SpeechRater™: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech

Klaus Zechner, Derrick Higgins, Xiaoming Xi

Educational Testing Service
Princeton, NJ, USA
{kzechner,dhiggins,xxi}@ets.org

Abstract

This paper presents an overview of the SpeechRater™ system of Educational Testing Service (ETS), a fully operational automated scoring system for non-native spontaneous speech employed in a practice context. This novel system stands in contrast to most prior speech scoring systems which focus on fairly predictable, low entropy speech such as read-aloud speech or short and predictable responses.

We motivate our approach by grounding our work in the TOEFL® iBT speaking construct ("what constitutes a speaker's ability to speak comprehensibly, coherently and appropriately?") and rubrics ("what levels of proficiency do we expect to observe for different score levels in different aspects or dimensions of speech?").

SpeechRater consists of three main components: the speech recognizer, trained on about 30 hours of non-native speech, the feature computation module, computing about 40 features predominantly in the fluency dimension, and the scoring model, which combines a selected set of speech features to predict a speaking score using multiple regression.

On the task of estimating the total score for a set of three responses, our best model achieves a correlation of 0.67 with human scores and a quadratically weighted kappa of 0.61, which compares to an inter-human correlation of 0.94 and an inter-human weighted kappa of 0.93.

1. Introduction

While speech scoring systems for linguistically simpler tasks such as reading or providing a short response have been in operation for some time ([1] [2]), few attempts have been made to automatically score spontaneous, non-native speech. (With the term 'spontaneous' we refer to high entropy speech where a large-vocabulary continuous speech recognition (LVCSR) system needs to be used for recognizing speakers' utterances. The speech events are not spontaneous in the sense that they arise from speakers without being prompted, however.) One of the main reasons for this gap is the fairly high word error rate for spontaneous non-native speech, especially that produced by speakers spanning a wide range of proficiencies.

ETS has, after several years of research (see [3]), designed and implemented an operational system, SpeechRater™, for scoring spontaneous non-native speech in the context of the TOEFL® iBT Practice Online (TPO) Speaking practice program, which has been operational since October of 2006. The TPO contains the same sort of tasks (items) administered in the operational TOEFL® iBT.

The ultimate goal of SpeechRater is to generate scores for these types of spontaneous spoken responses (which are about one minute in length) using a wide range of features covering all aspects of the speaking construct and the scoring rubrics. (A construct is a representation of what is measured by a test, whereas a rubric is a scoring guide for human scorers where characteristics of speech typical of different scoring levels are provided.) In the currently operational Version 1, however, the main area of feature coverage is fluency, while other dimensions such as grammar, vocabulary, or pronunciation are covered only in rudimentary fashion.

For most higher-level features such as vocabulary use, grammar, or content, a substantially higher word accuracy would likely be needed than what is achievable at this point. (The speech recognizer used for SpeechRater has a word accuracy of about 50% for TPO responses.)

The architecture of the SpeechRater system is a concatenation of these three components: a LVCSR system trained on non-native speech, a feature computation module, and a multiple regression scoring module.

The organization of this paper is as follows: We first review some related work (section 2), followed by a description of the speaking construct and the scoring rubrics of TPO (section 3), Next, we first describe our corpus (section 4), then discuss the system components (section 5) and finally present the results of a system evaluation (section 6).

2. Previous work

There has been previous work to characterize aspects of communicative competence such as fluency, pronunciation, and prosody. [4] compare the learning effect of a pronunciation tutor (Fluency) with classroom instruction for non-native speakers of English. After several training sessions with the system, the non-native speakers in the first group reduced their human-rated pronunciation errors by 47.2%, as opposed to those receiving classroom instruction who reduced their errors by 37.5% (This difference is not significant though, partially due to the much higher variance in the Fluency group.)

[2] present a system for automatic evaluation of the pronunciation quality of both native and non-native speakers of English on a phone level and a sentence level (EduSpeak). Candidates read English texts and a forced alignment between the speech signal and the ideal path through the Hidden Markov Model (HMM) is computed. Next, the log posterior probabilities for pronouncing a certain phone at a certain position in the signal are computed to achieve a local pronunciation score. These scores are then combined with other automatically derived measures such as the rate of speech (number of words per second) or the duration of phonemes to yield global pronunciation scores.

[5] and [6] describe a system for Dutch pronunciation scoring along similar lines. Their feature set, however, is more extensive and contains, in addition to log likelihood Hidden Markov Model scores, various duration scores, and information on pauses, word stress, syllable structure, and intonation. In an evaluation, correlations between four human scores and five machine scores range from 0.67 to 0.92..

[1] presents a test for spoken English (SET-10) that uses the following types of items: reading, sentence repetition, sentence building, opposites, short questions, and open-ended questions. All types except for the last are scored automatically and a score is reported that can be interpreted as an indicator of how native-like a speaker's speech is. In [7], an experiment is performed to establish the generalizability of the SET-10 test. It is shown that the SET-10 test scores can predict different levels on the Oral Interaction Scale of the Council of Europe's Framework for describing oral proficiency of second/foreign language speakers with reasonable accuracy [8]. This paper further reports on studies done to correlate the SET-10 automated scores with the human scores from two other tests of oral English communication skills. Correlations are found to be between 0.73 and 0.88.

Previous work at ETS ([9], [3]) investigated, to the best of our knowledge for the first time, the area of automated scoring of unrestricted, spontaneous speech of non-native speakers. We focused on exploring a number of different fluency features for the automated scoring of short (one minute) responses to test questions in a TOEFL-related program, called LanguEdge®. We explored scoring models based on classification and regression trees (CART) as well as support vector machines (SVM). We found that the SVM models were more useful for a quantitative analysis, whereas the CART models allow for a more transparent summary of the patterns that underline the data. For the current study we adopt multiple regression which performs about at par with the other methods and has the advantage of being more easily interpretable and explainable to the outside world. Another major difference between the previous work and the current study is that we use feature normalization and transformation to obtain statistically more meaningful input variables for the scoring model. In addition, we do not use the whole set of features, but a carefully selected subset that has the properties of being good predictors for human scores (high correlation, e.g.) and also of being as broad a representation of the speaking construct and rubrics as possible.

3. TOEFL iBT construct and rubrics

The TOEFL iBT Speaking Practice test assesses test takers' speaking proficiency in an academic environment. Specifically, it measures their ability to speak about campus life topics and academic course content in a comprehensible, coherent, and appropriate manner. The scoring rubric for human scoring represents the construct of speaking that is of interest to the TOEFL iBT Speaking Practice test. This construct of interest is the basis on which the construct-appropriateness of SpeechRater's automated scoring is evaluated.

The TPO Speaking test consists of six speaking tasks: two independent tasks, where test takers respond to questions on familiar topics, and four integrated tasks, where test takers first read and/or listen to some materials and then respond to a question with reference to what they have heard and/or read.

While applying the rubrics in evaluating speaking responses, raters issue a global, "impressionistic" score for each response on a score scale from 1-4, considering the combined impact of three key categories of performance: Delivery, Language Use, and Topic Development.

Delivery refers to the pace and clarity of the speech. In assessing delivery, raters consider the speaker's pronunciation, intonation, rhythm, rate of speech, and degree of hesitancy. *Language use* refers to the range, complexity, and precision of vocabulary and grammar use. Raters evaluate candidates' ability to select words and phrases and their ability to produce structures that appropriately and effectively communicate their ideas. *Topic development* refers to the coherence and fullness of the response. When assessing this dimension, raters take into account the progression of ideas, the degree of elaboration, the completeness, and, in the case of integrated tasks, the accuracy of the content.

4. Corpus

In building and evaluating the models described below, we made use of two data sets: the *TPO* data and the *iBT field study* data.

The TPO data, from the TOEFL iBT Practice Online practice program, contained 4162 spoken responses from four distinct test forms. We set aside 1907 of these responses (from 320 speakers) for the training of the speech recognizer, and partitioned the remaining data into a train and test set for the scoring model. The scoring model train and test sets consist of a set of responses with human scores in the range 1-4, and there is no overlap in speakers or topics between these sets. (Responses which are anomalous in some way may receive a score of 0, or of TD—technical difficulty. These score classes are handled by a filtering model not discussed in detail here.)

The second data source we used in our experiments was the TOEFL iBT Field Study, a pilot undertaken before the official roll-out of the TOEFL iBT test. While we were primarily interested in our models' performance on TPO data, we used the field study data in doing evaluation runs because the conditions under which the field study data was scored were closer to best practice than they were with the TPO data sets. The field study data contained 3,502 responses from a single TOEFL iBT Speaking test form: Since we used a previously trained recognizer for this data, we could use all of the data for the scoring model train and test sets. The iBT data sets used for scoring model training and testing were again defined so that there was no overlap between the two in speakers or topics. The training set included 1750 responses from 311 different speakers, and the test set included 1752 responses from 315 different speakers

Human agreement in scoring the iBT Field Study data was quite high, with a weighted kappa of 0.77 for single items, and 0.93 for aggregate scores of three items. For the TPO data, though, it was somewhat lower: the weighted kappa for single items was 0.55, and 0.68 for sets of three items. (We consider the sum of scores assigned to multiple tasks by the same candidate in calculating agreement, because such aggregates are often more stable than scores on individual items.) The Pearson coefficient of correlation shows the same difference between the two data sets: the correlation between independent human ratings on the iBT Field Study data was 0.77 for single items, and 0.94 for the total score over three items, while the

corresponding figures for the TPO data are 0.56 and 0.70, respectively.

This difference is largely because of the lack of variation in scores observed in the TPO set. Students self-select, so that the population falls predominantly at the higher end of the ability scale. The average human score assigned to responses in the evaluation set of the TPO data is 2.73, with a standard deviation of 0.69, which compares to 2.48 (1.00) for the field study evaluation set.

5. System components

5.1 Speech recognizer

The speech recognition system is a gender-independent fully continuous Hidden Markov Model system, trained on about 30 hours of non-native speech from the TPO program. For language model training, a larger corpus of non-native speech (about 100 hours) was used, and mixed with a large general domain model (Broadcast News [10]).

The speech recognizer serves as the front-end of the SpeechRater system. It accepts digitized speakers' responses to test questions and yields a first-best hypothesis including information on word identity, timing, and confidence scores. For each response, summary scores for acoustic model and language model representations are given as well.

The word accuracy on unseen TPO data was found to be around 50%. While this would be low in general terms, it is a realistic number for the fact that we are scoring speech from many different native languages and a wide range of speaking proficiency levels.

5.2 Feature computation module

The feature computation module takes the speech recognizer's output as its input and uses information on words and their timing information to compute features. The majority of features are related to the dimension of fluency, such as "words per minute" or "average pause length". In addition, there are features related to word types (as opposed to word tokens) that can represent the range or variety of vocabulary in the speaker's response, e.g., "types per second". There are also features representing the recognizer's language model and acoustic model internal scores. A total of about 40 features are computed and then, in some instances, transformed and/or normalized to yield a distribution with more desirable statistical properties. A full account of features used in earlier prototypes of SpeechRater can be found in [3] and [9].

5.3 Scoring model

A multiple regression model was used to assign a score to a response on the basis of a selected set of five features which represent fluency, vocabulary diversity, pronunciation, and grammatical accuracy. These features were selected based on input from the Content Advisory Committee (CAC), a group of content-area specialists convened to ensure the construct-appropriateness of our scoring model. One aim of model building was to obtain high agreement with human raters, but an additional goal was to structure the model so that its use of the predictive features was in conformance with content experts' understanding of the speaking construct. Toward this end, the regression equation was reduced so that only two parameters needed to be learned from the data: the slope parameter μ, and the intercept β:

$$Score = \mu \sum_i \alpha_i f_i + \beta \, .$$

The feature-specific weights α_i were specified in consultation with the CAC, who changed the feature weights derived from an optimal empirical model.

Note also that feature values were standardized so that the CAC weights assigned were comparable across all features.

6. Training and evaluation

The weights (α) of the standardized features were set to the values defined by the CAC, and the slope and intercept parameters of the regression model were set to values minimizing the least-squares error on the TPO scoring model training data.

This model was then applied to the test set of data from the TPO program, and the results of this evaluation are shown in the first column of the table below. We provide aggregate statistics as well as statistics related to accuracy of human score prediction on individual items, because ultimately the most important consideration is the accuracy of the total Speaking section score reported to the examinee. Because of the way the data was partitioned for this project, there were not enough examinees with six complete tasks (one full test form) in the TPO evaluation set for us to perform this evaluation on a set of six aggregated tasks.

The most important statistic for evaluating the quality of this scoring model is the correlation of the predicted scores with the human-assigned scores, which ranges from 0.37 for single items to 0.51 for the total score on three items for TPO data. (We report correlations for unrounded machine scores with integer human scores.) These correlations are not as strong as we might hope for; however, one major reason for this is the lack of score variability in the TPO data set mentioned above.

In order to establish the quality of the scores provided by our regression model independently of the restriction of score range encountered with the TPO data, we carried out the same training and evaluation procedure on the iBT Field Study data. Because of the larger number of responses, the wider ability range of examinees, and the possibility of aggregating complete sets of six items, the predicted scores have a considerably higher correlation with the human-assigned scores on this data set, as shown in the table below. On single items, the correlation is 0.61, and on a full form of six items, it reaches 0.68.

A large gap still remains between the level of human agreement cited in Section 3 above and the agreement of the automated scoring system with human raters. The development of features to cover more of the construct relevant to these TOEFL iBT Speaking tasks promises to narrow this gap somewhat. The current version of SpeechRater relies primarily on features related to the *delivery* of speech, and an improved treatment of *language use* and *topic development* would be expected to yield improvements in model performance.

Evaluation set	*TPO*	*iBT Field Study*
Single scores	**N=520**	**N=1752**
Weighted κ	0.32	0.51
Mean (SD) of predicted score	2.78 (.33)	2.45 (.61)
Correlation	0.37	0.61
Total score (3 tasks)	**N=163**	**N=555**
Weighted κ	0.44	0.61
Mean (SD) of predicted score	8.38 (.83)	7.43 (1.63)
Correlation	0.51	0.67
Total score (6 tasks)	**N=0**	**N=254**
Weighted κ		0.61
Mean (SD) of predicted scores		15.13 (3.04)
Correlation		0.68

Table 1: Scoring accuracy on two evaluation sets

7. Conclusions and future work

We have presented an automatic system for scoring non-native, spontaneous speech that is based on considerations of the TOEFL® iBT speaking construct. An evaluation of two different corpora shows correlations with human scores in the range of 0.51-0.67 for the total score on three items.

In future work, we plan to substantially extend our feature set to cover additional dimensions of the construct such as language use and content. The development of these features will serve the dual purpose of more fully addressing the set of properties which comprise speaking proficiency, and improving the accuracy of our scoring model.

In order for these features to yield useful information, ahigher word accuracy than currently obtainable will be needed. Therefore another major goal is to improve the speech recognizer's accuracy by, for instance, language model and acoustic model adaptation.

We also will experiment with using two distinct recognizers for different purposes: one adapted to non-native speech will optimize recognition accuracy, whereas another one trained on native or near-native speech will yield appropriate estimates of pronunciation accuracy.

References

[1] J. Bernstein, "PhonePass testing: Structure and construct," Ordinate Corporation, Menlo Park, CA May 1999.

[2] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, and J. Butzberger, "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning," presented at InSTiLL-2000 (Intelligent Speech Technology in Language Learning), Dundee, Scotland, 2000.

[3] K. Zechner, I. I. Bejar, and R. Hemat, "Towards an understanding of the role of speech recognition in non-native speech assessment," Educational Testing Service, Princeton, NJ RR-07-02, 2007.

[4] L. M. Tomokiyo, L. Wang, and e. al, "An empirical study of the effectiveness of speech recognition-based pronunciation training," presented at ICSLP-00, Beijing, China, 2000.

[5] C. Cucchiarini, S. Strik, and L. Boves, "Automatic evaluation of Dutch pronunciation by using speech recognition technology," presented at IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, CA, 1997.

[6] C. Cucchiarini, H. Strik, and L. Boves, " Using speech recognition technology to assess foreign speakers' pronunciation of Dutch," presented at Third international symposium on the acquisition of second language speech: NEW SOUNDS 97, Klagenfurt, Austria., 1997.

[7] J. Bernstein, J. DeJong, D. Pisoni, and B. Townshend, "Two experiments in automatic scoring of spoken language proficiency," presented at InSTIL2000, Dundee, Scotland, 2000.

[8] B. North, *The Development of a Common Framework Scale of Language Proficiency*. New York, NY: Peter Lang, 2000.

[9] K. Zechner and I. Bejar, "Towards Automatic Scoring of Non-Native Spontaneous Speech," presented at HLT-NAACL-06, New York, NY, 2006.

[10] Linguistic-Data-Consortium, "HUB-4 Broadcast News corpus (English)," 1997.