Automatically Assessing the ABCs: Verification of Children's Spoken Letter-Names and Letter-Sounds

MATTHEW P. BLACK and ABE KAZEMZADEH, Signal Analysis & Interpretation Laboratory, University of Southern California
JOSEPH TEPPERMAN, Rosetta Stone Labs
SHRIKANTH S. NARAYANAN, Signal Analysis & Interpretation Laboratory, University of Southern California

Automatic literacy assessment is an area of research that has shown significant progress in recent years. Technology can be used to automatically administer reading tasks and analyze and interpret children's reading skills. It has the potential to transform the classroom dynamic by providing useful information to teachers in a repeatable, consistent, and affordable way. While most previous research has focused on automatically assessing children reading words and sentences, assessments of children's earlier foundational skills is needed. We address this problem in this research by automatically verifying preliterate children's pronunciations of English letter-names and the sounds each letter represents ("letter-sounds"). The children analyzed in this study were from a diverse bilingual background and were recorded in actual kindergarten to second grade classrooms. We first manually verified (accept/reject) the letter-name and letter-sound utterances, which serve as the ground-truth in this study. Next, we investigated four automatic verification methods that were based on automatic speech recognition techniques. We attained percent agreement with human evaluations of 90% and 85% for the letter-name and letter-sound tasks, respectively. Humans agree between themselves an average of 95% of the time for both tasks. We discuss the various confounding factors for this assessment task, such as background noise and the presence of disfluencies, that impact automatic verification performance.

Categories and Subject Descriptors: I.5.4 [**Pattern Recognition**]: Applications—*Signal processing*

General Terms: Algorithms, Experimentation, Languages, Performance, Verification

Additional Key Words and Phrases: Automatic literacy assessment, children's read speech, letter-names, letter-sounds, pronunciation verification

ACM Reference Format:
Black, M. P. Kazemzadeh, A., Tepperman, J., and Narayanan, S. S. 2011. Automatically assessing the ABCs: Verification of children's spoken letter-names and letter-sounds. ACM Trans. Speech Lang. Process. 7, 4, Article 15 (August 2011), 17 pages.
DOI = 10.1145/1998384.1998389 http://doi.acm.org/10.1145/1998384.1998389

1. INTRODUCTION

Education is one area in which technology has already made a profound impact by providing an engaging learning experience to children [Eskenazi 2009]. Computer games have helped children develop problem-solving skills [Yildirim et al. 2011], and virtual peers have helped encourage creative thinking and children's use of imagination

This work was supported in part by the National Science Foundation.
Authors' addresses: M. P. Black, A. Kazemzadeh, S. S. Narayanan, Signal Analysis & Interpretation Laboratory (SAIL), University of Southern California, 3740 McClintock Ave., Los Angeles, CA 90089; email: Matthew.black@usc.edu; J. Tepperman, Rosetta Stone Labs, Boulder, CO 80302.

DOI 10.1145/1998384.1998389 http://doi.acm.org/10.1145/1998384.1998389

[Cassell and Ryokai 2001]. Literacy tutors have been developed to track children's reading and offer helpful feedback [Mostow et al. 1994; Hagen et al. 2007]. These technologies have been designed for a range of ages and developmental levels, and for children with special needs [Cosi et al. 2004].

While much research has focused on developing interactive educational technology, relatively fewer studies have concentrated on ways to use computer technology to help educators and teachers directly. We tackle this problem in the context of literacy assessment for young children from a diverse bilingual background that are learning to read English. Assessment of reading skills is an important aspect of early education [Black and Wiliam 1998]. Experts agree that one of the most effective assessment frameworks is *formative assessment*, where children are repeatedly assessed as they are taught. This pedagogical framework helps keep the teachers' goals and the children's progress aligned and prevents children from being left behind [Heritage 2007]. Unfortunately, formative assessment is challenging for a number of reasons. First, assessment often requires one-on-one time, which teachers may not be able to provide, especially in large classrooms. Second, formative assessment requires an adaptive approach to teaching, where teachers are continually adjusting their lesson plans based on the children's rate of learning.

Technology can help with this process in a number of ways. First, computers can be used to administer the various reading assessment tasks in a consistent, repeatable manner. Second, pronunciation verification systems can be developed to automatically assess the children's speech using objective signal-based methods (e.g., automatic speech recognition). And third, these results can be analyzed and displayed to teachers, so they can track the children's reading proficiency over time, and adjust their lesson plans accordingly. In this paper, we concentrate on the second point for two important English reading tasks: 1) reading the names of the English letters ("letter-names"), and 2) producing the sounds each letter represents ("letter-sounds"). Whereas most previous work in automatic literacy assessment has concentrated on children already reading words and sentences [Mostow et al. 1994; Hagen et al. 2007; Duchateau et al. 2007; Cincarek et al. 2009; Tepperman et al. 2011; Black et al. 2011], early assessment of children's foundational reading skills and early interventions are critical for children to develop into competent readers [Paratore and McCormack 2007].

Children start learning to read English with the alphabet and by producing the different sounds each letter makes. Letters are the building blocks for written words in English, and teachers will often refer to them during their lessons. For there to be successful communication between teachers and students, it is necessary that the students know the letter-names and are able to read them aloud [McBride-Chang 1999]. Knowing and successfully producing a language's letter-sounds is an integral part of learning to read. The concept of *phonemic awareness*, that is, understanding that words can be broken into individual sound units, is crucial for a child to successfully decode words when reading [National Reading Panel 2000]. The letter-sound task assesses a child's knowledge of phonemic awareness and a language's specific letter-to-sound rules [Blachman et al. 1994]. Multiple language effects, such as when children hailing from bilingual or multilingual households are learning to read, can complicate the matter. Children are oftentimes balancing different sets of letter-to-sound rules, since they may be learning to read in more than one language.

In our previous work, we used automatic speech recognition with assessment-constrained grammars and letter-specific lexicons to automatically verify letter-name and letter-sound pronunciations [Black et al. 2008, 2009]. We compared monophone acoustic hidden Markov models trained on isolated *word*-reading data vs. held-out in-domain letter-name and letter-sound data. This paper builds upon our previous work by: 1) using different data splits and cross-validation to ensure speaker-disjoint folds and to maximize the amount of labeled data we have, 2) comparing other automatic pronunciation verification methods and validating them with both item-level and child-level metrics, and 3) providing in-depth error analysis to help gain insight into the challenges of this verification task.

Section 2 describes the corpus we are using in this research. In Section 3, we discuss how we trained acoustic models for this assessment task. Section 4 discusses the various pronunciation verification methods we tried. Section 5 describes our results and provides a discussion. We conclude with future work in Section 6.

2. CORPUS

We utilized speech data from the Technology-Based Assessment of Language and Literacy (TBALL) Project, which was formed to create automatic literacy assessment technology for young children in early education from a diverse background [Alwan et al. 2007; Price et al. 2009]. The TBALL Project's main goal was to provide a technological assessment framework to help inform teachers and track children's progress on age/grade-specific reading tasks (e.g., from reading letter-names and letter-sounds aloud, to syllable-blending, word recognition, and reading comprehension tasks). These reading tasks were administered to children in actual kindergarten to second grade classrooms in Northern and Southern California. About half of the children were native American English speakers, with the other half non-native or bilingual speakers of English from a Mexican-Spanish linguistic background. The children's demographics (native language, grade, and gender) were obtained by forms filled out by assenting parents; most of the demographic information about the children was unknown, since the parents were not required to provide this information. The TBALL Corpus consists of speech recorded from a close-talking headset microphone [Kazemzadeh et al. 2005]. Since the reading tests were administered in real classrooms, the background noises included typical classroom sounds, such as other children's voices and the teacher's voice.

For this work, we analyzed a subset of the data from the letter-name reading task (recorded in mid-to-late 2007) and the letter-sound reading task (recorded in late 2005 and early 2006). There is no overlap in children between the two tasks, so comparisons of individual children's performance across the letter-name and letter-sound tasks are not possible. When administering these reading tests, one lower-case English letter was displayed on a computer screen at a time, and the children had up to five seconds to say the letter-name/letter-sound aloud before the next letter was shown. The children also had the option of advancing to the next letter before this five-second limit by pressing a button. The children were tested on all 26 English letters; the order in which the letters were displayed was random, but this random order was maintained for each child. During the data collection process, a trained research assistant listened beside the child, and if the child mispronounced three letters in a row, the assistant manually stopped the session. This was done to prevent the children from getting too frustrated.

The transition times between letters for each child were automatically recorded and used to split each child's audio into single-letter utterances. (For the letter-sound data, we do not have any utterances for the target letter "e" due to a systematic recording problem.) These single-letter utterances included both the speech from the child and silence/noise before and after. As part of this work, we automatically predicted the overall performance of each child by computing the fraction of letters the child pronounced correctly. To make this average calculation meaningful, we ignored all data from children that had fewer than 8 single-letter utterances.

After the considerations above, the remaining data consist of 3431 letter-name utterances from 168 children and 3507 letter-sound utterances from 153 children. We manually verified all utterances (accept/reject) using a dictionary of acceptable

Table I.
The acceptable letter-name and letter-sound dictionaries, constructed with the help of an expert linguist and teacher. “()” denotes optional phonemes; “|” denotes options

	Phonetic Spelling(s)			Phonetic Spelling(s)	
Letter	*Name*	*Sound*	*Letter*	*Name*	*Sound*
a	EY	AE	n	EH N	N (AH)
b	B IY	B (AH)	o	OW	AA
c	S IY	K	p	P IY	P
d	D IY	D (AH)	q	K Y UW	K W (AH)
e	IY	EH	r	AH R	R (AH) | ER (AH)
f	EH F	F	s	EH S	S
g	JH IY	G (AH)	t	T IY	T
h	EY CH	HH	u	Y UW	AH
i	AY	IH	v	V IY	V (AH)
j	JH EY	JH (AH)	w	D AH B AH (L) Y UW	W (AH)
k	K EY	K	x	EH K S	K S
l	EH L	L (AH)	y	W AY	Y (AH)
m	EH M	M (AH)	z	Z IY	Z (AH)

Table II. Human Evaluation Statistics for the Two Reading Tasks

Statistic	*Letter-Name*	*Letter-Sound*
Fraction Accepted Utterances	0.7546	0.7305
Fraction Disfluent Utterances	0.0827	0.1690
Avg. Pairwise Evaluator Agreement	0.9538	0.9490

phonetic spellings, constructed with the help of an expert linguist and teacher (Table I). Acceptable letter-name pronunciations were straightforward to produce, since there is a one-to-one mapping between an English letter and its correct name pronunciation. However, for the English letter-sounds, vowels have multiple pronunciations, and some letters' pronunciation depends on word context (e.g., the “c” in “face” vs. “cat”). In the letter-sound reading task, the children were instructed to say the “short” vowel sounds (a: /AE/, e: /EH/, i: /IH/, o: /AA/, u: /AH/) and the letter's primary pronunciation (e.g., c: /K/, g: /G/). Only these pronunciations were considered acceptable. For all voiced consonants in the letter-sound reading task, the children could end with the phoneme /AH/ (e.g., the letter “b” had two acceptable pronunciations: /B/ and /B AH/).

For both reading tasks, the data were split up and verified by three evaluators, with 260 utterances common to all. For the letter-sound data, the evaluators were the first three authors, each of whom has several years of experience working on children's literacy assessment research. For the letter-name data, the first author and two trained researchers manually verified the data. During these human evaluations, the speech data were organized by child, so the evaluators could adapt to the speaking style of the children. Evaluators manually verified each letter pronunciation (accept/reject) and also marked utterances that included disfluencies (e.g., repetitions, false starts, repairs). When manually verifying disfluent utterances, evaluators were instructed to accept or reject the final pronunciation only. Inter-evaluator agreement statistics were computed by having all three evaluators evaluate 10 randomly chosen utterances for each letter. The chosen agreement metric is called *accuracy* and is the fraction of utterances in agreement (Equation (1)). Table II shows the average statistics from the human evaluations.

$$A = \frac{\text{number of utterances in agreement}}{\text{number of utterances}}. \tag{1}$$

In our previous work [Black et al. 2008, 2009], we split the data into two disjoint partitions: a test set with 30 utterances per letter and a train set with the remaining data. We trained in-domain acoustic models, optimized our automatic verification methods

Table III.
Data partitions used in this study, each containing about 20% of the data for the reading task

Partition				
1	*2*	*3*	*4*	*5*
–	$\mathrm{DEV}_{1,2}$	$\mathrm{TR}_{1,2}$	$\mathrm{TR}_{1,2}$	$\mathrm{TR}_{1,2}$
–	$\mathrm{TR}_{1,3}$	$\mathrm{DEV}_{1,3}$	$\mathrm{TR}_{1,3}$	$\mathrm{TR}_{1,3}$
–	$\mathrm{TR}_{1,4}$	$\mathrm{TR}_{1,4}$	$\mathrm{DEV}_{1,4}$	$\mathrm{TR}_{1,4}$
–	$\mathrm{TR}_{1,5}$	$\mathrm{TR}_{1,5}$	$\mathrm{TR}_{1,5}$	$\mathrm{DEV}_{1,5}$
TEST_1	TR_1	TR_1	TR_1	TR_1
⋮	⋮	⋮	⋮	⋮
$\mathrm{DEV}_{5,1}$	$\mathrm{TR}_{5,1}$	$\mathrm{TR}_{5,1}$	$\mathrm{TR}_{5,1}$	–
$\mathrm{TR}_{5,2}$	$\mathrm{DEV}_{5,2}$	$\mathrm{TR}_{5,2}$	$\mathrm{TR}_{5,2}$	–
$\mathrm{TR}_{5,3}$	$\mathrm{TR}_{5,3}$	$\mathrm{DEV}_{5,3}$	$\mathrm{TR}_{5,3}$	–
$\mathrm{TR}_{5,4}$	$\mathrm{TR}_{5,4}$	$\mathrm{TR}_{5,4}$	$\mathrm{DEV}_{5,4}$	–
TR_5	TR_5	TR_5	TR_5	TEST_5

Table IV.
Letter-name and letter-sound human evaluation and demographic statistics for each partition. Native language (L1) categories include: English (E), Spanish (S), bilingual (B), unknown (U); grade categories include: kindergarten (K), first (1), second (2), unknown (U); gender categories include: female (F), male (M), unknown (U)

	Partition				
Letter-Name	*1*	*2*	*3*	*4*	*5*
Number Utterances	682	683	691	687	688
Fraction Accepted	0.761	0.725	0.758	0.786	0.743
Number Speakers	33	33	34	34	34
L1 (E:S:B:U)	9:12:0:12	12:10:0:11	14:7:0:13	10:10:0:14	11:13:1:9
Grade (K:1:2:U)	29:4:0:0	32:1:0:0	28:5:0:1	32:2:0:0	29:5:0:0
Gender (F:M:U)	17:16:0	20:12:1	12:21:1	15:18:1	22:11:1
	Partition				
Letter-Sound	*1*	*2*	*3*	*4*	*5*
Number Utterances	708	704	706	694	695
Fraction Accepted	0.687	0.733	0.725	0.759	0.760
Number Speakers	31	31	31	30	30
L1 (E:S:B:U)	6:16:1:8	11:10:0:10	10:15:0:6	7:13:0:10	8:13:2:7
Grade (K:1:2:U)	2:15:9:5	7:8:12:4	4:10:13:4	0:10:15:5	3:16:9:2
Gender (F:M:U)	15:12:4	10:17:4	12:15:4	12:13:5	6:22:2

on the train set, and tested these methods on the held-out test set. We showed in Black et al. [2008] that we needed approximately 70 utterances per letter for the acoustic models to converge and performance to level off.

In this paper, we split the data into five approximately equal-sized speaker-disjoint partitions. We made use of all the labeled data through a cross-validation technique to form train, test, and development sets (Table III). For example, when testing data from partition 1 (TEST_1), we used optimized parameter settings and acoustic models trained on the other four partitions (TR_1). These optimal parameters were found by averaging performance across the four development sets ($\mathrm{DEV}_{1,2}$, $\mathrm{DEV}_{1,3}$, $\mathrm{DEV}_{1,4}$, and $\mathrm{DEV}_{1,5}$), which in turn used acoustic models trained on data from the other three partitions ($\mathrm{TR}_{1,2}$, $\mathrm{TR}_{1,3}$, $\mathrm{TR}_{1,4}$, and $\mathrm{TR}_{1,5}$, respectively). We attempted to make the partitions as balanced as possible, with each having a similar number of children and single-letter utterances, fraction of accepted utterances, and demographic distribution (Table IV).

3. ACOUSTIC MODELS

For all pronunciation verification methods, we used acoustic methods trained on children's speech. This was important since children's speech has different acoustic

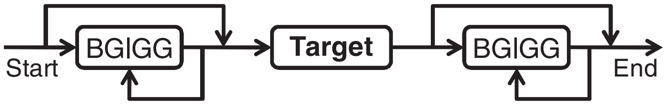

Fig. 1. Grammar used to endpoint single letter utterances (BG = background, GG = garbage).

properties and higher variability than adult speech [Lee et al. 1999]. We investigated two sets of acoustic models in this study: *generic* acoustic models and *in-domain* acoustic models. Generic left-to-right hidden Markov models (HMMs) were trained on 19 hours of held-out speech from word-reading and picture-naming tasks, also included as part of the TBALL Corpus. A total of 38 monophone models were trained on the speech regions. In addition, a phone-level filler "garbage" model was trained on all speech segments, and a "background" model was trained on all silent and background noise regions.

For the in-domain acoustic models, we trained phoneme-level HMMs for each of the 25 train sets (Table III) using an iterative bootstrap training procedure that is based on [Young et al. 2006] and explained in the following sentences. Beginning with prior knowledge of the target phoneme sequence for each audio file, but without segment-level boundaries, initial models were trained using a flat-start initialization and the Baum-Welch embedded reestimation algorithm. With these preliminary models we decoded each target letter-name and letter-sound in the dataset. The resulting phoneme segmentation times were used to train new HMMs from scratch, this time using the hypothesized segmentation for model initialization with Viterbi alignment and then embedded reestimation on each isolated phoneme (rather than over the whole sequence). Then the target sequences were decoded once more, and the new segmentation times were used again to train new acoustic models. This process of decoding and retraining was repeated for five iterations. For the utterances that were accepted by the evaluators, the target phoneme sequence was assumed to be one of the acceptable pronunciations, plus optional sequences of background and garbage preceding or following it (see Figure 1, where "Target" was the acceptable letter-name or letter-sound pronunciation). For the utterances that were rejected by the evaluators, we used a decoding grammar consisting of one or more repetitions of the background and garbage model.

All HMMs were trained on 39-dimensional Mel-frequency cepstral coefficients, with 3 hidden states and 16 Gaussian mixtures per state, using HTK [Young et al. 2006]. The window length was 25 ms, and the frame rate was 10 ms. Cepstral-mean subtraction was applied across each single-letter utterance.

4. VERIFICATION METHODS

This section will describe the four pronunciation verification methods we investigated. Table V provides a brief description of each, with full details in Sections 4.1-4.4. For all methods, we optimized any parameters across each fold's four development sets by maximizing agreement between the human evaluations and the automatic verification hypotheses (Equation (1)). We then applied these methods with optimal parameter settings to the held-out test data set and pooled all results to characterize performance at both the utterance-level and child-level. To compute performance at the utterance-level, we used accuracy (Equation (1)), and we also borrowed metrics commonly used in detection theory and binary classification tasks: precision (Equation (2)), recall (Equation (3)), balanced F-score (Equation (4)). In these equations, a "true positive"

Table V.
Abbreviations and short descriptions of the four pronunciation verification methods we investigated for both reading tasks

Abbreviation	*Short Description (Accept utterance if . . .)*
Recognition	target letter recognized
Lexicon	acceptable pronunciation recognized from letter-specific dictionary
Duration	duration is longer than letter-specific threshold
GOP	Goodness of Pronunciation score is higher than letter-specific threshold

is correctly rejecting a pronunciation.

$$P = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{2}$$

$$R = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{3}$$

$$F = \frac{2 \cdot P \cdot R}{P + R}. \tag{4}$$

We also wanted to make sure that our assessments were accurate at the child-level. To estimate how well a child performed overall, we computed the fraction of utterances the automatic method accepted for the child. To characterize system performance at the child-level, we computed Pearson's correlation coefficient between two vectors: one formed by the fraction of utterances accepted by the *automatic method* for each child, and the other corresponding to the fraction of utterances accepted by the *evaluators* for each child. It should be noted that this is a relatively simple metric to quantify how well the automatic method can predict children's overall performance. See Duchateau et al. [2007] and Black et al. [2011] for examples of automatic literacy assessment research that directly models evaluator's perception of children's overall reading performance.

4.1. Recognition Method

As a baseline method for pronunciation verification of the letter-names and letter-sounds, we ran automatic letter recognition on each single-letter utterance using the acoustic models described in Section 3, the dictionary of acceptable pronunciations (Table I), and a constrained grammar that allowed for any letter in the dictionary to be recognized (in addition to optional background/garbage). Therefore, we used the grammar depicted in Figure 1, where "Target" was "a | b | . . . | z". We found this grammar structure was best at endpointing the target pronunciation, even in disfluent and/or noisy utterances. We accepted the pronunciation if the target letter was recognized; otherwise, we rejected the pronunciation. We refer to this baseline method as "recognition," since it essentially reduces the letter-name/letter-sound *verification* task to a letter-name/letter-sound *recognition* task.

4.2. Lexicon Method

The recognition method is not optimal for this verification task for a number of reasons. First, it is designed to detect substitution errors, which is only one of the common types of errors a child might make. Second, many of the substitution errors are very unlikely (such as the student confusing "a" with "z"). To remedy this undergeneration and overgeneration of possible cohort pronunciations, we created another dictionary with foreseeable unacceptable pronunciations for each letter and for both reading tasks, based on our previous work [Black et al. 2008, 2009]. These unacceptable pronunciations fit into, and were assigned to, five pronunciation categories: 1) *alternative pronunciations*, that is, replacing a letter-name/letter-sound with an alternative pronunciation of

Table VI.
Description of the five unacceptable pronunciation classes, with corresponding average cardinality per letter (N) and example entries

	Letter-Name		*Letter-Sound*	
Pronunciation Class	N	Example(s)	N	Example(s)
Alternative pronunciations	0.96	f: /IH F/	1.35	c: /S/, g:/JH/
Visual confusions	0.81	b-d, p-q, h-n	1.08	b-d, p-q, h-n
Auditory confusions	1.65	f-s, m-n, c-z	2.42	f-s, m-n, c-z
Spanish-related confusions	1.46	j: /HH EY/	0.81	u: /UW/
Task confusions	1.73	k: /K/	1.04	k: /K EY/

a phoneme, 2) *visual confusions*, that is, saying a letter-name/letter-sound for a letter that closely resembles the target letter's shape, 3) *auditory confusions*, that is, saying a letter-name/letter-sound for a letter that is phonetically similar to the target letter, 4) *Spanish-related confusions*, that is, saying the letter-name/letter-sound correctly in Spanish or saying one or more phonemes with a common Spanish substitution [You et al. 2005], 5) *reading task-related confusions*, that is, saying the English letter-name during the letter-sound task (and vice versa). Example entries and the average cardinality of these unacceptable pronunciation classes are provided in Table VI. Note that individual unacceptable pronunciations can belong to more than one category.

The second pronunciation verification method we employed is referred to as the "lexicon" method, since it finds an optimal dictionary for each letter. For each of the five cross-validations, we found the combination of unacceptable pronunciation classes that optimized agreement (Equation (1)) across the development sets, for each letter individually. We again used the grammar shown in Figure 1. However, in this case, "Target" included only the acceptable and unacceptable pronunciations of the target letter, along with the background model, instead of all possible letters as in the case of the recognition method. This background model was included since it helped detect pronunciation errors caused by the child saying something unexpected or nothing at all; the background model was found to perform better than the garbage model for this purpose. This optimized dictionary was then applied to the test data for that fold, and this process was repeated for all five folds. Separate optimized dictionaries were found for both sets of acoustic models.

4.3. Duration Method

Another pronunciation verification method we tried was based on an empirical observation of forced alignment. If a pronunciation is force-aligned to audio in which there is a poor match (due to either an inaccurate transcript or mismatched acoustic conditions), the pronunciation will oftentimes be aligned to a small temporal region. Related work in computer-assisted language learning has exploited this phenomenon to extract duration-based features for utterance verification [van Doremalen et al. 2010]. We hypothesized that utterances that were rejected by evaluators would be poor matches to acceptable pronunciations of the target letter. Therefore, if we force-aligned only acceptable pronunciations to rejected utterances using the grammar in Figure 1, the letter would be more likely to be aligned to a small temporal region, as compared to utterances that were accepted by evaluators. This is due to the fact that the phoneme boundaries are chosen to maximize the log-likelihood of the utterance, and so the background/garbage model dominates the utterance, since it is a better match. Figures 2 and 3 demonstrate this phenomenon when using the in-domain acoustic models. Indeed, in utterances that were rejected by evaluators, the letter-name's/letter-sound's duration was shorter, on average.

The third automatic verification system we devised, referred to as "duration," rejects pronunciations below a letter-specific duration threshold, and accepts pronunciations

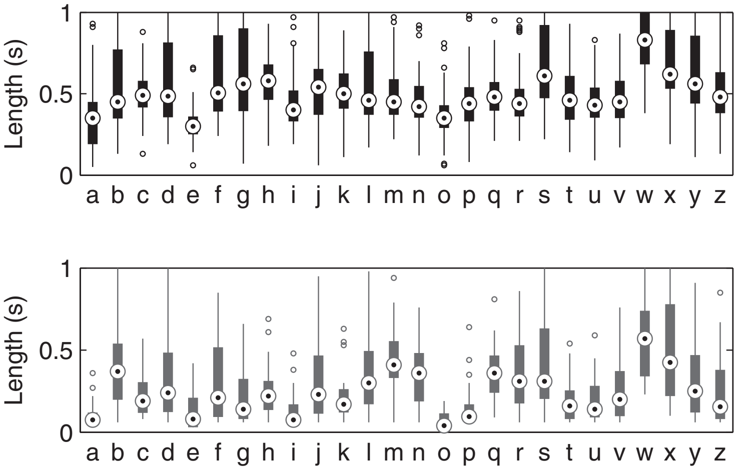

Fig. 2. Box plot of letter-name durations for manually accepted (top) and rejected utterances (bottom), based on forced alignment endpointing of acceptable pronunciations using in-domain acoustic models. The box plot shows the distribution of durations for each letter-name (the middle circle denotes the median, and the surrounding bar is bounded by the lower and upper quartile); in general, the rejected utterance durations are shorter than the accepted ones.

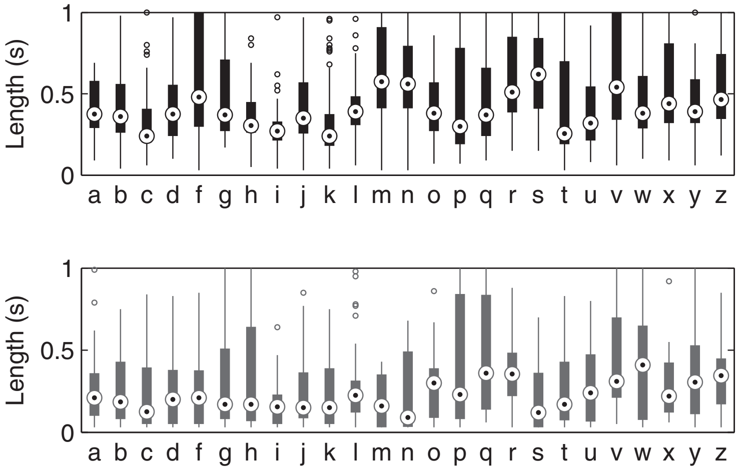

Fig. 3. Box plot of letter-sound durations for manually accepted (top) and rejected utterances (bottom), based on forced alignment endpointing of acceptable pronunciations using in-domain acoustic models. The box plot shows the distribution of durations for each letter-sound (the middle circle denotes the median, and the surrounding bar is bounded by the lower and upper quartile); in general, the rejected utterance durations are shorter than the accepted ones.

above this threshold. See Equation (5), where the duration thresholds, $T_{\mathrm{dur}}(l)$, for each letter l were chosen to optimize the agreement (Equation (1)) across the development sets for that particular cross-validation fold.

$$\mathrm{Reject}(l) \equiv \begin{cases} 1, & \mathrm{Duration}(l) \leq T_{\mathrm{dur}}(l) \\ 0, & \mathrm{Duration}(l) > T_{\mathrm{dur}}(l). \end{cases} \tag{5}$$

4.4. GOP Scoring Method

The final pronunciation verification method we tried was a variation of Goodness of Pronunciation (GOP) scoring. It was first introduced in Witt and Young [2000], and has been successfully used in many pronunciation evaluation applications [Rasmussen et al. 2009; Tepperman et al. 2011; Zheng et al. 2007]. In this technique, a word (or in our case, a letter-name or letter-sound), is decoded using a dictionary with only acceptable pronunciations. The resulting output will contain phoneme hypotheses, phoneme boundaries, and log-likelihoods for each decoded phoneme. An unconstrained phone loop is then decoded across each hypothesized phoneme region. A GOP score for each phoneme is then computed by subtracting the normalized log-likelihood of the recognized phoneme to the normalized log-likelihood of the phone loop. Higher GOP scores pertain to phonemes that are more likely to be pronounced correctly, and lower GOP scores are correlated with poorly pronounced phonemes. A threshold can be set to reject recognized phonemes that have GOP scores that are too low. Originally, this technique was used to find phoneme errors within words that were mispronounced. We have found that "word-level" GOP scores (or in our case "letter-level") produced better results. Letter-level GOP scores are computed by taking the mean GOP score across all recognized phones.

Equation (6) shows how to compute the GOP phoneme score (O is the acoustic observation, p is the phone, PL is the phone-loop, and N is the number of frames of phone p). Equation (7) shows how to compute the GOP letter-level score, by calculating the mean of the GOP phoneme scores for the letter l. Equation (8) shows how we thresholded the GOP letter-level score to reject or accept the utterance's pronunciation. This letter-dependent GOP threshold, $T_{\text{GOP}}(l)$, was chosen to maximize the agreement (Equation (1)) across the development sets for each fold.

$$\text{GOP}(p) \equiv \frac{1}{N} \log \frac{P\,(O|p)}{P\,(O|PL)} \tag{6}$$

$$\text{GOP}(l) \equiv \frac{1}{|p \in l|} \sum_{p \in l} \text{GOP}(p) \tag{7}$$

$$\text{Reject}(l) \equiv \begin{cases} 1, & \text{GOP}(l) \le T_{\text{GOP}}(l) \\ 0, & \text{GOP}(l) > T_{\text{GOP}}(l). \end{cases} \tag{8}$$

5. RESULTS & DISCUSSION

Results are shown in Table VII for the four verification methods. The performance of the four methods is higher for the letter-name task, implying this is an easier verification task, as compared to the letter-sound task. This makes intuitive sense, since the letter-sounds are temporally shorter and have greater variability in their pronunciations (and are less natural to produce in isolation). One of the principal results is that the lexicon method performed best, in terms of most of the chosen metrics, for both reading tasks. We also see that, in general, the generic acoustic models (trained on word-level TBALL data), performed slightly better than the in-domain acoustic models for most verification methods; this may be due to the fact that the generic models were trained on more hours of speech.

To test whether the differences in performance are statistically significant, we used two statistical tests: a one-sided difference in proportions test on the accuracy metric (to test differences in utterance-level performance), and a difference in correlation coefficient test on the child-level correlation metric. For both reading tasks and all four pronunciation verification methods, the performance between the two sets of acoustic

Table VII.
Letter-name and letter-sound performance using the four proposed pronunciation verification methods and the two different sets of acoustic models. We compare performance against *human agreement* (Table II) and a simple *voting baseline* method, which accepted all letter-name and letter-sound utterances

Letter-Name Methods	*A*	*P*	*R*	*F*	*Corr*
Human Agreement	**0.9538**	–	–	–	–
Voting Baseline	0.7546	0.0000	0.0000	–	–
Recognition (generic AMs)	0.7709	0.5189	**0.9121**	0.6615	0.8549
Recognition (in-domain AMs)	0.7473	0.4919	0.9002	0.6362	0.8313
Lexicon (generic AMs)	**0.8965**	**0.8322**	0.7245	**0.7746**	**0.9207**
Lexicon (in-domain AMs)	0.8895	0.8293	0.6924	0.7547	0.9104
Duration (generic AMs)	0.8630	0.7934	0.5974	0.6816	0.8023
Duration (in-domain AMs)	0.8566	0.8081	0.5451	0.6511	0.8269
GOP (generic AMs)	0.8825	0.7585	0.7648	0.7617	0.9195
GOP (in-domain AMs)	0.8802	0.7964	0.6876	0.7380	0.8974

Letter-Sound Methods	*A*	*P*	*R*	*F*	*Corr*
Human Agreement	**0.9490**	–	–	–	–
Voting Baseline	0.7305	0.0000	0.0000	–	–
Recognition (generic AMs)	0.6812	0.4549	0.9238	0.6096	0.8399
Recognition (in-domain AMs)	0.7149	0.4848	**0.9270**	0.6366	0.8326
Lexicon (generic AMs)	**0.8466**	0.7573	0.6339	**0.6901**	0.8561
Lexicon (in-domain AMs)	0.8415	**0.7653**	0.5937	0.6687	**0.8757**
Duration (generic AMs)	0.8186	0.7433	0.4995	0.5975	0.7864
Duration (in-domain AMs)	0.8004	0.7590	0.3799	0.5063	0.7578
GOP (generic AMs)	0.8340	0.7055	0.6593	0.6816	0.8530
GOP (in-domain AMs)	0.8358	0.7204	0.6381	0.6768	0.8571

models was not significant (all $p > 0.1$). This implies that similar system performance can be achieved in two common training scenarios: 1) where we have a sufficient amount of representative children's speech reading words aloud (not necessarily letter-names/letter-sounds), and 2) where we have a sufficient amount of in-domain data of children reading letter-names/letter-sounds.

We also tested to see if there is a significant difference between the best performing method (the lexicon method with the generic acoustic models) and the other verification methods. For the letter-name reading task, this best method has significantly higher accuracy, as compared to the voting baseline method, the recognition method, and the duration method (all $p < 0.001$); there is no significant difference between the GOP method ($p > 0.05$). The lexicon method still performs significantly worse than the interevaluator agreement ($p < 0.001$). Looking at the child-level metric, the lexicon method has a significantly higher correlation coefficient, as compared to the duration method ($p < 0.001$) and the recognition method ($p < 0.005$); there is no significant difference between the lexicon method and the GOP method in terms of child-level correlation ($p > 0.1$). For the letter-sound reading task, we see similar trends when comparing the accuracy utterance-level metric. The lexicon method performs significantly better than the baseline voting, duration, and recognition methods (all $p < 0.001$), significantly worse than human agreement ($p < 0.001$), and has no significant difference in performance with the GOP method ($p > 0.1$). Comparing the child-level metric for the letter-sound task, the lexicon method only significantly outperforms the duration method, with all other $p > 0.1$. This suggests that even when utterance-level performance suffers, these automatic methods are still able to estimate the overall children's performance (as captured by the fraction of utterances that were deemed acceptable pronunciations) with a high correlation to human evaluators.

The recognition method had the highest recall but suffered from the lowest precision; therefore, it was able to detect pronunciation errors well, but rejected many acceptable pronunciations. As mentioned before, this is most likely due to the fact that there are too many cohort pronunciations in the dictionary for this method. Conversely, the duration method performed competitively in terms of accuracy and precision, but it performed worse in terms of recall and correlation with the evaluators. The poor recall statistic means this method had many missed detections and was unable to reject a large fraction of the unacceptable pronunciations. The lower child-level correlation is most likely due to inherent differences in the children's speaking rate, thus producing several more errors for some children, leading to a lower correlation at the child-level with human evaluators.

The GOP method of verification was competitive with the lexicon method (with no significant difference in performance as described previously). One advantage of GOP scoring is that it does not involve the creation of a specialized lexicon to detect errors, but the downside is that a threshold must be explicitly found. Importantly, the resulting GOP output may be less instructive than the lexicon method. Whereas the GOP method only quantifies overall "goodness of pronunciation," the lexicon method explicitly shows the pronunciation error (and automatically classifies it according to the dictionary entry, e.g., Spanish-related confusion). This kind of high-level information might be very useful for a teacher, since it provides insight into the types of categorical errors the children are making. For example, for the letter-name task (when using the lexicon method and the generic acoustic models), the breakdown in the unacceptable pronunciation classes was as follows: 66.3% visual confusions, 19.2% task-related confusions, 7.6% Spanish-related confusions, 5.0% auditory confusions, and 1.9% alternative pronunciations. For the letter-sound task, the breakdown in the recognized errors was as follows: 45.4% visual confusions, 29.4% alternative pronunciations, 18.2% auditory confusions, 3.7% task-related confusions, and 3.4% Spanish-related confusions.

The large disparity between human agreement and the automatic methods is most likely due to a number of reasons. The letter-names and letter-sounds are usually less than half a second in duration (Figures 2 and 3), which means the verification is based on very little acoustic evidence. In addition, since we are verifying individual letter pronunciations, there is no contextual information we can use, as is oftentimes exploited in automatic speech recognition (e.g., with language models). Also, while the evaluators had the ability to adapt to the speaking style of the children, there was no analogous adaptation for the automatic verification methods. Incorporating acoustic model adaptation for each child can be explored in the future. Lastly, since the audio was recorded in real classrooms, it has variable noise sources, which make automatic verification challenging. We analyzed the effect that noise played on automatic performance by estimating the signal-to-noise ratio (SNR) for each utterance using Equation (9), where A_{target} is the root mean square (RMS) amplitude within the endpointed target word and $A_{background}$ is the RMS amplitude within the regions aligned to the background model.

$$\mathrm{SNR} = 20 \log_{10} \frac{A_{target}}{A_{background}}. \tag{9}$$

Table VIII shows that utterances in which the automatic system (lexicon method with generic acoustic models) erred had a mean SNR that was significantly lower than utterances in which the automatic system agreed with the evaluators. This means that either the background regions had higher relative RMS energy in these erred utterances and/or that the target word pronunciation was not properly endpointed. Robustness to noise and accurate endpointing of children's speech is an area of future research. We can also see in this table that the average SNR for the letter-sound

Table VIII.
SNR statistics comparing when the automatic system agrees with human evaluators ("correct") and when it does not agree ("error"). The mean SNR for "error" utterances is significantly lower than for "correct" utterances, using a one-sided unpaired t test

	System		*SNR statistics*		
Reading Task	*correct/error*	*N*	Mean	SD	*p*
Letter-Name	correct	3076	18.38	7.806	< 0.001
	error	355	14.69	8.548	
Letter-Sound	correct	2969	17.25	8.498	< 0.001
	error	538	12.45	9.780	

Table IX.
The probabilities that the automatic system is correct, conditioned on the fluency of the utterance. The system is more likely to err on disfluent utterances, although this difference is only significant with the letter-sound task (using a one-sided difference in binomial proportions test)

Reading Task	$Pr(correct\|fluent)$	$Pr(correct\|disfluent)$	*p*
Letter-Name	0.8984	0.8759	> 0.1
Letter-Sound	0.8615	0.7723	< 0.001

utterances was lower than the average SNR for the letter-name utterances; this difference was significant with $p < 0.001$. This implies that noisier conditions may be another contributing factor in the relatively lower automatic performance for the letter-sound task.

We also analyzed the effect that disfluencies had on system performance. As mentioned in Section 2, a portion of the children's utterances contained disfluent responses, mostly due to the children repeating themselves or self-correcting. It should be noted that there were twice as many disfluent utterances in the letter-sound task (Table II). These disfluencies can be viewed as noise for this verification task, since evaluators were instructed to ignore any disfluencies and only rate the final pronunciation. We used the grammar shown in Figure 1 to help endpoint these disfluencies by allowing for repetitions of the garbage or background model to be recognized. To test whether these disfluencies significantly affected system performance, we compared the probability of a system error, given the utterance was fluent vs. the probability of a system error, given the utterance was disfluent. Table IX shows that disfluencies had no significant effect on performance for the letter-name task, which implies that the disfluencies were either innocuous (e.g., repetitions) or were successfully filtered out with the grammar. On the other hand, system performance was significantly worse for disfluent letter-sound utterances, suggesting disfluencies were not filtered out by using this grammar and may have been incorrectly endpointed as the target word for some utterances. One explanation for this difference could be that disfluencies are inherently more acoustically similar (and hence, more confusable) to letter-sounds, as compared to letter-names.

Lastly, Tables X and XI show both human evaluator statistics (the fraction of utterances accepted) and system performance (lexicon method with generic acoustic models) across letters and demographics for the letter-name and letter-sound verification tasks, respectively. We see in the letter-name task that children performed best on the letter "o" and worst on the letter "q" (oftentimes confusing this letter with the letter "p" due to the similar shape). Interestingly, automatic performance was best, in terms of accuracy and F-score, for the letter "q." Some of the worst performance statistics for the letter-name task occurred for the letters "m" and "n," which may be due to the acoustic similarities of the phonemes /M/ and /N/. For the letter-sound task, children performed best for the letter "s" and again performed worst for the letter "q." Automatic performance was best, in terms of accuracy, for the letter "s," and in terms of F-score

Table X.
Letter-name performance across letters and demographics when using the "lexicon" pronunciation verification method and generic acoustic models. *N* is the number of utterances; *Frac Accept* is the fraction of utterances accepted by the evaluators; *A*, *P*, *R*, and *F* is the automatic performance (Equations (1)–(4))

Letter	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
a	129	0.7984	0.9380	0.9091	0.7692	0.8333
b	130	0.7154	0.8615	0.8065	0.6757	0.7353
c	121	0.8264	0.8926	0.7500	0.5714	0.6486
d	127	0.6614	0.8189	0.8333	0.5814	0.6849
e	134	0.8358	0.9254	0.7308	0.8636	0.7917
f	145	0.8690	0.9379	0.9167	0.5789	0.7097
g	136	0.5515	0.8971	0.9273	0.8361	0.8793
h	120	0.7250	0.8667	0.8148	0.6667	0.7333
i	136	0.7794	0.9559	0.9286	0.8667	0.8966
j	122	0.6803	0.8279	0.7647	0.6667	0.7123
k	126	0.8651	0.9524	0.8667	0.7647	0.8125
l	142	0.5986	0.9014	0.8772	0.8772	0.8772
m	159	0.8491	0.8805	1.0000	0.2083	0.3448
n	132	0.8106	0.7727	0.3913	0.3600	0.3750
o	137	0.9635	0.9416	0.3846	1.0000	0.5556
p	133	0.7895	0.9248	0.8462	0.7857	0.8148
q	140	0.3214	0.9857	0.9895	0.9895	0.9895
r	127	0.7795	0.8268	0.6667	0.4286	0.5217
s	144	0.8542	0.8403	0.4444	0.3810	0.4103
t	138	0.6522	0.8406	0.7826	0.7500	0.7660
u	132	0.7500	0.9545	0.9091	0.9091	0.9091
v	127	0.7402	0.9134	0.9583	0.6970	0.8070
w	117	0.7692	0.8889	0.8500	0.6296	0.7234
x	119	0.8824	0.9244	1.0000	0.3571	0.5263
y	131	0.7405	0.9466	0.9091	0.8824	0.8955
z	127	0.8268	0.8819	0.6667	0.6364	0.6512
L1	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
English	1113	0.8347	0.9263	0.7802	0.7717	0.7760
Spanish	1042	0.7428	0.8964	0.8509	0.7239	0.7823
Grade	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
Kindergarten	3028	0.7526	0.8937	0.8280	0.7196	0.7700
First	378	0.7540	0.9127	0.8659	0.7634	0.8114
Gender	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
Female	1776	0.7337	0.8913	0.8483	0.7209	0.7794
Male	1562	0.7785	0.9040	0.8101	0.7399	0.7734

for the letter "g." Automatic performance was worst, in terms of accuracy, for the letter "i," and in terms of precision, recall, and F-score, for the letter "v"; this was most likely due to a lack of training data for the phoneme /V/.

Comparing children's performance across demographics, the Spanish-speaking children performed significantly worse in both English reading tasks (both $p < 0.001$). There was no significant difference across grades for both reading tasks (both $p > 0.1$). For this particular subset of children, males significantly outperformed females for both reading tasks (both $p < 0.005$). Comparing system performance across the children's demographics, there was only one bias: the automatic system (lexicon method with generic acoustic models) erred significantly less for native English speakers, compared to Spanish-speaking children for the letter-name task ($p < 0.05$), with all other $p > 0.1$. Training separate sets of acoustic models for native English and Spanish speakers may help reduce this performance difference. But the fact that the automatic system was largely unbiased to demographic information is most likely due to the fact that the acoustic models were trained on speech from the TBALL Corpus, so each demographic appearing in the data was represented well.

Table XI.
Letter-sound performance across letters and demographics when using the "lexicon" pronunciation verification method and generic acoustic models. *N* is the number of utterances; *Frac Accept* is the fraction of utterances accepted by the evaluators; *A*, *P*, *R*, and *F* is the automatic performance (Equations (1)–(4))

Letter	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
a	116	0.5345	0.8448	0.9091	0.7407	0.8163
b	145	0.7931	0.9310	0.9167	0.7333	0.8148
c	139	0.6259	0.8273	0.7188	0.8846	0.7931
d	139	0.7770	0.8345	0.7500	0.3871	0.5106
f	148	0.8716	0.8986	0.7500	0.3158	0.4444
g	142	0.6338	0.9085	0.8421	0.9231	0.8807
h	147	0.7483	0.8163	0.5926	0.8649	0.7033
i	125	0.5680	0.7520	0.8108	0.5556	0.6593
j	137	0.7007	0.8613	0.8438	0.6585	0.7397
k	141	0.8440	0.8794	0.6316	0.5455	0.5854
l	145	0.6966	0.7655	0.6316	0.5455	0.5854
m	151	0.9139	0.9272	0.6250	0.3846	0.4762
n	150	0.9000	0.9133	0.5714	0.5333	0.5517
o	119	0.6218	0.8067	0.8438	0.6000	0.7013
p	144	0.7986	0.8681	0.7083	0.5862	0.6415
q	141	0.4681	0.7801	0.8548	0.7067	0.7737
r	152	0.7368	0.7961	0.7647	0.3250	0.4561
s	150	0.9400	0.9467	0.6000	0.3333	0.4286
t	141	0.8085	0.9078	0.8889	0.5926	0.7111
u	123	0.5122	0.7886	0.8542	0.6833	0.7593
v	142	0.8521	0.7676	0.2500	0.2857	0.2667
w	145	0.7793	0.8069	0.5714	0.5000	0.5333
x	139	0.6906	0.7842	0.6970	0.5349	0.6053
y	140	0.5000	0.8500	0.9455	0.7429	0.8320
z	146	0.7945	0.8630	0.6667	0.6667	0.6667
L1	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
English	966	0.7805	0.8727	0.7432	0.6415	0.6886
Spanish	1538	0.7217	0.8498	0.8069	0.6051	0.6916
Grade	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
First	1364	0.7449	0.8512	0.7474	0.6293	0.6833
Second	1346	0.7615	0.8685	0.7791	0.6262	0.6943
Gender	*N*	*Frac Accept*	*A*	*P*	*R*	*F*
Female	1230	0.6675	0.8415	0.8497	0.6357	0.7273
Male	1854	0.7799	0.8635	0.7183	0.6250	0.6684

6. CONCLUSION & FUTURE WORK

Our research aims to enable automated, formative assessment of children's reading abilities that will inform teachers and educators. Early interventions of children's reading proficiency requires assessment of children's foundational reading skills. In this paper, we automatically verified children reading letter-names and letter-sounds aloud, two critical reading tasks that children need to master before they learn to read words and sentences.

We investigated four automatic verification methods. The first method, "recognition," used a speech recognition approach where the letter names and sounds were recognized from a loop grammar. The second method, "lexicon," used a specific dictionary for each target item that included variants labeled as correct and incorrect. The third method, "duration," used the segmentation from forced-alignment to empirically separate correct utterances, which tended to have longer durations, from incorrect utterances, which tended to be shorter. The fourth method, "GOP" (Goodness of Pronunciation), implemented a variation of the GOP scoring method from previous studies that was adapted to our task. We found that the lexicon and GOP methods performed best, while the recognition and duration methods performed worse, but showed some strong points among our different metrics. We found no significant difference in performance when using acoustic models trained on words spoken by children, as compared to models trained directly on letter-name/letter-sound speech.

Automatic performance was significantly lower than interevaluator agreement for both reading tasks. As discussed in Section 5, there are many actions we can take to improve verification performance: increased noise robustness, improved target word endpointing, explicit disfluency detection, and child-adapted language-dependent acoustic models. Fusion of the four verification methods explored in this paper is also an area of future research. We tried combining the methods by cascading classifiers and incorporating duration information into the GOP score, but none of these fusion techniques were able to outperform the lexicon method. We feel a method that does not rely on automatic speech recognition might provide orthogonal information; possible features include prosodic cues (e.g., pitch, energy) and letter-specific properties (e.g., voice onset time [Kazemzadeh et al. 2006]). These acoustic information sources, along with the children's demographics, could also be used to train a cognitive model that assesses children's performance on the reading task, such as the one used in Tepperman et al. [2011]. This future work might further reduce the gap between automatic verification performance and interevaluator agreement and could provide additional high-level information that is useful for educational purposes.

Automating the assessment of other widely administered reading tasks for young children (e.g., word recognition, reading comprehension) and analyzing children's performance across reading tasks is another area of current and future research. Finally, greater emphasis on integrating automated reading assessment and reading tutors into classroom reading instruction is needed for continued progress in automated literacy assessment research.

ACKNOWLEDGMENTS

We would like to thank the entire TBALL Project team for their help in this effort. Special thanks to Professor Sungbok Lee and Patti Price for their research contributions.

REFERENCES

ALWAN, A., BAI, Y., BLACK, M. P., CASEY, L., GEROSA, M., HERITAGE, M., ISELI, M., JONES, B., KAZEMZADEH, A., LEE, S., NARAYANAN, S., PRICE, P., TEPPERMAN, J., AND WANG, S. 2007. A system for technology based assessment of language and literacy in young children: The role of multiple information sources. In *Proceedings of the International Workshop on Multimedia Signal Processing (MMSP).*

BLACHMAN, B. A., BALL, E. W., BLACK, R. S., AND TANGEL, D. M. 1994. Kindergarten teachers develop phoneme awareness in low-income, inner-city classrooms. *Read. Writ. 6,* 1, 1–18.

BLACK, M. P., TEPPERMAN, J., KAZEMZADEH, A., LEE, S., AND NARAYANAN, S. 2008. Pronunciation verification of English letter-sounds in preliterate children. In *Proceedings of Interspeech*.

BLACK, M. P., TEPPERMAN, J., KAZEMZADEH, A., LEE, S., AND NARAYANAN, S. 2009. Automatic pronunciation verification of English letter-names for early literacy assessment of preliterate children. In *Proceedings of the IEEE International Conference on Accustics, Speech, and Signal Processing (ICASSP)*.

BLACK, M. P., TEPPERMAN, J., AND NARAYANAN, S. 2011. Automatic prediction of children's reading ability for high-level literacy assessment. *IEEE Trans. Audio Speech Lang. Process. 19*, 4, 1015–1028.

BLACK, P. AND WILIAM, D. 1998. Assessment and classroom learning. *Assess. Educ. Princ. Policy Pract. 5,* 1, 7–74.

CASSELL, J. AND RYOKAI, K. 2001. Making space for voice: Technologies to support children's fantasy and storytelling. *Pers. Ubiq. Comput. 5,* 3, 169–190.

CINCAREK, T., GRUHN, R., HACKER, C., NÖTH, E., AND NAKAMURA, S. 2009. Automatic pronunciation scoring of words and sentences independent from the non-native's first language. *Comput. Speech Lang. 23,* 1, 65–88.

COSI, P., DELMONTE, R., BISCETTI, S., COLE, R. A., PELLOM, B., AND VUREN, S. 2004. Italian literacy tutor-tools and technologies for individuals with cognitive disabilities. In *Proceedings of the InSTIL/ICALL Symposium on Computer Assisted Learning*.

DUCHATEAU, J., CLEUREN, L., VAN HAMME, H., AND GHESQUIÈRE, P. 2007. Automatic assessment of children's reading level. In *Proceedings of Interspeech*.

ESKENAZI, M. 2009. An overview of spoken language technology for education. *Speech Comm. 51,* 10, 832–844.

HAGEN, A., PELLOM, B., AND COLE, R. 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Comm. 49,* 12, 861–873.

HERITAGE, M. 2007. Formative assessment: What do teachers need to know and do? *Phi Delta Kappan 89,* 2, 140–145.

KAZEMZADEH, A., TEPPERMAN, J., SILVA, J., YOU, H., LEE, S., ALWAN, A., AND NARAYANAN, S. 2006. Automatic detection of voice onset time contrasts for use in pronunciation assessment. In *Proceedings of Interspeech*.

KAZEMZADEH, A., YOU, H., ISELI, M., JONES, B., CUI, X., HERITAGE, M., PRICE, P., ANDERSON, E., NARAYANAN, S., AND ALWAN, A. 2005. TBALL data collection: The making of a young children's speech corpus. In *Proceedings of Interspeech*.

LEE, S., POTAMIANOS, A., AND NARAYANAN, S. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soci. Am. 105,* 3, 1455–1468.

MCBRIDE-CHANG, C. 1999. The ABCs of the ABCs: The development of letter-name and letter-sound knowledge. *Merrill-Palmer Quart. 45,* 2, 285–308.

MOSTOW, J., ROTH, S. F., HAUPTMANN, E. G., AND KANE, M. 1994. A prototype reading coach that listens. In *Proceedings of AAAI*.

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. Tech. rep. 00-4769, National Institute for Child Health and Human Development, National Institute of Health, Washington, D.C.

PARATORE, J. AND MCCORMACK, R. 2007. *Classroom Literacy Assessment: Making Sense of What Students Know and Do*. Guilford Press, New York, NY.

PRICE, P., TEPPERMAN, J., ISELI, M., DUONG, T., BLACK, M. P., WANG, S., BOSCARDIN, C. K., HERITAGE, M., DAVID PEARSON, P., NARAYANAN, S., AND ALWAN, A. 2009. Assessment of emerging reading skills in young native speakers and language learners. *Speech Comm. 51,* 10, 968–984.

RASMUSSEN, M. H., TAN, Z. H., LINDBERG, B., AND JENSEN, S. H. 2009. A system for detecting miscues in dyslexic read speech. In *Proceedings of Interspeech*.

TEPPERMAN, J., LEE, S., ALWAN, A., AND NARAYANAN, S. 2011. A generative student model for scoring word reading skills. *IEEE Trans. Audio Speech Lang. Process. 19*, 2, 348–360.

VAN DOREMALEN, J. J. H. C., CUCCHIARINI, C., AND STRIK, H. 2010. Optimizing automatic speech recognition for low-proficient non-native speakers. *EURASIP J. Audio Speech Music Process.* 1–13.

WITT, S. M. AND YOUNG, S. J. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Comm. 30,* 2-3, 95–108.

YILDIRIM, S., NARAYANAN, S., AND POTAMIANOS, A. 2011. Detecting emotional state of a child in a conversational computer game. *Comput. Speech Lang. 25,* 1, 29–44.

YOU, H., ALWAN, A., KAZEMZADEH, A., AND NARAYANAN, S. 2005. Pronunciation variations of Spanish-accented English spoken by young children. In *Proceedings of Interspeech*.

YOUNG, S., EVERMANN, G., GALES, M., HAIN, T., KERSHAW, D., LIU, X., MOORE, G., ODELL, J., OLLASON, D., POVEY, D., VALTCHEV, V., AND WOODLAND, P. C. 2006. *The HTK Book* (version 3.4). Cambridge University Engineering Department.

ZHENG, J., HUANG, C., CHU, M., SOONG, F. K., AND YE, W. 2007. Generalized segment posterior probability for automatic Mandarin pronunciation evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Received June 2010; revised October 2010; accepted January 2011