Pronunciation Verification of English Letter-Sounds in 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

Correctly reading letter-sounds is an essential first step towards reading words and sentences. Pronunciation assessment of letter-sounds is an important component of preliterate children's education, and automating this process can have several advantages. We propose a method to automatically verify the pronunciations of a letter-sound task administered to kindergarteners and first graders in realistic noisy classrooms. We compare different acoustic models, decoding grammars, and dictionaries to help differentiate between acceptable and unacceptable pronunciations. Our final system achieved 88.0 percent agreement (0.702 kappa agreement) with expert human evaluators, who agree themselves 94.9 percent of the time (0.886 kappa agreement).

Terms: children's speech, pronunciation verification, automatic reading assessment, letter-sounds

1. Introduction

Children are taught to read beginning with letters, the most basic units of written words. Both the names of the letters and the sounds they represent (called "letter-sounds" in this paper) are the focus of much attention in early literacy education. They are necessary first steps in learning to read words and phrases. Furthermore, the degree of "phonological awareness" a child acquires through explicitly mapping letters to sounds is an excellent predictor of long-term reading ability in the years to come [1].

This study focuses on the task of modeling and automatically verifying the correctness of these letter-sounds when produced by young children in a classroom setting. Implicit here is the application to general automatic assessment of a child's reading skills [2], in which letter-sound production would be one of the test components. Such an automatic assessment tool would be valuable to teachers for fine-grained analysis of pronunciation, documentation of long-term trends, and standardization of the assessment process over a large number of students, and over time.

Administering the letter-sound test automatically has not been widely explored, and is not a straightforward thing to implement even without automated methods. Though teachers generally agree as to what the single correct sound for each letter should be, the similarity of many letter-sounds (for example, "f" and "s") and the lack of a word or phrase context makes judging correctness quite a challenge, even for trained listeners. Because English does not have a one-to-one mapping from orthographic characters to the phonemes they represent, we must expect many variants in production which will usually be very close to one another in perceptual space. The variability in children's speech is generally quite high relative to that in adults [3], and the letter-sound task introduces its own peculiar variations. For example, some children hyper-articulate and lengthen the letter-sounds, confuse the letters' sounds with their names, or cannot remember the sound they were taught to associate with each letter. Additional variation can come from many sources, including sight confusions between characters (for example, "o" and "c"), and the influence of the child's native language (e.g., Spanish), with its own letter names and letter-to-sound rules. Speech technology for the letter-sound task must account for all these elements, without recourse to the visual cues teachers might use to help determine, for example, stop consonant identity. (Combining computer vision and speech technology is certainly a potential future direction [4,5]).

Speech technology can do things that even well-trained listeners may have trouble doing, such as detect a very subtle difference in production and quantify that difference, or compare a specific list of recognition hypotheses and rank them accordingly. With its close discrimination between similar acoustic expectations, the letter-sound task is a challenge toward shrewdness in designing a speech verification algorithm. In this paper we investigate several variations on implementing automatic letter-sound verification, including using different acoustic models, decoding grammars, and dictionaries for modeling pronunciation. Our goal is to automatically verify letter-sounds produced by children with accuracy approaching the inter-listener agreement in rating our corpus, and to reveal the unique challenges in modeling letter-sounds as a guide for future work in this very new area.

2. Corpus

The data used in this study was from the Tball Project [6], collected in kindergarten and first grade classrooms in the greater Los Angeles area. The speech was recorded with standard headset microphones in realistic classroom environmental conditions, where typical background noise included spontaneous speech from multiple children and the teacher [7]. The majority of the children in the corpus were from Mexican-American families where Spanish was the first language; this made Spanish-related errors quite prevalent [8].

2.1. Letter-Sound Reading Task

Children were tested on all 26 English alphabet letter-sounds. One lowercase letter at a time was displayed on a computer screen by an animated character, and the child had up to five seconds to read the appropriate letter-sound. The transition times between letters were automatically recorded and used to segment the sessions into single letter-sound utterances.

An expert linguist and teacher created a dictionary with all the acceptable phonemic spellings for each letter-sound. We used the convention that all vowel sounds were supposed to be "soft" (a: /ae/, e: /eh/, i: /ih/, o: /aa/, u: /ah/). Additionally, for consonants with multiple pronunciations depending on word context, only the primary pronunciation was accepted (e.g., for the letter "c," /k/ was accepted, and /s/ was rejected). For all voiced letter-sounds, the child could end with the phoneme /ah/ (e.g., the letter "b" had two correct pronunciations: /b/ and /b ah/).

2.2. Corpus Statistics

Using this dictionary, we manually verified 3685 letter-sounds from 150 children. 72.2 percent were marked as acceptable pronunciations, and 16.9 percent were labeled as *disfluent*, meaning that the utterance had repetitions and/or repairs. While utterances marked as disfluent were three percent more likely to also be marked as unacceptable pronunciations, this difference was not significant (p>0.1). The children performed significantly worse on vowels (47.8%) compared to consonants (75.6%), with p<0.001, most likely due to their multiple alternative pronunciations.

Using the manual verification annotations, we created a test set with 780 files (30 files per letter-sound) and a train set with the remaining 2905 files (approximately 110 files per letter-sound). 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-sound. To compute human agreement statistics, three expert annotators verified the same 260 files (10 files per letter-sound), randomly selected from the test set. The mean agreement between the three annotators across this subset was 94.9%, with Fleiss kappa agreement of 0.886.

3. Acoustic Modeling

For these experiments, we extracted the first 12 MFCCs plus energy with a 25 ms Hamming window at a frame rate of 100 Hz with HTK [9]. The delta and delta-delta coefficients formed the final 39-dimensional feature vector. Cepstral mean normalization, subtracted at the utterance-level, made these features more robust to classroom noise. We chose to model the letter-sounds at the phoneme level since unacceptable pronunciations usually differed from acceptable pronunciations by a single phoneme. “Whole-word” letter-sound models were less suitable due to the variations in the acceptable pronunciations (e.g., b: /b/ or /b ah/). For each phoneme, we trained a 3-state monophone HMM with 16 Gaussian mixtures per state and a diagonal covariance matrix. The number of Gaussian mixtures and the use of monophone versus triphone representations were empirically chosen.

Initial baseline monophone models were trained on 12 hours of isolated word-reading data (without letter-sounds) recorded for the Tball Corpus. Background models were trained on silent and background noise portions of the utterances, and generic phone-level “garbage” models were trained on all speech segments. The letter-sound train set was then automatically transcribed using these baseline models, and this set was then used to train new acoustic models from the letter-sound data alone. For this automatic transcription, files annotated as “acceptable” were decoded with a network of correct pronunciations in the dictionary, and files annotated as “unacceptable” were decoded with an open “garbage” and background loop. With these new letter-sound-derived monophone models, we did five iterations of decoding the letter-sound train set and then retraining the models based on the new transcripts. A comparison among the initial baseline acoustic models and the five subsequent letter-sound acoustic models will be made in Section 5.

4. Automatic Verification

Automatic pronunciation verification requires that we have representative models and methods to detect both acceptable and unacceptable pronunciations. While the acoustic models play a crucial role in modeling the speech, the choice of dictionary and decoding grammar significantly impact the automatic verification performance. In this paper, we compare several dictionaries and grammars.

4.1. Dictionaries

We created dictionaries that included phonemic spellings of acceptable *and* unacceptable pronunciations for the letter-sounds. The acceptable pronunciations were already known from the dictionary used to manually verify the data. We used this dictionary, which contained 44 entries (an average of 1.69 entries per letter-sound), to run baseline recognition experiments on the letter-sound data, making every acceptable letter-sound pronunciation an *unacceptable* pronunciation for the other letter-sounds.

This recognition dictionary is not ideal for pronunciation verification for many reasons. First, we are ignoring the fact that we know what letter-sound the child is prompted to say. This causes the dictionary to contain a lot of superfluous entries. For example, we would not expect a child to confuse the letter-sound for “z” with the letter-sound for “a”. The recognition dictionary is also not ideal because it does not contain many of the speaking errors that a child could make. For instance, when prompted with the letter “j,” they might say the letter-name /jh ey/, or when prompted with the letter “o,” they might say the phoneme /ow/. These types of errors are typically not in the recognition dictionary but are nevertheless possible.

To combat this problem, we constructed six additional dictionaries that included unacceptable letter-sound pronunciations from foreseeable categorical errors. These dictionaries were made with the help of an expert teacher and linguist and are described in Table 1. We then produced 64 sets of *verification* dictionaries through all combinations (2^6) of the six unacceptable pronunciation categories: { }, LN, PE, PR, … , LN-PE, LN-PR, … , LN-PE-PR, ... Each dictionary set contains 26 dictionaries (one for each letter-sound) and includes the acceptable pronunciation(s) and the appropriate unacceptable pronunciation(s) for the letter-sound. The rationale behind creating these verification dictionary sets is to experiment with detecting specific types of categorical errors. Thus, the LN-PR dictionary specializes in detecting English letter-names and alternative letter-sound pronunciations. If these errors are prevalent in the data and the acoustic models are representative, then this dictionary will perform well. The verification dictionary set that includes only acceptable pronunciations (referred to as the “none” set) can be thought of as the opposite of the recognition dictionary. The verification dictionary set that includes the acceptable pronunciations and all six unacceptable pronunciation categories (referred to as the “all” set) is capable of detecting all foreseeable mispronunciations.

Label	*Description*	*Size (per letter)*	*Examples*
LN	English letter-names	38 (1.46)	k: /k ey/
PE	Auditory confusions	75 (2.88)	f-s, m-n, k-t
PR	Alternative prons	48 (1.85)	c: /s/, g: /jh/
SI	Sight confusions	25 (0.96)	b-d, p-q, h-n
SP	Spanish confusions	21 (0.81)	u: /uw/
SPLN	Spanish letter-names	35 (1.35)	k: /k aa/

Table 1: *Description of the six unacceptable pronunciation dictionaries with corresponding size and example entries*

4.2. Grammars

The role of the grammar is to constrain the speech recognition to the probable segment-unit sequences. Ideally, the grammar will allow for proper endpointing of the letter-sound pronunciation(s). We tried eight grammar structures (where | means "or", { } means zero or more repetitions, [ ] means one or more repetitions, BG is background/silence, GG is "garbage," and TARGET depends on the dictionary):

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}
5. {BG | GG} [TARGET] {BG | GG}
6. {BG | GG} [TARGET | BG] {BG | GG}
7. {BG | GG} [TARGET | GG] {BG | GG}
8. {BG | GG} [TARGET | BG | GG] {BG | GG}

For the recognition dictionary, TARGET is A | B | C | … | Z, where the capital letters represent their corresponding acceptable letter-sound pronunciation(s). For the verification dictionaries, TARGET is constrained to only acceptable and unacceptable pronunciations for the letter being tested. For example, if the child was prompted to speak the letter-sound for "m," and we were using the LN verification dictionary set, TARGET is M | M_{LN}*, where M_{LN}* represents the English letter-name pronunciation(s) for the letter "m."

Grammars 1-4 constrain the recognition decoding to a single instance of TARGET, with grammar 2 also allowing recognition of background/silence in place of TARGET, grammar 3 allowing "garbage," and grammar 4 allowing both. Grammars 5-8 are the same as grammars 1-4 except that they allow for repetitions. These grammars were implemented since a large percentage of the data contained disfluencies.

4.3. Verification Method

There are six sets of acoustic models, 65 dictionaries (1 recognition, 64 verification sets), and eight grammars. To help determine which grammars and verification dictionaries were best, we tried all 3120 combinations on the *training* data and used a simple decision scheme to automatically determine the correctness of the pronunciation: if an acceptable pronunciation was recognized, the utterance was deemed acceptable; otherwise, the pronunciation was rejected. For grammars that allowed for repetitions (grammars 5-8), the final recognized pronunciation was used for verification, with all other recognized pronunciations in the utterance ignored.

We quantified performance of each combination using percent agreement with the manual verification annotations. Due to space constraints, we cannot list all the results, but the best combination for all acoustic models occurred when using decoding grammar 2 with the LN-PR verification dictionary. Grammars allowing repetitions (5-8) always performed worse than the grammars not allowing repetitions (1-4), although this difference was not significant ($p>0.1$). Grammar 2 outperformed or matched the performance of the other grammars in the majority of the cases. Grammar 2 allowed background/silence to be an unacceptable pronunciation and may have performed best since it was able to detect pronunciation errors occurring when the child said nothing.

While the LN-PR verification dictionary attained the highest agreement, several other verification dictionaries had similar performance. This is most likely because the types of errors the children made were letter-sound specific (e.g., sight confusion between the letters "b" and "d", and Spanish-related errors for vowels). Thus, different verification dictionary sets were more suited for particular letter-sounds. To account for this, we created a final *optimized* verification dictionary set that included the individual letter-sound dictionaries that resulted in the highest agreement on the training data. From these findings, we chose to use only grammar 2 and a select set of dictionaries (recognition, none, all, LN-PR, optimized) on the test data. We used two metrics to quantify the performance of the various combinations: percent and kappa agreement with the manual verifications.

5. Results & Discussion

Figure 1 plots kappa agreement statistics on the test data as a function of the acoustic model iteration, and Table 2 lists both accuracy (percent agreement) and kappa agreement statistics for the initial baseline and first iteration acoustic models. Table 3 shows the confusion matrix on the test data for the best overall combination (acoustic models iteration 1, grammar 2, and the optimized verification dictionary).

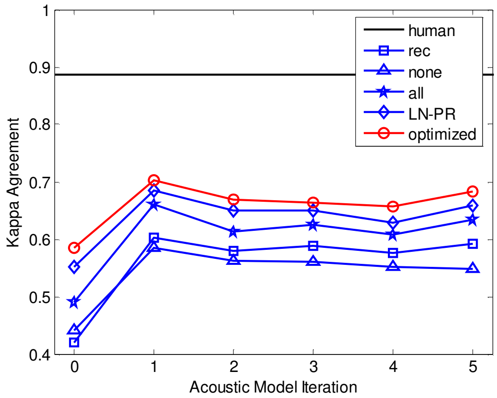

Figure 1: *Kappa agreement statistics on the test data when using different acoustic models and dictionaries ("human" = human kappa agreement, "rec" = recognition dictionary, and the other plots show the results from select verification dictionary sets).*

Method	Baseline Models		Iteration #1 Models	
	Acc (%)	Kappa	Acc (%)	Kappa
chance	71.67	0.0000	71.67	0.0000
recognition	70.13	0.4193	81.28	0.6022
none	80.13	0.4411	84.87	0.5855
all	75.64	0.4896	85.13	0.6599
LN-PR	82.44	0.5509	87.18	0.6852
optimized	83.97	0.5859	**87.95**	**0.7024**

Table 2: *Performance using the baseline and first iteration acoustic models for various dictionaries on the test data using grammar 2*

CONFUSION		*Manual Verification*		*Accuracy (%)*
		Reject	*Accept*	
Automatic Verification	*Reject*	173	46	79.00
	Accept	48	513	91.44
Accuracy (%)		78.28	91.77	**87.95**

Table 3: *Confusion matrix for best combination*

Figure 1 shows that the best results came from the first iteration acoustic models when using the optimized verification dictionary. Subsequent acoustic model iterations did not improve results, which means the mismatched monophone models trained on isolated words were able to provide a good alignment with the letter-sound data. As shown in Table 2, the best performance had 87.95 percent agreement and 0.7024 kappa agreement with the manual verification annotations. While this agreement is considered high, it is still significantly worse than human agreement of 94.9 percent (0.886 kappa), with $p<0.05$.

To see the effect of the number of training files on performance, we re-trained the first iteration of acoustic models with smaller subsets of training data and re-tested the resulting acoustic models on the test set. Figure 2 shows that the models appear to converge (in terms of kappa agreement) at around 80 files per letter-sound, and this plot suggests that we would not benefit greatly from more training data using our current modeling and decoding techniques. The large performance gap between the train and test data shows the inherent variability of letter-sounds and suggests there is room for acoustic modeling improvement.

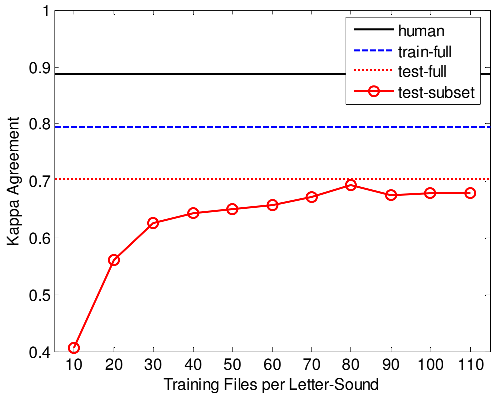

Figure 2: *Kappa agreement statistics ("train-full" and "test-full" show performance on the train/test set when training with the full training set, while "test-subset" shows performance on the test data as a function of the number of training files).*

Automatic verification performance was significantly worse on vowels (81.33%) than consonants (89.36%), with p<0.05. This may be due to the fact that the best combination maximized percent agreement with humans, which resulted in a lopsided confusion matrix. As shown in Table 3, the system is far better at accepting good pronunciations (91.77% accuracy) than rejecting bad ones (78.28% accuracy). Since vowels are spoken incorrectly more by children (section 2.2), the system tends to make more mistakes on vowels. Also, there is a shortage of acceptable pronunciations for vowels in the train set, and thus, less representative samples for HMM training for these sounds. Additionally, the Spanish-related pronunciation variations in the data may have brought the vowels closer in the MFCC feature space, making them less discriminable.

Automatic verification performance was also significantly worse on disfluent files (81.68%) than fluent files (89.21%), with p<0.05. This can be attributed, in part, to the use of grammar 2, which did not allow for repetitions of TARGET. In instances where the utterance had multiple pronunciations, grammar 2 had to choose one, which is not a robust endpointing procedure. Even though the grammars allowing repetitions fit the data better, they were not as reliable, falsely detecting so many disfluencies that overall performance was degraded.

Figure 1 also displays the effect that dictionary size has on kappa agreement. Performance suffers when the dictionary has too few unacceptable pronunciations ("none") and when it has too many ("rec" and "all"). It is best to limit the entries to the most relevant categorical mispronunciation types ("LN-PR" and "optimized"). While we initially designed the verification dictionaries to help detect relevant unacceptable pronunciations, it should be noted that these dictionaries also serve another important purpose: they classify the types of errors the child is making. For example, if the automatic verification system detected that the child made several sight confusion errors, perhaps the child has not learned the shapes of the sounds well enough yet. Thus, the verification dictionaries are not only able to differentiate between acceptable and unacceptable pronunciations but also provide information relating to why the child is making mistakes. This information could then be used by the teacher to help address these problems.

6. Conclusion & Future Work

We showed that accurate assessment of a letter-sound task can be accomplished through acoustic modeling of the letter-sounds at the phoneme level. We used generic children's monophone models trained on isolated words to help attain a rough alignment of letter-sound data and trained letter-sound specific monophone models in a single iteration. Our final system agreed with expert human manual verification 87.95 percent of the time (kappa = 0.7024) when an appropriate grammar was used with dictionaries optimized for detecting the most relevant letter-sound pronunciation errors. Future work includes the development of more advanced methods to detect disfluencies, adaptation of the isolated word monophone acoustic models to the letter-sound data, the use of language-specific acoustic models, and incorporating n-best recognition lists and log-likelihood thresholding.

In this study, we concentrate solely on pronunciation verification of letter-sounds using audio signals. However, a complementary source of information regarding what phoneme a person is saying is conveyed visually [4,5]. For example, the lip configuration can help differentiate /f/ from /s/. In the future, we want to explore how we can use this visual information robustly to aid in assessing letter-sounds to achieve performance that nears human agreement.

7. Acknowledgments

This work was supported in part by the National Science Foundation. Special thanks to Patti Price for her help in making the dictionaries and to the entire Tball Project team for the data collection and design of the letter-sound task.

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] 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.

[3] 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.

[4] S. Chu and T.S. Huang, "Bimodal speech recognition using coupled hidden Markov models," Proc. ICSLP, Beijing, 2000.

[5] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: summary from 2006 JHU summer workshop," Proc. ICASSP, Hawaii, 2007.

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

[7] 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.

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

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