INVESTIGATING AUTOMATIC ASSESSMENT OF READING COMPREHENSION IN YOUNG CHILDREN

M. Gerosa^{+} *and S. Narayanan*^{x}

($^{+}$) FBK-Irst Centro per la Ricerca Scientifica e Tecnologica, 38050 Pantè di Povo, Trento, Italy
(^{x}) Department of Electrical Engineering, University of Southern California,
Los Angeles, CA 90089, USA

gerosa@itc.it, shri@sipi.usc.edu

ABSTRACT

This paper describes a preliminary investigation into automatic assessment of reading comprehension in young children. In particular we studied the feasibility of automatic scoring of answers to open-ended questions related to the contents of a passage read by a child. Data from 70 children in grades 1 and 2 were used in this work. An automatic speech recognition system, especially trained for children's speech, was used for tracking the read passage, and two methods for automatic assessment were tested and compared with scores assigned by elementary school teachers. Automatic assessment showed a high kappa statistics agreement with evaluation scores obtained from teachers' scores, K=0.62, comparable to the inter-teacher agreement, K=0.64.

Index Terms— Children's speech, Reading comprehension, Automatic Speech Recognition, Literacy assessment

1. INTRODUCTION

There is a growing need for reliable and objective reading assessments in US schools. In fact, the 2000 Report of the National Reading Panel [1] advocated the use of classroom-based assessments to inform reading instruction and enable teachers to gather data about a large number of discrete skills. However, the investment of teacher time and intellectual energy needed to assess students individually often limits, if not precludes, the widespread use of extensive classroom-based assessments.

In the last few years, in response to this growing need, significant research has been done on automatic assessment of children's language skills. These studies often make use of Automatic Speech Recognition (ASR) based techniques and focus mostly on the assessment of reading abilities [2], the detection of mispronunciations and reading miscues [3] and pronunciation verification [4].

However, while the assessment of skills such as word decoding or reading ability is a fairly well defined task, assessment of reading comprehension is still not well understood and needs further investigation both from a pedagogical point of view and technology facilitation point of view. In fact when assessing answers to open-ended questions, often teachers will not agree on a strict definition of what constitutes a "right" answer in every case, for every type of student [5].

For this reason often reading comprehension is not assessed directly but derived from the assessment of word reading skill. In fact, for most children with reading difficulties poor reading comprehension is a secondary problem [1, 6], since most of these children comprehend spoken material as well as the average readers but they struggle with inaccurate or slow word reading. However, there are also children that have problem understanding text material even if their word reading skills are at normal level [5]. These children often have problems comprehending main ideas and making inferences, even in spoken material. An automatic system able to assess reading skills and comprehension skills separately would be really useful to single out children with this behavior, that are often missed by teachers in lower grades. So, though there is most likely a correlation between reading fluency and comprehension, it is dangerous to use it diagnostically for children in lower grades.

In this paper we describe a preliminary investigation into automating assessment of reading comprehension in young children, by studying the feasibility of automatic assessment of answers to open-ended questions connected to a text read by a child. This work was carried out in the context of the Technology-Based Assessment of Language and Literacy (TBALL) project (http://diana.icsl.ucla.edu/Tball/assess_frame.html). This project aims at automatically assessing the English literacy skills of children in grades K-2, both native talkers of American English and those that are English language learner or bilinguals with Mexican Spanish background.

Data from 70 children in grades 1 and 2 were manually annotated, scored and analyzed. Two methods for automatic assessment were evaluated and their respective results were compared with manual scores given by 5 elementary school teachers. Results showed that our automatic assessment demonstrates a high agreement and correlation with respect to expert teacher assessments.

The paper is organized as follows. The speech corpus used for this study is described in Section 2. Section 3 presents description and analysis of the manual scoring of answers to open-ended questions. Section 4 describes the ASR setup and the experiments on automatic assessment of comprehension skills, comparing the results achieved with the reference manual scores. Final remarks are given in Section 5 which concludes the paper.

2. SPEECH CORPUS

The speech data used in this study come from recordings collected in Los Angeles public schools in the context of the TBALL Project. These data were collected from children from kindergarten through grade 2, in a classroom environment with close talking microphones. 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) [7, 8]. Children's responses were elicited

This work was partially financed by the National Science Foundation, IERI project # 0326214. The autors would like to thank the members of the TBALL project for their help and many useful suggestions.

through the use of a multimedia interface for presenting stimuli in audio, text, and graphics, suitable for the child's grade level and for the task at hand. This interface is part of a prototype of an automatic system for assessing and evaluating the language and literacy skills of young children [9]. Among the assessments we have reading 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 a set of questions about it.

In this paper we used speech data collected from the reading comprehension task. In this task each child is first asked to read aloud a short paragraph. The paragraph was the same for all children within a certain grade (only children in grades 1 and 2 were tested). After reading the paragraph each child was asked to provide answers to 8 yes/no questions and 3 open-ended questions to test the child's comprehension of what was read. Data from 70 speakers were used in this work, for a total of 2h:40m of speech. Table 1 reports the partitioning of the speakers by grade and gender.

Grade / Gender	Male	Female	Total
1^{st}	14	19	33
2^{nd}	20	17	37
Total	34	36	70

Table 1. Details about speakers grade and gender.

3. MANUAL ANALYSIS AND ASSESSMENT

Manual transcription of the read passages and of the answers given by the children, including manual annotation of truncated words and spontaneous phenomena, was carried out. Seven different spontaneous phenomena were annotated, including lip and breath noise, filled pause, unintelligible speech, non verbal sounds, laughter and microphone noise.

The answers to the 3 open-ended questions were scored independently by 5 elementary school teachers. The teachers came from different language backgrounds and had different classroom experiences, however all of them had experience with bilingual issues. Each teacher was provided an answer-key containing a total of 36 sentences as examples of acceptable, partially acceptable and unacceptable answers to each question. About 30% of the children's answers were exactly predicted in the answer-key. The teachers provided a score on a 3-point scale, rating each answers as correct (1), partial (0.5) or wrong (0). Table 2 reports the average scores of each teacher, considering partial score as 0.5, on each question, together with the kappa statistic (K) agreement.

	$1^{st}grade$			$2^{st}grade$			Total
Question	1a	2a	3a	1b	2b	3b	Total
Teacher 1	0.67	0.91	0.41	0.50	0.59	0.32	0.57
Teacher 2	0.69	0.91	0.39	0.48	0.89	0.56	0.66
Teacher 3	0.71	0.94	0.39	0.49	0.42	0.39	0.55
Teacher 4	0.56	0.92	0.40	0.42	0.67	0.39	0.54
Teacher 5	0.70	0.85	0.38	0.51	0.62	0.35	0.56
Total	0.67	0.90	0.39	0.48	0.64	0.40	0.58
K	0.78	0.75	0.88	0.50	0.26	0.59	0.64

Table 2. Average of teachers' scores and kappa statistic (K) agreement for each question.

As we can see the teachers' scores are very consistent, even if there is a clear difference between different questions. The three questions regarding the 1st grade passage are more constrained and easier to score than the three 2nd grade questions. As a consequence we can see that the K agreements obtained for the 2nd grade questions are significantly lower than the ones obtained for the 1st grade questions. As we can see from the table, the average K agreement computed over all 70 speakers and between each teacher pair is K=0.64.

3.1. Predicting comprehension from reading ability

A basic assumption common in education research is that reading comprehension can be thought of as the joint product of printed word identification and listening comprehension [10]. 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 [11], but when kids move beyond the stage of learning to read, the correlations between reading comprehension and spoken language increase, and by college level the correlation reaches 0.90 [12].

In principle, it can be assumed that for children in grades 1-2 there is a correlation between reading skill and comprehension, however it's not automatic that readers who struggle but in the end succeed in reading something, understand it less well than fluent readers.

In this work we analyzed the correlation between reading fluency and comprehension. For each child, we measured the number of disfluencies in the read passage and divided it by the total number of words in the passage. By "disfluencies", here we mean the sum of truncated words, hesitations, non verbal sounds and filled pauses. We represented the comprehension as the sum of the correct answers on the 8 yes/no questions and the average of the teachers' scores on the 3 open-ended questions, obtaining results over a 11-point scale. As expected, we have a negative correlation between the relative number of disfluencies and comprehension, the correlation coefficient is $c = -0.49$. Figure 1 shows the relation between disfluencies and total score for each speaker.

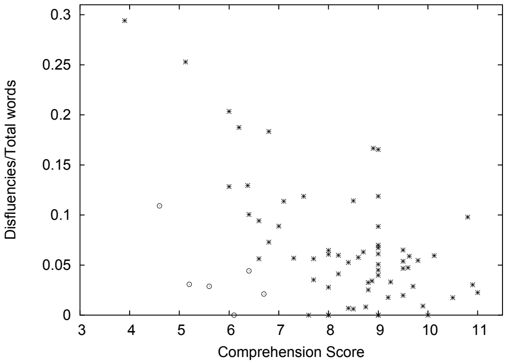

Fig. 1. *Correlation between relative number of disfluencies and comprehension score (from teachers) for each speaker. We marked differently readers with a high correlation between reading and comprehension skills (marked with a star), and readers that show poor comprehension but good reading skills (marked with a circle).*

We can see that while there is a clear correlation between the number of disfluencies and comprehension ($c = -0.49$) there are some children that show different behavior. In addition to 'fluent readers' who also demonstrate good comprehension and 'poor readers' who have poor comprehension, we can also see some fluent readers with poor comprehension skills. A practical significance of this result underscores the need for not only recognizing robustly what the child spoke in the presence of disfluencies, but also the ability to localize and identify the disfluecy [13].

4. ASR BASED ANALYSIS AND RECOGNITION

A set of recognition experiments was carried out with the aim of investigating the feasibility of automatic assessment of reading comprehension. In particular we focused on reliable assessment of open-ended questions, since the recognition of yes/no question is a well studied and documented task.

4.1. ASR setup

To be able to present results on all 70 speakers, we adopted a "leave-one-out" strategy. With this strategy we partitioned the database described in Section 2 seven times, each time selecting 10 speakers as test speakers, while the data from the remaining 60 speakers, about 2h:15m - 2h:20m, were used for training the system. In addition to these training data, we used about 48 hours of speech from the Colorado reading tutor corpus [2] and the OGI "Kid's Speech" corpus [14]. The trained system was then used in the experiments concerning the test speakers selected. This procedure was repeated 7 times allowing each speaker to appear once among the test speakers and 6 times in the training sets. On average, each system was trained exploiting about 50 hours of speech.

For the recognition experiments we used the IRST Hidden Markov Model (HMM) software package employing state-tied, cross-word triphone HMMs [15]. In particular, a Phonetic Decision Tree (PDT) was used for tying the states of triphone HMMs. Output distributions associated with HMM states were modeled with mixtures with up to 16 diagonal covariance Gaussian densities. "Silence" was modeled with a single state HMM. In addition, 7 models for the common non-verbal phenomena in our data were trained. The total number of Gaussian densities of each system was about 80000.

Each speech frame was parameterized into a 39-dimensional observation vector composed of 13 mel frequency cepstral coefficients (MFCCs) plus their first and second order time derivatives. Cepstral mean subtraction was performed on static features on an utterance-by-utterance basis.

A baseline bigram Language Model (LM), trained using the IRST LM Toolkit [16], was estimated for each of the 7 systems. The LMs were estimated using the text from the 2 passages and the manual transcriptions of the read passages from the 60 training speakers. We made use of an extended lexicon that took into account both the L1 accent (Chicano dialect) and non-native L2 Spanish influenced pronunciation variations and the most frequent truncated words present in the training set.

To better fit with the characteristics of Mexican-English accented speakers, each set of acoustic models was adapted by exploiting furthermore the data from the 60 training speakers. Maximum likelihood linear regression (MLLR) adaptation was performed, using a regression class tree for dynamic definition of regression classes during the adaptation process. Gaussian means of each regression class were adapted by using a full transformation matrix while variances were not adapted.

Using the above-described AM and LM to automatically recognize the 70 read stories, we achieved 15.2 % WER. This value is consistent with the values reported literature for similar conditions [2].

4.2. Automatic assessment of reading comprehension

We investigated two methods to reliably assess answers to open-ended questions.

The first one makes use of a manually built grammar based on rules derived from the answer-key. The grammar is largely meant to spot single words and short phrases, with the aim of detecting the correct answer. If the grammar detects all the keywords needed the answer is scored as "Correct", while if only a part of them is recognized the score is considered "Partially correct". The keyword list is determined from the possible "Correct" answers contained in the answer-key.

The second method we used makes use of the information provided in the training set. For each question we built three different language models, by adapting the baseline LM estimated on the story texts and transcriptions. For each question we created a LM for correct, wrong and partially correct answers, adapting the LM with a set of sentences that included:

- the appropriate set of sentences taken from the answer key;
- the answers given by the speakers in the training set with an average teachers' score below 0.25 for "Wrong", over 0.75 for "Correct" and between 0.25 and 0.75 for "Partially correct".

Each answer is recognized with each of the 3 grammars. Then, the automatic score is given by the grammar that provides the highest likelihood. To test the validity of this method, we computed the correlation between teachers' scores and the likelihood ratio obtained using the different LMs. In particular, for each answer we computed the average teachers' score and the likelihood ratio L, computed as $L = \frac{likelihood_0}{likelihood_1}$, where $likelihood_1$ and $likelihood_o$ are obtained while recognizing the utterance with the LMs computed with the "Correct" and "Wrong" sentences. Figure 2 shows the relation between likelihood ratio and average score for each utterance .

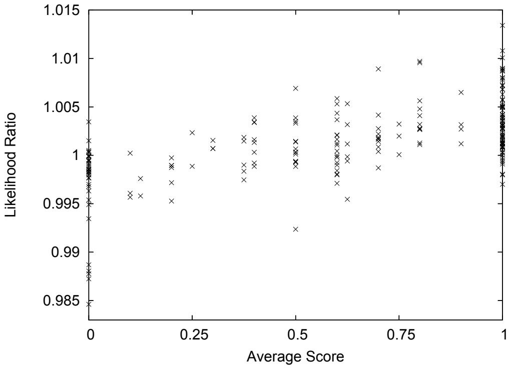

Fig. 2. *Correlation between likelihood ratio and average score for each answer.*

The correlation coefficient between likelihood ratio and average score is $c = 0.61$, suggesting a good correlation between likelihood vales obtained with the three different LMs and teachers' scores.

Table 3 reports, for each scoring method, the average scores for each question and average K agreement between the method and each of the 5 teachers. As a reference the mean value computed over the 5 teachers are reported too (from Table 2).

As we can see, using a manually built grammar provides a significantly lower agreement than the one achieved between the teachers. The main problem we faced with this method is that, as can be expected, children show a high language variability when answering open-ended questions, often uttering sentences not foreseen in the answer key. In fact, we can note how the average score for each question is significantly lower than the average teachers' score. This means that when the child answers something not foreseen in the

	$1^{st} grade$			$2^{st} grade$			Total
Question	1a	2a	3a	1b	2b	3b	Total
Manual grammar							
Avg.	0.28	0.81	0.25	0.12	0.11	0.17	0.28
K	0.30	0.58	0.61	0.05	0.05	0.30	0.34
LMs comparison							
Avg.	0.67	0.92	0.39	0.51	0.58	0.39	0.57
K	0.70	0.76	0.85	0.48	0.31	0.51	0.62
Teachers							
Avg.	0.67	0.90	0.39	0.48	0.64	0.40	0.58
K	0.78	0.75	0.88	0.50	0.26	0.59	0.64

Table 3. Average scores ("Avg.") and kappa statistics agreement ("K") for the the two automatic scoring methods, reported for each question. As a reference the mean value computed over the 5 teachers is also reported.

answer-key the grammar is not able to handle it and scores it as wrong.

On the other hand, using the information provided by the training data to adapt the baseline LMs gives good results. Both the average values and the K agreement are very similar to the ones computed on teachers' scores. In fact K agreement is higher than the one obtained among teachers for 2 questions out of 6. This high agreement value may also be due to the fact that we adapted the LM with answers labeled based on the average teachers' scores, a promising venue for further investigation.

5. CONCLUSIONS

In this paper we investigated automating assessment of reading comprehension in young children. This work focused on automatic scoring of answers to open-ended questions based on material read by children.

Manual transcription of the read passage and manual scoring of each answer was carried out for a set of 70 speakers in grades 1 and 2. Each answer was scored by 5 different elementary school teachers with different language backgrounds and classroom experiences. Analysis of manual score showed a high consistency over all speakers, with a K agreement of 0.64. Agreement is significantly higher for 1^{st} grade questions than for 2^{nd} grade questions, probably due to the higher complexity of the latter.

We showed that there is a significant correlation between fluency (measured as the relative number of disfluencies in the passage) and comprehension (represented as the sum of the scores on each question), c=0.49, but there is a clear evidence of speakers who are good readers but still have poor comprehension skills.

Experiments on automatic scoring of open-ended questions show that using a static grammar based on an answer key is not flexible enough to model the high variability of children's answers. In fact, often children formulate their answers in a unique way that is very difficult to predict beforehand. On the other hand using a few training samples to adapt simple bigram language models seems to be adequate for a reliable scoring. The correlation between average score and the likelihood ration obtained with the LMs computed with the "correct" and "wrong" sentences is significant, with a correlation coefficient of c=0.61. Moreover, the kappa statistics agreement between this method and the teachers' scores, K=0.62, is in fact almost as high as the inter-teacher agreement, K=0.64.

This result opens new prospects for the development of applications for automated assessment of reading comprehension.

6. 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] A. Hagen, B. Pellom, and R. Cole, "Children's Speech Recognition with Application to Interactive Books and Tutors," in *Proc. of IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, St. Thomas, US Virgin Islands, Dec. 2003.

[3] S. Banerjee, J. E. Beck, and J. Mostow, "Evaluating the Effect of Predicting Oral Reading Miscues.," in *Proc. of EUROSPEECH*, Geneva, Switzerland, Sept. 2003.

[4] 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," in *Proc. of INTERSPEECH/ICSLP*, Antwerp, Belgium, 2007.

[5] B. Wise and L. Snyder, *Identification of Learning Disabilities: Research to Practice*, Lawrence Erlbaum, Mahwah, NJ, 2002.

[6] G. R. Lyon, "Towards a definition of dyslexia," *Annals of Dyslexia*, vol. 22, pp. 3–30, 1995.

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

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

[9] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A System for Technology Based Assessment of Language and Literacy in Young Children: the Role of Multiple Information Sources," in *Proc. of International Workshop on Multimedia Signal Processing*, Chania, Creete, GREECE, Oct 2007.

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

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

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

[13] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, "Automatic Detection and Classification of Disfluent Reading Miscues in Young Children's Speech for the Purpose of Assessment," in *Proc. of INTERSPEECH/ICSLP*, Antwerp, Belgium, 2007.

[14] K. Shobaki, J.P. Hosom, and R.A. Cole, "The OGI KIDS' Speech Corpus and Recognizers," in *Proc. of ICSLP*, Beijing,China, Oct. 2000.

[15] F. Brugnara, "Context-dependent Search in a Context-independent Network," in *Proc. of ICASSP*, Hong Kong, Apr. 2003.

[16] M. Federico and M. Cettolo, "Efficient Handling of N-gram Language Models for Statistical Machine Translation," in *Proc. of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 2007, pp. 88–95.