Automatic Assessment of Oral Reading Fluency for Spanish Speaking ELs

Daniel Bolaños¹, Patricia Elhazaz Walsh², Wayne H. Ward³ and Ronald A. Cole¹

¹Boulder Language Technologies (BLT), Boulder, Colorado (USA)  
²Department of English Philology, San Pablo CEU University, Madrid, Spain

dani@bltek.com, pelhazaz@ceu.es, whw@bltek.com, rcole@bltek.com

Abstract

This article presents an approach to the automatic assessment of the oral reading fluency (ORF) of children in Spain who are learning to read English. We compared different acoustic modeling configurations and adaptation methods to determine the most accurate means of estimating reliable children’s oral reading fluency scores using the standard metric of words correct per minute (WCPM). We addressed the problem of identifying word errors by extracting a series of features in order to learn how the human experts are actually annotating individual words.

Experimental results show that the difference between WCPM scores produced by the proposed system and two human judges on the same dataset is smaller than the average difference between the scores produced by the two judges. In addition, the system scored individual words in texts as correctly or incorrectly read with an accuracy similar to that of human annotators.

Index Terms: reading fluency assessment, non-native speech recognition, ELs, children’s speech, Spanish

1. Introduction

Currently in Spain substantial efforts and resources are devoted to teaching the English language in primary and secondary education. In fact, there exists a number of bilingual schools in which both Spanish and English are used to teach academic content.

In the context of these bilingual education programs, students are taught to read fluently in English as the vehicle to understand written academic content and develop mental structures that incorporate the meaning extracted from the text. Assessments of oral reading fluency are an important component of reading instruction and are used in the schools, along with other measures, to evaluate an individual’s reading level and proficiency.

Oral reading fluency (ORF) is frequently used, along with other measures, to assess an individual’s reading level and proficiency. While oral reading fluency does not measure comprehension directly, there is substantial evidence that estimates of ORF predict future reading performance and correlate strongly with comprehension [1, 2]. Because oral reading fluency can be measured quickly (usually less than 5 minutes), and with good validity and reliability, it is widely used to screen individuals for reading problems and to measure reading progress over time.

ORF is defined as the ability to read a grade level text accurately, at a natural speaking rate, and with proper expression [3]. Of the three components of oral reading fluency — accuracy, rate and expressiveness — accuracy and rate have been most studied. In fact, these two components have been combined into a single measure, Words Correct Per Minute (WCPM), which is widely used to assess individuals’ reading ability.

There is substantial research devoted to the automatic assessment of ORF in children’s speech. In the last decade, research conducted by Jack Mostow and his colleagues in Project Listen at Carnegie Mellon University has demonstrated the effectiveness of speech recognition for improving reading fluency and comprehension for both native [4, 5] and non-native speakers of English [6, 7, 8]. In addition, technology has been incorporated into commercial products for assessing children’s reading fluency [9].

In previous work [10, 11] we presented preliminary results showing that it is possible to produce reliable assessments of ORF using automatic speech recognition and machine learning techniques. Experimental results using our assessment system called FLORA (FLuent Oral Reading Assessment) [10] showed that the proposed techniques produced WCPM scores that were within 3 to 4 words of human scorers across students in different grade levels and schools. We also showed [11] that computer-generated ratings of expressive reading agreed with human raters better than the human raters agreed with each other.

In this article, we address the assessment of ORF on grade-leveled texts by children in Spain who are learning English in school. To the best of our knowledge, this represents the first attempt to automatically assess ORF in this setting. Our intention is to develop a cost-effective solution that can be used to generate data that enables the study of the ORF skills of this population of students.

Based on previous work on recognizing non-native speech [12, 13], we have experimented with different acoustic modeling techniques and adaptation methods in order to model the acoustic variability of non-native children’s speech. We have studied the connection between recognition errors and errors in the estimation of ORF scores, and have proposed a method to maximize the reliability of these scores in the presence of a high word error rate. Finally, in order to classify words in the text as correctly or incorrectly read (needed for informing reading instruction), we have proposed and evaluated three sets of word and phone level features extracted from the output of complementary speech recognition systems. Features proposed are intended to help making fine-grained phonetic distinctions by dealing with the lack of phonetic confusability. For each target word a set of competing distractors was automatically generated using knowledge from the domain [14].

The rest of the article is organized as follows: section 2 describes the speech corpus collected for this study, section 3 describes the speech recognition set-up utilized for the experi-
In this section we describe the speech corpora used for the experiments in this study. Speech corpora from two different sources have been utilized for the experiments described in this paper: about 10 hours of read speech from native Spanish children, which constitutes the domain data, and about 106 hours of read speech from three different corpora of native English children [15, 16, 17].

The speech corpus from native Spanish children developed specifically for this study was collected in two primary and secondary schools in Madrid (Spain). Students from 5th, 6th (primary education) and 1st (secondary education) grade participated in the study. The corpus comprises 595 one-minute reading sessions from 191 native Spanish children reading stories in English. About one fourth of the data was collected in the first school (primary education only) while the rest of the data were collected in the second school (primary and secondary education). All but a few of the 191 speakers read three stories. Table 1 summarizes the corpus. Grade-level English text passages used for the study were obtained from [18], which are freely available for non-commercial use. Passages originally designed for first and second graders were selected in order to match the English proficiency level of the native Spanish students who participated in the data collection.

FLORA [10], our web-based tool for assessing oral reading fluency was used to collect the speech material in the classroom environment. During the data collection stage, the student was seated before a laptop computer and wore a set of Sennheiser headphones with an attached noise-cancelling microphone. The experimenter observed or helped the student enroll in the session, which involved entering the student’s gender, age, and grade level. FLORA then presented a text passage, started the one minute recording at the instant the passage was displayed, recorded the student’s speech.

All recorded text passages were transcribed by a trained bilingual speaker; in addition WCPM scores and word-level reading errors were marked by two native English experts. Given that annotating non-native word pronunciations of children is not an easy task, a training session was scheduled before the actual annotation with the expectation that both human experts would develop a satisfactory level of agreement.

For the experiments in this study the corpus was partitioned into three subsets (see table 2), for each subset speakers and recordings were balanced by gender and grade.

Table 1: Summary of the corpus of native Spanish children.

<table>
<thead>
<tr>
<th>grade</th>
<th>age</th>
<th># speakers</th>
<th>#recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>male</td>
<td>female</td>
</tr>
<tr>
<td>fifth</td>
<td>10-11</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>sixth</td>
<td>11-12</td>
<td>48</td>
<td>37</td>
</tr>
<tr>
<td>first</td>
<td>12-13</td>
<td>45</td>
<td>22</td>
</tr>
</tbody>
</table>

3. Speech recognition system

The speech recognizer used in this study is a large vocabulary continuous speech recognition (LVCSR) system developed by Daniel Bolaños, supported jointly by BLT and the University of Colorado (CU). Acoustic modeling is based on Hidden Markov Models and Gaussian Mixture Models. The accuracy of this system has been evaluated on different tasks comprising read and conversational speech and found to be comparable to other state-of-the-art ASR systems.

Speech data were parameterized using Mel-Frequency Cepstral Coefficients (MFCC) and cepstral mean subtraction was applied for noise robustness. Acoustic models were trained under the Maximum Likelihood criterion using Baum-Welch re-estimation.

Trigram language models were trained from the text passages accounting for the possibility of word-repetitions and word-skips. The vocabulary was always limited to words in the stories read. Multiple pronunciations were considered.

Each of the one-minute reading sessions was automatically segmented into the corresponding utterances using our Speech Activity Detection (SAD) system. The SAD system consisted of two 5-state Hidden Markov Models (left-to-right without skip), one to model silence and another to model speech. About 20 mixture components for both speech and silence were tied across the respective five states. A penalty parameter was used to control the detection of silence regions within the audio.

4. Producing automatic assessments of WCPM for non-native speech

As noted above, WCPM, the number of words an individual reads out loud from a leveled text passage in one minute, is a standard measure of oral reading fluency. Norms have been collected from tens of thousands of students across the U.S. in the beginning, middle and end of each school year in grades 1 through 5 using grade-level text passages. Thus, WCPM scores can be used to assign students to an ORF percentile to determine their reading ability, and whether they may be at risk for learning to read [19]. In this section we describe our proposed method to produce WCPM scores for Spanish children reading leveled English text passages.

4.1. Problem description

Ideally, in order to compute an accurate WCPM score for a one-minute reading session we would first generate a hypothesis containing the sequence of words that the kid read using the speech recognizer. Then, we can label words in the text passage as correctly or incorrectly read by aligning it against the hypothesis. The WCPM estimate would just be the total number of words labeled as correct.

Error rates of speech recognition systems are known to increase for children with accents or dialects if their speech patterns are not modeled well by the speech data used to train the system. Errors in the word recognition hypothesis translate into
two types of classification errors:

- False alarm (FA): a word in the reference text is marked as incorrect but was read correctly.
- False positive (FP): a word in the reference text is marked as correct but was not read correctly or not read at all.

An important observation is that the distance between the machine generated score (\(WCPM_M\)) and the human score (\(WCPM_H\)), which is the quantity we need to minimize, can be expressed as a function of those two errors:

$$\text{distance}(WCPM_M, WCPM_H) = |#FA - #FP| \quad (1)$$

Thus, rather than trying to build a speech recognition system that minimizes the word error rate (WER), we are interested in a system that, on average, produces hypotheses with a balanced number of FA and FP.

### 4.2. Experiments and results

Using the two speech corpora described above collected in the U.S. and Spain, three sets of acoustic models were trained. First we trained models on the corpora of native English speakers, these models were based on context dependent triphones with a variable number of Gaussian components per state, the total number of components was around 100k for about 4k HMM-states. Then we obtained another set of acoustic models by doing Maximum A Posteriori (MAP) adaptation on the non-native speech corpus (training partition). The optimal prior knowledge weight for MAP adaptation \(\tau\) was found to be 2, only the Gaussian means were adapted. Finally we trained acoustic models just from the corpus of non-native speech (around 7 hours of speech), these models had only around 8k Gaussian distributions and were context independent since not enough data was available to successfully model context dependency.

During recognition, unsupervised adaptation of the means and variances of each of the Gaussian distributions was carried out using Maximum Likelihood Linear Regression (MLLR) before the second and third recognition passes. The regression tree used to cluster the Gaussian distributions comprised 50 base-classes and the minimum occupation count to compute a transformation was set to 3500 feature frames. K-means clustering was used to cluster the Gaussian means. Table 3 shows for each set of acoustic models the WER and the average difference between WCPM scores produced by the system (denoted as \(M\)) and those from each of the two human scorers (\(H_1\) and \(H_2\)) and the human average (\(H_{avg}\)). Recognition parameters (language model weight and insertion penalty) were optimized separately in order to minimize WER and in order to balance classification errors. In general, the language model weight was set high in order to minimize WER (optimal values were always around 30 to 40). However, smaller values were needed for in-domain and domain-adapted models in order to achieve a good balance between classification errors (acoustic likelihood is significantly higher which favors false-positives). All parameters were optimized in the development set.

It can be seen that a considerable reduction in WER can be attained by using MAP adaptation. However, the best models in terms of WER are those trained solely on domain data. Speaker adaptation only produced a substantial improvement for models trained on native English speech. The best system in terms of WCPM scores was the one using the acoustic models with lowest WER. This system produced WCPM scores for the 90 recordings in the test set that were only 4.85 words different respect to the average human WCPM score, and 5.68 and 4.96 words respect to WCPM scores produced by the first and second human annotators respectively. These numbers are below the inter human agreement, which was 5.92 words for the same recordings (about 115 words per recording on average). This is a very encouraging result that can be partially attributed to the relatively low inter-human agreement. In previous work we showed that trained human annotators can produce WCPM scores for one-minute reading sessions of native English children with an inter-scorer reliability of around 1.5 words. The gap between 1.5 and 5.92 can be explained by the high pronunciation variability of non-native speech that increases the difficulty of the annotation.

Finally it can be observed that reductions in WER directly translate into better WCPM scores. Comparing the worst and the best systems it can be seen that WER is reduced by 34% relative, while the WCPM difference respect to the average human-score is around 35% relative.

### 5. Improving word-level classification

WCPM is a widely used metric for assessing ORF from the perspective of reading rate and accuracy, unfortunately this metric by itself does not provide any information about the degree in which automaticity (reading speed) and accuracy are independently developed. For example, if two children have a very similar WCPM score it indicates that they are reading below grade level, however it is perfectly possible that the source of their fluency and comprehension difficulties is very different. While one child has a very well developed automaticity which allows him to read at a fast pace, he may have significant problems decoding. On the other hand, the other child reads slow because she has not developed yet good automaticity. These two kids have fundamentally different reading difficulties, yet the WCPM score fails to characterize them. For this reason, and in order to address through instruction the source of the fluency difficulties, it is very important to provide the teacher with fine-grained information about a student’s accuracy and automaticity. This information can be very relevant in the context of formative assessment.

In order to diagnose the nature of insufficient reading fluency it is very useful to know whether each word in the text was read correctly. For example, if a student reads many words but makes also many errors it means that they may have problems decoding, on the other hand if a student does not read enough words but makes a small number of errors, insufficient automaticity is most likely the source of the problem. In addition, knowing which words the student cannot decode properly can be very informative for instructional purposes.
For this reason we decided to build into our assessment system a mechanism to annotating individual words in the text passages as correctly or incorrectly read. Recall from the previous section that producing accurate estimates of WCPM can be achieved by balancing out the two kinds of assessment errors (false alarms and false positives), which does not necessarily mean producing the best classification (correct/incorrect) of individual words in the text. Thus, estimating WCPM scores and identifying individual word errors need to be seen as different tasks.

In this section we describe our approach to classifying words in the text passages as correctly or incorrectly read. Our approach consists of generating three complementary sets of features extracted from word and phone based recognition systems trained on the U.S. and Spanish corpora and combining them using a Support Vector Machine (SVM) classifier. We start with the word-classification obtained from our best system in terms of WER (the speaker adapted system trained only on non-native speech, see table 3). The word classification produced by aligning the hypothesis from this system and the reference text presents two desirable properties: it has the best classification accuracy we could get (90.68%) and classification errors are strongly biased towards false-positives, which is the kind of error that our approach for improved classification can address.

5.1. Features proposed

The first set of features is intended to provide word-level discriminative information by de-weighting the language model. The motivation for these features is that, due to the use of a heavily weighted language model (which we found necessary in order to produce the best classification error rate) the recognizer is prone to wrongly hypothesize words in the text. These features are listed below with the prefix ‘w’ and were extracted from the output of a word decoding pass in which only the story text was used to train the language model. Features w1 and w2 are extracted from a word hypothesis using a weaker language model while w3 is extracted from a lattice by computing posterior-probability based confidence measures [20] doing acoustic scaling.

Given that the decoding lexicon contains a very limited number of words (only those in the text passage to read) we do not expect these features to help the classifier to make fine-grain phonetic distinctions that are needed in order to discriminate between correct and incorrect pronunciations of the same word that can be close in the acoustic space. This motivated the definition of the remaining two sets of features.

The second set of features deals with the lack of phonetic confusability exhibited by the word-based system. These features have the prefix ‘p’ and were extracted from the output of a phone decoder using a phone-based language model trained on the whole set of stories (few thousand words). These features are based on comparing the phonetic transcription of the target word with other phone sequences that are generated on a quasi-unconstrained fashion.

The third set of features is intended to provide discriminative information resulting from the comparison of the target phonetic transcription with phonetic transcriptions of potential mispronunciations generated from rules. These features have the prefix ‘c’ and are extracted by running a constrained phone-based recognition pass in which only the phonetic transcription of the correct word plus phone sequences playing the role of distractors are possible. Non-native speakers may pronounce some English phonemes differently than native speakers. Some of these mispronunciations may be due to unsuccessful acoustic phonetic or orthographical knowledge transfer between the mother tongue and the target language. A typical example of unsuccessful knowledge transfer when transitioning from Spanish to English is pronouncing the $v$ and $b$ phonemes like $h$. In this work distractors are generated starting from the phonetic transcription of the target word and considering pronunciation variations that are commonly found in Spanish accented English as described in [14].

The following list summarizes the features extracted for each word in the text passage:

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td>Is word in best-path (slightly weighted trigram)?</td>
</tr>
<tr>
<td>w2</td>
<td>Is word in best-path (unigram)?</td>
</tr>
<tr>
<td>w3</td>
<td>Word confidence estimate (lattice computed)</td>
</tr>
<tr>
<td>p1</td>
<td>Does phone hypothesis match the phonetic transcription of target word?</td>
</tr>
<tr>
<td>p2</td>
<td>Minimum edit distance between phonetic transcription of target word and phone hypothesis from best path.</td>
</tr>
<tr>
<td>p3</td>
<td>Is phonetic transcription of target word in lattice?</td>
</tr>
<tr>
<td>c1</td>
<td>Does phone hypothesis match the phonetic transcription of target word?</td>
</tr>
<tr>
<td>c2</td>
<td>Minimum edit distance between phonetic transcription of target word and phone hypothesis from best path.</td>
</tr>
<tr>
<td>c3</td>
<td>Is phonetic transcription the best path in the lattice?</td>
</tr>
</tbody>
</table>

For each feature two different values were computed, one using the acoustic models trained on non-native speech and one using the acoustic models trained on native English readers. Features were only extracted for words tagged as correct by the baseline classification.

5.2. Classification method

In order to classify the words in the reference text as correct or incorrect using the features described previously, we used the LibSVM implementation of Support Vector Machine classifiers (SVMs) [21]. A linear kernel was used to train the SVM, the optimal value of the $C$ hyperparameter (cost attached to misclassifying a training sample) was estimated for each fold independently, using cross-validation on the training partition. Before training and evaluating the classifiers, all the feature values were scaled to the interval $[-1, 1]$.

5.3. Results

Table 4 shows the baseline classification accuracy and the accuracy resulting from the features proposed. The baseline accuracy (90.68%) comes from the system with the best WER, which was used to generate the initial word hypotheses that serve as input to the classification process. It can be seen that each of the features contributes to improving the baseline classification accuracy except ‘p’ features extracted from the system trained just on native speech, which slightly degrades performance. The combination of all the features produces a 1.59% absolute improvement relative to the baseline system. We note that the two human annotators produced the same word label 92.35% of the time. A naive classification, consisting of labeling all the words as correct, produces 91.46% accuracy. In sum, the experiments show that the proposed features result in a classification accuracy similar to that from human scorers. Accuracy on the proposed features was computed using both human classifications as reference and averaging afterward.
Table 4: Word classification accuracy.

<table>
<thead>
<tr>
<th>feature set</th>
<th>classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>90.68</td>
</tr>
<tr>
<td>w</td>
<td>91.33 91.22</td>
</tr>
<tr>
<td>p</td>
<td>90.37 91.61</td>
</tr>
<tr>
<td>c</td>
<td>91.28 91.67</td>
</tr>
<tr>
<td>w+p+c</td>
<td>91.53 91.92</td>
</tr>
<tr>
<td>all</td>
<td>92.27</td>
</tr>
</tbody>
</table>

6. Conclusions and future work

This research demonstrates that the ORF of children in Spain reading English texts can be estimated at a level of accuracy similar to agreements among human judges. The average disagreement among human judges for this corpus of 191 speakers reading 595 stories was 5.92 words. This compares to an average agreement of between 1 and 2 words for trained judges scoring students in grades 2, 3 and 4 in U.S. schools with approximately 80% native English speakers [10]. Since oral reading fluency is known to be a fast, valid and reliable way to measure students’ reading ability, and correlates well with comprehension and future reading performance, automatic measures may soon prove to provide an fast and cost-effective way to estimate their progress in response to reading instruction.

These preliminary results can be extended and improved in several ways: collecting more data from the domain, doing a more detailed analysis of the features proposed, using discriminatively trained models in order to better discriminate between the target word and distractors, using a language model that better models reading disfluencies, improving inter-annotator agreement, etc.

7. References