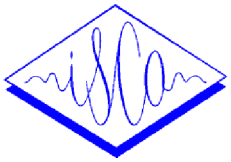

Speech and Language Technology
in Education (SLaTE2007)
Farmington, PA, USA
October 1-3, 2007

Automatic Evaluation of Children's Performance on an English Syllable Blending Task

Shizhen Wang[1], Patti Price[2], Margaret Heritage[3] and Abeer Alwan[1]

[1]Department of Electrical Engineering, University of California, Los Angeles
[2]PPRICE Speech and Language Technology Consulting
[3]National Center for Research on Evaluation, Standards, and Student Testing, UCLA
szwang@ee.ucla.edu, pjp@pprice.com, mheritag@ucla.edu and alwan@ee.ucla.edu

Abstract

In this paper, speech recognition techniques are applied to automatically evaluate children's performance in a 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 likelihoods 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 correlates well with those of teachers, but requires no human supervision.

Index Terms: speech recognition, syllable blending

1. Introduction

Phonemic awareness, related to developing reading and writing skills, is important for children to master to become proficient readers [1]. One assessment of phonemic awareness is syllable blending which tests children's ability to orally blend syllables into a whole word, such as ***ta*** + ***ble*** = table. Human evaluation of children's syllable blending performance is time-consuming. To reduce teachers' efforts while maintaining the instructional utility of the assessments, we are developing an automatic evaluation system for assessing syllable-blending skills.

Recently, a considerable number of studies has been devoted to automatic pronunciation assessment using acoustic parameters and/or prosodic features [2-4]. Such studies show that spectral likelihood and duration scores correlate well with human evaluations. Automatic evaluation of children's performance on blending tasks, 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 than adults.

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 acceptable pronunciations (including dialectal variations) for each task word. The automatic evaluation system, once constructed, requires no human supervision: it employs word verification to determine if the child's utterance is a 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.

2. The Blending Task and Teachers' Evaluations

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 five Kindergarten classrooms in Los Angeles. The schools were carefully chosen to provide balanced data from children whose native language was either English or Mexican Spanish [5]. 173 children were asked to orally blend eight two-syllable words: ***bamboo***, ***napkin***, ***nova***, ***peptic***, ***stable***, ***table***, ***wafer*** and ***window***. The reason for the choice of those words that might be unfamiliar to young children is we wanted them to focus on listening to the phonemes without bringing in any knowledge of existing words to help them. Among the 173 children, 11 are five or seven years old, and all the rest are six years old. The distribution of children by native language and gender is shown in Table 1.

Before the recording, children first practiced with some examples to be familiar with the task. During data collection, there is an about one second silence between the first and the second syllable in the prompts, and 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. Since the pause between syllables (or the lack thereof) is critical to blending skills, we focus on inter-syllable pause durations.

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.

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 (Eval-I) or child by child (Eval-II).

Native language	English	Spanish	Unknown
Boy	29	32	23
Girl	30	38	21
Total	59	70	44

Table 1. *Distribution of children by native language and gender*

2.2. Inter-correlation of Teachers' Evaluation

Nine teachers' assessments are used to calculate the inter-correlation at the word and speaker level. Word-level inter-correlation is calculated on the results from Eval-I, and speaker-level inter-correlation is calculated based on Eval-II results.

Teachers' evaluations are reasonably consistent at both levels. Average inter-correlations between teachers regarding the overall quality are 81.6% and 86.7% at the word and speaker level, respectively. The higher correlation at the speaker level shows that evaluations based on several words from a speaker (and thus with more speaker specific information) are more reliable than those based on single words. 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 ***wick*** for ***rick*** may indicate an articulation issue and not a phonemic awareness 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. Such speaker-specific information, however, may lead to biased evaluations since dialect or accent, if any, is highly subjective and thus people may perceive it differently.

Detailed analysis of the assessments on pronunciations and 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. 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.

2.3. Choice of Words in Syllable Blending Task

From teachers' comments, we also find that children's background knowledge of the task words greatly affects their performance. For syllables of unfamiliar words, it usually takes longer for a child to give the answer. For example, many children are unfamiliar with ***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/ can be acoustically 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 tended to pronounce 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 the Spanish words ***España*** or ***esperanza***.

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 Dictionary

An ASR system is developed to evaluate children's performance in the blending task. We use log likelihood to evaluate pronunciation quality, syllable duration scores to assess 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 audio prompts, prior information of the expected answer can be used in ASR, 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 applied to generate the syllable segmentations, produce the corresponding log likelihoods and detect the inter-syllable pause, if any.

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.

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:

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

where s_i is the log likelihood of the i th 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 larger weights. To treat all segments equally, which is more desirable in this syllable blending task, local normalization is defined as:

$$S_l = \frac{1}{N}\sum_{i=1}^{N} \frac{s_i}{d_i} \quad (2)$$

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

$$S_g > t_g \quad \text{or} \quad S_l > t_l \quad (3)$$

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

3.3. Blending Smoothness Measurement

Syllable duration ratios are used to measure the smoothness. Since rate of speech (ROS) is typically applied in continuous sentences, and it is not well defined for isolated words, we use duration ratios in a normalized sense as

$$r_i = d_i \Big/ \left(\sum_{j=1}^{N} d_j\right) \quad (4)$$

It is not surprising that blended words with long inter-syllable pauses are unacceptable since the target syllables 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 an important 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. 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, unacceptable syllable duration in this case.

Fig. 1 shows the histograms of duration ratios relative to the whole word for stressed, unstressed syllables and inter-syllable pause obtained from 300 native speakers 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

$$S_d = \sum_{i=1}^{N} \log \mathcal{N}\left(r_i; \mu_i, \sigma_i\right) \quad (5)$$

where $\mathcal{N}\left(\cdot; \mu, \sigma\right)$ 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.

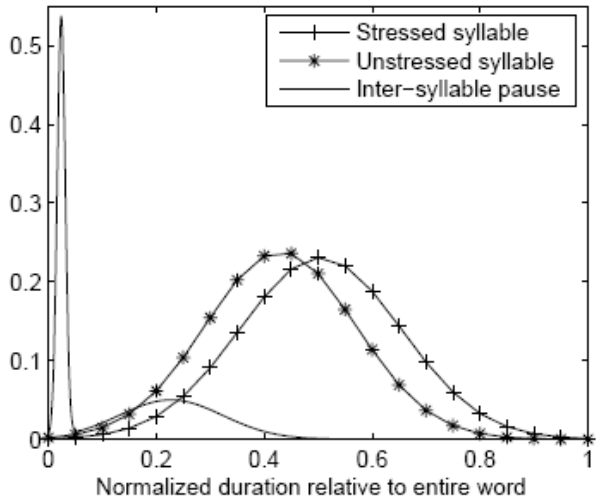

Figure1. *Histograms of duration ratios for stressed (win), unstressed (dow) syllables and inter-syllable pause for the word "window"*

3.4. Overall Evaluation and Optimization

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,

$$S = w * S_l + (1 - w) * S_d \quad (6)$$

where S is the overall quality score. 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.

In the task, some background information about a child may be available to optimize the automatic evaluation performance. 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

$$\tilde{d}_i = d_i * \frac{ROS_{\text{test speaker}}}{ROS_{\text{average}}} \quad (7)$$

where ROS_{average} is the average rate of speech over the training set.

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

4. Experimental Results

To evaluate the system's performance, evaluations from nine teachers are used as the reference. The performance is tested on 1200 utterances from children. Table 2 shows the correlation between automatic and teachers' evaluation at both the word and speaker levels. For word-level evaluation, no speaker information is assumed and thresholds are all speaker-independent, while speaker-level evaluation applies both speaker-specific thresholds and ROS normalization.

Speaker-level correlations are better than word-level correlations in all evaluations. This is reasonable because at the speaker level speaker-specific information is exploited to optimize the performance, 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 likelihoods 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. But it doesn't improve much when combining global and local scores. So equally weighting all syllables seems to be a good strategy.

For the smoothness measurement, duration ratio scores achieved comparable performance to the average inter-correlation between teachers, especially at the speaker level. The overall 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 than pronunciation in the blending task.

A detailed examination shows that the correlations between automatic and teachers' evaluations are about 3% lower for nonnative English speakers than for native speakers. This may be due to the models which are trained on native English data.

	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 2. *Correlation between automatic and teachers' evaluations at word and speaker level (in percent)*

5. Summary and Discussion

We propose an automatic evaluation system to assess children's performance on a syllable blending task. The system makes use of a pronunciation dictionary for word verification and forced alignment to generate syllable 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 information such as dialect and rate of speech can be used to optimize performance. Compared to teachers' assessments, the optimal system 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 English speakers, further work may be needed to investigate the pronunciation issues imposed by cross-language differences.

6. References

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

[2] H. Franco et al, "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 et al, "TBall Data Collection: The Making of a Young Children's Speech Corpus", in *Proc. Eurospeech*, pp. 1581-1584, 2005

[6] L. Neumeyer et al, "Automatic Text-independent Pronunciation Scoring of Foreign Language Student Speech", in *Proc. ICSLP*, pp. 1457-1460, 1996