AUTOMATIC PRONUNCIATION VERIFICATION OF ENGLISH LETTER-NAMES FOR EARLY LITERACY ASSESSMENT OF PRELITERATE CHILDREN

Matthew Black, Joseph Tepperman, Abe Kazemzadeh, Sungbok Lee, and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA
{matthepb,tepperma,kazemzad,sungbokl}@usc.edu, shri@sipi.usc.edu

ABSTRACT

Children need to master reading letter-names and letter-sounds before reading phrases and sentences. Pronunciation assessment of letter-names and letter-sounds read aloud is an important component of preliterate children's education, and automating this process can have several advantages. The goal of this work was to automatically verify letter-names spoken by kindergarteners and first graders in realistic classroom noise conditions. We applied the same techniques developed in our previous work on automatic letter-sound verification by comparing and optimizing different acoustic models, dictionaries, and decoding grammars. Our final system was unbiased with respect to the child's grade, age, and native language and achieved 93.1% agreement (0.813 kappa agreement) with human evaluators, who agreed among themselves 95.4% of the time (0.891 kappa).

Index Terms— Children's speech, pronunciation verification, automatic reading assessment, letter-names

1. INTRODUCTION

Children's future reading proficiency and their ability to learn effectively through reading has been shown to be correlated with the mastery of reading the names of the letters (letter-names) and producing the sounds of the letters (letter-sounds) at an early age [1]. Assessing children's skills in these reading tasks is an important element of early education to confirm that the children are learning.

Automatic assessment of letter-sounds and letter-names can have several advantages. The personalized assessment required to properly score a child's reading level takes one-on-one time, which a teacher may not always be able to provide. Furthermore, an automatic system may remove some of the personal biases inherent in the judgment of the child's reading level and standardize the grading process. In addition, automatic systems can provide teachers with a fine-grained analysis of the child's pronunciation, offering them insight for future instructional planning.

This paper concentrates on automatically verifying letter-names spoken by preliterate children, complementing our previous work addressing the letter-sound task [2]. Please note that the letter-name verification task is not reduced to one of letter-name recognition. That is, we are not interested in specifying which letter-name the child said, but rather whether the letter-name pronunciation was read acceptably. In most letter-name recognition research (an application that arises, for example, when a person spells aloud an out-of-vocabulary word), the intended letter is not known ahead of time, but the assumption is that it is spoken correctly [3-6]. For this paper, we know what letter-name the child was prompted to say. The difficulty is robustly detecting the innumerable ways a child could produce an unacceptable pronunciation, while not penalizing a child for acceptable pronunciation variations (such as nonnative accent).

There are numerous engineering challenges in automatic letter-name verification for children. Children's speech has high variability within and between speakers [7], and the data used in this research was collected in noisy classrooms from children with multiple language backgrounds. These conditions make it difficult to train representative acoustic models. Furthermore, many of the letter-names are acoustically similar (e.g., /eh m/ and /eh n/), and almost all of them share at least a common phoneme (e.g., /b iy/, /c iy/, /d iy/, /iy/, /jh iy/, /p iy/, /t iy/, /v iy/, and /z iy/). In addition, there is no word or letter context for this isolated letter-name reading task, so we cannot train language models, as is typically done in letter-name recognition tasks when the speaker is spelling real words [4].

We experimented with different acoustic models, dictionaries, and decoding grammars with the goal of attaining automatic letter-name verification with accuracy that neared human agreement. Section 2 describes the data we analyzed. Section 3 briefly describes our verification method, which builds upon our previous work on automatic letter-sound verification [2]. Section 4 shows the experimental results, with a discussion following in Section 5, including an in-depth error analysis and comparison to the letter-sound task and results. We conclude in Section 6.

2. CORPUS

We used data from the Technology-based assessment of language and literacy (Tball) Project [8,9]. The Tball corpus [10] was recorded in kindergarten to second grade classrooms in the greater Los Angeles area. Typical

background noise included speech from other children and the teacher. The corpus contains both native English and Spanish speakers; thus, we can expect certain pronunciation trends, as described in [11]. All 26 English alphabet characters were tested for the letter-name reading task. One lowercase letter was displayed on a computer screen for a maximum of five seconds before the next letter was shown. These transition times were automatically recorded and used to segment the files into single letter-name utterances.

We manually verified (accept/reject) 3508 letter-name utterances, of which 25.1% were rejected. 23.4% of these rejected utterances were due to the child saying nothing. 8.27% of all the utterances were marked as having at least one disfluency (fillers, repetitions, and/or repairs). Table 1 shows performance across various demographics that were provided for some of the children. Using the manual annotations, we created a test set with 780 files (30 files per letter-name) and a train set with the remaining 2728 files (approximately 105 files per letter-name). The data were partitioned so that the proportion of acceptable to unacceptable pronunciations was the same between the train and test set for each letter-name. To compute human agreement statistics, three trained native speakers verified the same 260 files (10 files per letter-name), randomly selected from the test set. Mean pairwise evaluator agreement was 95.38%, with kappa agreement of 0.8914.

Demographic		*Number*	*% Accepted*
Gender	Female	1820	**72.36**
	Male	1582	**77.81**
Grade	K	3012	**75.13**
	1st	420	**70.48**
Age	5	1897	78.12
	6	556	76.80
Native Language	Spanish	1203	**72.98**
	English	1151	**82.71**

Table 1. Children performance (based on manual verification) across various demographics. Bold numbers indicate the difference in proportion is statistically significant ($p < 0.05$).

3. AUTOMATIC VERIFICATION

This section outlines how we trained different acoustic models and designed various dictionaries and decoding grammars for this verification task. These three components each played a key role in attaining performance nearing human agreement. We used the same verification method in our previous work on automatic letter-sound verification, which provided promising results [2].

3.1. Acoustic Modeling

We extracted the first 12 MFCCs plus energy and their corresponding delta and delta-delta coefficients as features. We modeled the letter-names at the phoneme-level since this was the linguistic unit where most of the pronunciation errors occurred. We trained 3-state monophone HMMs with 16 Gaussian mixtures per state. Initial baseline models were trained on 12 hours of isolated word-reading data (without letter-names), also recorded for the Tball Project. A background model was trained on silent and background noise portions of the utterances, and a single generic phone-level "garbage" model was trained on all speech segments. Five sets of acoustic models were iteratively trained directly on the letter-name train set, as described in [2]. All feature extraction and model training was done with HTK [12].

3.2. Dictionaries

A *recognition* dictionary that included all the acceptable letter-name pronunciations served as a baseline dictionary. This dictionary was not ideal since it did not take into account the fact that we knew what letter-name the child was prompted to say. For this reason, we also constructed five additional dictionaries that included unacceptable letter-name pronunciations from foreseeable categorical errors (Table 2). We then produced 32 sets of *verification* dictionaries through all 2^5 combinations of the five categories (none, LS, PE, SI, …, LS-PE, LS-SI, …, all). Each dictionary set contained a dictionary for each letter with acceptable letter-name pronunciations and appropriate unacceptable ones. We refer to the verification dictionary set that did not include any unacceptable pronunciations as the "none" set, and the one that included all the unacceptable pronunciations as the "all" set.

Label	*Description*	*# of Entries*	*Examples*
LS	English letter-sounds	45	v: /v/, /v ah/
PE	Perceptual confusions	43	m-n, f-s, c-z
SI	Sight confusions	21	b-d, p-q, o-c
SP	Spanish confusions	14	j: /hh ey/
SPLN	Spanish letter-names	28	d: /d ey/

Table 2. Description of the five unacceptable pronunciation categories, with the corresponding number of entries and examples

3.3. Grammars

The difficulty in designing a good decoding grammar is finding the correct balance between grammar simplicity and decoding coverage. Ultimately, the goal of the grammar is to constrain the recognizer and endpoint the letter-name pronunciation. We tried four different grammar structures in this paper (where | means "or," { } means zero or more repetitions, BG is background/silence, GG is garbage, and TARGET is one of the entries in the dictionary being used):

1. {BG | GG} TARGET {BG | GG}
2. {BG | GG} TARGET | BG {BG | GG}
3. {BG | GG} TARGET | GG {BG | GG}
4. {BG | GG} TARGET | BG | GG {BG | GG}

3.4. Verification method

We tried all combinations of the verification dictionaries and grammars on the *train* set to determine which were best. If an acceptable pronunciation was recognized, the utterance

was deemed acceptable; otherwise, the pronunciation was rejected. Grammar 2 and the SI-SPLN verification dictionary achieved the highest percent agreement with manual annotations. Grammar 2 performed best, since it was able to detect when the child said nothing, which happened in almost a quarter of the rejected files. Several verification dictionaries performed nearly as good as the SI-SPLN dictionary set because the errors the children made were letter-name specific. For this reason, we created a final letter-specific verification dictionary set that included the individual letter-sound dictionaries with the best performance. From these findings on the train set, we chose to only use a select set of dictionaries (recognition, none, all, SI-SPLN, letter-specific) and grammar 2 on the *test* data.

4. RESULTS

Figure 1 plots kappa agreement for all combinations, with Table 3 providing percent and kappa agreement statistics for the baseline and fourth iteration models.

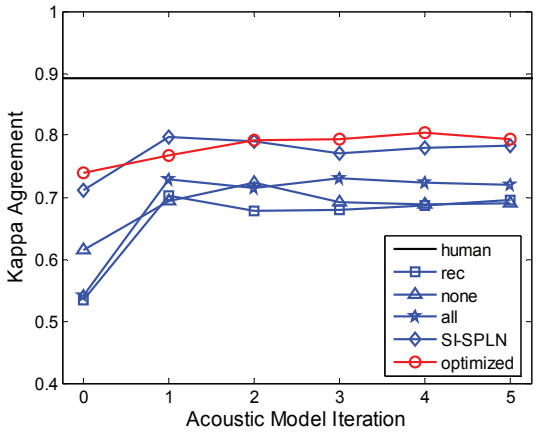

Fig. 1. Kappa agreement statistics on the test data when using different acoustic models and dictionaries ("human" = human kappa agreement, "rec" = recognition dictionary, and the other plots show the results for select verification dictionary sets).

Method	*Baseline Models*		*Iteration #4 Models*	
	Acc (%)	Kappa	Acc (%)	Kappa
chance	75.13	0.0000	75.13	0.0000
recognition	78.21	0.5341	86.79	0.6852
none	87.44	0.6146	89.62	0.6918
all	78.85	0.5421	88.72	0.7306
SI-SPLN	89.62	0.7117	92.31	0.7837
letter-specific	90.51	0.7389	**93.08**	**0.8128**

Table 3. Performance using baseline and fourth iteration acoustic models for various dictionaries on the test data using grammar 2

5. DISCUSSION

The acoustic models trained directly on letter-names (1-5) performed better than the baseline acoustic models, which suggests that the phonemes in spoken letter-names are different than ones within words. However, the baseline acoustic models were able to provide a good alignment with the letter-name data, since acoustic models 1-5 show very little difference in performance. The dictionaries with the most relevant unacceptable pronunciations (SI-SPLN, letter-specific) performed best; having too many unacceptable pronunciations (recognition, all) or too few (none) is less ideal. The best combination (fourth iteration acoustic models, letter-specific dictionary, and grammar 2) achieved 93.08% agreement with human labels, significantly higher than chance (75.13%) but still significantly lower than inter-evaluator agreement (95.38%), with both $p<0.01$.

5.1. Error analysis of letter-name results

Table 4 is the confusion matrix for the best overall combination; the system was better at verifying good pronunciations (95.9% accuracy) than rejecting bad ones (84.5% accuracy). Table 5 shows the system was unbiased with respect to grade, age, and native language. However, the system agreed with human labels more often with males than females, which may be due to the fact that the females did worse than males for this particular corpus (Table 1).

CONFUSION		*Manual Verification*		*Accuracy* (%)
		Reject	Accept	
Automatic Verification	Reject	164	24	87.23
	Accept	30	562	94.93
Accuracy (%)		84.54	95.90	**93.08**

Table 4. Confusion matrix for the best combination

Demographic		*# in test*	*Accuracy (%)*
Gender	Female	389	**91.0**
	Male	368	**94.8**
Grade	K	678	93.1
	1st	81	91.4
Age	5	404	93.3
	6	120	90.0
Native Language	Spanish	248	91.5
	English	253	93.7

Table 5. System performance (according to automatic results) across several demographics. Bold numbers mean the difference in proportion is statistically significant ($p < 0.05$).

To analyze how background classroom noise affected system performance and inter-evaluator agreement, we estimated the signal-to-noise ratio (SNR) of each test utterance using Equation 1, where $\{A_s\}$ is the set of amplitudes corresponding to the samples endpointed within the letter-name pronunciation ("signal"), and $\{A_n\}$ is the complementary set of amplitudes ("noise"). Table 6 shows the resulting SNR statistics. The mean SNR for utterances in which all three evaluators had the same manual verification was not significantly different than the mean SNR for utterances in which they disagreed ($p>0.1$).

However, the mean SNR for utterances in which the system erred (disagreed with the manual verification) was significantly lower than the mean SNR for utterances in which the system was correct ($p<0.01$). This implies that noise did not affect human evaluator agreement but adversely affected automatic verification performance.

$$\mathrm{SNR} = 10\log_{10}\frac{\frac{1}{S}\sum_{s=1}^{S}A_s^2}{\frac{1}{N}\sum_{n=1}^{N}A_n^2} \quad (1)$$

Partition of test data		*# in test data*	*SNR Statistics [dB]*	
			mean	std. dev.
Inter-evaluator	agree	193	9.335	3.712
	disagree	33	8.632	3.292
System	correct	648	**9.623**	3.533
	error	42	**7.810**	3.796

Table 6. SNR statistics comparing the effect of noise on inter-evaluator agreement and system performance. Bold numbers means the difference in means is statistically significant ($p<0.01$).

5.2. Comparison between letter-names and letter-sounds

According to the manual verification, children performed significantly better on the letter-name task (74.9% accepted) than the letter-sound task (72.2% accepted), with $p<0.05$. This is probably because all letter-names have a one-to-one mapping for their pronunciations, while many of the letter-sounds have alternative pronunciations depending on word context. The letter-sounds are also shorter and less natural to pronounce aloud, which may have been a factor in the letter-sounds having twice as many disfluencies (16.9%), a significant difference with $p<0.05$. Human agreement statistics for both tasks were nearly identical.

We found the same trends in our automatic verification performance for both the letter-name and letter-sound tasks, in that the baseline models were worse than models trained on in-domain data, with grammar 2 and the letter-specific dictionary providing the best results. English letter-name substitutions and alternative pronunciations were the most common categorical errors for the letter-sound task, with sight confusions and Spanish letter-name errors dominating the letter-name task. Overall, we attained higher verification accuracy on the letter-name task (93.08% accuracy), compared to the letter-sound task (87.95% accuracy), with $p<0.01$. We feel this difference is mostly due to the acoustic models. Whereas HMMs using MFCC features model letter-name phonemes well, they seem to be less suited for the more noise-like letter-sounds. Future research on letter-sound specific features will hopefully help bridge this gap.

6. CONCLUSION

We showed that we could accurately verify letter-name pronunciations through acoustic modeling at the phoneme-level. We achieved the best results using a dictionary optimized for each letter separately. Our final automatic system agreed with humans 93.1% of the time (0.813 kappa), nearing inter-evaluator agreement of 95.4% (0.891 kappa), and was unbiased with respect to the child's grade, age, and native language. This system also performed significantly better than the one we previously developed to verify the more difficult letter-sounds [2]. In the future, we want to improve system performance in the presence of noise through improved acoustic modeling and/or by automatically detecting when there is too much background noise to reliably verify the utterance.

7. ACKNOWLEDGEMENTS

This works was supported in part by the National Science Foundation. Special thanks to Matthew Tan and Isaac Rottman for their help in transcribing the letter-name data.

8. REFERENCES

[1] National Reading Panel, "Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction," *NICHD*, NIH Publication 00-4769, Washington, DC, 2000.

[2] M. Black, J. Tepperman, A. Kazemzadeh, S. Lee, and S. Narayanan, "Pronunciation verification of English letter-sounds in preliterate children," *Proc. Interspeech*, Antwerp, Belgium, 2007.

[3] M. Fanty and R.A. Cole, "Spoken letter recognition," *Advances in Neural Information Processing Systems 3*, San Mateo, CA: Morgan Kaufmann, 1991.

[4] H. Hild and A. Waibel, "Recognition of spelled names over the telephone," *Proc. ICSLP*, Philadelphia, PA, 1996.

[5] P.C. Loizou and A.S. Spanias, "High performance alphabet recognition," *IEEE Trans. Speech and Audio Processing*, 4(6):430-445, 1996.

[6] M.E. Munich and Q. Lin, "Explicit modeling of common acoustic features for character recognition," *Proc. EUSIPCO*, Vienna, Austria, 2004.

[7] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: developmental changes of temporal and spectral parameters," *J. of Acoust. Soc. Am.*, 105:1455-1468, Mar. 1999.

[8] Tball. http://diana.icsl.ucla.edu/Tball/assess_frame.html.

[9] A. Alwan *et al*., "A system for technology based assessment of language and literacy in young children: the role of multiple information sources," *Proc. MMSP*, Greece, 2007.

[10] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball data collection: the making of a young children's speech corpus," *Proc. Eurospeech*, Lisbon, Portugal, 2005.

[11] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, "Pronunciation variations of Spanish-accented English spoken by young children," *Proc. Eurospeech*, Lisbon, Portugal, 2005.

[12] Cambridge University, HTK 3.2, htk.eng.cam.ac.uk.