Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 968–984
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Assessment of emerging reading skills in young native speakers and language learners

Patti Price [a,*], Joseph Tepperman [b], Markus Iseli [c], Thao Duong [d], Matthew Black [b], Shizhen Wang [e], Christy Kim Boscardin [f], Margaret Heritage [g], P. David Pearson [h], Shrikanth Narayanan [b], Abeer Alwan [i]

[a] PPrice Speech and Language Technology, 420 Shirley Way, Menlo Park, CA 94025, USA
[b] Dept. of Electrical Engineering, University of Southern California, EEB 400, 3740 McClintock Ave., Los Angeles, CA 90089, USA
[c] Dept. of Electrical Engineering, Henry Samueli School of Engineering and Applied Science 63-134 Engr. IV, UCLA, 405 Hilgard Ave., Box 951594, Los Angeles, CA 90095-1594, USA
[d] Graduate School of Education, 1501 Tolman Hall #5519, University of California, Berkeley, Berkeley, CA 94720-1670, USA
[e] Electrical Engineering, University of California, Los Angeles, UCLA, CA 90095, USA
[f] School of Medicine, Office of Medical Education, University of California, San Francisco, 521 Parnassus Avenue, Room C-254, San Francisco, CA 94143-0410, USA
[g] CRESST/UCLA, 10945 Le Conte Ave, Los Angeles, CA 90095-7150, USA
[h] Graduate School of Education, 1501 Tolman Hall #1670, University of California, Berkeley, Berkeley, CA 94720-1670, USA
[i] Dept. of Electrical Engineering, Henry Samueli School of Engineering and Applied Science 66-147G Engr. IV, UCLA, 405 Hilgard Ave., Box 951594, Los Angeles, CA 90095-1594, USA

Received 2 July 2008; received in revised form 26 December 2008; accepted 4 May 2009

Abstract

To automate assessments of beginning readers, especially those still learning English, we have investigated the types of knowledge sources that teachers use and have tried to incorporate them into an automated system. We describe a set of speech recognition and verification experiments and compare teacher scores with automatic scores in order to decide when a novel pronunciation is best viewed as a reading error or as dialect variation. Since no one classroom teacher is expected to be familiar with as many dialect systems as might occur in an urban classroom, making progress in automated assessments in this area can improve the consistency and fairness of reading assessment. We found that automatic methods performed best when the acoustic models were trained on both native and non-native speech, and argue that this training condition is necessary for automatic reading assessment since a child's reading ability is not directly observable in one utterance. We also found assessment of emerging reading skills in young children to be an area ripe for more research!

Keywords: Children's speech recognition; Reading assessment; Language learning; Accented English; Speaker adaptation

1. Introduction

Our work focuses on two key conclusions highlighted in a report of the National Literacy Panel on language-minority children and youth (August and Shanahan, 2006): (a) adequate assessments seem to be essential for gauging the individual strengths and weaknesses of language learners and (b) "existing assessments are inadequate to the need in most respects." Towards improving diagnostic

* Corresponding author. Tel.: +1 650 704 0728; fax: +1 815 366 8264.
E-mail addresses: pjp@pprice.com (P. Price), tepperma@usc.edu (J. Tepperman), iseli@ee.ucla.edu (M. Iseli), thaod@berkeley.edu (T. Duong), matthepb@usc.edu (M. Black), szwang@ee.ucla.edu (S. Wang), BoscardinCK@medsch.ucsf.edu (C.K. Boscardin), mheritag@ucla.edu (M. Heritage), ppearson@berkeley.edu (P. David Pearson), shri@sipi.usc.edu (S. Narayanan), alwan@ee.ucla.edu (A. Alwan).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.05.001

assessments of language and literacy, we have developed the TBALL (Technology-Based Assessment of Language and Literacy) system, which aims to automatically assess English literacy skills of English native speaker American and Mexican American children learning English in kindergarten, first and second grade (roughly age 5–7 years). Our goals were to (1) improve automatic speech recognition with a focus on bilingual children (by using child demographic background information, teacher feedback, and human linguistic experience, and by applying a pronunciation dictionary), and (2) maximize the correlation between automatic assessment scoring and teacher manual scoring (by using information from child, teacher, and language experts, and by applying scoring rules).

Not all competent readers can read aloud well or at all, and, for some, skill at reading aloud is not very representative of reading comprehension. In school, however, one way to gauge young children's reading development is through reading aloud, which can provide a teacher with information about the reading cues children are using as well as levels of fluency and accuracy. Being able to read silently is an important step in reading development and is one that students will typically acquire during the period of kindergarten through second or third grade. In our assessment framework reading aloud is only one of a suite of assessments to determine fluency and accuracy. Students are also asked to respond to comprehension questions about the text they have read. If fluency, accuracy or comprehension is weak, we have other measures to investigate potential sources for these weaknesses.

We hope that automatic assessment and diagnosis of language and literacy skills of young children can help teachers spend more time on effective teaching and less time on assessing students. In this section, after providing background on the challenges represented by the acoustics of child speech (Section 1.1) and of automating assessment (Section 1.2), we briefly describe the structure of the paper (Section 1.3).

1.1. Background on the acoustics of child speech

Automatic speech recognition for children poses even greater challenges than speech recognition for adults. For instance, the age-dependent acoustic and linguistic variability in children's speech (e.g., Lee et al., 1999) makes speech recognition more difficult compared to adults (Gerosa et al., 2007). Zue et al. (2000) found that the in-vocabulary word error rate for children was almost twice that for adult users. Although this is in part because the underlying models were trained with mostly adult speakers, it appears that adaptation and speaker normalization techniques, or equivalent methods, are required for dealing with such variability.

The acoustic characteristics of speech (especially, conversational speech) of young children have not been adequately described. The spectral and temporal characteristics of children's speech are different from those of adult speakers and are highly influenced by growth and other developmental changes. In a key study (Eguchi and Hirsh, 1969), later summarized by Kent (1976), age-dependent changes in formants (resonances of the vocal tract) and fundamental frequency (F0) of children ages three to thirteen were reported. Although the study was somewhat limited in scope and in the number/type of subjects, when compared to adult speech, important differences in the spectral characteristics of children voices were observed, including higher F0 and formant frequencies, and greater spectral variability. To deal with such variability, parametric models for transforming vowel formants of children speakers to the adult speaker space (vowel formant frequency normalization) have been considered (e.g., Goldstein, 1980; Lee et al., 1999). In recent years, research advances such as vocal tract length normalization, speaker adaptive training, language modeling, and pronunciation variation modeling have proved useful in improving speech recognition of child voices (Gerosa et al., 2007; Potamianos and Narayanan, 2003; Hagen et al., 2007).

A detailed comparison of temporal features and speech segment durations for children compared to adult speakers (Kent, 1976; Smith, 1992; Lee et al., 1999) also found distinct age-related differences: on average, the speaking rate of children is slower than that of adults and children display higher variability in speaking rate and vocal effort. Lee et al. (1999) reported variations in temporal and spectral parameters using speech data from older children, 6–18 years, and from adults. A key finding is that the areas of the second formant plotted as a function of the first formant two-sigma ellipses (representing the extent parameter variability in formant frequencies) were much larger for children compared to adults for most vowels. In addition, a systematic decrease in the values of the mean and variance of the acoustic correlates such as formants, pitch, and duration with age was observed, with values reaching adult ranges around 13 or 14 years. These differences are attributed mainly to anatomical and morphological differences in the vocal tract geometry, less precise control of the articulators, and a less refined ability to control suprasegmental aspects such as prosody. In sum, the acoustics of child voices and the large variability observed pose quite a challenge for speech recognition.

1.2. Background on computer technology in assessment

Computer technology has been used in literacy instruction and assessment (e.g., McCullough, 1995) with different types of interventions resulting in gains in different areas (e.g., Whitehurst and Lonigan, 1998). Some software appears to be effective for teaching phonological sensitivity skills (Barker and Torgesen, 1995; Barron et al., 1992; Farmer et al., 1992; Lovett et al., 1994), but research on technology-assisted learning systems shows mixed results that vary with the skill being taught, skill level of the learner, the system being evaluated, and age and grade level of

sample size (Khalili and Shashaani, 1994; McCullough, 1995; Van Dusen and Worthen, 1993; Wiburg, 1995). Most computer-assisted programs adopt a point and click approach, with children selecting the appropriate answers in response to prompts, e.g., "click on the number of syllables you heard," "click on the pair of words that rhyme". Our goal is to go beyond pointing and clicking by developing appropriate speech recognition and understanding techniques.

Recent advances in speech and multimedia technology have spurred worldwide deployment of prototype and commercial spoken dialogue applications (e.g., Buntschuh et al., 1998; Gorin et al., 1997; Lamel et al., 1997; Zue et al., 2000), most of which are exclusively targeted toward adults. The idea of voice interfaces for children is not a new one (Narayanan and Potamianos, 2002), but the scope of systems developed so far has been relatively limited. Examples of prototypes include aids for reading (e.g., Mostow et al., 1995) and pronunciation tutoring (e.g., Russell et al., 1996). In the pioneering Project LISTEN, CMU researchers are developing an automated tutor to analyze oral reading (grades 1–5). SRI's computer-based education efforts have resulted in the EduSpeak software, which includes speech recognition models for children. The "CU Animate" project at Colorado (Ma et al., 2002) targets spoken dialog interfaces to facilitate reading, especially for children with developmental disabilities. Another relevant project focused on young children is Watch Me! Read, developed by IBM's T.J. Watson Research Cente, in which teachers in the Houston School District found that students were motivated to use the software, and that the system improved oral presentation skills (Williams et al., 2000). Although we know of no study that has specifically targeted the speech of non-native English speaking children, speech recognition has been shown to be successful for non-native adults in foreign language pronunciation training (e.g., Dalby and Kewley-Port, 1999; Eskenazi, 1999).

Research also exists in designing multimodal interfaces that combine speech with a variety of other input modalities such as text and touch (DiFabrizzio et al., 1999; Sharma et al., 1998; Takezawa and Morimoto, 1998). Results of these investigations suggest that the use of multiple modalities, rather than a single modality, leads to more efficient and natural interaction and enhances the overall user experience (e.g., Cohen et al., 1998). The efforts at OGI by Oviatt and colleagues focus on multimodal interactions (using pen and voice inputs) as exemplified by the "I SEE" project (e.g., Coulston et al., 2002; Darves et al., 2002; Oviatt and Adams, 2000; Xiao et al., 2002). A key finding is that the disfluency rate in child-machine communication was significantly lower (by three times) than in interpersonal communication. Perhaps children treat animated characters differently from humans and may be careful when communicating with them. Similarly, Cassell and colleagues at MIT are investigating conversational interfaces in the context of interactive story telling (e.g., Cassell and Ryokai, 2001). Our work aims at a unified analysis of the speech of young children from a variety of linguistic and socioeconomic backgrounds to create speech processing algorithms that enable tools to measure and assess longitudinal academic development.

1.3. Guide to the paper

In this paper we argue that knowledge sources beyond the acoustics of the signal are required to provide adequate scoring and instructionally valid diagnostic information. Since we build significantly on the TBALL assessment system, we first summarize the TBALL goals and challenges as well as the progress to date, highlighting results particularly relevant for the present study (Section 2). As outlined above, recognizing child speech, especially those who are not native speakers poses significant challenges. There are also significant challenges in assessment methodology given that teachers do not always agree on what represents a correct reading. In Section 3, we outline these issues and challenges and describe our approach generally as well as in specific reading assessment tasks. Once we had a working system in use by teachers, we began working closely with six teachers to help evaluate our work and to help maximize the system's impact, as described in Section 4. Finally, in Section 5, we describe our progress in speech technology and reading assessment and how it has led to increased understanding of new challenges that could lead to interesting possibilities for future directions.

2. Overview of the TBALL project

The TBALL assessment system provides teachers of kindergarten through second grade (teaching children of ages about 5 through 7 years old) with information about student language skills and helps them to plan individualized or group instruction. The system consists of three main parts: (1) a child-friendly student interface, which presents the assessment tasks using a graphical user interface and which records audio data and test performance data, (2) a speech recognition module that analyzes and scores assessment responses, and (3) a teacher interface that provides teachers with feedback on student performance. The assessments are situated in an original and comprehensive framework that addresses critical reading skills children need to acquire in the early grades to become proficient readers, including assessments of phonological awareness and phonics knowledge, word identification, oral and written language comprehension, and accuracy and rate in reading text. The assessment framework was developed in collaboration with expert reading teachers and was designed to be embedded in an instructional framework so that the results could be used to plan instruction and to minimize the amount of time required from teacher and child while maximizing the informational benefit of the assessment. Further details of the

TBALL effort appear in (Alwan et al., 2007; Jones et al., 2007); a brief summary of the TBALL effort is included below, including goals and challenges (Section 2.1) and progress (Section 2.2), highlighting the results relevant for the present study.

2.1. TBALL goals and challenges

A major TBALL goal has been to benefit teachers by automating assessment of early literacy skills, making it easier, less time-consuming, and more consistent, compared to the usual paper assessments. The system can display test results so teachers can group students by scores for tailored individual or group instruction and more easily find and track indicators of later performance. Distinguishing characteristics of the TBALL project include: (1) a multidisciplinary character, combining engineering, computer science, education, and linguistics; (2) a focus on speech from young children, whose voices when compared to adult speech are known to yield less accurate recognition results; and (3) the inclusion in the effort of many children who are bilingual or who speak Spanish-accented English.

For these goals, acoustic issues combined with assessment issues made it difficult to determine what aspects of assessment could be automated or semi-automated given the current state of the art. Achieving reliable, fair, and consistent automatic diagnostic scores in this context has posed some significant challenges: (1) the interface for children must be robust to whatever happens when a child uses it – or plays with it; (2) robustness is essential; in addition to specific acoustic challenges outlined in Section 1.2, other acoustic challenges arise when a child touches the microphone, moves it close to the mouth, speaks loudly or varies greatly in amplitude levels, hyperarticulates, whispers or mumbles responses, stutters or repeats and/or makes self-corrections; (3) automatic assessment is required to save time, but is challenging because for many assessments human raters do not agree well; it is hard to predict what a child might say, and why, and whether or not it is evidence of competence.

2.2. Summary of TBALL progress

Major accomplishments of the TBALL project to date include:

- *Initial data*: we collected initial data to train and develop automatic speech recognition algorithms suitable for children's speech (both native and non-native English speakers), including recordings of more than 300 children amounting to a total of more than 12 h of recorded speech, most of it read speech. Data has been recorded in-class with a Logitech Premium USB Headset 350 microphone using a sampling frequency of 16 kHz with 16 bit precision. Using a close-talking microphone helped to reduce classroom background noise.
- *Speech recognition*: we developed pronunciation modeling and talker adaptation techniques for young children, including two major algorithms for rapid speaker adaptation that work especially well with limited amounts of adaptation data. One is regression-tree based spectral peak alignment and the other is subglottal resonance-based rapid speaker adaptation. The algorithms perform better than previous state-of-the-art techniques such as Vocal Tract Length Normalization (VTLN) and Maximum Likelihood Linear Regression (MLLR).
- *Database design*: a database (based on open-source MySQL) organizes the acoustic and other data, including demographic data, information about the recording session, test results by item from hand-scoring as well as automated scores, summary scores, and additional annotations and transcriptions. The database system adheres to current SQL standards to enable rapid porting to commercial DBMS, if needed.
- *System architecture*: an overall architecture of the TBALL system and the user interface helps ensure that the primary clients (students and teachers) can and will use the platform. This includes interfaces for both students being assessed and for teachers viewing and analyzing the results and patterns of results.
- *Assessment content*: the 'content' of the assessment system includes the assessment framework, with components that are core recurring benchmark assessments as well as diagnostic subroutines called as needed. The assessment content also includes the actual tests themselves, including some rather novel formats and a set of usage guidelines.
- *The DAM assessment paradigm*: we developed a new assessment of words and nonwords, the DAM (Decodable Assessment Measure). The key concept of this assessment is that one child's nonword might be another child's rare word and therefore there is not a clean separation between real words and nonsense words. We take advantage of this to diagnose targeted reading issues in rare words and nonsense words.
- *The BARLA assessment paradigm*: the BARLA (Bay Area Reading and Language Assessment) was developed to separate language understanding aspects (vocabulary, syntax, and morphology) from the 'decoding skills' (knowledge of letter to sound rules for a given language orthography) required to understand written language and to enable comparable tests that could be used for either reading comprehension or listening comprehension. We generalized and adapted a format involving a forced choice among images representing answers to a spoken or text prompt. For more information, see Price, 2007.

In the remainder of this section we describe TBALL results of particular interest for the present study, focused on knowledge sources that can help diagnose reading issues, for example: systematic pronunciation variations arising from language learning, separation of language

learning from other developmental issues, orthographic influences, determining when a variant pronunciation is a reading error, diagnosing causes of reading errors, and improving the correlation between automated scores and teacher hand-scoring.

When learning to speak English, non-native speakers may pronounce some English sounds differently from native speakers, which can degrade speech recognition performance unless properly modeled. Although the focus of this project is Spanish-accented English in young Mexican–Americans, the rule-based and data-driven techniques used can be generalized to other language pairs. Successful modeling of accented speech, especially in young children, would significantly advance the state of the art in speech processing.

As a first step in modeling pronunciation variation, we analyzed pronunciation variation in 4500 word tokens spoken by eighteen 5- to 7-year-old Mexican–American English language learners whose first language was Spanish (You et al., 2005). A set of linguists derived rules to create expected pronunciations based on substitutions that could arise from: (1) phonetic differences, as happens when the American English phone does not exist in Spanish (e.g., /dh/ >/d/), (2) phonological differences, as happens when the sounds exist but are not in contrast to distinguish meaning as for /ih/ vs. /iy/, (3) phonotactic differences, for example, when sound variation that depends on the phonetic context differ between the two languages as for /b/ intervocalically which remains /b/ in English but becomes /v/ in Spanish, and (4) letter to sound rule differences for those who can read Spanish or who commonly hear English spoken by Spanish talkers influenced by Spanish orthography (e.g., in Spanish the letter **y** is pronounced /jh/ in some dialects making the pronunciation of the word 'yes' sometimes occur as /jh eh s/ even though a /y/ exists in Spanish). These hypothesized sources of pronunciation variability were used to expand the lexicon to include additional pronunciations.

The data-driven analysis was based on phonetic transcription. We used the basic set of phonemes for English augmented with other symbols to capture accented English based on the data. A dynamic programming-based analysis algorithm was then designed to find common pronunciation variations in Spanish-accented English as spoken by these children. This analysis confirmed some of the linguistic hypotheses: e.g., the variant pronunciations of /v/,/z/,/dh/ were observed. Some predicted pronunciation variation patterns, however, were not observed in this analysis (e.g., /v/ > /b/ and /r/ > /rr/). The data-driven approach added some new variation patterns that had not appeared in the list of predictions. An important issue our analyses reveal is one also facing teachers: when is an unexpected pronunciation a reading error and when is it dialect variation? Since no one classroom teacher can be expected to be familiar with as many dialect systems as might occur in an urban classroom, making progress in automated assessments in this area can improve the consistency and fairness of reading assessment.

3. Speech recognition issues and results

As described above, the state of the art in automatic speech recognition is not perfect for the adult population and spoken dialogue applications for children pose even greater challenges. Differences between adult and child speech reflect physiological and anatomical changes associated with the development of articulators and the effects of linguistic differences during a child's growth. A pragmatic solution requires quantification of specific issues through systematic analyses of realistic data and deriving appropriate algorithms. Our approach to acoustic and language modeling is two-pronged: (1) systematic analyses of intra- and inter-speaker speech variability along several dimensions such as age, gender, and linguistic background, and (2) informed by these analyses, creation of robust algorithms and interface design strategies that take into account the strengths and limitations of the technology.

Automatic pronunciation evaluation is similar to automatic reading assessment, but with some important differences. Pronunciation evaluation assumes a closed set of target pronunciations (reference models) against which a metric is defined relative to the reference and all pronunciations falling within some predetermined distance are accepted. Reading assessment, however, is not so simple. Reading skills are correlated with pronunciation and production skills, but what constitutes a 'correct' pronunciation is determined uniquely by each speaker. For example, if a child reads the word "car" as /k aw/ ("cow"), without other evidence we cannot know if this was a reading mistake or a result of the child's potential difficulty in pronouncing the phoneme /r/. However, if we have more data indicating that all of this child's /r/ sounds in this context are produced like /w/, then we can infer that probably the child has read the word "car" correctly. Similarly, for a child who does not show any other /r/ substitutions, such a pronunciation probably signifies a reading error. More generally, to know what word was spoken, one needs to have evidence beyond a single word since many factors may affect pronunciation. The example above is an ideolect issue in which /r/ is not pronounced correctly. Other sources of variation include:

- *Dialect*: a speaker might reliably separate /ih/ and /eh/ but speak from a dialect that draws the line between the two at a place different from the hearer so that a teacher speaking another dialect would hear "pin" when the child's intention was "pen". It is not appropriate to penalize a child in reading simply because of dialect.
- *Foreign accent*: a speaker whose first language is Spanish might not reliably distinguish /iy/ and /ih/, or, as with the dialect issue, may distinguish the two but the phonetic realizations of the two may both be within the /iy/ category for native speaker–hearers. In this case, the hearer/assessor might hear "seat" when "sit" was intended. Although a language teacher might want to

work on pronunciation, it is probably not appropriate to penalize a child in reading for what is really a dialect or accent issue.

- *Other orthographies*: a child who is also learning to read in Spanish (whether a native speaker of English or of Spanish) may also have produced something more like "seat" than "sit" using Spanish letter to sound rules. In this case, the correct letter to sound rule in Spanish has been used, but not the correct one for English. How the teacher responds to this depends a good deal on what was done recently in the classroom and whether the classroom is bilingual or not.

These examples are intended to show that any given pronunciation in isolation is difficult or impossible to score for *reading* accuracy without knowing the intended word, and that a given pronunciation may arise from many different causes, so diagnosis is impossible without knowing the larger pattern. With a *pattern* of pronunciation instances, however, it is possible to estimate the intended word (just as we accommodate to new dialects and accents over time). Knowing these patterns, we have a chance at separating reading issues from pronunciation issues. Teachers should not assume that a non-native speaker (or a speaker of a dialect different from one's own) is a poor reader just because of an 'accent' – the 'correct' pronunciation when reading depends on the speaker's style and pronunciation in general. There may also be other variables that testify to a student's reading skills and are in some ways independent of segment-level pronunciation, such as speaking rate or fluency. In order to provide appropriate intervention, it is important to diagnose the causes of pronunciation divergence from 'canonical' forms – whether resulting from phonological difficulties common in second-language acquisition, or reading skill development related to word decoding (knowledge of letter-to-sound rules). How to score a child's reading may also depend on knowledge of the child's vocabulary and the teacher's recent curriculum.

In sum, it is not always clear exactly how reading assessment should be done automatically – i.e., which pronunciation cues or measures should be used and how much knowledge beyond an individual test item should be considered. Teachers do not always agree in their assessments, and rarely know explicitly how their perception of reading ability functions, so they can only inform automatic methods to a point. Ingenuity is required to generate a final score that approximates expert reading assessment. In the sections below we outline our general methodology (Section 3.1) relevant to all of the studies conducted, each of which is described in a subsequent subsection: letter-sound task (Section 3.2), phonemic awareness tasks (Section 3.3), and word tasks (Section 3.4).

3.1. General methodology

A number of standard speech recognition and machine learning techniques allow us to integrate prior knowledge into our automatic assessment modules. We can train acoustic models on speech from both native and non-native speakers, to cover as much of the variability in pronunciation as is observed in the data. When using these acoustic models for an unknown speech signal, we can constrain the results to a closed set of pronunciations that reflect common variants made by speakers in the target population. Finally, in synthesizing a binary reading score from a set of acoustic cues, we can condition our decision on prior knowledge of the child's demographics – native language, grade level, and other factors teachers might use when making the same assessment. These building blocks provide methods to assess students fairly, regardless of their native languages or pronunciation idiosyncrasies.

We found that automatic methods performed best when the acoustic models were trained on both native and non-native speech. Both accented and unaccented speech can demonstrate adequate reading ability, so recordings from both are needed for training – and, as a general rule, using more in-domain data tends to lead to more robust acoustic models.

We began with a small subset of data transcribed by linguists at the phone level, and derived models used to automatically transcribe future data that, in most cases, was only scored as an acceptable or unacceptable reading at the item-level. The 'ground truth' scores were provided either by experienced reading teachers, or by other transcribers who were found to agree with the teachers as well as the teachers agreed among themselves. All of these scores could be used as reference labels for train and test data in future verification experiments and the 'acceptable' items were used for training acoustic models using the automatic transcriptions.

The automatic transcription used two components that would be useful later: dictionaries of common pronunciation variants, and grammars of expected responses to assessment prompts. Pronunciation variants were derived from a number of sources, but primarily Spanish-accented phoneme substitution rules based on statistical evidence in the corpus (You et al., 2005). Task-specific examples of these rules and the pronunciation variants derived from them can be found below, in Sections 3.2, 3.3 and 3.4. Further pronunciation possibilities came from common reading errors suggested by experts, and also in the form of test- and item-specific pronunciation peculiarities (e.g., "frog" pronounced as "forg" – reversals such as this are common for many beginning readers). Automatic recognition grammars were based partly on the format of the TBALL assessments, and partly on the speech patterns the child users exhibited. In a sequence of responses, each test item was assumed to be preceded and followed by optional silence (or, more commonly, background classroom noise). The grammars also allowed for the inevitable repetitions, disfluencies, and out-of-domain speech the children regularly produced, either by decoding such things as generic 'garbage' speech (an acoustic model of unspecified phonetic properties), or as fragments of the target item (the

form that disfluent reading will often take). We developed additional task-dependent variations described in the sections below, but these are the common characteristics in the dictionaries and grammars overall.

The use of child demographic information in the modeling or automatic classification was somewhat limited, for various reasons. Objective information about each child's 'accentedness' was not available to us, and the child's 'native language' does not really predict pronunciation traits, especially for bilinguals. Ideally, if the acoustic models and pronunciation dictionary covered all potential variation, prior knowledge of the child should not offer much improvement (and this will be the case in the word-level section below). Furthermore, based on the overall trends observed at the time of the test, in judging reading skills, teachers who did not know a child's background showed high agreement with teachers who did – *provided they were experienced with other language learners of similar background.* When demographics were used, they were incorporated as additional features in the verification scheme, alongside acoustic scores. Demographics of the test set were examined in connection with the automatic results, to measure any potential system bias. For the word-level experiments (Section 3.4), demographics were found to offer no improvement in agreement between automatic and human scores.

3.2. Letter-sounds

Producing the 'sounds' of the English alphabet letters (called here 'letter-sounds') is a crucial first step toward reading words, phrases, and sentences. A child's ability to successfully map letters to sounds has been shown to be a good indicator for future reading ability (National Reading Panel, 2000). Teachers consider it important that children are assessed on this vital skill in their early education stages. Automatic pronunciation evaluation of English letter-sounds is a quite challenging task. Many of the letters have multiple pronunciations depending on word context (e.g., **c** can be /k/ or /s/), and some sounds have similar acoustic properties (e.g., **m**-**n**, **t**-**k**, **s**-**f**) and are frequently misheard, especially in noisy environments or over the telephone, for example. Stop consonants pose a particular problem since their 'hold' state is essentially silent with much of the information as to their identity being in transitions from and to surrounding sounds. With the large percentage of native Spanish-speaking children, the TBALL corpus (Kazemzadeh et al., 2005) has even more variability in children's speech, blurring the boundaries between acceptable and unacceptable pronunciations of letter-sounds.

Children in the TBALL project were tested on all 26 English alphabet letter-sounds. One lowercase letter at a time was displayed on a computer screen, and the children were given up to five seconds to say the appropriate letter-sound. The transition times between letters were automatically recorded and used to segment the sessions into single letter-sound utterances. We transcribed over 4000 letter-sounds but kept 3685 of them for our research, omitting files that contained excessive noise and/or had cut-off speech from the children. For these retained files, we annotated the pronunciation as acceptable or not, based on majority decision of the teachers and linguists doing the ratings. Vowel sounds were supposed to be 'short sounds' (**a**: /ae/, **e**: /eh/, **i**: /ih/, **o**: /aa/, **u**: /ah/), and consonants with multiple pronunciations were supposed to be pronounced with their primary pronunciation. For all voiced consonants, the child could optionally end with the phoneme, /ah/.

We created a test set of 30 files per letter-sound (780 utterances total), with the remaining 2905 utterances (~110 files per letter-sound) used to form the train set. The test and train sets were split so that the ratio of acceptable to unacceptable pronunciations for each letter-sound was even in the train and test set (but otherwise, the selection of files for the test set was random). Baseline acoustic monophone models were trained on 12 h of word-level speech in the TBALL corpus (Tepperman et al., 2007; Black et al., 2008). In-domain acoustic monophone models were trained directly on the letter-sound train set by first transcribing the data at the phoneme level using the baseline acoustic models. To establish human agreement statistics, we randomly selected 10 files per letter-sound from the test set, and three of the speech researchers each listened to the responses and marked them as acceptable or unacceptable sounds for the stimulus letter. Average pair wise human agreement was 94.9%, a statistic that served as an upper limit on performance for this letter-sound verification task.

We modeled variations in pronunciations at the phoneme-level since, for this task, unacceptable pronunciations usually differed from acceptable ones by a single phoneme. We first constructed a dictionary that included only acceptable phonemic spellings of each letter-sound (as defined by an expert teacher and linguist). This dictionary was used in a baseline recognition experiment, where each test utterance was decoded as any of the 26 letter-sounds (or silence). If a test utterance was recognized as the target letter, then the system classified the pronunciation 'acceptable', and if a different letter-sound or silence was recognized, the system classified the pronunciation 'unacceptable'.

This baseline recognition dictionary, however, contains superfluous entries (e.g., we would not expect a child to confuse the letter-sounds for **z** and **a**) and does not contain many unacceptable pronunciations that might occur. To remedy this over- and under-generation, we constructed six sets of dictionaries that included predicted pronunciation errors for the letter-sounds: (1) English letter-name pronunciations (e.g., **k**: /k ey/), (2) perceptual auditory confusions (e.g., **m**-**n**), (3) alternative English pronunciations (e.g., **c**: /s/, **g**: /jh/), (4) sight confusions (e.g., **b**-**d**, **p**-**q**), (5) Spanish-related confusions (e.g., **a**: /aa/, **z**: /s/), and (6) Spanish letter-name pronunciations (e.g., **k**: /k aa/). The letter-name confusions were added since some children confuse letter-sounds with letter-names, and the Spanish-related errors were added since many confused

English letter-sounds with corresponding Spanish letter-sounds.

We optimized the dictionary for each letter-sound independently using the letter-sound train set. For each letter-sound, we explored using all combinations of the six categorical unacceptable pronunciation types and formed a final optimized dictionary that yielded the highest agreement with human annotations on the training data. We found that all the optimized vowel letter-sound dictionaries included the Spanish-confusable letter-sound pronunciations, proving that the addition of these unacceptable pronunciations in the dictionary was helpful in detecting pronunciation errors in non-native speakers. If a child repeatedly makes Spanish-related pronunciations, the system will detect this class of errors, and the teacher can infer from the automatic results that the child may be confusing the letter-sound pronunciations of the two languages.

Table 1 shows agreement statistics between the human labels and automatic verification results for the different acoustic models and dictionaries on the test data. The best automatic system agreed with human annotators 87.95% of the time and used the optimized dictionary set and the in-domain acoustic models trained directly on letter-sounds. The performance gain when using the in-domain acoustic models implies that the phones in letter-sound productions differ significantly compared to those spoken within words. This difference may arise from the lack of co-articulatory effects in the isolated letter-sound productions and/or hyper-articulation of the letter-sounds. While the best automatic system outperformed the baseline system by 18%, a significant difference with $p < 0.05$, this system is still significantly worse than the 94.9% human agreement on this corpus ($p < 0.05$). Future research will try to bridge this gap by using sound-specific features extracted at the sub-phoneme-level.

Table 2 shows that there was no significant difference in system performance ($p > 0.1$) between native English- and Spanish-speaking children, even though English native speakers performed significantly better according to the manual annotations ($p < 0.05$). This is most likely due to the fact that the acoustic models were trained using data from both native English and Spanish speakers, thus making the system unbiased.

Table 1
Comparison of system performance on letter-sound test data between baseline acoustic models trained on words and in-domain acoustic models trained on letter-sound and between the baseline dictionary and optimized dictionary.

Agreement	Baseline models (%)	In-domain models (%)
Baseline dictionary	70.13	81.28
Optimized dictionary	83.97	87.95

Table 2
Comparison of children performance on letter-sound task (based on manual human annotations) and system performance between native English- and Spanish-speaking children.

	Native language	
	English (%)	Spanish (%)
Children production accuracy	78.63	70.21
System labeling accuracy	88.64	86.73

3.3. Phonemic awareness blending

Our three phonemic awareness blending tasks were designed to test ability to blend syllables, onsets and rimes, or phonemes into whole words: e.g., /t ey SIL b el/ = "table" (blending syllables), /p SIL aa t/ = "pot" (blending onset and rime), and /p SIL ih SIL k/ = "pick" (blending phonemes). The blending tasks are intended to assess both pronunciation accuracy and blending smoothness, or fluency.

Of these two evaluation criteria, fluency is more subjective, especially in these short utterances. However, in this case it may be more important than accuracy since the point of these assessments is to determine whether the child can show awareness of the subunits of words by fluently combining them. Typically, syllable awareness is developed, then onset/rime (important for understanding rhymes), and, finally, phonemes, essential for learning to 'decode', i.e., interpret the sounds of words based on an alphabetic writing system. In any case, accuracy and fluency can be judged somewhat independently since words can be pronounced smoothly and accurately, not smoothly but accurately, smoothly and inaccurately or not smoothly and inaccurately. The overall pronunciation is accepted if the word is smooth and accurate, otherwise the word is deemed unacceptable. Similar strategies to those in (Wang et al., 2007) were used to automatically evaluate children's performance on these three blending tasks.

To assess phonemic blending, word verification is performed to filter out utterances pronounced incorrectly. For valid words, forced alignment is applied to generate phoneme segmentations and produce the corresponding HMM log likelihood scores. Normalized spectral likelihoods and duration ratio scores are combined to assess the overall quality of the child's productions. Speaker-specific information is incorporated to optimize performance, with a dictionary containing acceptable pronunciations (including dialect variations) for each word to address possible pronunciation variations.

When evaluating children's responses, teachers tend to take into account some speaker-specific information (accent, speaking style, etc.) to decide whether a response is acceptable or not; the more speech from a child the rater hears, the more familiar the rater is with the system of contrasts used by the child, and thus the more reliable the manual assessment tends to be. Not knowing a child's pronunciation habits may lead to biased evaluations. To incorporate speaker-specific information, all items from one child were taken as reference pronunciations to detect the child's dialect or accent, speaking rate, etc., based on a carefully tagged dictionary.

Besides canonical pronunciations for each word, the dictionary contained entries for non-canonical but correct pronunciations from different dialects common in the Los Angeles area. For example, many speakers do not distinguish "cot" and "caught", pronouncing both as /k aa t/. Therefore, /k aa t/ and /k ao t/ would both be considered correct pronunciations. The dictionary also included /iy/-/ih/ substitutions, flagged as 'Spanish-accented', since many Spanish-speaking learners of English do not differentiate between them consistently. Pronunciations based on Mexican Spanish letter to sound (LTS) rules were not applied in this dictionary, since LTS rules are based on reading evaluations and this task's prompts are aural not written. Although it is possible that these rules may have some effect (since children imitate the speech of literate adults who may be influenced by Spanish LTS rules when speaking English), such instances appeared to be rare relative to the increase in dictionary size that would be needed to cover them comprehensively.

The pronunciations in the dictionary had tags for these various categories (Spanish-accented pronunciation, canonical pronunciation, phonological development issue, etc.). In this way, 'accent' or 'dialect' or 'idiolect' could be detected in a simple way: the likelihood for each pronunciation was calculated and the pronunciation category with the highest likelihood, if non-canonical, was declared as the 'idiolect' for the speaker for that word. A pattern of many words through the Spanish-accented path would confirm a speaker as having Spanish-accented speech. A constraint for detecting dialect is that the speaker must produce a consistent dialect – i.e., the dialect, if detectable, must be the same in most of the task words. In this way, we can model the dialect as a system of distinctions, which is linguistically appropriate.

A common pattern in children's responses in blending tasks is to repeat what they heard and make several attempts before giving a final answer. In such cases teachers made their decisions based on only the last attempt. To be consistent with teachers, our system also used only the final answer and ignored all attempts before that. As a complementary result, however, our system can output recognition results for all attempts, providing some diagnostic information as to a child's answer-making process – such information can be important in detecting emerging skills.

Another common error pattern is that perceptually similar phonemes (**p/b**, **f/v**, etc.) may be pronounced incorrectly but blended smoothly, and thus show strong blending skills without manifesting accurate pronunciation. Many children pronounced "pot" as /b aa t/ because the letter shapes are very confusable. The letter sounds are also very confusable, especially when spoken in isolation. The confusion is particularly strong for Spanish-speaking children learning English, since Spanish/p/ is acoustically very similar to English/b/. In the blending task, some substitutions or deletion/insertion patterns may occur when the syllables to be blended do not exist in the child's native language. For example, children from Spanish linguistic backgrounds tended to pronounce the word "stable" as "estable" or "estaple" because no words begin with the sound ***st*** in Spanish; **s + t** clusters always have a vowel preceding the cluster, such as the Spanish words "*estar*" or "*estacion*". Different teachers in different contexts may choose to flag these as errors or not. However, if we add these pronunciations to the dictionary, we have a better chance at recognizing what the child said, and if we model patterns of pronunciations via the tags, we can hypothesize a source of a variation pattern (e.g., Spanish accent) – and leave it to the teacher to make a decision on whether these instances are to be considered accurate or not. Instances of a pronunciation variation without a confirming pattern in other words would be a reading error (as opposed to a pronunciation variant). In addition to substitution errors, a common blending error is to pause between the blending components, or, more rarely, children will lengthen the first syllable or phoneme component.

Table 3
Comparison of speech recognition performance and inter-teacher correlation on three blending tasks.

Task	Human–human correlation (%)	Machine–human correlation (%)
Blending syllables	86.7	87.5
Blending onset/rime	85.3	80.8
Blending phonemes	84.7	79.8

To be consistent with the goals of the blending task, the final automatic score is based on both the pronunciation correctness and the blending smoothness, i.e., a word can be acceptable only when the pronunciation accuracy and the blending smoothness are both acceptable. The inter-teacher correlation and the correlation between the automatic system and the teachers' assessments are shown in Table 3. For the syllable blending task, the optimal system achieves a correlation slightly better than the average inter-teacher correlation, while for phoneme and onset/rime blending tasks, the performance is a little worse. The performance difference between syllable blending and onset/rime blending is statistically significant ($p < 0.05$), while the performance difference between onset/rime and phoneme blending tasks are not statistically significant. A possible reason is that the syllable blending task is easier than the other two tasks since the confusion between syllables is lower than that between phonemes – and, linguistically, it is a more natural task than the others. A detailed examination shows that the machine–human correlation is about 3% lower for non-native English speakers than for native speakers. For non-native English speakers, further work may be needed to investigate the pronunciation issues imposed by cross-language differences.

3.4. Word-level tasks

The words used for assessment of reading of isolated words were drawn from the high-frequency and 'BPST'

(beginning phonics skills task) word lists (Harris and Jacobson, 1982; Shefelbine, 1996). Productions of these words exemplify many of the issues with foreign-accented speech already mentioned. Variants seen in the data are of three main types:

- acceptable variations of the canonical version expected from native-speaking children, e.g. the word "frog" read /f r aa g/ or /f r ao g/,
- variants showing the influence of Mexican Spanish letter-to-sound rules, e.g., "frog" pronounced /f r ow g/, and
- common reading mistakes, for native or non-native speakers, such as making a vowel say its name, e.g., "frog" pronounced /f r ow g/ (note that this is the same as the accented pronunciation listed above), or "lip" pronounced /l ay p/.

With prior knowledge of these categories of pronunciation variants, we can expand the recognition pronunciation dictionary to include them all, with descriptive tags for each of the three subsets. Acceptable variants for native speakers were defined as those seen in a standard American English pronunciation dictionary. The Mexican Spanish-accented pronunciations were made by applying foreign letter-to-sound rules, to predict the pronunciation of the word as a Mexican reader would say it. For example, the character *z* always maps to the phoneme /s/, as dictated by Spanish orthography. Variants coming from common reading mistakes were either predicted by rule (e.g., a vowel is pronounced as its name) or by duplicating idiosyncratic pronunciations seen in the training data (e.g., "lip" pronounced /p ih l/). This process of adding pronunciations increases the size of the recognition dictionary by roughly a factor of four.

As explained above, it is not simply a matter of comparing all defined pronunciation paths and mapping the best hypothesis to an assessment score, since it is not the case that any pronunciation should be categorically accepted or rejected as evidence of adequate reading skills. Even the expected reading mistakes, though clearly indicating poor word decoding on a child's part, are often so close to the canonical pronunciations as to be difficult to detect reliably – this is indeed true with the overlap between the second and third pronunciation categories in the above example.

Our solution in (Tepperman et al. (2007)) was to decode each pronunciation category separately and use recognition-based features from all three in a Bayesian network classifier. Features derived from recognition results included, for example, likelihood scores for each category of pronunciation variant, or the number of times each category appeared in the list of n-best recognition results. Such a classifier would infer an overall item-level score based on all these acoustic observations from the recognition results, and also unite those features in a generative framework that reflects the decoding and feature extraction process. In mathematical terms, it would select the value of the binary accept/reject variable Q that maximizes its joint probability with all the recognition-based features $X_1, \ldots, X_n$:

$$P(Q, X_1, \ldots, X_n) = P(Q)\Pi_{i=1}^{n} P(X_i | Pa(X_i))$$

where $Pa(X_i)$ refers to the generative 'parents' of X_i among the other features, defined by our hypothesized network of inter-feature dependencies. For example, in maximum likelihood ASR decoding, the likelihood scores can be thought to 'generate' the n-best list of results – these features are then treated as conditionally dependent for purposes of Bayesian inference on the reading score variable Q. Other features included as conditional factors in this classification were child demographics and other item-level information such as word length in letters, as a rough measure of item difficulty – factors we would expect to influence a teacher's perception of a child's reading skills.

Automatic classification experiments on the best hypothesized Bayesian Network structure indicated that using features derived from all three pronunciation categories yielded the best results: 0.68 item-level Kappa agreement and 0.92 speaker-level correlation, compared to an inter-teacher agreement of 0.85 Kappa on the item-level and 0.95 correlation on the speaker-level. As seen in Table 4, omitting features related to the canonical variants degraded performance the most, followed by omitting Spanish-accented pronunciations and then omitting those resulting from common reading errors. Note that eliminating the demographic variables did not degrade classifier performance. These findings speak to the importance of incorporating many acoustic cues to pronunciation categories, but show that prior knowledge of the child's grade level or native language does not offer further improvement. One partial explanation for this is that a child's native language, though correlated with accentedness, does not necessarily indicate the presence of an accent. We had no independent measure of accentedness, and it's possible that many of the non-native speakers manifested no accent that might cause the automatic methods to perform differently. Furthermore, over our test set of 11 native and 11 non-native speakers (166 and 164 test items, respectively), we found that the difference in classifier performance between native and non-native students was not statistically significant at the 95% confidence level, indicating that this method is unbiased even when demographic information is used.

In all cases, speaker-level correlation was much closer to the inter-teacher agreement than the item-level Kappa statistic was. Our assumption was that accuracy in item-level decisions would not be as important to teachers as overall test-level scores would, but this was something that the case study was intended to help determine. Another thing we wanted to learn from the case study was whether these pronunciation categories are used implicitly by the teachers, and if reporting scores related to these categories in the machine scores would be useful to teachers in diagnosing

Table 4
Classifier performance with various pronunciation and feature categories individually omitted.

	Omitted feature category					Inter-teacher agreement
	Native	Accent	Errors	Demographics	None	
Kappa	0.55	0.61	0.64	0.68	0.68	0.85
correlation	0.88	0.89	0.92	0.92	0.92	0.95

a child's reading difficulties. This feedback from teachers using the system under real classroom conditions would then be an essential ingredient in future improvements to all the speech recognition tasks described above.

4. The case study

With the TBALL system implemented and field-tested, the focus of the project shifted towards providing teachers with the feedback they expected from the system through the teacher interface. We approached this task through use of a one-year case study (June 2007 until May 2008) involving six teachers at a Los Angeles school and six teachers at a San Francisco Bay Area school. Involving the teachers and collaborating with them was crucial in solving many problems, but introduced some new challenges. By using the system to assess and analyze reading skills, teachers' views of the purpose of assessment changed and they saw the value of more frequent assessments to guide day-to-day instruction after using the system while being coached by TBALL researchers. It was also observed that instructional practices changed as a result of using the assessment system to target specific learning needs in reading and that finding reliable and consistent right/wrong scoring criteria suitable for each teacher and student was a challenge. Our evidence suggests it is more helpful to provide teachers with more fine-grained analysis by categorizing reading mistakes.

4.1. Goals of the case study

In the case study we wanted to learn what did and did not work, and why, by working closely with a small group of the teachers. An important motivation from an automatic assessment point of view was the need for feedback from teachers in refining speech recognition models and algorithms. From analysis of pronunciation trends observed in recordings from our data collection (You et al., 2005), we knew what phoneme-level variations to expect in Spanish-accented children's speech, and how different types of pronunciations were correlated with demographics and performance on the assessments. We still did not know, however, exactly which cues teachers used in evaluating student responses, and how these cues interacted and led to an overall perceptual impression. Of the various pronunciations modeled, which are most relevant to assessment, under which conditions and how should they be combined to generate an automatic score?

Since hand-scoring forms the basis of 'truth' in model creation, it should be accurate and reliable. Our goal to minimize the difference between the automatic and manual scores models teacher behaviors, but what types of results would be needed and how would they be used? For example, beyond simple accept/reject and overall test performance, were item-level scores or phoneme-level recognition results useful, or would they just confuse things? What level of detail might teachers want, and how could subtler detection capabilities of speech recognition modeling be best exploited and reported? Furthermore, how much should be reported to ensure that teachers feel confident in the system's performance? The case study was designed with all of these questions in mind, although only portions of it are relevant for this study.

4.2. Selecting assessments for the case study

Although we had anecdotal evidence from the teachers concerning which assessments were most useful, we also wanted some quantitative data on which assessments might be helpful in predicting later performance. In a 2006–2007 study, we had determined the predictive validity of TBALL measures on reading outcomes. We used regression analyses to examine the relationship between assessments given at the beginning of the academic school year and the performance on standardized assessment of reading at the end of the school year. TBALL assessments administered at the beginning of the academic year were: Letter-names and letter-sounds, blending (syllables, onsets/rime, and phonemes) and reading words (words with regular letter to sound rules, nonsense words and irregular high frequency words).

An important finding from this study is the *lack* of performance gap between English language learners and proficient English speakers at the beginning of the school year in kindergarten, when most students are still at the emergent pre-competent reading stage. The performance gap increases in Grades 1 and 2 between these two groups as reading development progresses and the reading assessment tasks become more demanding and closer to actual reading rather than precursor measures. In particular, the results showed that the **word reading** tasks were a significant predictor (p- values ranging from.001 to.009) of all subtests on a standardized reading assessment. For Kindergarten, the **letter-names** task was also a significant predictor of year end performance.

When shown these results, three first grade teachers (two Spanish/English bilingual teachers and one English-only teacher) who were involved in both the earlier research studies agreed that a further look into the word reading

tasks and assessment items was appropriate in order to diagnose and place students with targeted and effective word reading instruction. We hoped to inform speech recognition modeling from the results and also to deepen teacher knowledge and possibly answer questions as to why the reading gap exists between native speakers and language learners for word reading beyond Kindergarten. Although the letter-names task was significant at the Kindergarten-level, it was felt that the reason for significance was that it was a sign of parental involvement rather than because this was more crucial than, for example, letter-sounds in beginning reading. The word reading tasks were particularly important: they were significant predictors at all grade levels, all teachers emphasized the importance, they provided a better target for speech recognition being longer units, and they are more natural production units than the sub-word units of the other tasks.

4.3. Methodology

Our two testing sites explored different variations on the common theme of assessing system utility: The Southern California team focused on supporting teachers to make optimal practical use of the new student feedback data available from the TBALL system, while the Northern California site focused on how teacher knowledge could improve recognition, particularly for the word reading tasks. The Northern California results therefore are more relevant to the goals of this paper, and we will limit the discussion to these data, and, in particular, to those tasks particularly predictive of reading skills and for which we had both speech recognition results and sufficient hand-scoring from teachers: the word reading tasks.

For the word reading tasks, teachers performed manual scoring of words and nonsense words to inform speech recognition and to detect emergent reading skills. With a TBALL researcher, teachers listened to the recordings of student productions and discussed any discrepancies among the various scores. The teachers quickly realized that before developing appropriate reading intervention lessons, they had to diagnose, or hypothesize causes for the predominant student errors. That is, to know what to do about an error, one needs to know more than the fact that an error occurred. To expect the recognition system to give further detail, however, would require hand labeling. This was not an easy task, and took a great deal of reflection, discussion and negotiation. What counts as correct can depend on many factors, as we have already argued.

Teachers struggled to include their combined knowledge in how letters and sounds connect in both English and Spanish. For the English-only teacher, using knowledge of letter to sound rules in assessing reading in a language learner is different from the bilingual teachers who use both Spanish and English. When a student reads the word 'rub' as /r uw b/, the English-only teacher might interpret it as 'incorrect letter-to-sound' rule, whereas the bilingual teachers might interpret it as the incorporation of Spanish letter-to-sound rules. Knowing that a student is using Spanish letter-to-sound rules systematically indicates that the principle of regularity in letter-to-sound rules has been acquired by the student, an important precursor to skilled reading. A teacher might design an intervention to address differences between Spanish rules and English ones in this case, while a more random set of incorrect answers may mean that the student has not yet learned this principle and a different intervention might be needed. Three teachers collaborated to provide a score of 'right' or 'wrong' for a sample of their first grade students. We have included here only those students for whom we have completed manual as well as automated scores: 18 students from two different Spanish–English bilingual classrooms, and 8 children from an English-only classroom in data collected in spring 2008. Although the children in the English-only classroom were taught only in English, most of them were native speakers of Spanish and spoke Spanish at home.

In addition to the 'right/wrong' judgment, teachers added some additional tags for items that were not clear. Items could be unclear for several reasons: The child could be mumbling or drowned out by background noise, or perhaps the 'next' button was pressed by accident, or the response could be in between two sounds (e.g., between /iy/ and /ih/). Two common areas of uncertainty involved Spanish influence on English and were the most common recurring patterns, and a topic we are anxious to incorporate in future work:

1. *Phonology*: a child might pronounce a word correctly *except* for mapping English sounds to Spanish ones. If the child regularly makes such sound substitutions, the production is accurate with respect to reading issues, but inaccurate with respect to English phonology. A teacher may sometimes want to count these incorrect and sometimes correct, depending on what was being taught and how the child was instructed in doing the reading assessment. These were labeled as a phonological issue.
2. *Letter to sound rules*: a child might pronounce a word using Spanish letter to sound rules. This is usually incorrect (except for cases when English letter to sound rules produce the same result, e.g., "part" /p aa r t/), and a teacher may want to detect such instances so that a specific intervention can be designed. These were labeled as a letter to sound issue.

Not all the teachers gave the whole test, so we narrowed the analysis to those words that all students completed: We examine here the 15 beginning reading words and 15 corresponding nonwords (see Table 6). These words have only the English 'short vowel' sounds and the later words start to include two consonant letters in a row. In these data, we can use the scores of 'correct' and 'incorrect' to assess system performance. In future work, we hope to also take advantage of the diagnostic labels.

4.4. Results comparing hand-scoring and automated scoring

Overall the agreement between the manual and automatic scores was on the order of 60–70%, depending on the condition. Though there are no doubt ways to improve the correlation, we need to consider what this might mean. Although teacher ratings are important since that is how reading is currently judged, it is quite possible that there are biases that affect judgments. Further, since many teachers have little or no background in linguistics or the cross-language comparisons that might be needed, they might agree on an answer, but perhaps for the wrong reasons from a pedagogical point of view. An automatic scoring system can overcome the potential bias issue, though it is not readily adaptable to automatically knowing information such as what was just taught that day.

It is clear there is more to do in both finding valid measures of reading performance and in producing automated systems to create them. In the meantime, let's turn to what information teachers might glean from the current state of the art. If we compare the labels made automatically with those made by the teachers, we see in Table 5 that at the class level, the teachers have a similar view of the students using the hand scores compared to the automatic scores. That is, for either scoring method, we see that (1) the real words are more accurately produced than nonsense words, and (2) those in the bilingual classroom are slightly less accurate than those in the English only classroom. It appears that the automatic system might be a little more 'strict' a grader compared to the teachers. More research is needed to determine possible biases in the teachers when rating words compared to nonwords, vs. biases and noise in the automated rating system.

Table 6 shows results by individual item. Although there are some notable discrepancies in the scores, teachers who got either the manual or the automatic scores would conclude: (1) some words are read more accurately than others, (2) consonant vowel consonant words are more accurately read than those with two consonants together, (3) there are factors, perhaps word familiarity that seem to lead to differences between the accuracy of the words vs. nonwords (e.g., 'frog' is read much more accurately than 'rub'), and (4) the vowels indicated by the letters **e** and **u** seem to be more difficult than the others.

Table 5
Percent correct for words and nonsense words, for English language learners in a bilingual classroom (ELL) and for students in an English-only classroom (EO), scored manually by the teachers (Manual) and scored automatically (Auto) by the system.

	Manual		Auto	
	Words (%)	Nonsense (%)	Words (%)	Nonsense (%)
ELL	79	63	72	69
EO	83	68	75	68

5. Discussion and future directions

Although we were aware that our goal of providing automated assessment for young children reading presented significant challenges, our progress in speech technology and reading assessment (Section 5.1) has been punctuated by increased understanding of new challenges (Section 5.2), which leads to interesting possibilities for future directions (Section 5.3).

5.1. Lessons learned

In this investigation we had some expected findings, for example: it is necessary to train acoustic models on in-domain data. The pronunciation of any speaker, adult or child, varies depending on the task, and this was especially true of the children in our study. For example, acoustic models trained on word-internal pronunciations are not the best references for the hyper-articulated and elongated realizations seen on the letter-sound task. Models trained from the letter-sound recordings alone yielded improved

Table 6
Percent correct by word, comparing manual and automatic scores.

	Words							
	MAP (%)	RIP (%)	MET (%)	RUB (%)	MOP (%)	LIP (%)	LOT (%)	ZAP (%)
Manual	96	73	73	65	85	85	96	92
Auto	96	85	65	31	88	65	96	100
	Nonsense							
	MAB (%)	RIT (%)	MEP (%)	RUP (%)	NOP (%)	MIP (%)	ZOT (%)	FAP (%)
Manual	52	68	56	64	84	72	64	76
Auto	92	92	44	48	88	72	76	92
	Words							
	ZELL (%)	LUT (%)	MAFT (%)	NUST (%)	FROP (%)	FLIG (%)	SNECK (%)	
Manual	60	60	72	68	80	56	40	
Auto	88	36	80	84	68	64	12	
	Nonsense							
	FELL (%)	NUT (%)	LEFT (%)	MUST (%)	FROG (%)	FLIP (%)	SNACK (%)	
Manual	69	77	73	81	92	69	69	
Auto	69	23	62	46	100	73	92	

performance in automatically assessing that task. Contrary to this expectation, however, we found that training on both native and non-native speakers improved recognition performance for both groups – though it is not easy to separate native speakers from non-native speakers in an increasingly multilingual world. Most of the students in the schools in our study vary along a continuum of English–Spanish bilinguality, which may have affected these results.

More significantly, we have built on earlier evidence that combining a data-driven with a theory-driven approach can lead to significant performance improvements. Although there is no absolute truth in automatic scoring of reading skills, we can integrate various sources of knowledge, using categorical tags that describe pronunciation variants and information from teachers on scoring 'paradigms' and tune to their manual scoring results. These techniques can improve overall scoring reliability and increase the level of detail of the scoring results. For example, we have improved recognition of letter-names through optimizing the dictionary for each letter-sound independently. More generally, by using dictionaries containing relevant pronunciation variations of the target words and by applying garbage models for unexpected speech data, the speech recognition task can be reduced to the simpler and more accurate task of speech verification. By adding specific expected 'accented' pronunciations to the dictionary, we can remove a built-in bias against children who speak with an accent so that we can better separate reading errors from accentedness, and, for some tasks (syllable blending) we achieve machine scoring comparable to human scoring. For other tasks, the automatic scores provide results comparable to hand-scoring when aggregated over sets larger than one child reading one word. For example, aggregating across the class for individual words can determine which words are hardest for the class, or aggregating over all words spoken by each child can determine which children are doing the best and the worst.

Simple machine recognition of pronunciation variants is not sufficient to inform judgment of a child's reading skills. Not only were many of the pronunciation variants too close to reliably distinguish automatically (or manually, for that matter), but each pronunciation's context in the child's overall pronunciation trends also played a large part in how a particular instance might be judged. Because one word alone, out of context, is not sufficient to make a reliable judgment, more sophisticated pronunciation verification techniques were needed. Toward this end, we found it beneficial to augment the recognition lexicon with tags for general pronunciation categories, as explained in Section 3. These pronunciation tags, automatically verified along with their corresponding target words, can be used to further enhance the verification feature set and to provide meaningful feedback to teachers. Although it may be impossible to perform a reliable diagnosis on the basis of one utterance, an accumulation of evidence through use of these tags can improve reliability.

Teachers, too, need to take a child's learning background into account when evaluating performance and thus may have varying decision criteria from child to child. For example, if a Kindergarten-level Spanish-accented child blends ***p*** + ***ot*** as ***bot***, teachers may consider this a demonstration of the child's emerging skills in the phonological transfer from Spanish to English. While for a first or second grade child, such a blending response may not be acceptable. This may partly explain the performance differences between the machine ratings and those of the teachers. To improve performance and also diagnostic utility for teachers, future work needs to address the issue of emerging skills.

One way we have incorporated additional information into the dictionary is by predicting classes of expected pronunciations. We have done this for predictions based on a mapping of Spanish phonology to English and we have also predicted pronunciations based on Spanish letter to sound rules. We have added constraints that help restrict the recognized utterances to a single style. The resulting dictionary not only improves performance, but also enables use of a diagnostic tag to an 'incorrect' label. This is important in assessing pronunciations of native speakers of various dialects, but it is crucial in assessing reading by language learners. For English language learners, reading a word in English is confounded by phonological knowledge of a first language. These children seem to negotiate and hesitate between two sound systems and two different letter-to-sound orthographies. Teachers need to know more than whether or not the students read a word correctly; they also need to identify the phonological and orthographic knowledge of the students in both Spanish and English in order to plan for instruction. A student who applies Spanish letter to sound rules to English words shows emerging reading skills in English, whereas a student who seems to guess or supply random words may still be struggling with the concept of mapping letters to sounds. These two students require different interventions.

5.2. Challenges

We knew that speech recognition of young children in school settings, particularly those learning a new language, would be a challenging task. Through this investigation we have found significant new challenges. For example:

- Teachers do not always agree on what should be counted as correct, and count the same response as correct or incorrect depending on a variety of factors, including the purpose of the test, the instructions given preceding the test, their knowledge of the child's language and literacy development, their knowledge of the assessment, etc.
- Objective independent ratings of degree of foreign accent are typically not available and yet they are an important aspect of rating reading skill based on oral

reading. School ratings of language level, at least in California, are focused on reading and writing skills, and not on speaking skills.

- Although teachers would like a diagnosis of the source of errors, this is a complex issue since errors are often a melding of various factors including languages being learned, orthographic systems being learned, letter shape confusions, letter-sound confusions, reversals of letter sequences, biases toward more familiar or recently used words, etc. Further, more than one error source can give rise to the same error, and error sources can be combined to create a quite large number of possible pronunciations.
- Many of the errors marked by teachers in student productions did NOT belong to an easily detectable recurring pattern; some errors were clearly the result of a particular influence (e.g., letter reversals, using Spanish letter-to-sound rules), but a great many of them were not clearly a part of a systematic pattern.

5.3. Future directions

There are some obvious future directions: analyzing more of the data already collected (more students, more grades, more assessments, video data, etc.) and applying more noise robust techniques. A particularly compelling area for future research involves the diagnostic tags. We have begun to explore how best to use them to inform teachers and to improve recognition, and whether, for example, responses on a scale rather than binary *correct*/*incorrect* labels are useful or not. Improving tag usefulness first requires developing an easy way for teachers to provide input. Most teachers are NOT phoneticians, so it is not easy for them to be reliable and explicit about what they heard. One approach might take advantage of sorting the words the child has produced in, for example, a picture naming task in which the teacher could listen to a child naming the number '6' and a picture of 'keys' to help determine if there is a reliable difference between the vowels in these words, and how the distinction is made. Listening to those two words and then deciding if a reading of the word 'dip' has the vowel more like 'six' or more like 'keys' might be easier than transcription. If the child reliably separates /ih/ and /iy/ in other words, and uses /iy/ in 'dip', it is likely NOT a phonology issue, but a letter to sound rule application, since Spanish letter to sound rules would predict this pronunciation.

Another alternative would be to apply more than one tag: a production of /d iy p/ for 'dip' could be a phonological error or a letter-to-sound error. Although this complicates the modeling, as with any diagnosis problem, it is realistic to have multiple tags: Any given symptom could be caused by many different diseases, and the diagnosis is guided by the frequency/rarity of the disease in the given population, but also by the notion that a set of symptoms should cluster together. One can also take advantage of the fact that some instances of tags are more diagnostic than others; e.g., variants predicted by only one hypothesized source are better evidence for that source than those that have more than one possible source, as in the 'dip' example.

In view of either of these directions, it is probably useful to include in a picture naming task some minimal pairs (e.g., 'mitt'/ 'meat') to obtain a better model of the child's phonology, and also to assess the child's knowledge of the *meanings* of items, since reading is after all a mapping between sound and meaning and it might not make sense to assess reading of words the child does not know. This task, or subparts of it, could be repeated every now and then to get an idea of phonological development. Eventually, one could imagine an adaptive system that proposes a targeted picture naming task or game in order to determine baseline phonology. These directions could also benefit from inclusion of test items that are more diagnostic themselves, in the sense of providing a single hypothesis for variant pronunciations. Provided we can find an efficient and consistent means of getting teacher ratings, another interesting direction is to generalize the notion of the tags to include additional properties of a speaker, e.g., dialect, issues such as dyslexia, or certain speech disorders, such as lisping or stuttering.

Another future direction involves the incorporation of yet more knowledge sources into the modeling, for example, timing and inter-assessment information. Using information about a student's previous assessment performance (same assessment or different) should yield better recognition accuracy and more informed scoring. For example, a student who pronounced /r/ as /w/ before may be more likely to make this substitution again compared to a student who has not done so.

An important direction for future work is generalizing the techniques to other languages and to cases in which there may be no information about the child's first language. One possibility would be to repeat the approach here, combining human expert and data-driven techniques embodied as diagnostic tags in the dictionary for each desired language. A teacher could select for each child or for each class which languages other than English should be used. It might be possible to load in all pronunciation sets for all languages available and do one or more recognition passes after which tag sets that score poorly are removed from consideration. Alternatively, English only could be used in a first pass and if scores are below a certain threshold more information is obtained, either by asking a human to enter the other languages to be used or by instigating probes to determine a first language. Given dialect variability and the fact that the distinction between a dialect and a language is not clearly defined, the multiple hypotheses with constraints and pruning seems to be a better approach. One could also imagine schemes based on picture naming tasks that try to map out the phonology of the speaker and then modify the dictionary based on that system, with no need to know anything about dialect or native language – which may be similar to the process

people might use in dealing with a dialect or accent they have never heard before. In any case, this clearly is an interesting area for further work.

In sum, we have investigated the types of knowledge sources that teachers use, and have tried to incorporate them into an automated system. We hope that our results can contribute to the consistency and fairness of reading assessment. We have found that knowledge sources beyond the acoustics of the signal are required to provide adequate scoring and diagnostic information. This is obvious in the sense that we already know that language is a system not of absolute values of the signaling mechanism, but, rather, a system of contrasts relative to the context. What is not so obvious is how we can adequately model all aspects of the context important to determining when a word was read incorrectly, but also how best to diagnose the cause of the error in order to correct it. We found assessment of emerging reading skills in young children to be an area ripe for more research!

Acknowledgements

We gratefully acknowledge assistance from Kimberly Reynolds and Barbara Jones for the work reported in this paper. We deeply appreciate the cooperation of the many students and teachers who graciously collaborated with us in this study and the support through the National Science Foundation (NSF) of IERI Project # 0326214. Opinions are those of the authors and not of NSF.

References

Alwan, A., Bai, Y., Black, M., Casey, L., Gerosa, M., Heritage, M., Iseli, M., Jones, B., Kazemzadeh, A., Lee, S., Narayanan, S., Price, P., Joseph, J., 2007. A system for technology based assessment of language and literacy in young children: the role of multiple information sources. In: Proc. IEEE Internat. Workshop Multimedia Signal Processing, Greece, pp. 26–30. <http://diana.icsl.ucla.edu/Tball/publications/tball_mmsp07.pdf>.

August, D., Shanahan, T. (Eds.), 2006. Developing Literacy in Second-Language Learners: Report of the National Literacy Panel on Language-Minority Children and Youth. In: Erlbaum Lawrence, Mahweh, N.J., Executive. <http://www.cal.org/projects/archive/nlpreports/Executive_Summary.pdf>.

Barker, T.A., Torgesen, J.K., 1995. An evaluation of computer assisted instruction in phonological awareness with below average readers. J. Edu. Comput. Res. 13, 89–103.

Barron, R.W., Golden, J.O., Seldon, D.M., Tait, C.F., Marmurek, H.H., Haines, L.P., 1992. Word recognition in early reading: a review of direct and indirect access hypotheses. Cognition 24, 93–119.

Black, M., Tepperman, J., Kazemzadeh, A., Lee, S., Narayanan, S., 2008. Pronunciation verification of English letter-sounds in preliterate children. In: Proc. InterSpeech ICSLP, Brisbane, Australia, pp. 2783–2786.

Buntschuh, B., Kamm, C., DiFabbrizio, G., Abella, A., Mohri, M., Narayanan, S., Zeljkovic, I., Wright, J., Marcus, S., Sharp, R.D., Duncan, R., Wilpon, J., 1998. VPQ: a spoken language interface to large scale directory information. In: Proc. ICSLP, Sydney, Australia, pp. 2863–2867.

Cassell, J., Ryokai, K., 2001. Making space for voice: technologies to support children's fantasy and storytelling. Personal Technol. 5 (3), 203–224.

Cohen, P., Johnston, M., McGee, D., Oviatt, S., Clow, J., Smith, J., 1998. The efficiency of multimodal interaction: a case study. In: Proc. ICSLP, Vol. 2, Sydney, Australia, pp. 249–252. <http://www.research.att.com/~johnston/papers/cohenetalicslp98.pdf>.

Coulston, R., Oviatt, S.L., Darves, C., 2002. Amplitude convergence in children's conversationals with animated personas. In: Proc. 7th Internat. Conf. Spoken Language Process., Vol. 4, Denver, CO, pp. 2689–2692.

Dalby, J., Kewley-Port, D., 1999. Explicit pronunciation training using automatic speech recognition technology. CALICO J. 16 (3), 425–445.

Darves, C., Oviatt, S.L., Coulston, R., 2002. Designing effective conversational interfaces for next-generation educational software. In: Proc. 7th Internat. Conf. Spoken Language Process., Denver, CO, pp. 16–20.

DiFabrizzio, G., Ruscitti, P., Narayanan, S., Kamm, C., 1999. Extending computer telephony and IP telephony standards for voice-eEnabled services in a multi-modal user interface environment. In: Proc. Interactive Dialogue Multi-modal Systems, Kloster Irsee, Germany, pp. 9–12.

EduSpeak. <http://www.speechatsri.com/products/eduspeak.shtml>.

Eguchi, S., Hirsh, I.J., 1969. Development of speech sounds in children. Acta. Otolaryng. 257 (Suppl.), 1–51.

Eskenazi, M., 1999. Using a computer in foreign language pronunciation training: what advantages? CALICO J. 16 (3), 447–469.

Farmer, M.E., Klein, R., Bryson, S.E., 1992. Computer-assisted reading: effects of whole-word feedback on fluency and comprehension in readers with severe disabilities. Remedial Special Edu. 13 (2), 50–60.

Gerosa, M., Giuliani, D., Brugnara, F., 2007. Acoustic variability and automatic recognition of children's speech. Speech Comm. 49, 847–860.

Goldstein, U.G., 1980. An Articulatory Model for the Vocal Tracts of Growing Children, Ph.D. Thesis, MIT, Cambridge, MA.

Gorin, A., Riccardi, G., Wright, J., 1997. How may I help you? Speech Comm. 23, 113–127.

Hagen, A., Pellom, B., Cole, R., 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. Speech Comm. 49 (12), 861–873.

Harris, A.J., Jacobson, M.D., 1982. Basic Reading Vocabularies. Macmillan Publishing Company, New York.

Jones, B., Heritage, M., Boscardin, C.K., Min, H., 2007. Bringing it all together: articulating an early reading assessment framework for English learners. In: Presented at American Educational Research Association (AERA), Chicago. <http://diana.icsl.ucla.edu/Tball/publications/AERA%202007%20Jones.pdf>.

Kazemzadeh, A., You, H., Iseli, M., Jones, B., Cui, X., Heritage, M., Price, P., Anderson, E., Narayanan, S., Alwan, A., 2005. TBALL data collection: the making of a young children's speech corpus. In: Proc. Eurospeech/INTERSPEECH-2005, Lisbon, Portugal, pp. 1581–1584. <http://diana.icsl.ucla.edu/Tball/publications/tball_data.pdf>.

Kent, R.D., 1976. Anatomical and neuromuscular maturation of the speech mechanism: evidence from acoustic studies. J. Speech Hear. Res. 19, 421–447.

Khalili, A., Shashaani, L., 1994. The effectiveness of computer applications: a meta-analysis. J. Res. Comput. Edu. 27, 48–61.

Lamel, L., Bennacef, K., Rosset, S., Devillers, L., Foukia, S., Gangolf, J.J., Gauvain, J.L., 1997. The LIMSI RailTel system: field trial of a telephone service for rail travel information. Speech Comm. 23, 67–82.

Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: developmental changes of temporal and spectral parameters. J. Acoust. Soc. Amer. 105, 1455–1468.

Lovett, M.W., Barron, R.W., Forbes, J.E., Cuksts, B., Steinbach, K.A., 1994. Computer speech-based training of literacy skills in neurologically impaired children: a controlled evaluation. Brain Language 47, 117–154.

Ma, J., Yan, J., Cole, R., 2002. CU Animate tools for enabling conversations with animated characters. In: International Conference on Spoken Language Processing (ICSLP), Vol. 1, Denver, CO, USA, pp. 197–200. <http://cslr.colorado.edu/beginweb/cuanimate/cuanimate_paper.html>.

McCullough, C.S., 1995. Using computer technology to monitor student progress and remediate reading problems. School Psychol. Rev. 24, 426–439.

Mostow, J., Hauptmann, A.G., Roth, S.F., 1995. Demonstration of a reading coach that listens. Proc. ACM Symp. User Interface Software Technol., 77–78.

Narayanan, S., Potamianos, A., 2002. Creating conversational interfaces for children. IEEE Trans. Speech Audio Process. 10 (2), 65–78.

National Reading Panel, 2000. Teaching Children to Read: An Evidence-based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction, NICHD, NIH Publication 00-4769, Washington, DC. <http://www.nationalreadingpanel.org/Publications/subgroups.htm>.

Oviatt, S., Adams, B., 2000. Designing and evaluating conversational interfaces with animated characters. In: Cassell, J., Sullivan, J., Prevost, S., Churchill (Eds.), Embodied Conversational Agents. MIT Press, pp. 319–343.

Potamianos, A., Narayanan, S., 2003. Robust recognition of children's speech. IEEE Trans. Speech Audio Process. 11, 603–616.

Price, P., 2007. Multimedia technologies and solutions for educational applications: opportunities, trends and challenges. In: Proc. IEEE Internat. Workshop Multimedia Signal Process., Greece. <http://diana.icsl.ucla.edu/Tball/publications/mmsPrice07Final.pdf>.

Project LISTEN. <http://www.cs.cmu.edu/~listen/>.

Russell, M., Brown, B., Skilling, A., Series, R., Wallace, J., Bonham, B., Barker, P., 1996. Applications of automatic speech recognition to speech and language development in young children. In: Proc. ICSLP, Philadelphia, PA, pp. 176–179.

Sharma, R., Pavlovic, V., Huang, T., 1998. Toward multimodal human computer interface. Proc. IEEE 86 (5), 853–869.

Shefelbine, J., 1996. BPST – Beginning Phonics Skills Test. <http://www.sandi.net/staff/circulars/20012002/ac/acpdf/ac.0066.attach.pdf>.

Smith, B.L., 1992. Relationships between duration and temporal variability in children's speech. J. Acoust. Soc. Amer. 91, 2165–2174.

Takezawa, T., Morimoto, T., 1998. A multimodal-input multimedia-output guidance system: MMGS. In: Proc. ICSLP, Sydney, Australia, Paper 0958.

Tepperman, J., Black, M., Price, P., Lee, S., Kazemzadeh, A., Gerosa, M., Heritage, M., Alwan, A., Narayanan, S., 2007. A Bayesian network classifier for word-level reading assessment. In: Proc. InterSpeech, Antwerp, Belgium, pp. 2185–2188. <http://diana.icsl.ucla.edu/Tball/publications/tepperman_tball_icslp07.pdf>.

Van Dusen, L., Worthen, B., 1993. Factors that facilitate or impede implementation of integrated learning systems. In: Bailey, G. (Ed.), Computer-Based Integrated Learning Systems. Educational Technology Publications, Englewood Cliffs, NJ, pp. 35–48.

Wang, S., Price, P., Heritage, M., Alwan, A., 2007. Automatic evaluation of children's performance on an English syllable blending task. In: Proc. SLaTE, Farmington Pennsylvania, pp. 120–123.

Watch me! Read. <http://www.ibm.com/ibm/ibmgives/grant/education/programs/reinventing/watch.shtml>.

Whitehurst, G.J., Lonigan, C.J., 1998. Child development and emergent literacy. Child Dev. 69 (3), 848–887.

Wiburg, K., 1995. Integrated learning systems: what does the research say? The Comput. Teacher 22 (5), 7–10.

Williams, S.M., Nix, D., Fairweather, P., 2000. Using speech recognition technology to enhance literacy instruction for emerging readers. In: Fishman, B., O'Conner-Divelbiss, S. (Eds.), Proc. 4th Internat. Conf. Learn. Sci.. Erlbaum, Mahwah, NJ, pp. 115–120.

Xiao, B., Girand, C., Oviatt, S.L., 2002. Multimodal integration patterns in children. In: Proc. 7th Internat. Conf. Spoken Language Process., Vol. 1. Denver, CO, pp. 629–632.

You, H., Alwan, A., Kazemzadeh, A., Narayanan, S., 2005. Pronunciation variations of Spanish-accented English spoken by young children. In: Proc. Eurospeech, Lisbon, pp. 749–752. <http://diana.icsl.ucla.edu/Tball/publications/SpanishSpeech2005.pdf>.

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., Hetherington, L., 2000. Jupiter: a telephone-based conversational interface for weather information. IEEE Trans. Speech Audio Process. 8, 85–96.