AUTOMATIC EVALUATION OF CHILDREN'S PERFORMANCE ON ENGLISH SYLLABLE BLENDING TASK

Shizhen Wang[1]*, Abeer Alwan*[1] *and Patti Price*[2]

[1]Department of Electrical Engineering, University of California, Los Angeles

[2]PPRICE Speech and Language Technology Consulting

szwang@ee.ucla.edu, alwan@ee.ucla.edu and pjp@pprice.com

ABSTRACT

In this paper, speech recognition techniques are applied to automatically evaluate children's performance in an English syllable blending task. Word verification is performed to filter out utterances pronounced incorrectly. For valid words, forced alignment is applied to generate syllable segmentations and produce the corresponding HMM log likelihood scores. Normalized spectral likelihood and duration ratio scores are combined to assess the overall quality of children's productions. Speaker-specific information is further incorporated to optimize performance. Experimental results show that the automatic system results correlate well with those of a teacher, but requires no human supervision. The optimal system exhibits a 87.5% correlation compared with teachers' assessments, slightly better than the average inter-teacher agreement of 86.7%.

Index Terms— syllable blending, pronunciation evaluation, smoothness evaluation, speech recognition

1. INTRODUCTION

Phonemic awareness, which is related to developing reading and writing skills, is an important ability that children need to acquire to become proficient readers [1]. One assessment of children's phonemic awareness is the syllable blending task, which tests children's ability to orally blend syllables into a whole word, such as ***ta*** + ***ble*** =table. Human evaluation of children's

This work was supported in part by NSF Grant No. 0326214 and by a fellowship from the Radcliffe Institute for advanced study to Abeer Alwan.

syllable blending performance is time-consuming and subjective. To reduce teachers' efforts while maintaining the instructional utility of the assessments, we are developing an automatic evaluation system for assessing children's blending skills.

In recent years, a considerable number of studies have been devoted to automatic pronunciation assessment using acoustic parameters and/or prosodic features [2–4]. Such research efforts show that spectral likelihood and duration scores correlate well with human evaluations. The automatic evaluation of children's performance on the syllable blending task, however, is more difficult than pronunciation assessment in that the assessment of syllable blending performance needs to address both the pronunciation quality and the blending smoothness. In addition, children's speech demonstrates larger inter- and intra-subject acoustic variability.

In this paper, we use normalized HMM log likelihoods for pronunciation scoring and a duration ratio score for smoothness evaluation. The weighted summation of log likelihood and the duration score is used to assess the overall blending performance. Pronunciation variations are addressed with a dictionary containing possible and acceptable pronunciations (including dialect variations) for each task word. The automatic evaluation system, once constructed, requires no human supervision to use: it employs word verification to determine if the child's utterance is the target word and, for valid target words, to generate syllable segmentations and produce log likelihood scores, the basis used to evaluate the child's blending skills.

The remainder of this paper is arranged as follows: in Section 2, we briefly review the syllable blending task and analyze the inter-correlation of teachers' evaluations; in Section 3, we present the automatic evaluation algorithm including the pronunciation variation dictionary, global and local pronunciation scores, and blending smoothness scores. Performance optimization using speaker-specific information is also discussed; experimental results are demonstrated in Section 4; and Section 5 includes a summary and conclusions.

2. SYLLABLE BLENDING TASK AND TEACHERS' EVALUATION

2.1. Syllable Blending Task

The syllable blending task for children learning to read English is designed to assess both pronunciation accuracy and blending skills (smoothness). In the task, audio prompts present the syllables of a two-syllable word separately, and a child is asked to orally blend them into a word. A child is said to be proficient in this task provided:

- The child reproduces all the sounds of the original syllables in the final word.

- The child can smoothly blend the two syllables together to make one word.

The database was collected in Kindergarten classrooms from five elementary schools in Los Angeles. The schools were carefully chosen to provide balanced data from children whose native language was either English or Spanish [5]. 173 children were asked to orally blend eight two-syllable words: ***bamboo***, ***napkin***, ***nova***, ***peptic***, ***stable***, ***table***, ***wafer*** and ***window***.

During data collection, a timer with expiration time of 3 seconds was used to for maximum pause between the prompt and the answer. That is, if a child didn't respond within 3s after the prompt, the child would be classified as unable to answer and the prompt for the next word would be presented. Since the pause between syllables (or the lack thereof) is critical to blending skills, we focus in this paper on the time duration of the inter-syllable pause.

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

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

For each question, two choices were presented to classify the quality: acceptable or unacceptable. Teachers also provided comments for their decisions. Audio samples from children were grouped in two ways: word by word (samples of the same words were put in the same webpage, Eval-I) or child by child (samples of the same child were put in the same webpage, Eval-II).

2.2. Inter-correlation of Teachers' Evaluation

Since teachers' assessments are used as the reference to test the automatic evaluation system, we need to measure the consistency or the inter-correlation between teachers. Nine teachers' assessments are used to calculate the inter-correlation at the word and speaker level. Word-level inter-correlation is calculated based on the evaluation results from Eval-I, and speaker-level inter-correlation is calculated based on the results from Eval-II.

Teachers' evaluations are reasonably consistent at both word level and speaker level. The average inter-correlations between teachers about the overall quality are 81.6% and 86.7% at word and speaker level, respectively. The higher correlation at speaker level shows that evaluations based on several words from a speaker (and thus with more speaker specific information) are more reliable than based on single word. This is because the more speech from a child the rater hears, the more familiar the rater will

be with the system of contrasts used by the child. For example, hearing a child say ***cow*** for ***car*** may indicate an articulation issue and not a reading issue. For the speaker-level evaluations, all samples from the same child can be taken as references to the child's dialect or accent, speaking-rate, etc.

Detailed analysis of the assessments of pronunciations and blending smoothness reveals that the average inter-correlation in evaluating pronunciation, about 97.5%, is much higher than that in evaluating blending smoothness, about 85.3%. This makes sense because compared to pronunciation accuracy, smoothness evaluation is more subjective especially in short utterances. It may be that smoothness is more important than accuracy in the syllable blending task because that is the goal of a blending assessment. In any case, it is an orthogonal judgement because words can be smooth and accurate, not smooth and accurate, smooth and inaccurate or not smooth and inaccurate.

2.3. Choice of Words in Syllable Blending Task

From teachers' comments, we also find that children's background knowledge of the syllables they're blending greatly affects performance. For syllables of unfamiliar words, it usually takes longer for a child to give the answer. For example, many children are completely unfamiliar with word ***peptic*** and with the unusual occurrence of /p/ and /t/ sounds together. In this case, there will typically be long pauses between the end of prompt and a child's answer, and also between the two syllables to be blended.

Another issue is for syllables of confusable words: children tend to pronounce them incorrectly but blend them smoothly, and thus show "strong blending" skills. For the word ***stable*** many children pronounced it as ***staple*** because the two words are very confusable especially when spoken in isolation without any context. The confusion is particularly strong for Hispanic children learning English, since Spanish /p/ is acoustically very similar to English /b/.

There are also some 'language-driven' errors. That is, substitution or deletion/insertion errors can occur when the syllables to be blended do not exist in the child's native language. For example, children from Spanish linguistic backgrounds pronounced the word ***stable*** as ***estable*** or ***estaple*** because no words begin with the sound ***sp*** in Spanish and they always have a vowel preceding the consonant cluster, such as ***Espana*** or ***esperanza***.

In our evaluations, however, to be consistent with the goals of this syllable blending task, the final decision is based on both the pronunciation correctness and the blending smoothness, i.e., a word can be acceptable only when the pronunciation accuracy and the blending smoothness are both acceptable.

3. AUTOMATIC EVALUATION ALGORITHM

3.1. Pronunciation Variation Dictionary

We developed a speech recognition system to automatically evaluate children's performance in a syllable blending task. In the system, we use log likelihood scores to evaluate the pronunciation quality, time duration scores to assess the blending smoothness and the weighted summation to judge the overall acceptability.

Since the task is designed to evaluate a child's language learning skills based on his/her responses to the audio prompts, prior information of what is expected from the child can be used in speech recognition, making the recognition actually a verification task. That is, we know what the child is supposed to say in this case. Word verification is used to verify the target words. For those words that pass the verification filtering, forced alignment is then applied to generate the syllable segmentations and produce the corresponding log likelihood for each segment and to determine the inter-syllable pause, if any.

The dictionary used in word verification needs to consider possible pronunciation variations. Besides canonical pronunciation for each word, the dictionary also contained entries for non-canonical but correct (and common in kids) pronunciations from different dialects that are 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 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 audio sounds. 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.

The pronunciations in the dictionary have tags for these various pronunciations (Hispanic accented pronunciation, canonical pronunciation, phonological development issue, etc.) In this way, "accent" or "dialect" or "idiolect" can be attributed in a simple way: the likelihood for each pronunciation is calculated and the pronunciation with the highest likelihood, if non-canonical, is declared as the "idiolect" for the speaker for that word. A pattern of many words through the Hispanic accented path would confirm a speaker as having Hispanic accented speech. A constraint for detecting dialect is that the speaker must produce a consistent dialect, that is, the dialect, if detectable, must be the same in most of the task words. In this way, we can model the dialect as a system of distinctions, which is linguistically much more appropriate than simply adding more pronunciations and probabilities.

3.2. Pronunciation Quality Evaluation

The HMM log likelihood of a given word, which measures the similarity between the testing speech and the training native speech, is used to evaluate the pronunciation qualities. In the HMM framework using the Viterbi algorithm, the log likelihood highly depends on the length (time duration) of the test utterance. To compensate for the effects of duration, two normalization methods are applied [6]. One is global normalization, defined as

$$S_g = \left(\sum_{i=1}^{N} s_i\right) \Big/ \left(\sum_{i=1}^{N} d_i\right)$$

where s_i is the log likelihood of the ith segment (syllable or inter-syllable pause), d_i is the corresponding time duration in frames, and the summation is over all the N segments. It is straightforward to show that the above defined global normalization biases long duration segments with heavier weights. To treat all segments equally, which is more desirable in this syllable blending task , local normalization is defined

$$S_l = \frac{1}{N} \sum_{i=1}^{N} \frac{s_i}{d_i}$$

The pronunciation is declared as acceptable if either global or local likelihood scores satisfy:

$$S_g > t_g \quad \text{or} \quad S_l > t_l$$

where the thresholds t_g and t_l can be speaker-independent empirical values or speaker-specific values to take into consideration of individual speaker's acoustic characteristics.

3.3. Blending Smoothness Measurement

Syllable durations of were used in the measure of blending smoothness. Due to co-articulation, exact phonetic boundaries are much more dynamic than stable in continuous speech, which makes the syllable duration unreliable. For this reason, some studies proposed the idea of substituting whole syllable duration with syllable nucleus (the essential vowel) duration [4, 7]. It is also the vowel that varies much more than the consonant with rate of speech. This idea, however, is not suitable for our task. In syllable blending, there are clear inter-syllable pauses in most of children's responses, so it is feasible to obtain reliable boundaries on either side of the pause for each individual syllable.

Syllable durations are obtained from forced alignments with the most likely pronunciations. Since rate of speech (ROS) is typically applied in the context of continuous sentences, and it is not well defined for the case of isolated words, we normalize

syllable durations by setting the whole word length to unit one, that is the duration ratio is defined as

$$\bar{d}_i = d_i / \sum_{i=1}^{N} d_i$$

It is not surprising that blended words with long inter-syllable pauses are unacceptable since the target syllables sound have a perceptible pause between them instead of being smoothly blended together. On the other hand, however, concatenating syllables tightly with no or short inter-syllable pauses doesn't necessarily make the blended word acceptable. Teachers' comments show that prosodic awareness or position of stress plays a critical role in the acceptability of the blended word. This is because English tends toward stress-timing: the stressed syllables tend to occur at more or less even intervals with the unstressed syllables being shortened and reduced. Equal stress (equivalent to no stress) or incorrect stress position makes a word sound strange and unacceptable. According to teachers' evaluations, about half of the unacceptable words were because of stress issues. Since syllablic stress is typically achieved by increased duration and intensity, syllable duration is considered a fundamental parameter for stress.

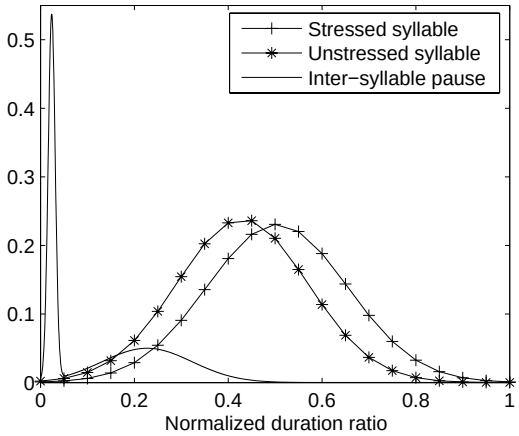

Fig. 1. Histograms of duration ratios for stressed (win), unstressed (dow) syllables and inter-syllable pause of word 'window'

Fig. 1 shows the histograms of duration ratios for stressed, unstressed syllables and inter-syllable pause obtained from a native speaker saying the two-syllable word ***window***. It is clear that stressed syllables on average have longer duration than unstressed ones. The histogram of inter-syllable pauses confirms this observation. That is, the duration of the pause, though an important contributor, is not sufficient for the determination of the word's overall acceptability in the blending task.

Gaussian mixture models (GMM) are used to approximate the distribution of syllable duration ratios for each task word. The

log likelihood of given duration ratios against the GMM is used as smoothness scores S_d. If S_d is greater than the smoothness threshold t_d, the blending smoothness is acceptable.

3.4. Overall Quality Evaluation

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. That is, taking the local pronunciation score S_l for example,

$$S = w * S_l + (1 - w) * S_d$$

where S is the overall quality score. Similar to pronunciation evaluation, the threshold T (to decide the acceptability of the overall quality) can be speaker-independent or speaker-specific.

3.5. Performance Optimization with Speaker-specific Information

In the syllable blending task, some background information about a child may be available to optimize the automatic evaluation performance for a specific speaker. With one or two enrollment utterances from the child, rapid speaker adaptation can be applied to the HMM models and thus produce more reliable likelihood scores and syllable durations after forced alignments with the adapted speaker-dependent models. The accent of a nonnative speaker child, if detected from the enrollment utterances, can be quantized and used as a bias to adjust the thresholds. Finally, rate of speech can be estimated from the enrollment utterances to "normalize" each syllable's duration d_i to

$$\tilde{d}_i = d_i * ROS$$

This normalization will have no effect on the smoothness scores since we use duration ratios instead of absolute durations in the calculation of these scores. Normalization will, however, adjust the pronunciation scores (both globally and locally), introducing a speaker-dependent factor to incorporate the specific speaking rate. In this paper, we use ROS normalization and speaker-dependent thresholds to optimize the evaluation performance for each individual speaker.

4. EXPERIMENTAL RESULTS

To evaluate the performance of our automatic evaluation algorithm, the decisions from nine teachers' evaluations are used as the reference. Table 1 shows the correlation between automatic and teachers' evaluation at both the word and speaker levels. For

	Scores	Correlation	
		word-level	speaker-level
Pronunciation	S_g	88.3	92.6
	S_l	91.7	94.5
	S_g, S_l	92.1	94.7
Smoothness	S_d	83.6	84.8
Overall	S_g, S_d	78.5	85.9
	S_l, S_d	80.3	87.5

Table 1. Correlation between automatic and teachers' evaluations at word and speaker level (percent)

word-level evaluation, no speaker information is assumed and the thresholds are all speaker-independent, while speaker-level evaluation applies both speaker-specific thresholds and ROS normalization.

In all evaluations, the speaker-level correlation is better than the word-level correlation. This is reasonable because at the speaker level speaker-specific information is exploited to optimize the performance for each speaker, while at the word level this is not possible. It can also be noted that speaker-specific information has more influence on pronunciation evaluations than on smoothness evaluations.

For pronunciation quality evaluation, both global and local HMM log likelihood correlate well with teachers' assessments, indicating that acoustic similarity between a test utterance and the training native speech is a good measure of pronunciation acceptability. The correlation using local score S_l, reaching 94.5%, is better than global score S_g. It doesn't improve much when combining global scores and local scores. So equally weighting all syllables in a word seems to be a good strategy.

For the blending smoothness quality, duration scores achieved comparable performance to the average inter-correlation between teachers, especially at the speaker level. The overall quality evaluation using a weighted summation of pronunciation and smoothness scores obtained a correlation of 87.5%, slightly better than the average inter-teacher correlation (86.7%). The weight of the optimal performance is $w = 0.15$, which means that smoothness is more important in blending syllables into a whole word.

5. SUMMARY AND DISCUSSION

In this paper we proposed an automatic evaluation system to assess children's performance on a syllable blending task. The system makes use of a pronunciation variation dictionary for word verification and forced alignment to generate syllable segmentations and produce HMM likelihood scores. The weighted summation of normalized likelihood and duration scores is used to evaluate the overall quality of children's responses. Speaker specific information such as dialect and rate of speech can be used to optimize performance. Compared to teachers' assessments, the optimal system performs achieves a correlation slightly better than the average inter-teacher correlation.

As to the choice of words in designing the syllable blending task, it would be helpful to exclude confusable and unfamiliar words since we are more interested in children's blending ability than with their familiarity of the words. For nonnative speakers, further work is needed to investigate the pronunciation issues imposed by cross-language differences for sounds not in their native languages.

6. REFERENCES

[1] R. Sensenbaugh, "Phonemic awareness: An important early step in learning to read", KidSource Online. Available: http://www.kidsource.com/

[2] H. Franco, L. Neumeyer, Y. Kim and O. Ronen, "Automatic Pronunciation Scoring for Language Instruction", in *Proc. ICASSP*, pp. 1471-1474, 1997

[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] 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*, pp. 1581-1584, 2005

[6] L. Neumeyer, H. Franco, M. Weintraub and P. Price, "Automatic Text-independent Pronunciation Scoring of Foreign Language Student Speech", in *Proc. ICSLP*, pp. 1457-1460, 1996

[7] A.W. Howitt, "Automatic Syllable Detection for Vowel Landmarks", PhD Thesis, MIT, 2000