A System for Technology Based Assessment of Language and Literacy in Young Children: the Role of Multiple Information Sources.

Abeer Alwan†, Yijian Bai†, Matt Black*, Larry Casey‡,
Matteo Gerosa*, Margaret Heritage‡, Markus Iseli†, Barbara Jones‡, Abe Kazemzadeh*,
Sungbok Lee*, Shrikanth Narayanan*, Patti Price§, Joseph Tepperman*, Shizhen Wang†
*SAIL, Department of Electrical Engineering, University of Southern California
†SPAPL, Department of Electrical Engineering, University of California, Los Angeles
‡CRESST, Department of Electrical Engineering, University of California, Los Angeles
§PPRICE Speech and Language Technology

***Abstract*— This paper describes the design and realization of an automatic system for assessing and evaluating the language and literacy skills of young children. This system was developed in the context of the TBALL (Technology Based Assessment of Language and Literacy) project and aims at automatically assessing the English literacy skills of both native talkers of American English and Mexican-American children in grades K-2. The automatic assessments were carried out employing appropriate speech recognition and understanding techniques.**

In this paper, we describe the system focusing on the role of the multiple sources of information at our disposal. We present the content of the assessment system, discuss some issues in creating a child-friendly interface, and how to provide a suitable feedback to the teachers. In addition, we will discuss the different assessment modules and the different algorithms used for speech analysis.

I. Introduction

In this paper we describe an automated assessment system developed in response to the growing need for reliable and objective reading assessments in US schools. The 2000 Report of the National Reading Panel [1] advocated the use of classroom-based assessments to inform reading instruction, enabling teachers to gather data about a large number of discrete skills including phonological awareness, alphabet knowledge, word decoding, sight recognition of familiar words, fluency, vocabulary, and comprehension skills. A system designed to robustly meet these broad demands must make use of multiple information sources when eliciting responses from the children, automatically processing these responses, and reporting assessment scores to the teachers.

Our system provides teachers of grades K-2 with an on-demand technologically-sophisticated tool that allows them to efficiently gather data about their students' language skills from reliable classroom-based assessments in order to plan individualized instruction tailored for each child's needs. The developed system consists of 3 main parts:

- a multimedia interface for presenting stimuli in audio, text, and graphics, and for collecting data over various sources and modes;
- an assessment module that makes use of Automatic Speech Recognition (ASR) technology and is capable of analyzing and scoring the students' responses in a reliable, fair, and efficient manner;
- an interface for teachers, connected to a data mining system that monitors students' progress, creating a query-based database for students, groups, and classes.

In eliciting reading assessment responses from children through the multimedia interface, it is clear that no one testing modality - sound, graphics, text, or otherwise - is appropriate for all assessment tasks. There are many assessments - that of phoneme blending, for example - for which audio prompts illustrate the question's intent much more clearly than text-based ones. In many cases the test's fairness or effectiveness can be improved by presenting multiple modes of information to the student - for example, the assessment of reading comprehension by the combined use of text and accompanying audio. The shortcomings of each information source individually can be mitigated by a careful combination of these testing modalities.

US schools are very diverse - in the Los Angeles Unified School District (LAUSD), for instance, over 40% of students are considered English Learners. For this reason, our ASR processing of the tested child's speech input accounts for the possibility of accented pronunciations (typically that of Mexican Spanish), so that nonnative speakers' reading abilities can be judged without bias. To this end, we incorporate multiple information sources into the assessment module in the form of prior knowledge of child demographics and the target text and its expected pronunciation variants, in a Bayesian statistical framework along with the speech signal itself.

Often teachers will not agree on a strict definition of what constitutes a "right" answer in every case, for every type of student - this is true, for instance, when it comes to open-ended reading comprehension questions. In a case such as pronunciation assessment of words read out loud, we might say that the correct answer depends on a wide variety of interdependent sources, the exact relationship among which is not always clear. This further explains our motivation in using multiple information sources in speech technology assessment, but also suggests that the teacher interface for reporting test results could benefit from access to all knowledge obtained through elicitation and speech recognition, including test prompts, pronunciation scores, and audio recordings. Our intention is to translate as much of the useful variability over all modalities directly from the student to the teacher, while providing tools with which to tease apart the interactions among the information sources, making each student's unique needs more evident to the teacher.

This paper is organized as follows. Section II presents the children's interface. Section III presents the content of the system and the assessment tasks. The corpus collected as a first step of the project is briefly described in Section IV. Section V describes the ASR technology used in the different assessment tasks. The teacher interface is described in Section VI and conclusions and final remarks are reported in Section VII.

II. The children's interface

The first module of the system is the children's interface, which provides appropriate stimuli to the children and records and collects their responses. By "appropriate" we mean stimuli suitable for the child's grade level and for the task at hand. For example, certain assessments cannot be reliably administered to younger students without audio prompts to substitute for fluent reading. Other assessments incorporate certain types of stimuli because they are demanded by the nature of the test and the type of response it is designed to elicit.

In addition to the choice of modality, a number of different aspects must be taken into account when deciding on the appropriateness of the stimuli. Important factors are, for example, the amount of time for which the stimuli is presented, the voice type in the case of audio stimuli, and the character style in the case of written stimuli. The combination of all these factors determines the appropriateness of the stimuli for a certain grade level and task.

When dealing with young children the quality of the interface is critical, especially if the system is designed so that children interact with it on a one-on-one basis without direct supervision, as in our case. The system must provide clear directions, be sufficiently stimulating to maintain focus while limiting distractions, and motivate the children to complete the assessment task. Prior research has shown that children as young as 4 years old can interact comfortably with computerized agents [2].

To enable interactive and child-friendly assessment systems, we have designed a mixed-initiative conversational interface with multimodal capabilities (i.e. supporting visual display, touch-screen input, and auditory stimuli). Each assessment has a specific character associated with it and the character presents the task (see Figure 1 for an example character). The children's interface enables students to independently access the assessments and to independently follow directions and complete tasks, thus minimizing the time spent by teachers in acquiring information about individual students.

Fig. 1. *Example of an animated character presenting a word.*

In most of the assessments the interface goes beyond pointing and clicking by using appropriate speech recognition and understanding techniques to assess the children's literacy skills. These assessments are not limited to reading words or sentences - for example, a child can be asked to blend two syllables into a single word, to answer questions about a story read to them, or to identify the sound represented by an alphabet character. Analysis not only takes into account the speech uttered by the child, but combines other data inputs such as the speed of the response and the child's demographic information (age, grade, and language background).

During the development of the interface several usability studies were conducted with children and teachers. A critical issue that emerged from these studies was the timing of the assessment tasks. For example, more proficient children lost interest if the assessment task progressed too slowly for them, while less proficient children often needed more time between items in the individual assessments. Furthermore, some teachers wanted to be able to adjust the timing for English language learner students to accommodate differences in language proficiency, while others wanted the same speed to apply to all students.

To solve this issue, in the final version of the interface, the child must press a key after completing each item in an assessment to proceed to the next item. In this way the child has some control over the speed of the assessment and highly proficient children don't feel the system is too slow for them. However, following the suggestion of teachers and educators, we set also a maximum allowed time (usually 5 seconds) to complete an item. If after that the child still has not pressed the key to continue, the system goes on to the next item automatically. If this automatic timeout happens three times in a row, the assessment stops as we assume the child is not ready for it. This is another example of an additional information source used in these assessments.

III. Content and Assessment Tasks

The assessment system was designed to provide child-friendly, computer-based individual screening and to give monitoring assessments that offer maximum flexibility and utility, are easy to use and interpret, and provide teachers with ongoing information for instructional planning.

The individual assessments are situated in an original and comprehensive framework that addresses the specific, critical skills, identified in the reading literature [3], [4], [1], [5], [6], [7], [8], [9] that children need to acquire in the early grades of schooling to become proficient readers. These include assessments of phonological awareness and phonic knowledge, word identification, oral and written language comprehension, accuracy, and rate in reading text.

The assessment framework was developed in collaboration with a group of expert reading teachers. All the assessments were designed to be embedded in an instructional framework so that the results could be used to plan instruction. In addition, we designed the content so as to minimize the amount of time employed by the teacher and child while maximizing the informational benefit of the assessment.

Following these guidelines, we designed a hierarchical rather than a uniform approach to assessment. All students take benchmark assessments as a check on progress, and particular students take “drill down” assessments related to specific skills on an as-needed basis. Using the information from the core assessments, teachers make determinations about whether sufficient information has been gained to plan the next instructional steps, or whether certain “drill down” assessments are necessary for diagnostic purposes. An important feature of the framework is that assessments are organized around teacher decision points. This organization enables teachers to know what assessments they need to target for diagnostic purposes and when instructional interventions may be necessary.

In designing the content of the assessments we had to consider also the children’s interface and how we elicited the children’s response. In fact, the content and the elicitation modes cannot be considered separately: a particular paragraph can be suitable as the input for a listening comprehension task but may be too complicated for a reading comprehension task. The elicitation mode was particularly important when designing assessments for children in grade K, since most of them are not ready to read. In fact the tests have to be keyed to a specific grade and only a few can be presented to children in all grades K-2. For this reason, we have different versions of the same assessment task for different grades. Among the assessments we have lists of words, recognizing the name and sound of alphabet letters, blending syllables into whole words and reading or listening to a short paragraph and then answering to a set of questions about it.

In summary, the system includes more than 10 different assessment tasks. Some of these are keyed to children in a certain grade, while other can be presented to children in all grades. What follows is a brief description of the assessment tasks where the child has to utter a word or letter, and the ASR-based technologies that are used in the evaluation of student responses.

IV. Collection and analysis of in-domain children’s data

ASR for children is a difficult task: it is well known that acoustic and linguistic characteristics of children’s speech are widely different from those of adult speech and vary rapidly with age [10], [11], [12], [2]. For recognition of children’s speech, age-specific acoustic models (AMs) trained on speech collected from children of the target age, or group of ages, should be adopted to ensure good recognition performance [13], [14].

For this reason, the first step in the TBALL project was the collection of a corpus of children’s speech from about 256 students in grades K-4 in the Los Angeles Unified school district [15]. These students come from diverse socio-economic backgrounds and their number includes not only speakers of English as a second language, but also children who are acquiring English as a first language but with the accent/pronunciation characteristics of Los Angeles Chicano English (a dialect of English spoken by Mexican-descended Americans) [16], [17].

A “Wizard of Oz” interface simulating the fully-deployed system was used to collect the data. In this way precious information could be collected regarding children’s behavior and response to the prompts, their interest in the various assessments, and in general their response to the system. These data were analyzed by teachers and experts in educations and were used in the design of the final children’s interface.

A data-driven analysis of the pronunciation variations in Spanish-accented English was carried out on the corpus [18]. The analysis showed that there is a significantly large pronunciation variation for phonemes that do not exist in Spanish, and it confirmed some of the variations derived by knowledge-driven pronunciation modeling. These rules were used to form an extended lexicon that was used in all the ASR-based assessments.

V. Automatic assessment of children’s speech

Even if several modalities can be used to assess children’s language and literacy, speech is the most used mode of collecting students’ response and is the one that usually allows a teacher insight into the most information. In our system, the speech recognition module captures students’ responses on the assessment tasks and scores those responses according to standards that the research team and the collaborating teachers have established.

a) Word Verification: Arguably, the most important part of automatically assessing a new reader’s literacy level is verifying his or her pronunciation of read-aloud target words. However, pronunciation evaluation is a difficult task, and especially difficult in children, non-native speakers and pre-literates. In fact, to accurately judge a child’s literacy level by the pronunciation of read-aloud words, a teacher must first distinguish between simple allowed variants in pronunciation and true reading mistakes that betray a lapse in comprehension. In addition, prior knowledge of the child’s age, native language or regional dialect may influence what one believes to be an acceptable pronunciation [19]. For example, to pronounce ”can” as the Spanish-influenced /k aa n/ may for some children be an acceptable variant, but to render it as /k ey n/ (the common mistake of making the vowel say its name) shows a more evident misunderstanding of letter-to-sound decoding rules.

This task is usually addressed by formulating it as a traditional pronunciation verification problem: given an utterance we compare the likelihood it was drawn from some target model distribution to the likelihood that it comes from a filler model of expected reading mistakes [20], [21]. However, considering the complexity of the task, a likelihood ratio or even a pruned decision tree will not consider all available and necessary cues in making an automatic literacy assessment decision.

The word verification assessment of this system makes use of a Bayesian network classifier that models the generative story among speech recognition based features, treating pronunciation variants and reading mistakes as distinct but not independent cues to an overall qualitative perception of reading. With this novel method we are able to assess speaker’s overall literacy level with accuracy approaching the agreement between expert evaluators. The details of the automatic assessment algorithm and the results achieved are reported in [22].

b) Syllable Blending: Syllable blending is one assessment that measures phonemic awareness, an important skill to master to become proficient readers [23]. In particular it tests children's ability to orally blend a syllable into a whole word, such as **ta + ble = table**. The syllable blending task is designed to assess both pronunciation accuracy and blending skills (smoothness). In fact, a child is considered proficient in this task if:

- he/she reproduces all the sounds of the original syllable in the final word;
- he/she smoothly blends the two syllables together to form the word.

The automatic evaluation system is composed of 2 steps: first the system makes use of an extended pronunciation dictionary for word verification and forced alignment to generate the syllable segmentation. Then a weighted summation of normalized likelihood scores and duration scores is used to evaluate the overall quality of the child's response. The system achieves a correlation with teacher scores at least as good as the average inter-teacher correlation. The details of the automatic assessment algorithm and the results achieved are reported in [24].

c) Letter Naming: The letter naming task is an assessment where the child is simply asked to say the alphabet name of a given letter. This task is a well-known one that has been studied extensively in the past, but not for young children learning to read. In the reading verification paradigm, the focus is on determining whether the correct letter has been said or not, rather than recognizing which letter was said as in the case of alphabet recognition.

The accept/reject decision was made taking into account recognition results, the likelihood of a forced alignment with the target, and a confidence score. About 80% verification accuracy was obtained in this task.

d) Letter Sound: In the letter sound task the child is asked to produce the sound that a given letter represents. In contrast to letter naming, speech recognition for the letter sound task is not as well studied. In particular, it involves recognizing speech sounds out of any word or syllabic context. Moreover, a given letter can represent several different sounds (e.g., 'c' can be both /s/ and /k/), unlike the one-to-one mapping of letters and their names.

For the recognition of this task, we tried both recognizing the utterance with a grammar to capture disfluencies and also using speech/silence segmentation as a preprocessing step. However, this task is particularly challenging given the acoustic similarity of many sounds. We are currently investigating the possibility of using articulatory information, time domain features, and a two-step verification process in which the first step has the goal of identifying the phoneme class.

e) Reading Comprehension: A basic assumption common in education is that reading comprehension can be thought of as the joint product of printed word identification and listening comprehension [25]. In the beginning stages of reading development, the limiting factor in reading comprehension is primarily decoding ability. At the beginning of the literacy acquisition process, the correlations between reading and spoken language are small [26], [27], but when kids move beyond the beginnings of learning to read, the correlations between reading comprehension and spoken language increase, and by college level the correlation reaches 0.90 [28].

In this assessment, we were interested in assessing comprehension, and not reading fluency. In fact, for children in grades 1-2, it's not clear whether readers who struggle but in the end succeed in reading something, understand it less well than fluent readers. For this reason, fluency might not be a good predictor of comprehension.

In this assessment, ASR technology was used to recognize simple yes/no questions. Simple word spotting was used with an extended lexicon that takes into account both the Mexican-Spanish pronunciation variations and the high linguistic variability of children when speaking with a computer. Answers to open-ended questions were not analyzed in this version of the system.

Simple word spotting was used with an extended lexicon that takes into account

VI. The Teacher Interface

The teacher interface enables monitoring of students' progress and presents the results of the automatic assessments to the teachers. It is implemented as a Microsoft Excel sheet using Visual Basic Macros and can graphically display ASR results and generate useful reports for teachers. The interface connects in real-time to a MySql database, which stores ASR results.

Using a query-based data mining system, a teacher can retrieve and organize in any way all the information needed for assessment and instruction needs. Different levels can be displayed. For example, on one level, ASR results of different assessments for one class are shown.

last_name	first_name	Reading Comprehension	Syllable Blending	Letter Naming
ESPINOZA	PAULINA	1	8	15
EVANS	DYLAN	1	8	14
GONZALEZ	JAYBE	2	7	10
HERNANDEZ	SAMANTHA	3	5	19
JONKER	KIMBERLY	4	6	19
MADULE	JOSEPH	5	5	20
MANJARREZ	CHRISTIAN	6	7	20
MARTINEZ	NICOLAS	7	7	18
MUNOZ	MAYA	8	8	19

Fig. 2. *Example of a particular view of the top level teacher interface.*

Fig. 2 shows a simplified screenshot of the top level with the assessments "Reading Comprehension", "Syllable Blending", and "Letter Naming" arranged horizontally and the student names listed vertically. Each number represents an ASR score per student and assessment. The ASR scores are color-coded, representing different levels of proficiency.

In addition to ASR, the teacher interface can access all the information derived from the various knowledge sources gathered during the assessments, including test prompt elicitation mode, pronunciation scores, audio recordings, response speed, and child demographic information. The goal is not only to transfer this information directly from the students to the teacher, but also to provide tools with which to organize interactions and assessments with respect to different information sources, pointing out to the teachers what each individual student needs.

While some of the information sources are used directly in the ASR automatic assessments, others are just presented to the teachers so that they can independently decide how much weight to give them and make a more conscious decision about possible instructional interventions. It is important to remember that this automatic evaluation system is not meant as a teacher replacement, but as an instructional aid to the teachers and students to improve the quality of instruction by gathering information on children's skills without investing a large amount of the teacher's time.

VII. Conclusions

This paper described the design and realization of an automatic system for assessing the language and literacy skills of young children in grades K-2. A multimedia interface, capable of presenting stimuli in auditory, text and graphical format, was developed to provide appropriate stimuli to the children.

Several assessment modules that make use of ASR-based technology and analyze students' responses in a reliable and efficient manner were developed. The system makes use of different information sources derived from the automatic assessments (i.e. acoustic scores, response speed), child background information (i.e. age, gender, grade, native language) and history of the child's previous assessments (i.e. number of times the child repeated the assessment, scores in similar tests). While only some of this information is used directly by the ASR algorithms, all information is accessible by the teacher to make each student's unique needs more evident when the teacher plans the next steps of the instruction. Information is provided to the teacher in a multimedia interface that shows students' scores, allows them to listen to children's responses, and organizes information from different sources in an efficient way.

In the future, we plan to develop ASR technology that can deal reliably with discourse-level phenomena so that the system can include assessments like children's retelling of stories, for example. In addition, we also propose to extend the system to assess other skills such as mathematical and scientific reasoning. Finally, we plan to expand the current system to other grade levels and language pairs.

Acknowledgment

This work was supported in part by the National Science Foundation, IERI project # 0326214

References

[1] National Reading Panel, "Teaching children to read: An evidence-base assessment of the scientific research literature on reading and its implication for reading instruction," Tech. Rep. 00-4769, National Institute for Child Health and Human Development, National Institute of Health, Washington, DC, 2000.

[2] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness and Frustration Language in Child-Machine Interactions," in *EUROSPEECH*, Aalborg, Denmark, Sept. 2001, pp. 2675–2679.

[3] M. J. Adams, *Beginning to read: Thinking and learning about print*, The MIT Press, 1990.

[4] D. K. Dickinson and P. O. Tabors, *Beginning literacy with language: Young children learning at home and school*, Brookes Publishing, 2001.

[5] B. K. Gunn, D. C. Simmons, and E. J. Kameenui, "Emergent literacy: Research base," in *What reading research tells us about children with diverse learning needs: Bases and basics*, pp. 19–50. Erlbaum, Mahwah, NJ, 1998.

[6] G. R. Lyon, "Overview of reading and literacy research," in *The Keys to literacy*, S. Patton and M. Holmes, Ed., pp. 1–15. Council for Basic Education, Washington, DC, 1998.

[7] H. Scarborough, "Connecting early language and literacy to later(dis)abilities: Evidence, theory and practice," in *Handbook of early literacy research*, S. B. Neuman and D. K. Dickinson, Ed. Guildford Press, New York, 2001.

[8] C. Snow, S. Burns, and P. Griffin, *Prevention of reading difficulty in young children*, Brookes Publishing, Washington, DC, 1998.

[9] G. Whitehurst and C. Lonigan, "Child development and emergent literacy," *Child Development*, vol. 69, pp. 848–872, 1997.

[10] S. Lee, A. Potamianos, and S. Narayanan, "Acoustic of children's speech: Developmental changes of temporal and spectral parameters," *Journal of Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, March 1999.

[11] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of Children, Women and Men: the Effect of Vocal Intensity Variation," *Journal of Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1532–1542, September 1999.

[12] S. Narayanan and A. Potamianos, "Creating Conversational Interfaces for Children," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78, Feb. 2002.

[13] J.G. Wilpon and C.N. Jacobsen, "A Study of Speech Recognition for Children and Elderly," in *Proc. of ICASSP*, Atlanta,GA, May 1996, pp. I–349–352.

[14] A. Hagen, B. Pellom, and R. Cole, "Children's Speech Recognition with Application to Interactive Books and Tutors," in *in IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop, St. Thomas, USA*, December 2003.

[15] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball data collection: the making of a young children's speech corpus," in *Proc. of INTERSPEECH/EUROSPEECH*, Lisboa, Portugal, Sept. 2005, pp. 1581–1584.

[16] C. Fought, *Chicano English in context*, Palgrave MacMillan, 2003.

[17] T. Veatch, "Los Angeles Chicano English," *Ph. D. Thesis, University of Pensylvania*, 1991.

[18] H. You, A. Kazemzadeh, A. Alwan, and S. Narayanan, "Pronunciation variation of Spanish accented English spoken by young children," in *Proc. of INTERSPEECH/ICSLP*, Lisboa, Portugal, Sept. 2005, pp. 749–752.

[19] W. Labov and B. Baker, "What is a Reading Error?," in *http://www.ling.upenn.edu/ wlabov/Papers/WRE.html*, 2003.

[20] J. Tepperma, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation Verification of Children's Speech for Automatic Literacy Assessment," in *Proc. of ICSLP*, Pittsburg,PA, Sept. 2006.

[21] D. Willet, A. Worm, C. Neukirchen, and G. Rigoll, "Confidence Measures for HMM-Based Speech Recognition," in *Proc. of ICSLP*, Sydney, Australia, 1998.

[22] J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan, "A Bayesian Network Classifier for Word-Level Literacy Assessment," , Belgium, 2007.

[23] M. J. Adamas, B. R. Foorman, I. Lundberg, and T. Beeler, "The Elusive Phoneme: Why Phonemic Awareness Is So Important and How To Help Children Develop It.," *American Educator*, vol. 22, pp. 18–29, 1998.

[24] S. Wang, P. Price, M. Heritage, and A. Alwan, "Automatic Evaluiation of Children's Performance on an English Syllable Blending Task," in *In Proc. Workshop*, 2007.

[25] P. B. Gough and W. E. Tumner, "Decoding, reading, and reading disability," *Remedial and Special Education*, vol. 7, pp. 6–10, 1986.

[26] M. Curtis, "Development of components of reading skill," *Journal of Educational Psychology*, vol. 72, pp. 656–669, 1980.

[27] T. Sticht and J. James, "Listening and reading," in *Handbook of reading research*, P. Pearson, Ed. Longman, New York, 1984.

[28] M. A. Gernsbacher, *Language comprehension as structure building*, Erlbaum, Hillsdale, NJ, 1990.