TBALL DATA COLLECTION: THE MAKING OF A YOUNG CHILDREN'S SPEECH CORPUS

Abe Kazemzadeh[*], *Hong You*[+], *Markus Iseli*[+], *Barbara Jones*[+], *Xiaodong Cui*[+], *Margaret Heritage*[+], *Patti Price*[^], *Elaine Anderson*[*], *Shrikanth Narayanan*[*], *and Abeer Alwan*[+]

University of Southern California[*], University of California Los Angeles[+], and PPRICE Speech and language Technology[^]

ABSTRACT

In this paper we describe the data collection for the TBALL project (Technology Based Assessment of Language and Literacy) and report the results of our efforts. We focus on aspects of our corpus that distinguish it from currently available corpora. The speakers are children (grades K-4), largely non-native speakers of English, and from diverse socio-economic backgrounds, who are learning to read. We also describe how we adapted our methodology to accommodate these differences: our recording setup, data collection methodology, and transcription scheme. We also discuss the task this corpus was designed to serve and our research approach.

1. INTRODUCTION

In an education system that is increasingly focused on standards to ensure the quality of education, there is a growing need for reliable, objective assessment. Speech and language technology further promises to automate this assessment, provided that key challenges are overcome. In speech processing research these challenges can present themselves as technological obstacles, yet frequently the true challenge is one of logistics; this is especially true in the case of corpus building. This paper examines such challenges in the construction of a corpus for non-native children's speech.

John Steinbeck once wrote, "Some people there are who, being grown, forget the horrible task of learning to read. It is perhaps the greatest single effort that the human undertakes, and he must do it as a child"[0]. A corollary of this statement is that it is also difficult to teach a child to read. An important part of teaching a child to read is assessing their ability, especially in the case of non-native speakers, so that they may receive the appropriate instruction at this critical time. Assessment of a child's reading abilities requires personal attention by teachers, which is a time-consuming process that decreases actual time spent on instruction. This situation is exacerbated when there are fewer qualified teachers available and when teachers must meet the needs of children who are not only learning to read, but also learning to speak English. Automated voice-recognition assessment technology has the potential to partially alleviate teachers' workload by assessing students on the oral reading skills which necessitate a one-to-one format.

These demands are what motivates our long-term research goals, namely, to automate literacy assessment measures using speech and language technology. Recently, there have been several notable projects that have used speech technology in educational applications for children. Examples of spoken dialogue system prototypes include word games for preschoolers [1], aids for reading [2] and pronunciation tutoring [3]. In the LISTEN project [4], CMU researchers are developing an ASR-based tutor to analyze student's oral reading (grades 1-5.) An effort at SRI aimed at computer-based education has lead to the development of the Eduspeak software [5] that includes speech recognition models for children. The "CU Animate" project at Colorado [6, 7] aims at creating spoken dialog interfaces to facilitate reading especially for children with developmental disabilities. Another relevant project is Watch Me! Read (WM!R), developed by IBM's T.J. Watson Research Center, which was implemented in the Houston School District [8]. Our project aims to continue such research efforts and, in addition, to validate the effects of educational technology by relating our automatically derived literacy measures to later reading performance. In addition, we will disseminate our corpus data. This contribution will add to existing children's speech databases [9,10,11,12].

What follows is a review of the first stages of this project in which we constructed a corpus for our task. Below we describe our methods for data collection (2), initial results (3), transcription of the resulting data (4) and conclusions (5).

2. METHODOLOGY

One of the main challenges we faced in building our corpus was how to prompt children in the target age group (who are just beginning to read), to say what we wanted them to say for the process of ASR training, while at the same time to get the variability of input representative of a fully-deployed system. We treat this issue in the following subsections: 2.1 describes the "Wizard of Oz" interface we used to prompt the children, 2.2 covers the recording setup, and 2.3 deals with the testing battery used to prompt the children.

2.1. Wizard of Oz interface

Since our modified Wizard of Oz interface used to collect data was designed to be similar to the requirements of our target deployable system, similar specifications were used. The interface was designed in Java to have the capacity to present the child with engaging stimuli of readable material (letters, numbers, words, and sentences) as well as pictures, which were especially useful in collecting data from preliterate children. This interface interacted with a database containing

information about the children. The children's speech input was recorded and saved in lieu of the automated literacy analysis component of our target system, and the flow of the evaluation tasks was controlled by human operators instead of by software. There was one operator who gave the student instructions and monitored him/her and another operator that controlled the presentation of stimuli.

There was an effort to make the stimuli sufficiently engaging for the children in order to keep their attention and to provide a pleasant experience for them. To do this we included animations in the presentation of words, for example dinosaurs that pulled the words out of the ocean or frogs that ate the words after the child answered. This stimuli carrier received mixed reviews from our subjects. For older children, the novelty of the animations wore off and led to boredom. Younger kids enjoyed the animations, but were also more likely to get distracted by them. Another disadvantage of the animations was that they slowed the pace of the interaction and gave the children the impression of a narrower window for their response. In the future we will correct this approach by providing animations only at the beginning of the testing and before individual sections of the testing battery. One successful way of engaging the children while recording was to play their voices back to them. The children almost always enjoyed this and it also served as a way to double-check our recording process.

The database of information about the child contained the child's age, grade, English language development level, native language, language used at home, language used with friends, and the parents' native languages and birthplaces. This information was gathered along with the consent forms that the parents needed to sign for their child to participate in the project.

The two operators who ran the Wizard of Oz system introduced themselves to the classes at the beginning of the recording sessions and prompted the students to ask any questions they had. This proved to be an effective way of rapidly gaining rapport with the children and making them at ease during the experiments.

The operator who gave instructions to the child and monitored him/her was a native speaker of English with Spanish fluency. He/she was also the one who walked the students to and from their classrooms. This operator was seated next to the child and followed a protocol of introducing the student to the other operator and explaining the task. The other operator controlled the presentation of the stimuli, adjusted the rate of stimuli being presented to the child, monitored the recording, entered child data, and interrupted the program if the child missed more than three answers in a row. He/she was seated at a separate table to the side of the child. Contrary to the standard Wizard of Oz experiment, no attempt was made to conceal the operator with a curtain. The rationale for this was that there was limited benefit to be gained from hiding the operator, but the potential to make the child suspicious, distracted, or anxious.

2.2. Recording setup

A laptop computer was used to run the Wizard of Oz interface, while a second LCD screen was used to present the child with the stimuli. This computer recorded the children's speech to the hard drive, but a DAT recorder was used as a preamplifier and backup recorder.

The recordings were done in Los Angeles area schools in available rooms made available for the project. We used a close-talking headset microphone to reduce environmental noise, which included traffic (the schools were in an urban area) as well as the school sounds (bells, chairs moving in upstairs classrooms, etc.). Special care was necessary to prevent children from playing with the headset and other cords.

The speech was recorded at 44,100 Hz to enable voice source research as well as our recognition studies. The recording sessions with each student were conducted to last less than 20 minutes so the child would maintain maximum concentration and miss as little class as possible. At this pace, each child's session took up approximately 40 MB. Assuming good progress of 15 children per day, this resulted in around 600 MB or approximately 1.9 hours of speech per day.

2.3. Recording Materials

Our recording materials paralleled the testing battery of our target system. This consisted of the set of reading and picture naming tasks we gave the children during the Wizard of Oz, which were designed to test children at the different reading levels for the ages we recorded and to provide balanced examples of speech sounds for recognition and pronunciation modeling.

For the early readers, we had picture naming, color naming, number reading, and alphabet tasks. For the pictures, there were generally several different responses than the ones we had planned. This was partly due to the nature of the task, but also to the diversity of backgrounds of the children. For example, being in sunny California, children had a tendency to say "jacket" instead of "coat" despite having mittens and a hood as clues. Children from families of lower socioeconomic status who might live in apartments may not recognize a picture of a garage while children from more affluent families may tend to say "jogging" instead of "run" when presented with a picture of a person running. For older children we had lists of words and sentences of different levels of difficulty.

Through planning and experience we settled on a testing routine in which we gave the children the picture-naming task first and used their performance on that and subsequent tasks to choose following tests. After twenty minutes, we ended the session and repeated the scenario for each new child.

3. RESULTS AND OBSERVATIONS

Overall, we recorded 256 children, mainly ages five to eight and roughly evenly divided by gender. Of these, 69% were native speakers of Spanish, 24% were native speakers of English, and 5% were native speakers of both English and Spanish. There is 13 GB of speech data in almost 30,000 recordings totaling just over 40 hours. In the following subsections, first we consider the performance of the children on the various tests, with respect to age/grade and language background. Then we discuss the salient characteristics of the

children's speech that need to be accommodated by future ASR development.

3.1 Age/grade effects

As would be expected, children from higher grades generally did better than younger children, given the same task. We did not anticipate the size of the effects of the position within the semester on the children's ability. Since this data collection took place during the fall when children were back to school or starting for the first time, the effect was dramatic and we had to use tests designed for lower grades.

Another effect of age that we noticed was that younger children were more timid on average, which manifested itself in several ways. Some of the shy children just needed more encouragement, but others would mumble, whisper, or not talk at all, and a few started crying. Some possible remedies for this would be to avoid kindergarteners in the first month of classes who have not had exposure to English or computers, having a movie clip or game at the beginning of the recording, and letting them feel comfortable to speak in Spanish if they do not know English.

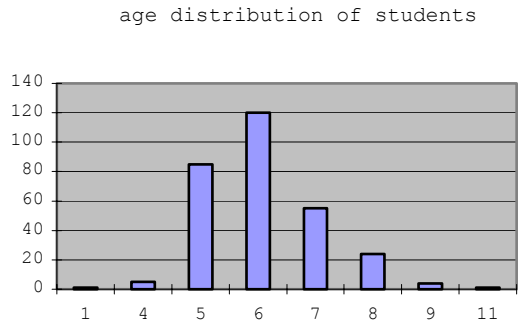

Figure 1: Age distribution of speakers.

3.2 Language background effects

The children's language background had effects on performance that ranged from difficulty associating a word with a pictures or a pronunciation with letters to only slight differences in pronunciation. Sometimes when encouraged, the child could think of the English word that they could not retrieve at first. With a couple of children who could read well in Spanish, but not in English, there was a seemingly contradictory observation that the child could read sentences better than words in isolation or from pictures. Also, children who could read Spanish but not English sometimes could sound out words with Spanish pronunciation, without understanding the meaning.

3.3 Pronunciation variation

One main difference in the children's speech when reading was that the speaking rate was slower, although many of the following observations may apply to the picture-naming task as well. Some of the phenomena we noticed were: long breaks between fricatives combined with stops (e.g. "s-tart"); longer liquids, nasals, and fricatives, which in the case of liquids, may become syllabic (e.g., "light"-> "llight", "close"->"cllose", "might"->"mmight", and "fire" -> "ffire"). Also, syllables may be more spread apart in time (e.g., "a-long") and final consonants may be delayed ("par-t") or dropped/glottalized ("par-"). These are all common for children learning to sound out worlds.

Words that one rarely sees in isolation like "an" and "am" gave children more difficulty than we expected. This could be simply because these words never occur in isolation or because they do not signify a physical object. Also, phonologically, words without an onset are *marked* (less common) and so they tend to be harder to pronounce.

Besides these phenomena, which seemed involuntary, the children also sometimes talked with funny or exaggerated speech, perhaps like voices from cartoons. One child only spoke out of the side of his mouth towards where the microphone was placed.

3.4 Reading tactics

Sounding out words generally helped the children who used it as a reading tactic and a few children sounded out all the letters in a word before actually saying it. Using this tactic could also result in speech that would be perceptually slower than normal speech. It also could produce the observations we noted where children would mispronounce a word by confusing the sounds of a portion of the word with another word ("once" pronounced like "on" and "using" pronounced like "us") or perhaps just confusing the different sounds an orthographic symbol may have ("now" pronounced like "no").

3.5 Higher level phenomena

One of the observations we noted with the picture naming task was that some children had a tendency to use the determiners "a/an/some" when naming the pictures. Though untested, this may be due to differences in the grammatical usage of determiners in Spanish and English.

We also made a number of observations when the children read sentences. One thing we noticed was that sometimes children would change the verb in the sentence to a different tense. Also, children sometimes formed contractions from words that commonly appear in contractions (but not vice versa, i.e., they never substituted the longer version for a contraction that they were reading). Another phenomenon with sentences was that the children would reanalyze the sentence after something later in the sentence clued them into something they had missed earlier on.

4. TRANSCRIPTION

The challenge we faced when transcribing was that we needed a scheme that would allow us to transcribe both native and non-native speech. Also, although we had some a priori expectations about the pronunciation variation we would encounter, we needed a framework that was sufficiently flexible to cover unexpected variation. In this section we explain how we addressed this challenge.

We used the ARPABET symbols (used by many dictionaries, including the CMU pronunciation dictionary) as a

starting point for our transcriptions and added extensions for dealing with non-native speech. These extensions were largely based on an exploratory analysis of approximately 60 utterances and subsequently modified as transcribers agreed to new additions. For the consonants, these included: dental variants of /t/ and /d/, /lt/ and /ld/ respectively; unaspirated voiceless consonants, /pb/, /td/, and /kg/; negative voice onset time consonants (pre-voiced) /mb/ and /nd/; /ths/ for a lispy /s/; /tq/ for a glottalized /t/; /ff/ for especially long frication in /f/; /rr/ for trill; and also the syllabic consonants /en/ and /el/ for the sounds in "button" and "bottle".

For vowels, we faced more difficulty in enumerating all the possible variations. Instead we used a convention for naming non-native sounding vowels based upon the two nearest vowels in the articulatory vowel space already defined by the transcription symbols. To make this uniform, the higher vowel came first (e.g. /iy ih/ for a sound in between /iy/ and /ih/"). The goal in the transcription was a very broad phonetic transcription focused on sounds likely to be made by Hispanic native speakers and by young children. Note that we use the angle brackets indicating a phonemic level despite the fact we are trying to capture some phonetic aspects--in fact we are relatively neutral with respect to a phonemic or phonetic level since the phonological system of these children is quite fluid because of their age and because of their stage of learning English.

CONSONANTS								
	Bi lab.	lab den	den	Alv	post alv	pal	vel	g l
stop	p b		lt ld	t d			kg	
affric.					ch jh			
nasal	M			n			ng	
fricat.		f v	th dh	s z	sh zh			h
apprx				r		y	w	
flap				dx				
trill				rr				
Where symbols appear in pairs, the left is voiced. Use pb, td, and kg for "short" lag p, t, and k, respectively. Use tq for a glottalized t Use ff for a "true fricative" f Use mb and nd for a "prevoiced" b and d Use ths for a lispy s Use en, er, and el for vocalic n, r, and l, respectively								

Figure 2: Consonants transcription symbols.

For our transcription task, we averaged 82% inter-annotator phone agreement, as determined by NIST's SCLITE dynamic alignment program. The transcriptions were done by ear at a rate that yielded approximately 80 single word sound files per hour. This was accomplished using a user interface written in TCL/TK with the Snack audio library. The transcription process is still underway. Of the nearly 30,000 audio files, more than 10% are transcribed.

VOWELS			
	front	central	back
high	Iy		uw
	Ih		uh
mid	Ey		ow
	Eh	ax	ao
low	Ae	ah	aa
plus diphthongs /ay/, /aw/, and /oy/			

Figure 3: Vowel transcription symbols.

5. CONCLUSION

The construction of the TBALL corpus was the first step in a four-year project that seeks to apply a technology-based literacy assessment system longitudinally starting with kindergarten. We will develop and implement this system, which will make use of automatic speech recognition and understanding, and datamining. The system will be incorporated in the Los Angeles and Long Beach Unified School districts, as well as UCLA's University Elementary School. We will then conduct comparative studies between native English speaking children and Mexican Spanish Speaking English learners based on the impact of this technology. For more information about our research please see http://diana.icsl.ucla.edu/Tball/.

6. ACKNOWLEDGEMENTS

This work was supported in part by the NSF. This work would not be possible without the hard work of transcribers Daylen Riggs and Nathan Go; the patience and bilingualism of Kimberly Reynolds, and Blanca Martinez; the careful recordings of Erdem Unal, Vivek Rangarajan, Shiva Sundaram, Yirong Yang, Jinjin Ye, and Yijian Bai; and Larry Casey and Christy Boscardin.

7. REFERENCES

[1] E F Strommen and F S Frome "Talking back to big bird: Preschool users and a simple speech recognition system", *Educational Technology Research and Development*, vol. 41, pp. 5-16, 1993.

[2] J Mostow, A G Hauptmann, and S F Roth, "Demonstration of a reading coach that listens", *Proc. ACM Symposium on User Interface Software and Technology*, pp. 77-78, 1995.

[3] M Russell, B Brown, A Skilling, R Series, J Wallace, B Bonham, and P. Barker, "Applications of autiomatic speech recognition to speech and language development in young children", *Proc. ICSLP*, Philadelphia, PA, Oct. 1996.

[4] http://www-2.cs.cmu.edu/listen/.

[5] http://www.eduspeak.com.

[6] http://www.cslr.colorado.edu

[7] J Ma, J Yan, R Cole, "CU Animate tools for enabling conversations with animated characters", *Proc. ICSLP*, vol. 1, pp. 197-200, Denver, CO, Sept. 2002.

[8] S M Williams, D Nix, and P Fairweather, "Using speech recognition technology to enhance literacy instruction for emerging readers", In B Fishman and S O Conner-Divelbiss Eds., Fourth International Conference on Spoken Language Processing, vol. 1, pp. 629-632, Denver, CO, Sept. 2002.

[9] S Narayanan and A Potamianos, "Creating conversational interfaces for children", *IEEE Trans. Speech and Audio Proc*, 2002.

[10] J D Miller, S. Lee, R M Uchanski, A F Heidbreder, B B Richman, and J Tadlock, "Creation of two children's speech databases". *Proc. ICASSP*, pp. 849-852, 1996.

[11] K Shobaki, J-P Hanson, R A Cole, "OGI Kids Speech Corpus and Recognizers", *Proc. ICSLP,* Beijing, 2000.

[12] M S Eskenazi, "Kids: A database of children's speech", *Proc ASA*, Hawaii, 1996.

[0] J Steinbeck, "Some thoughts on Juvenile Delinquency", *Saturday Review* 38, 1955.