

# Overcoming Technical Barriers to a Speech-enabled Children's Reading Tutor

Peter Fairweather

Don Nix

Dan Oblinger

Bill Adams

Carla Laffra

{peterf,nix,oblinger,badams,carlal}@watson.ibm.com

IBM T.J. Watson Research Center

Yorktown Heights, New York

## Abstract

This paper discusses technical innovations needed to realize a practice environment to help children learn to read aloud. We describe the construction of a child-appropriate acoustic model structured as a hidden Markov model and built on an extensive collection of language samples. We further describe how this acoustic model was enhanced to discriminate among phonetically similar targets by including probable miscues in the active vocabulary. We also describe how we adapted traditional point-and-click interfaces to operate within a system where judgments of accuracy could not be guaranteed to be without error. These enhancements resulted in a commercially available set of systems that provide emergent readers with oral reading practice opportunities with feedback as well as a means to demonstrate and record their performance.

Keywords: human computer interaction, multimedia applications



# **Overcoming Technical Barriers to a Speech-enabled Children's Reading Tutor**

In this paper, we discuss a successful effort to develop a child-appropriate acoustic model to support speech recognition in a commercially available set of tutors for beginning reading instruction. Furthermore, we elaborate interface design decisions necessary to wring successful computer-learner interactions out of speech recognition technology with less-than-perfect accuracy.

Failure to learn to read often has a cascading effect, resulting in many subsequent failures, condemning youngsters to an agonizing school experience and later starving them of opportunities in the workplace. The size of this group plodding through life is stunning: 90 million American adults cannot read well enough to be considered functionally literate (Educational Testing Service, 1993).

Commercially available educational software often seems somewhat alien to learning to read, deaf as it is to children's oral productions. To develop word recognition skill, such software must resort to clever schemes to compensate for its inability to react to youngsters reading aloud. One common strategy calls upon the child to perform tasks that presume to exercise the same fund of skills that reading requires, with directions like, "Find the word on the screen that rhymes with this picture." These sorts of activities fail to give children much of a sense of the experience of reading—unsurprising, given their orientation toward isolated word recognition and their reliance on picture interpretation.

Another strategy involves having children record their productions while reading, permitting them to compare them to an aural model (e.g., Discus, 1991; Jostens Learning, 1994). However, these strategies often fail because, to work, they assume the child possesses a certain level of phonolog-

ical awareness (Ball & Blachman 1991; Yopp, 1988) to make the required comparison. To successfully compare his or her own production with a pre-recorded model and to derive information from it, the learner must already recognize that a spoken word consists of a sequence of individual sounds and must be able to readily analyze it — the very skills of phonological awareness that the programs are trying to teach.

Researchers have long sought to enable computer-based reading tutors to listen to learners' productions, but failed to deploy their efforts, betrayed by inadequate processing power or the lack of child-appropriate acoustic models. For example, Jack Mostow and his coworkers at Carnegie-Mellon University (Mostow, et al., 1993, 1994) have provided a stunning example of the potential of a speech-enabled reading tutor. Displaying a page of text on the screen, their tutor listens as a learner reads it, providing delayed feedback on word substitutions, omissions, mispronunciations or insertions. Unfortunately, the lack of capable continuous-speech models for young readers (five or six years old) has blocked the dissemination of this work.

This paper will explore the issues that vex those trying to build literacy tutors for young emergent readers and how we overcame them to create the technology to support commercially successful products (Edmark, 1997).

## **CHALLENGES OF SPEECH RECOGNITION WITH YOUNG CHILDREN**

A program trying to recognize what young children say, even in the context of a linguistically highly determined task such as oral reading, must overcome a range of difficulties (Nix, Fairweather, and Adams, 1998). Some of these include:

- Words read aloud by young children tend to be shorter and therefore less easily discriminated by speech recognition systems than those produced by adults.
- Children often respond to the directive to read a word by producing a syntactic frame that contains that word. For example, when asked to repeat the word *cow* that they may have just misread in the phrase *I am not a cow*, they often respond with “a cow.”
- Linguistic diversity, especially in urban areas, is increasing. For example, a reading tutor deployed in New York City would need an acoustic model that could deal with the well over 100 languages recognized as spoken there by the Board of Education.
- Children’s variable maturation rates force a reading tutor to deal with a wide range of articulatory competence.
- Beginning reading is usually a social affair where a learner interacts and shares reading tasks with other reader-speaker-listeners, such as parents, peers, or teachers. A successful interface cannot ignore the give and take of goal-directed *social* cognition.
- Oral reading—even by proficient readers—is replete with omissions, substitutions, regressions, and mispronunciations, all of which require a nimble interface.
- Emergent readers dealing with unfamiliar text that has been read aloud to them may repeat it without clearly forming discriminable word boundaries (the “*Iple-jallejuns* problem”). The recognition system must be able to recognize these conglomerations and respond appropriately.

- As one would expect, the pitch of young children's speech is higher than that of adults. The input channels on some speech recognition systems, having been tuned on samples of adult speakers, sometimes even clip these higher frequencies to limit bandwidth and improve performance for adults.

## **THE ANATOMY OF SPEECH RECOGNITION SYSTEMS**

To understand the technical issues surrounding the construction of speech-enabled tutors, it is helpful to grasp the basic operation of the recognition engines they sit atop. Having been built with dictation in mind, these engines profoundly affect how a tutor works.

All commercially viable speech recognition engines operate on principles of statistical pattern recognition first applied by Baker (1975) and Jelinek (1976) and their colleagues at IBM Research during the 1970s. This approach, still largely unchanged, turned the speech recognition community away from methods based on linguistic analyses.

The speech recognition process begins with the digitization of the speech input as a set of short ( $\cong$  10 milliseconds) chunks. These are transformed and filtered to:

- handle attenuation of signals caused by the operation of the speech mechanism, such as signal radiation from the lips,
- map the transformed sounds into a set of categories that approximate the frequency resolution of the human ear,
- decorrelate the transformed coefficients to provide independent information about the sounds,
- reduce the number of coefficients needed to represent each sound.

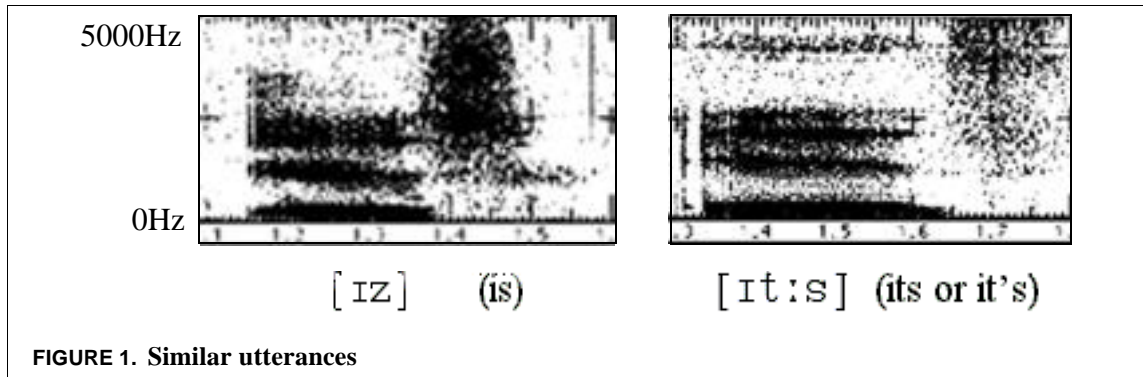
Vectors of coefficients representing each sound are fed to a set of hidden Markov models (HMM) that probabilistically represent both the sequence of sounds that make up a phoneme as well as those making up words as ordered sets of states. Not all sequences can occur, as phonotactic constraints rule out certain combinations altogether. Others are more or less likely, so the computational task becomes one of moving back and forth through the model to select the sequence that best matches the sound while maximizing the product of the transitional probabilities from state to state. Central to the success of this effort was the creation of a new HMM built out of 110,000 utterances gleaned from 1830 children from 20 sites all over the United States.

The process of computing the conditional probability of occurrence of particular phones given the presence of others is mirrored in the evaluation of the resultant acoustic hypotheses in terms of another HMM known as the *language model*. The language model can be thought of simply as a set of trigrams, each representing the probability of occurrence of a particular word given the occurrence of a two-word sequence before it. The availability of a language model for dictation permits one of a set of competing acoustic hypotheses to be considered most likely in terms of the conditional probabilities imposed by the preceding two words.

While crucial to dictation systems, the language model adds nothing to a tutor for beginning readers because the text so rigidly determines what the reader should say<sup>1</sup>. However, this may suggest that the point of diminishing returns for improving an acoustic model may differ for a dictation system than for a reading tutor. The acoustic model for dictation has a language model to further disambiguate its results. A reading tutor has no such court of appeal.

---

1. For more proficient readers who make meaning-preserving miscues such as reading the word “forest” for the word “woods,” it might be argued that a language model is important. The class of tutors discussed here might well be inappropriate for such advanced readers.



### DICTATION BIAS

A reading tutor built upon a speech recognition engine crafted for dictation must overcome certain problems stemming from the departure of the tutor from the original purpose of the engine. Consider the utterance spectra in Figure 1, noting the gross similarities of the vocalic areas (left half of each picture) and the differences between the two consonantal areas (right half). From an acoustic point of view, a dictation engine could classify these utterances identically, relying on the language model to resolve the ambiguity by using the fact that the locational distributions of “is” and “its” or “it’s” overlap little.

A dictation engine must overcome the subtle differences of acoustic differences that speaker-hearers do, treating, for example, both *butter* and *butter* as “butter.” Although a reading tutor must also be able to classify such variants as acceptable recodings for “butter,” it must be able to distinguish other, equally small, pronunciation differences and react to them if they make a difference in learning to read.

### DISCRETE AND CONTINUOUS SPEECH MODELS

Successful dictation systems use continuous speech models and, because even children just learning to read run familiar phrases together when reading aloud, we assumed at the outset of this development that a continuous model was required. Certainly we could not permit the use of a discrete system if it forced staccato, word-by-word reading on the children.

Fashioning a workable children's model that met our initial error criterion proved so daunting that we set aside the goal of using continuous speech. To handle children's "bursty" productions, we loaded the tutor's receptive vocabulary not only with the words to be read, but the power set of the contiguous vocabulary items with all of their phonemic variants. This means that to monitor a child's reading of "You are a cow," the vocabulary holds all the variants of "you", "are", "a", "cow", "you are", "are a", "a cow", etc. This pseudo-continuous model contributed strongly to bringing the error rates to our criterion for commercial use.

## **DESIGN GOALS**

*Watch-me!-Read* and Edmark's *Let's Go Read: An Island Adventure* are both designed to provide 5-7 year old emerging readers with a direct experience of reading. *Watch-me!-Read* displays a facsimile of one of a series of popular reading books chosen by the user and uses an on-screen companion in the form of a panda to help the child read the book (see Figure 2). The panda provides the same sorts of support as would a parent or teacher, modeling reading, highlighting text to be read by the child, giving encouragement, and providing feedback. Together, the child and the panda work to create a joint reading performance that the child can play back or demonstrate to others. It not only contains the contributions of both learner and panda, but any video or audio annotation the learner has chosen to add.



**FIGURE 2.** *Watch-me!-Read* panda listening for a child to read

Central to the evaluation of the design was the degree to which the tutor erred by either failing to accept a correct production or inappropriately failing to provide corrective feedback to an erroneous production. As one would expect, emergent readers tolerate false positive errors, generally because they are unaware of them. If a child were to produce “this” for “that” in the text and the tutor accepted it, for example, the child is usually unaware that anything out of the ordinary has happened. False negatives, on the other hand, frustrate children quickly if they occur more frequently than in 5% of the tutor’s responses to a learner’s productions.

## **EVALUATION OF THE TUTOR’S PERFORMANCE**

Efforts in two areas set this work apart from other attempts to build a speech-enabled reading tutor and led to its incorporation into commercial products: the creation of a child-appropriate acoustic model and the crafting of an interface to accommodate the inevitable uncertainty of such systems.

### **THE ACOUSTIC MODEL**

To evaluate the children's acoustic model, we built a data collection routine to capture what the child was asked to read, how the child responded, the application's judgment, what it recognized, as well as keyboard and timing events and measurements.

Response and judgment data were collected for 30 first and second grade children over four weeks during a summer program. Each child used the system once a day for 10 to 30 minutes, yielding approximately 20,000 classified responses. Multiple judges reviewing these data found that the computer misclassified the children's responses, 5% classified "wrong" when they should have been accepted (false negative) and 4% "right" when they should have been considered "wrong." (false positive).

However, these data contained the means to improve the performance of the model. We identified learner responses the application misjudged as correct. These "chaff" words were then included in the pool of words the tutor searched when monitoring learners' productions. For example, finding that the application misjudged "well" to be a correct response for "will," we added "well" as a chaff word to the active vocabulary when "will" was asked for. By including common, phonically similar miscues in the active vocabulary to be searched, we virtually eliminated false positive judgments. In effect, these known miscue words function analogously to the language model that disambiguates acoustic alternatives in a dictation system.

False negative judgments were attacked similarly. Although the acoustic model was built to accommodate a wide range of phonetic variation, the response data demonstrated that sometimes we had to add alternative forms of pronunciation to our dictionary to properly judge response. For example, because so many children pronounce “comfortable” as if it were spelled “comfterble,” that alternative was included.

### **INTERFACE ISSUES PECULIAR TO SPEECH-ENABLED TUTORS**

These approaches cannot yield completely accurate judgments. Children’s tolerance of occasional misjudgments was enhanced by tempering the way such judgments were signaled. For example, when the panda signals that the child has not read what was asked, she never says “no” or “incorrect” but instead presents a simpler task, asking, for example, a more focused “what’s this word?”

False negative judgments were particularly destructive of the sought-after “feelings of reading” when they occurred on the last word of a sentence. Under these conditions, early versions of the interface would wait for the correct response, prompting the learner to repeat his or her complete response. The application could “loop” in this state, the child unaware that only the last word in the production had been misjudged. To remedy this, an action was developed for the learner to indicate that his or her response was finished, paralleling the one to be used to signal the beginning.

### **SUMMARY**

Using speech recognition to help children read has been shown to work with three broad enhancements. First, a child-appropriate acoustic model structured as a HMM was built on an extensive collection of language samples. Second, the ability of the acoustic model to discriminate among phonetically similar targets can be enhanced by including probable miscues in the active vocabu-

lary. Third, the user interface was modified to adapt to the uncertainty of the system's judgments and to preserve the give-and-take of the conversation central to one-on-one episodes common to effective reading teaching. This resulted in a set of systems that provide emergent readers with oral reading practice opportunities with feedback as well as a means to demonstrate and record their performance.

## **REFERENCES**

- Adams, M. J. (1990) *Beginning to Read: Thinking and Learning About Print*. Cambridge, MA: Bradford Books.
- Baker, J. K. (1975) The Dragon System - an Overview. *IEEE Trans. ASSP*, VPr