Measuring Children's Phonemic Awareness through Blending Tasks

Shizhen Wang[†], *Patti Price*[‡], *Yi-Hui Lee*[†] *and Abeer Alwan*[†]

[†] Department of Electrical Engineering, University of California, Los Angeles
[‡] PPRICE Speech and Language Technology Consulting
szwang@ee.ucla.edu, pjp@pprice.com, yihuilee@ucla.edu and alwan@ee.ucla.edu

Abstract

In this paper, speech recognition techniques are applied to automatically evaluate children's phonemic awareness through three blending tasks (phoneme blending, onset-rhyme blending and syllable blending). The system first applies disfluency detection to filter out disfluent phenomena such as false-starts, sounding out, self-repair and repetitions, and to localize the target answer. Since most of the children studied are Hispanic, accent detection is applied to detect possible Spanish accent. The accent information is then used to update the pronunciation dictionaries and duration models. For valid words, forced alignment is applied to generate sound segmentations and produce the corresponding HMM log likelihood scores. Normalized spectral likelihoods and duration ratio scores are combined to assess the overall quality of the children's productions. Results show that the automatic system correlates well with teachers, and requires no human supervision.

1. Introduction

Increasing attention has been devoted to applying automatic speech recognition (ASR) techniques to children's speech for educational purposes. Many automatic assessment, tutoring, and computer aided language learning (CALL) systems have been developed. The Technology-based assessment of language and literacy (TBall) project [1] was designed to automatically evaluate English language learning and literacy skills of predominantly Mexican-American children in grades K-2 (ages 5-7 years). A critical component of the TBall project is assessment of phonemic awareness because of its key role in reading and writing, especially for the targeted age group. Since human evaluation of phonemic awareness is time-consuming, we have aimed to reduce teaching efforts while maintaining the instructional utility of the assessments by developing an evaluation system to automatically assess children's phonemic awareness using ASR techniques. In our earlier efforts [2], we focused only on the syllable blending task without taking into account disfluency and accent issues. Here, we extend that work to other blending tasks and address disfluency and accent.

Studies of automatic pronunciation assessment have used acoustic parameters and/or prosodic features [3, 4]. Such studies show that spectral likelihood and duration scores correlate well with human evaluations. Automatic evaluation of children's phonemic awareness in the TBall project, however, is more complex because of the children's young ages and their multi-lingual background. As part of the learning process, disfluencies such as repetitions, false starts, and self-repairs, etc. often occur in young children's speech. In addition, accents present another challenge for ASR. Therefore it is important to detect both disfluency and accent in the automatic evaluation system.

Disfluency detection is a common challenge for spontaneous-speech ASR, and many approaches have been proposed. A decision model was applied in [5] using prosodic features. The authors in [6] studied the combination of multiple knowledge sources including acoustic-prosodic features, language models and rule-based knowledge. A disfluency-specialized grammar structure was applied in [7] to detect disfluent reading miscues. An efficient hybrid word/subword unit recognition system was proposed in [8] which works well on children's speech.

Studies on accent detection usually employ prosodic features such as pitch, stress, and durations, etc [9, 10]. Knowledge-based and data-driven approaches have also been proposed to detect accent using phoneme-dependent accent discrimination models [11], Gaussian mixture models [12] or parallel phoneme recognizers followed by phoneme language models [13].

In this effort, we have used a partial-word recognizer for disfluency detection and combined knowledge-based and data-driven approaches for accent detection. Normalized HMM log likelihood is used for pronunciation accuracy measurement and a duration ratio score for smoothness evaluation. The weighted summation of log likelihoods and duration scores is used to assess the overall blending performance. The automatic evaluation system requires no human supervision.

2. Blending and Teacher Evaluations

2.1. Blending tasks for phonemic awareness

Phonemic awareness can be assessed through oral segmenting and blending tasks at various linguistic levels. Here we primarily focus on phoneme blending, onset-rhyme blending and syllable blending. Examples of each blending task are shown in Table 1. The blending tasks assess both pronunciation accuracy and smoothness of the target words. A child who can correctly reproduce all the sounds and smoothly blend them together to make one word is said to be proficient in blending.

Table 1: *An example of the TBall blending tasks: audio prompts are presented and a child is asked to orally blend them into a whole word. A one-second silence (SIL) is used within the prompts to separate each sound.*

Blending task	Audio Prompt	Target
Phonemes	/hh/ SIL /ae/ SIL /ch/	hatch
Onset-rhyme	/r/ SIL /ae m p/ (r+amp)	ramp
Syllables	/p eh p/ SIL /t I k/ (pep+tic)	peptic

The speech corpus was collected in five Kindergarten class-

Supported in part by NSF Grant No. 0326214

rooms in Los Angeles. The schools were carefully chosen to provide balanced data from children whose native language was either English or Mexican Spanish. Each blending task has eight words, most of which are unfamiliar words to young children. By choosing such words, we intend to reduce the likelihood that a child could guess the target answer without focusing on blending the components. Before the recording, children first practiced on examples to become familiar with the task. During data collection, a timer with expiration time of three seconds was used as the maximum pause between the prompt and the answer. If a child didn't respond within 3s after the prompt, the prompt for the next word would be presented. A total of 193 children were recorded, and Table 2 shows the distribution of children by native language and gender.

Table 2: *Speaker distribution by native language and gender.*

Native language	English	Spanish	Unknown
Boy	38	43	11
Girl	41	47	13
Total	79	90	24

2.2. Teachers evaluations

In previous work [2], we found that evaluations based on several words from a speaker are more reliable than those based on single words, since the more speech from a child the rater hears, the more familiar the rater becomes with the system of contrasts used by the child. Therefore audio samples were grouped by speaker to allow teachers to apply speaker-specific information (dialect or accent, speaking-rate, etc.) for judgment adaptation.

Teachers assessed both pronunciation accuracy and smoothness by responding to the following questions:

- Are the sounds correctly pronounced? (accuracy)
- Are the sounds smoothly blended? (smoothness)
- Is the final word acceptable? (overall)

For each question, two choices were presented to classify the quality: acceptable or unacceptable. Teachers also provided comments for their decisions.

Assessments from nine teachers were collected to calculate the inter-correlation between evaluators. As shown in Table 3, teacher evaluations are reasonably consistent for the three tasks. The inter-correlations in evaluating the overall quality are similar for all the tasks: about 85%. The inter-correlations on accuracy evaluations are significantly higher than those on smoothness. This is because, compared to pronunciation accuracy, smoothness evaluation is more subjective especially toward short utterances. However, smoothness may be more important than accuracy in the blending task because that is the goal of a blending assessment. In any case, it is an orthogonal judgment because words can be smooth and accurate, not smooth and accurate, smooth and inaccurate or not smooth and inaccurate. Of the three tasks, phoneme blending is the most difficult for children and draws much disagreement among teachers; while syllable blending is relatively easy.

3. Automatic Evaluation System

Our automatic evaluation system to measure children's performance on the blending tasks consists of four core components: disfluency detection, accent detection, accuracy assessment and smoothness assessment. The system flowchart is shown in Fig. 1, with detailed descriptions in subsequent sections. Disfluency detection uses the partial-word recognizer to filter out disfluent phenomena such as false-starts, sounding out, self-repair and repetitions, and to localize the target answer. Accent detection is then applied to the target word to detect possible non-native English pronunciations. The accent information is then used to update the pronunciation dictionaries and duration ratio models. Similar to the techniques we proposed in [2], normalized log likelihood and duration ratio scores are used to measure accuracy and smoothness, respectively. These two scores are combined together to get the final result. Since the tasks are designed to evaluate a child's language learning skills based on responses to audio prompts, prior information of the expected answer is available for use in ASR. Hence the automatic system can work in a supervised mode and exploit knowledge-based information derived by linguistic experts for better and more reliable performance.

Table 3: *Average inter-evaluator correlation on pronunciation accuracy, smoothness and overall evaluations on three blending tasks.*

Blending task	Accuracy	Smoothness	Overall
Phonemes	87.6	80.8	83.3
Onset-rhyme	91.3	82.4	84.1
Syllables	97.5	85.3	86.7

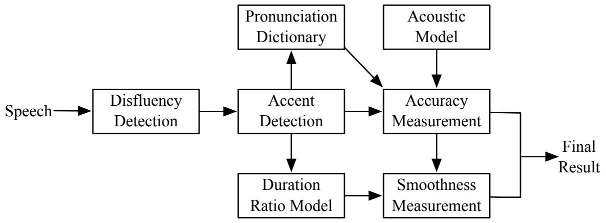

Figure 1: *Flowchart of the automatic evaluation system for the blending tasks.*

3.1. Disfluency detection

Generally speaking, disfluencies include everything spoken by the child that disrupts the natural flow of the target word pronunciation. Typical disfluencies found in our data are: fillers such as *uhhh* or *ummm*, partial- and/or full-word repetitions where syllables or phonemes within a word or a whole word are repeated, self-corrections, long pauses within a word, elongations where syllables or phonemes (usually the first one) are lengthened. These last two disfluencies (pauses and elongations) are related to the smoothness measure, and will be addressed using duration ratio models.

The first stage, disfluency detection, is used mainly to filter out fillers and repetitions, and to get the approximate beginning and ending times for the target answer. If the target word is repeated several times, only the last one is used for further evaluations in order to be consistent with teachers' decision-making protocols, where only the last answer is accounted for.

A partial-word recognizer (PWR) [8] is used to detect disfluency with sub-word units derived from the dictionary based on the task; sub-word units are phonemes, onsets or rhymes, or syllables depending on the blending task. An example of the detection network is shown in Fig. 2 for a syllable blending word *peptic*. A background/garbage model is used to consume background noises, fillers and out-of-vocabulary words. Long pauses are allowed between sub-word units. The PWR can be bypassed to whole word recognition (WWR) for disfluency-free speech. The WWR is a regular phoneme-based recognizer except that it allows repetitions. WWR can also be bypassed for the case where the child does not make an attempt to say the target whole word.

For computational efficiency, only one canonical dictionary pronunciation is used to generate sub-word units, and no accented alternatives are taken into account at this state. This is reasonable because here the disfluency detector is mainly used to localize the target answer of interest (not score it). Evaluated on a subset of the blending tasks data, the disfluency detector is able to filter out around 85% of the disfluent miscues. The subsequent process detects accent and uses that information to choose the pronunciation dictionary and duration models.

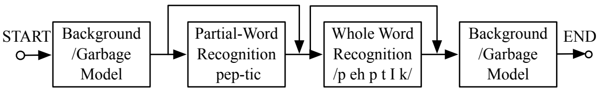
START
Background /Garbage Model
Partial-Word Recognition pep-tic
Whole Word Recognition /p eh p t I k/
Background /Garbage Model
END

Figure 2: *An example of the disfluency detection network for a syllable blending task word 'peptic', where START and END are the network enter and exit points, respectively.*

3.2. Accent Detection

The TBall data used here were collected from children with multi-lingual backgrounds, and thus contain foreign accented (mainly Spanish accented) English. An example of pronunciation variation for Spanish-accented English is the replacement of /dh/ (**there**) with /d/ (**dare**), since /dh/ does not exist in Spanish. Based on the analysis in [14], an algorithm is developed to automatically detect Spanish accent.

Given the pronunciation variation patterns, a simple but effective measure for accent detection is the occurrence ratio of such patterns in an utterance, defined as:

$$R_{ph1|ph2} = \frac{C(ph2 \rightarrow ph1)}{C(ph2)} \qquad (1)$$

where $R_{ph1|ph2}$ is the occurrence ratio of pronunciation change pattern from phoneme2 to phoneme1, which is denoted by $\{ph2 \rightarrow ph1\}$; $C(ph2)$ is the occurrence count (OC) of $ph2$, and $C(ph2 \rightarrow ph1)$ is the OC of pattern $\{ph2 \rightarrow ph1\}$. Since the system is running in a supervised mode with available transcriptions, the OCs can be easily calculated through forced alignment using a canonical pronunciation dictionary first and then an accented pronunciation dictionary. The two alignment outputs are analyzed and compared to calculate the OC of each pattern.

The average value of all occurring pattern ratios is a measure of the overall accent of a speaker, i.e.,

$$R = \frac{1}{M} \sum_{\{ph2 \rightarrow ph1\} \in \mathcal{P}} R_{ph1|ph2} \qquad (2)$$

where $\mathcal{P}$ represents all valid pronunciation change patterns, and M is the total number of patterns occurring in the utterance. To make reliable estimates, patterns with OCs of $C(ph2)$ below a threshold of 3, are not included in the calculation.

The speaker level accent measure in Eq. 2 treats all pronunciation change patterns equally. The statistical analysis of our data, however, shows patterns do not occur at the same probability, and some patterns occur much more frequently than others, e.g., the occurrence of pattern {/z/→/s/} has a probability of 73.6%, while the pattern{/v/→/f/} occurs only 21.5% of time. The occurrence probabilities can be viewed as the correlation between each pattern and the overall accent. The higher the probability, the more related the pattern is to accent. To take this into account, Eq. 2 is changed into the following equation:

$$R = \sum_{\{ph2 \rightarrow ph1\} \in \mathcal{P}} p(ph2 \rightarrow ph1) \cdot R_{ph1|ph2} \qquad (3)$$

where $p(ph2 \rightarrow ph1)$ is the probability of pattern $\{ph2 \rightarrow ph1\}$, which is normalized to make the summation of all pattern probabilities equal to 1.

The accent score from Eq. 3 is used to classify a speaker's accent level. The higher the score, the more accented the utterance is. Since our database does not have accent level information, a binary detection is performed to decide if a speaker is Spanish-accented or not. Given a threshold T_a (0.6 in our experiment), if the score R is greater than T_a, then the speaker has Spanish accent. The accent detector achieved 83% correctness on an evaluation dataset which was labeled for accent.

3.3. Pronunciation Dictionary

The dictionary used in accuracy assessment needs to consider possible pronunciation variations. Besides the canonical pronunciation for each word, the dictionary also contains entries for non-canonical but correct (and common in kids) pronunciations from different dialects common in the Los Angeles area. For example, many speakers do not distinguish **cot** and **caught**, pronouncing both as /k aa t/. Therefore, /k aa t/ and /k ao t/ are both considered as correct pronunciations. The dictionary also includes iy/ih alternations since Spanish learners of English often do not separate them well. Hispanic letter to sound (LTS) rules are not applied in the dictionary, since LTS rules are for reading evaluations while in our task the prompts are auditory. Although it is possible that these rules may have some effect (since they hear speech of adults who are literate and influenced by Hispanic LTS rules when speaking English), such instances appeared to be rare relative to the increase in size of the dictionary that would be needed to cover them comprehensively.

3.4. Accuracy and Smoothness Measurements

Techniques similar to those reported in our previous work [2] are applied for accuracy and smoothness measurements. Normalized HMM log likelihoods through forced alignments are calculated to evaluate the pronunciation qualities. Accent information from the accent detection component is used to choose appropriate entries from the pronunciation dictionary. Local normalization is applied to compensate for utterance length (time duration):

$$S_l = \frac{1}{N} \sum_{i=1}^{N} \frac{s_i}{d_i} \qquad (4)$$

where s_i is the log likelihood of the ith segment (phoneme, syllable or the pause between), d_i is the corresponding time duration in frames, and the summation is over all N segments. The pronunciation is acceptable if the log likelihood score $S_l > T_l$, where the threshold T_l can be speaker-independent empirical values or speaker-specific values to take individual speaker's acoustic characteristics into consideration.

Segment durations are used to measure the blending smoothness. The durations are obtained from forced alignments with the most likely pronunciations. To compensate for the effects of rate of speech, the durations are normalized as:

$$\bar{d}_i = d_i / \sum_{j=1}^{N} d_j \tag{5}$$

Gaussian mixture models (GMM) are used to approximate the distribution of syllable duration ratios for each task word. Two GMM models are constructed from the training data, one for native English and the other for Spanish-accented English. Information from the accent detection component is used to select the appropriate model. The log likelihood of given duration ratios against the GMM is used as smoothness scores S_d:

$$S_d = \sum_i \bar{d}_i \cdot \log \sum_m \mathcal{N}(\bar{d}_i;\ \mu_{im}, \sigma_{im}) \tag{6}$$

where $\mathcal{N}(\cdot;\ \mu, \sigma)$ is a Gaussian with mean μ and variance σ. If S_d is greater than the smoothness threshold T_d, the blending smoothness is acceptable.

3.5. Overall quality measurement

The overall quality is unacceptable if either pronunciation or smoothness is unacceptable. If the pronunciation and smoothness are both acceptable, the overall quality is evaluated based on the weighted summation of pronunciation scores and smoothness scores:

$$S = w \cdot S_l + (1 - w) \cdot S_d \tag{7}$$

A threshold T is used to decide the acceptability of the overall quality. Similar to pronunciation evaluation, T can be speaker-independent or speaker-specific.

4. Results

To test system performance, evaluations from teachers were used as references. Acoustic monophone models were trained on the TBall database (excluding the blending tasks) with approximately seven hours annotated recordings from both native and nonnative speaker. For each blending task, performance was tested on 1350 utterances. Speaker independent decision thresholds were used in all experiments. Table 4 shows the correlation between automatic and average teacher evaluations for the three blending tasks.

For pronunciation quality evaluation, normalized likelihoods correlate well with teacher assessments. For the smoothness measurement, duration ratio scores achieved comparable performance to the average inter-correlation between teachers. The overall evaluation using a weighted summation of pronunciation and smoothness scores obtained an average correlation around 88% over the three tasks, slightly better than the average inter-teacher correlation. The weight of the optimal performance is 0.35, which means that smoothness is more important than pronunciation in the blending task. Note that on the syllable blending task, overall performance is improved from 87.5% (in [2]) to 91.8% due to disfluency and accent detection.

Table 4: *Average correlation between ASR and teacher evaluations on pronunciation accuracy, smoothness and overall qualities for three blending tasks.*

Blending task	Accuracy	Smoothness	Overall
Phonemes	90.5	79.8	85.6
Onset-rhythm	93.2	83.1	87.9
Syllables	95.4	90.7	91.8

5. Summary and Discussion

An automatic evaluation system is developed to assess children's performance on three blending task. The system applies disfluency detection and accent detection for pre-processing and uses a pronunciation dictionary for forced alignment to generate sound segmentations and produce HMM likelihood scores. The weighted summation of normalized likelihoods and duration scores is used to evaluate the overall quality of children's responses. Speaker specific accent information is used to update the dictionary and duration ratio models. Compared to teachers' assessments, the system achieves a correlation better than the average inter-teacher correlation. Future work will aim to improve performance using additional features and speaker specific modeling.

6. References

[1] A. Alwan, et al., "A System for Technology Based Assessment of Language and Literacy in Young Children: the Role of Multiple Information Source," in *Proc. IEEE MMSP*, Greece, October 2007.

[2] S. Wang, et al., "Automatic evaluation of children's performance on an English syllable blending task," SLaTE Workshop 2007.

[3] R. Delmonte, "SLIM prosodic automatic tools for self learning instruction," *Speech Communication*, vol. 30, pp. 145-166, 2000.

[4] F. Tamburini, "Prosodic Prominence Detection in Speech," in *Proc. ICASSP*, pp. 385-388, 2003.

[5] E. Shriberg, R. Bates and A. Stolcke, "A Prosody-Only Decision-Tree Model for Disfluency Detection," in *Proc. Eurospeech*, pp. 2383-2386, 1997.

[6] Y. Liu, E. Shriberg and A. Stolcke, "Automatic Disfluency Identification in Conversational Speech Using Multiple Knowledge Sources," in *Proc. Eurospeech*, pp. 957-960, 2003.

[7] M. Black, et al., "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," in *Proc. Interspeech*, 2007.

[8] A. Hagen, B. Pellom and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, pp. 861-873, 2007.

[9] V. Storm, "Detection Of Accents, Phrase Boundaries And Sentence Modality In German With Prosodic Features," in *Proc. Eurospeech*, pp. 2039-2041, 1995.

[10] N. Minematsu, N. Ohashi and S. Nakagawa, "Automatic Detection of Accent in English Words Spoken by Japanese Students," in *Proc. Eurospeech*, pp. 701-704, 1997.

[11] K. Kumpf and R. King, "Foreign Speaker Accent Classification using Phoneme-Dependent Accent Discrimination Models and Comparison with Human Perception Benchmarks," in *Proc. Eurospeech*, pp. 2323-2326, 1997.

[12] T. Schultz, et al., "Speaker, accent and language identification using multilingual phone strings,". in *HLT-2002.*

[13] T. Chen, et al., "Automatic accent identification using Gaussian mixture models," in *ASRU Workshop*, 2001.

[14] H. You, et al., "Pronunciation Variation of Spanish-accented English Spoken by Young Children," in *Proc. Eurospeech*, pp. 273-276, 2005.