Spoken English Assessment System for Non-Native Speakers Using Acoustic and Prosodic Features

Qin Shi[1], *Kun Li*[2], *ShiLei Zhang*[1], *Stephen M. Chu*[3], *Ji Xiao*[2], *and ZhiJian Ou*[2]

[1]IBM China Research Lab, Beijing, China
[2]Tsinghua University, Beijing, China
[3]IBM T. J. Watson Research Center, New York, USA

{shiqin,slzhang}@cn.ibm.com, {likun06,xiaoj,ozj}@mails.tsinghua.edu.cn, schu@us.ibm.com

Abstract

The absence of real-time and targeted feedback is often critical in spoken foreign language learning. Computer-assisted language assessment systems are playing an ever more important role in this domain. This work considers the idiosyncratic pronunciation patterns of Chinese English speakers and uses both acoustic and prosody features to capture pronunciation, word stress, and rhythm information. The proposed system uses *a.* automatic speech recognition and alignment for pronunciation assessment, *b*. a set of special features with appropriate normalization for word stress detection, and *c*. a prosody phrase prediction model for rhythm assessment; and is shown to give immediate and accurate analyses to speakers to improve learning efficiency.

1. Introduction

The continuing growth in global trade and communication has tremendously boosted the demand for English learning in recent years, especially in developing countries such as China. Moreover, greater emphasis is now given to oral communication ability. Traditional classrooms that might be adequate for text-central English education often lack the necessary focus on individual students. Computer-assisted language assessment systems can provide many potential benefits to both students and teachers. These systems allow continuous feedback to the student without requiring the sole attention of the teacher, facilitate self-study, and encourage interactive use of the language in contrast to rote-learning.

The criteria for spoken English assessment can generally be grouped in to two levels. The first focuses on lower-level acoustics and includes *a*. the quality of pronunciation of individual words, and *b*. the quality of prosody of utterances. The second looks at higher-level linguistic events, and typically measures the degree of language proficiency in terms of vocabulary and grammar. In automatic assessment systems, the first level usually relies on some form of acoustic scores from an automatic speech recognition (ASR) engine; and the second level criteria can be measured using language models.

This paper focuses on the acoustic and prosodic level assessment. Most existing work uses posterior probabilities of phonetic units [1]-[4] to estimate the pronunciation quality. This approach, though useful in giving a coarse assessment of overall pronunciation quality, is usually unable to give meaningful insights of the problems. Therefore, we proposed to look into methods that track the perceptual capabilities of human listeners to grade speech quality. In particular, we will investigate specific dimensions including stress [5], intonation, speaking rhythm, and fluency to improve spoken language assessment. For pronunciation assessment, we use multiple acoustic models to give alternative alignments, which are used to generate fine-grained analysis of pronunciation problems. For word stress assessment, we describe an effective multi-level normalization recipe, as features and normalization are critical to the task. For rhythm assessment, the probability of prosody phrase boundary, which is extracted using a *text-to-speech* (TTS) system, is used to measure the speaker's rhythm. And finally, by combing the rhythm and speaking rate, the fluency score is given.

The paper is structured as follows. In Section 2, we describe the system architecture to give an overview of the assessment system. Section 3 introduces the method for pronunciation evaluation. Section 4 covers word stress, rhythm and fluency estimation. Section 5 describes the evaluation result, followed by conclusions and future work in Section 6.

2. System description

The assessment system consists of three stages. The first is English speech recognition for Chinese accented speakers. The second is the assessments of speech pronunciation and prosody characteristics using word posterior probability, pronunciation variation, word stress, speech rhythm, and fluency. In the third stage, a summary or the final score of the entire assessment is given by fusing the multiple features. The diagram of the system is illustrated in Figure 1.

2.1 Speech recognition with constrained LM

Speech recognition component is the core of the system, but the acoustic models built from native speakers cannot be directly used here, because English pronunciations of Chinese speakers often deviate far from the canonical models. To improve the acoustic model for the task, a database of typical Chinese accented English speech is collected. Combining the native English speech database with non-native speaker speech database, the acoustic model is built. The system also uses the Constrained LM to improve the system's performance. The language model which is built from testing sentences is combined with a general language model.

2.2 Assessment of pronunciation and prosody

In the system, word posterior probabilities are used as the main measurement of spoken language skill. They come from frame-based posterior probabilities. In order to give a direct instruction to the speaker, we apply a special forced alignment based on the recognition result with alternative pronunciations generated from a mapping containing most common pronunciation errors. The description will be given in Section 3.

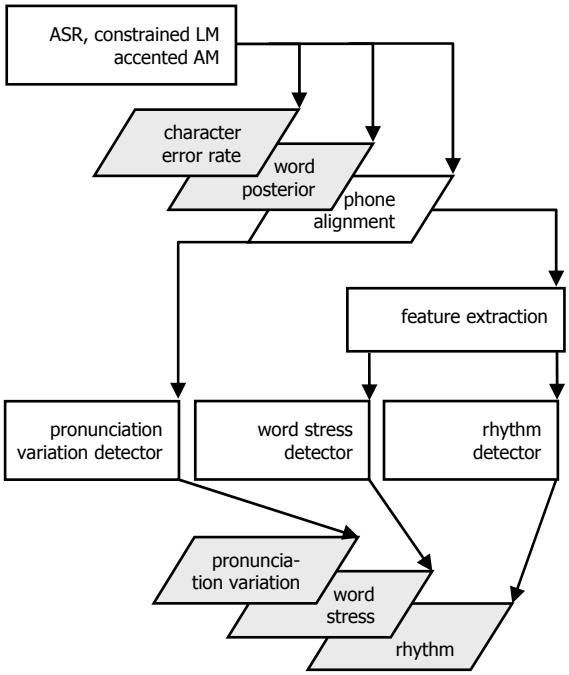

Figure 1: *Overview of the spoken English assessment system*

Word Stress is another problem for Chinese speakers. Utilizing the alignment result, the duration, energy and pitch information for each phone can be generated. A Bayesian classifier is used to determine which syllable is stressed. The result is compared with the dictionary for evaluation.

Rhythm is determined by the proper pause position in the speech. For automatic assessment, silences in speech can be detected by the ASR engine. We compute the probability of pause positions using the prosody phrase structure prediction model. An assessment of rhythm can then be given by comparing the speech silence and the probability of pause position. Fluency is measured by the speaking rate, and the number of improper pauses occurred in the speech.

2.3 Summarization of assessment

For self-learning applications, categorized feedback on pronunciation and prosody is sufficient. When used as an automatic language assessment agent, however, a final evaluation score becomes necessary. The score should be highly consistent with corresponding human evaluation score. In the proposed system, we use the support vector machine (SVM) to learn a mapping from features to scores. Features include word error rate, posterior probabilities, as well as the word stress, rhythm, and fluency scores.

3. Pronunciation assessment

3.1 Chinese accented English speech recognition

Speech recognition is the key for the system. IBM HMM-based context dependent speech recognition system [9] is used here. In order to obtain reasonable speech recognition for Chinese accented English, we apply two methods to improve the accuracy of speech recognition. The first is to combine Chinese accented English speech database with native English speech data to build acoustic model. The second is to apply constrained language modeling for decoding.

3.2 Word posterior probability

Like the typical articulation quality evaluation, the posterior probabilities of spectral observations are used as scores. For each frame in a segment corresponding to the phone q_i, we compute the frame-based posterior probability $P(q_i \mid y_t)$ of the phone q_i as in the following:

$$P(q_i \mid y_t) = \frac{p(y_t \mid q_i)P(q_i)}{\sum_{j=1}^{M} p(y_t \mid q_j)P(q_j)} \quad (1)$$

where y_t denotes the observation of phonetic segment. $p(q_i \mid y_t)$ is the probability density of the current observation y_t using the model corresponding to the q_i phone. The sum of j runs through all the phone candidates. The average of the log of the frame-based phone posterior probability over all the frames of the segment is defined as the posterior probability score for the *i*-th word segment:

$$\bar{\rho}_i = \frac{1}{d_i} \sum_{t=t_i}^{t_i+d_i-1} \log P(q_i \mid y_t) \quad (2)$$

3.3 Pronunciation variations

Table 1 lists the typical mispronunciations made by Chinese accented English speakers. According to the speech recognition result, we expand the possible phonetic spelling according to these rules. Forced alignment is then applied to find the best matching alternative. The best candidate reflects closest phonetic spelling according to the actual articulation [13]. This procedure in effect can be viewed as detectors of the common pronunciation mistakes.

Phone	Errors	Phone	Errors
/e/	/æ/ /a ɪ/	/iː/	/e ɪ/ /ɪ/
/ɜ ː/	/æ/	/ɔ ː/	/aʊ / /ɔ /ɔ/
/Uː/	/U/	/ɔ /ɔ/	/ɔ ː/ /aʊ /ɔ/
/U/	/Uː/	/æ/	/e/ /ɜ ː/
/ŋ/	/n/	/ɪ/	/iː/
/θ/	/s/	/v/	/ w/
/ð/	/z/	/ w/	/v/

Table 1. *A list of common phonetic-level pronunciation errors made by Chinese speakers.*

4. Prosody assessment

4.1 Word stress assessment

As discussed in [10], prosodic prominence includes two different features: word stress and sentence accent. In our system, we concentrate on the word stress assessment, which is more reliable to be realized without manual annotation. The performance of word stress assessment depends on the stress detector, and selecting a suitable feature set is the key factor.

Studies have shown that the prominence syllables may exhibit a longer duration, greater energy, and higher pitch, and the main correlates of syllable stress are syllable duration and energy [10]-[12]. In [10], the entire syllable duration is substituted by the syllable nucleus duration, which is rather difficult to obtain with automatic procedures. In our system, we use the vowel duration as the syllable duration. Syllable dura-

tion includes vowel and consonants, and the duration of consonants is generally exaggerated.

Energy, especially in the 500-2K Hz band, has a strong correlation with stress. In addition, loudness which is based on psychology and much closer to perception than energy, has been proposed in many studies [11],[12]. We adopt a method similar to [12] to compute the value of loudness.

Instead of using the average loudness in a syllable, we use the maximum loudness to detect stress. There are two reasons. First, the maximum loudness has the strongest stimulus in perception. Second, computing the mean loudness requires clear boundaries of syllables. Errors in the boundary information will inevitably reduce reliability of the feature. To test contribution of pitch in stress patterns, we also include the maximum semitone as an additional feature [11].

The above features vary considerably in dynamic range. To compensate this, mean normalization is applied for each feature at the utterance level. Multivariate Gaussian models are trained for stressed syllable and for unstressed syllables. Thus the stress value of a syllable can be defined as the posterior probabilities of the stressed model. For classification, the syllable with the maximum stress value in the word will be classified as stressed.

4.2 Rhythm assessment

The rhythm is mostly defined by the positions and lengths of silences. The *correct* silence positions are provided by prosody phrase analysis borrowed from the TTS domain. We use a manually labeled English TTS database that has prosody structure information to build the prosody phrase model. A decision tree is used to determine the probability of prosody phrase boundary for each word using the shallow parser information. The features used in the decision tree are:

1. *POS of the lexical word*
2. *Word position in the syntactic phrase*
3. *Type of the syntactic phrase*
4. *Syllable numbers of the words*

At run time, the distribution of each leaf gives the probability of the prosody phrase boundary. We compare the location and length of actual silences with the prosody phrase boundary probabilities indicated given by the TTS engine. Any words with low prosody phrase boundary probably but followed by a long silence are classified as errors.

5. Experiments

5.1 Chinese accented English speech database

To improve the performance of speech recognition in the language assessment system, we collected a special speech database containing English speech from Chinese speakers. The database consists of 100 subjects (college students studying in Beijing, but from different regions in China), 30K utterances, and more than 70 hours of audio.

A test set was constructed by randomly selecting 283 sentences from the 30K pool, and annotated manually. Three annotators (native English speakers) were asked to give an overall assessment of spoken English quality for each utterance, as *good*, *fair*, or *bad*. The average human label consistency rate is found to be 80.1%. Rhythm problems are also labeled by one of the three annotators.

5.2 Word posterior probability evaluation

The above acoustic training set is combined with 200 hours of native English speech data to build the acoustic models. The following table shows the relationship among the human labels, word error rate (WER) and the average word posterior probability.

human annotation	number of utterances	word error rate	average word posterior prob
bad	62	13.6%	75.6%
fair	192	5.2%	82.3%
good	29	2.0%	85.0%

Table 2. *The difference in average word posterior probability is large between utterances labeled by human annotators as bad and fair, but rather small between fair and good.*

From the result, we can see that the difference in average word posterior probability between *bad* and *fair* is large, but the difference between *fair* and *good* is small. The prosody level evaluation shown in the next section will provide further differentiation.

5.3 Word stress evaluation

Ground truth of word stress is generated based on a 90K-word dictionary with stress labels. The training set consists of 200 speakers. We used simple Gaussian models to model stressed and unstressed syllables. Figure 2 shows the stressed and unstressed syllables as a function of normalized syllable duration and normalized maximum loudness.

The test set is composed of 20 speakers with no overlapping with the training set, and contains a total of 24,542 syllables, excluding monosyllabic words or syllables without the stress labels.

Using syllable duration and maximum loudness as the features gives a classification rate of 79.71%. Adding maximum semitone to the features reduces the classification rate slightly, to 79.63%. The result is consistent with finding is in [11], that pitch only offers limited contribution to the word stress.

5.4 Rhythm evaluation

To build prosody phrase model, we use 3,232 sentences as training data (one speaker) and the prosody structures are manually labeled. A separate set of 355 sentences is used as testing data. A decision tree is built using the four types of features discussed in Section 4.2. With the threshold set at 0.5, the prosody phrase prediction accuracy rate is 83.9% and recall rate is 75.6%.

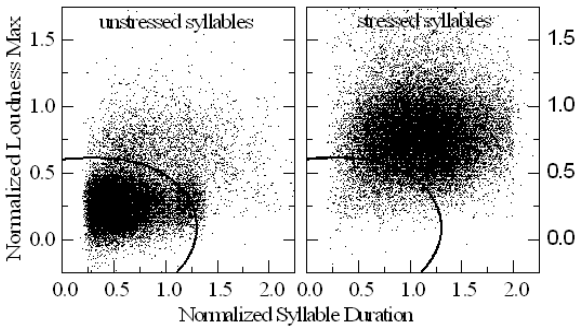

Figure 2: *Distribution of stressed and unstressed syllables as a function of normalized syllable duration and normalized maximum loudness*.

Based on the above model, we define that a rhythm error is recorded if the silence occurs at a position where the phrase predication boundary probability is lower than 0.2. Table 3 compares the automatic rhythm assessment result with human annotation. The automatic assessment results are measured as precision/recall rates against the manual labels.

human annotation	number of utterances	rhythm problem (# utt., manual)	rhythm assessment precision/recall
bad	62	40	96.1%/ 72.1%
fair	192	35	97.5%/ 70.2%
good	29	0	100.0%/100.0%

Table 3. *Comparing automatic rhythm assessment with human annotation. The automatic assessment results are measured as precision/recall rates against the manual labels*.

6. Conclusions

This paper describes our initial effort toward an automatic language assessment system aimed to provide interactive and detailed feedbacks to non-native speakers of English. We introduce an overall architecture and the components for pronunciation, word stress, and rhythm assessments. One limiting factor is the lack of data with detailed annotations for language assessment experiments. We have made the firststep by creating the Chinese accented English data set. In future work, we will continue to add richer annotations to the set and further develop the system.

References

[1] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, vol. 30, pp. 83-93, 2000.

[2] C. Cucchiarini, H. Strik, and L. Boves, "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms," *Speech Communication* vol. 30, pp. 109-119, 2000.

[3] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," in *Speech Communication*, vol. 30, pp. 121-130, 2000.

[4] S. M. Witt, "Use of Speech Recognilion in Computer assisted Languagc Learning," *Ph.D. Thesis*, the University of Cambridge, Nov.1999.

[5] A. Chandel, *et al*, "Sensei: spoken language assessment for call center agents," in *Proc. IEEE ASRU*, December 9-13, 2007.

[6] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation", in *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 16, No. 1, January 2008.

[7] H. You and A. Alwan, "Pronunciation variations of Spanish-accented English spoken by young children," *Proc. Eurospeech'2005*, Lisbon, Portugal, October 2005.

[8] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," in *Proc. Interspeech'2007*, Antwerp, Belgium, 2007.

[9] S. Chen, B. Kingsbury, L. Mangu, D. Povery, George Saon, H. Soltau, and G. Zweig, "Advances in speech transcription at IBM under the DARPA EARS program," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1596-1608, 2006.

[10] F. Tamburini, "Prosodic prominence detection in speech," in Proceedings of ISSPA, vol. 1, pp. 385-388, Paris, France, 2003.

[11] D. Wang, and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," IEEE transactions on audio, speech, and language processing, vol. 15 (2), Feb. 2007.

[12] C. Li, "Research on objective evaluation of pronunciation quality in an interactive language learning system," *Thesis*, Graduate School of Chinese Academy of Sciences, Beijing, 2007.