-while-listening, and independent reading with computer-based speech feedback. *Reading Research Quarterly*, *23*(2), 219-235.

S. F. Roth and I. L. Beck. (Spring 1987). Theoretical and instructional implications of the assessment of two microcomputer programs. *Reading Research Quarterly, 22*(2), 197-218.

G. D. Spache. (1981). *Diagnostic Reading Scales*. Del Monte Research Park, Monterey, CA 93940: CTB Macmillan/McGraw-Hill.

M. F. Spiegel. (January 1992). *The Orator System User's Manual - Release 10*. Morristown, NJ: Bell Communications Research Labs,

T. Umezaki. (1993). *Talking Trainer*. Japan: Gakken, (In Japanese). Educational software for training deaf children to speak.

J. A. L. Vacca, R. T. Vacca, and M. K. Gove. (1991). *Reading and Learning to Read (Second Edition)*. Harper Collins.

C. S. Watson, D. Reed, D. Kewley-Port, and D. Maki. (1989). The Indiana Speech Training Aid (ISTRA) I: Comparisons between human and computer-based evaluation of speech quality. *Journal of Speech and Hearing Research, 32,* 245-251.

B. Wise, R. Olson, M. Anstett, L. Andrews, M. Terjak, V. Schneider, J. Kostuch, and L. Kriho. (1989). Implementing a long-term computerized remedial reading program with synthetic speech feedback: Hardware, software, and real-world issues. *Behavior Research Methods, Instruments, & Computers, 21*, 173-180.