PRONUNCIATION VERIFICATION OF CHILDREN'S SPEECH FOR AUTOMATIC LITERACY ASSESSMENT

Joseph Tepperman[1], *Jorge Silva*[1], *Abe Kazemzadeh*[1],
Hong You[2], *Sungbok Lee*[1], *Abeer Alwan*[2], *and Shrikanth Narayanan*[1]

[1] Signal Analysis and Interpretation Laboratory, University of Southern California
[2] Speech Processing and Auditory Perception Laboratory, University of California, Los Angeles

ABSTRACT

Arguably the most important part of automatically assessing a new reader's literacy is in verifying his pronunciation of read-aloud target words. But the pronunciation evaluation task is especially difficult in children, non-native speakers, and pre-literates. Traditional likelihood ratio thresholding methods do not generalize easily, and even expert human evaluators do not always agree on what constitutes an acceptable pronunciation. We propose new recognition- and alignment-based features in a decision tree classification framework, along with the use of prior linguistic information and human perceptual evaluations. Our classification methods demonstrate a 91% agreement with the voted results of 20 human evaluators who agree among themselves 85% of the time.

Index Terms: children's speech, literacy, pronunciation

1. INTRODUCTION

Automatically assessing a child's literacy skills is a complex problem. Based on a read-aloud speech signal along with prior knowledge of the expected target utterance, we can infer quite a bit – the child's confidence in reading, his level of fluency or comfort, even the influence and degree of non-native phonetics – just as a real reading tutor could. But how we can distinguish a mispronunciation based on underdeveloped reading skills from one caused by a non-native accent or speaker-dependent speech production difficulties is another matter. Here the problem lies in teasing apart these various factors in an effort to provide an accurate and meaningful assessment of literacy, aside from expected variations in accent or pronunciation.

In an ASR task such as this, a given mispronunciation cannot be presumed to be the sole result of any one source. The child who, when prompted with the word /f ay n d/ ("find"), reads aloud something that sounds like /f ih n d/, probably does so because of an unfamiliarity with the orthographic conventions of English letter-to-sound rules. However, first-graders of Mexican-American background (as populate much of the Los Angeles public school system) might read the target word "two" more like "do" because of the shorter Voice Onset Time in Spanish-accented stops [5], but chances are this pronunciation variant would not be the product of poor reading skills, and therefore should not be assessed as such. Combine this ambiguity with an easily confusable wordlist (typical Grade 1 words: well/will, saw/so, etc.) and the high age-dependent variability of children's speech [6], and you have an evaluation problem for which simple word-level recognition grammars and traditional log-likelihood ratio thresholding will not suffice.

A reasonable solution, then, would be to customize the assessment algorithm to account for those predictable phonemic insertions, deletions, and substitutions which prior knowledge of the speaker set and target vocabulary deem acceptable for the literacy assessment task, as in the do/two case above [8]. Though this is in a sense possible (and is the method used in this study), it perhaps becomes an intractable problem for larger vocabularies demanding overspecified linguistic rule sets. Additionally, experts in child literacy don't always agree on what constitutes an acceptable mispronunciation by a speaker with a nonnative accent (see Section 3 for details on correlation among human annotators), so there is always a degree of uncertainty in these assigned class labels no matter how specific we allow the a priori pronunciation rules to be.

Our main concern in this study was to tackle the pronunciation evaluation problem as a preliminary but crucial step of this proposed literacy assessment - to compare our automatic results with those obtained from reliable human ears and to generalize our methods such that they might be easily adapted to suit the young reader's ever-expanding vocabulary.

2. PRONUNCIATION VERIFICATION

The likelihood that a given time-series of observations is consistent with a target model is given by the probability expression $P(O \mid \lambda_i)$. To generate a more refined score of the confidence that O belongs to class λ_i, the classic method [9] is to take a ratio of these likelihood probabilities:

$$\tau = \frac{P(O \mid \lambda_i)}{P(O \mid \lambda_f)}$$

where λ_f is a generalized "filler" model for all miscellaneous speech. Taking the log of τ we can turn the likelihood ratio into a difference of log-likelihoods. And choosing an appropriate threshold T we can empirically optimize the binary decision of accepting or rejecting the pronunciation: for $\tau \geq T$ we accept, for $\tau < T$ we reject.

This is how word-level verification is often performed, though it has been shown to be less than useful in all but the simplest of recognition tasks, since the threshold is not easily generalized for a large vocabulary; depending on their phonetic properties, certain words will require a verification threshold farther from the filler model than others. One suggested improvement [7] is to use a unique filler model for each target word, one that omits any instances of the target word in question during the training stage, though this necessitates retraining acoustic models each time the reading list vocabulary is changed (as it often will be).

Clearly this classification task demands more features, and perhaps a more complex classification algorithm. Sources such as [2] and [3] suggest deriving new acoustic scores based on a recognition grammar over the entire task vocabulary, rather than from fixed alignment of the target word and global filler model. Likelihood scores for linguistically close pronunciations will serve as a discriminative foil for the target pronunciation's acoustic model (a finer-grain and pronunciation-oriented version of the word-dependent filler model mentioned above), and the distant words need not be considered, and will not be recognized except in the case of an extreme mispronunciation. In this way we can then estimate the filler model dynamically without severe training overhead, and focus on improving performance in the case more commonly seen in pronunciation verification – that of false acceptance.

As for the classifier, a linear threshold is a good place to start, but the framework needs to be augmented to account for cases where, for example, the log-likelihood ratio is relatively high but the target likelihood component is not, or the target word is not recognized despite the high score computed upon alignment with the target model (both cases and many more idiosyncratic were well-represented in the data). For this reason, and because of its acceptance of both binary and continuous features and its easy interpretability, we decided to use a decision tree classifier for our pronunciation assessment. Details about this algorithm and its features are discussed in Section 4.

Now what about making use of the specialized prior knowledge alluded to in the Introduction? An effective and scalable way of incorporating the acceptable but non-canonical pronunciations into the verification task is to augment the recognition dictionary with all acceptable pronunciation variants, derived both from known linguistic rules of the students' native language or foreign accent [8,11] and from careful analysis of the transcriptions, as an indication of what to expect.

We also propose using human evaluation knowledge besides on the transcription level. This is more or less an unsupervised learning task – we do not know in advance the "true" class labels, acceptable or unacceptable, for any of the pronunciations, because even expert labelers cannot always agree when evaluating pronunciation. So to train an accurate decision tree, we'll need to somehow estimate the true class labels using human evaluations – the same human evaluations collected for purposes of comparison with our automatic classification results. Our method intends to demonstrate an improvement in the classification results when said classifier is informed by human evaluations.

3. HUMAN EVALUATIONS

The data used in this study comes from the Tball Corpus [4], which was gathered at Los Angeles area schools and motivated by a long-term goal to develop automatic literacy assessment software for elementary school teachers. The particular subset we used is the Grade 1 word list recordings, consisting of 2076 one-word utterances from roughly an equal number of boys and girls, ages 5-8, over a 51-word vocabulary typical of first grade reading ability.

Our evaluation consisted of a set of 102 utterances, two examples of each target word from the Grade 1 word list, representative of typical canonical and noncanonical pronunciations, respectively. Our 20 evaluators – 3 of them teachers, 8 of them native American English speakers, all of varying degrees of Spanish language fluency – were asked to mark each item as "acceptable" or "unacceptable" based on the child's pronunciation of the target word. "Unsure" was also given as a choice, though we mapped this response to "unacceptable," assuming that a live reading tutor would err on the side of rejecting a good pronunciation rather than accepting a bad one. A pairwise measurement of our evaluators' Kappa agreement resulted in a mean score of 0.69 with a standard deviation of 0.10 (in terms of percent agreement this was 85.1% with a standard deviation of 6%).

In addition to using our evaluation results as a metric for our automated procedure, we also wanted to explore the human raters' performance with respect to their English and Spanish abilities, and their teaching experience. Figure 1 shows how evaluator performance varied with some different groupings. Though the expert teacher agreement is numerically higher than the non-teacher class, we found with 95% confidence that the teachers did not have statistically higher inter-agreement than the non-teachers. The same was true of the native vs. non-native agreement means. This indicates that experts and native speakers do not necessarily perceive pronunciation with dramatically more agreement than anyone else.

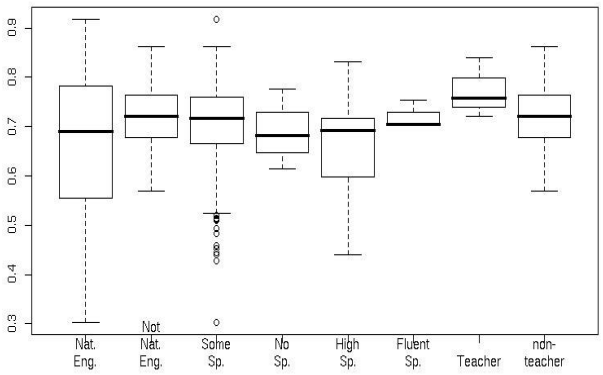

Figure 1. Distribution of evaluator agreement (Kappa), grouped as Native and Non-native speakers, varying Spanish ability, and teachers vs. non-teachers.

4. ALGORITHMIC DETAILS

4.1. Acoustic Model Training

The basic acoustic models used in all these experiments come from the Kindergarten Wordlist and Beethoven Elementary subsets of the Tball data (which are disjunct from the Grade 1 recordings mentioned above). All recordings were transcribed on the word level, then segmented on the word boundaries so that silence/background models could be trained separately from the phoneme models. After expanding the word transcriptions into canonical phone-level pronunciations, Hidden Markov Models for each phone were trained in HTK [1] using embedded re-estimation with MFCCs (plus delta, acceleration, and energy coefficients). Each monophone HMM had a standard three-state topology with 16 Gaussian mixtures per state. The silence/background model had 256 mixtures per state, which we found was necessary for accurate endpointing when measuring the students' response time (as an indication of their fluency). Using all the speech data, we also trained a generalized word-level filler model with three states and 16 mixtures per state - this word-level filler performed better than a comparable phone-level filler.

4.2. Feature Selection

The features chosen for pronunciation verification are all derived from word alignment and recognition likelihood scores. For baseline experiments with the traditional approach, we used only the traditional confidence measure in Section 2, calculated based on the likelihood score after alignment with the target word models normalized by two different filler models (for comparison): the general word-level filler, and a dynamic filler derived from the likelihood score of the recognized word. The G1 list contained many examples of phonetically close pronunciations, so the best result returned by the recognizer should serve as an indication of pronunciation variability. For the baseline threshold classifiers, these scores were obtained with the dictionary of canonical pronunciations – no prior linguistic rules were used.

Training of the decision tree classifier included all these baseline features – target word likelihood score, recognized word likelihood score, word-level filler, dynamic filler, and all combinations of likelihood ratios – though a dictionary of pronunciation variants was used to obtain target word alignment scores, and the dynamic filler was calculated by averaging the likelihood scores of the 20-best results. The best recognition result was included as a binary feature (1 if it matched the target word, 0 if it didn't), and the percentage of the 20-best results which matched the target was also included as a feature. The differences between comparable target and recognized word likelihood ratios were included as well. And the student's response time and word duration were also used as features, as they may be indicative of pronunciation fluency.

4.3. Experimental Setup

The threshold imposed on the baseline log-likelihood ratios was determined empirically over a wide range of possible thresholds, the best one being selected based on highest agreement with the human evaluations. As the true class labels were unknown for this task, we explored two techniques for assigning class labels to the decision tree's training set: one, take a majority "vote" of the human evaluations for what the true class should be (these voted class labels had 93% average agreement with the three expert teacher evaluators); two, use the word-level transcriptions so that if the transcribed word matches the target word, we put it in the acceptance class, otherwise it's in the rejection class. The latter method does not necessarily agree in the pronunciation verification case, since an acceptable mispronunciation might generate a different dictionary word as surface form (as with do/two); however, the transcription class labels were found to agree with the voted ones 95% of the time, so the method seemed a valid choice – we can think of the transcriptions as data from another expert human evaluator (and they are, in fact), with which to compare our automatic results. These decision tree training methods were compared using a leave-one-out crossvalidation procedure over the entire evaluation set. The decision tree was trained using the C4.5 algorithm implemented in the Weka toolkit [10].

5. RESULTS AND DISCUSSION

The results of our four classifiers are enumerated in Table 1, alongside statistics for inter-evaluator agreement. Each algorithm is compared to the expert teacher evaluators, the voted approximation of the "true" class labels, and an average with respect to all 20 evaluators. In the context of this work, the log-likelihood ratio threshold classifier with the general filler model (*threshold : general*) can be

		Kappa	*P(agreement)*
inter-evaluator	*teachers*	0.77	0.89
	voted	0.85	0.93
	all	0.69	0.85
threshold : general (baseline)	*teachers*	0.59	0.80
	voted	0.66	0.83
	all	0.55	0.78
threshold : dynamic	*teachers*	0.72	0.87
	voted	0.82	0.91
	all	0.67	0.84
tree : voting	*teachers*	0.72	0.86
	voted	0.80	0.90
	all	0.68	0.84
tree : transcripts	*teachers*	**0.72**	**0.86**
	voted	**0.82**	**0.91**
	all	**0.67**	**0.84**

Table 1. Mean Kappa and agreement statistics for human evaluators and four classifiers, compared with expert teacher evaluators, the voted class labels, and averaged over all evaluators.

considered as a baseline for automatic verification. And, as expected, it performed the poorest of the four algorithms. The other three all had very similar performance, and came well within the 6% standard deviation of the inter-evaluator agreement scores. The voted class labels (an approximation of what the "true" classes may be) agree with all human evaluators 93% of the time, on average, so a 90-91% agreement with the voted class labels indicates that these classification algorithms perform about as well as a human evaluator. And since the teachers agree among themselves 89% of the time, our 86-87% agreement with the teachers suggests these automatic methods can serve about as well as an expert evaluator. In outperforming the baseline, the other classifiers demonstrated that expert prior knowledge, in the form of human evaluations and acceptable pronunciation variants, can dramatically improve classifier performance, as we had hypothesized.

Of the three best classifiers, the simple threshold classifier with dynamic filler model (*threshold : dynamic*) performed as well or better than the more complex decision trees. However, to set the optimal decision threshold for both of the traditional classification schemes, we explored a range of thresholds and chose the one with the highest agreement with human evaluations. Consequently, we can say that we have an over-optimistic setting for the traditional threshold systems, because we used test set performance information to iteratively perfect the classification of the test set itself. Whereas the decision tree results are based on a leave-one-out crossvalidation procedure which keeps the training and test instances separate and relies on human evaluations only as training set labels. But the high *threshold : dynamic* results suggest that the large number of extraneous features may not be necessary. After pruning, the decision trees were found to branch only on the following three attributes: the binary recognition result, the percentage of the 20 best results which match the target word, and the target word likelihood score.

Using the transcript-based class labels to train the decision tree (*tree : transcripts*) resulted in a slightly better average agreement with the human evaluations. This seems to indicate that the human evaluators were choosing to reject a pronunciation if the variant resulted in a new dictionary word, and would accept only what they perceived to be a surface form variant of the target word that did not become an entirely different word, much like word-level transcribers. We can conclude, then, that to provide class labels for our decision tree's training set, we probably only need one expert evaluator: a transcriber.

6. ACKNOWLEDGMENTS

This work is supported by a grant from the National Science Foundation.

7. REFERENCES

[1] Cambridge University, *HTK 3.2*, http://htk.eng.cam.ac.uk/, 2002.

[2] J. Caminero, C. de la Torre, L. Villarrubia, C. Matin, and L. Hernandez, "On-line garbage modeling with discriminant analysis for utterance verification," in *Proc. of ICSLP*, Philadelphia, 1996.

[3] J.G.A. Dolfing and A. Wendemuth, "Combination of Confidence Measures in Isolated Word Recognition," in *Proc. of ICSLP*, Sydney, 1998.

[4] 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," in *Proc. Eurospeech*, Lisbon, 2005.

[5] P. Ladefoged, *A Course in Phonetics*, Fifth Edition, Thomson Wadsworth, Boston, 2006.

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

[7] P. Ramesh, C.-H. Lee, and B.-H. Juang, "Context Dependent Anti Subword Modeling for Utterance Verification," in *Proc. of ICSLP*, Sydney, 1998.

[8] A. Sethy, N. Mote, S. Narayanan, and L. Johnson, "Modeling and automating detection of errors in Arabic language learner speech," in *Proc. of Eurospeech*, Interspeech, Lisbon, 2005.

[9] D. Willett, Andreas Worm, Christoph Neukirchen, and Gerhard Rigoll, "Confidence Measures for HMM-Based Speech Recognition," in *Proc. of ICSLP*, Sydney, 1998.

[10] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

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