**Automated English proficiency scoring of unconstrained speech**

**Abstract**

This paper evaluates the performance of 17 machine learning classifiers in automatically scoring the English proficiency of unconstrained speech. Each classifier was tested with different sets of features drawn from a master set of segmental and suprasegmental measures founded on Brazil’s prosody model. The segmental and suprasegmental measures were calculated from the output of an Automatic Speech Recognizer (ASR) that recognizes phones instead of words and other software designed to detect the elements of Brazil’s prosody model. The sets of features were chosen by means of three different feature selection algorithms: take-out/add-in, simulated annealing, and genetic algorithm. The performance of the best classifier, Pairwise GentleBoost, was 0.68 in terms of the correlation between the computer’s calculated proficiency ratings and those scored by humans. This correlation is higher than other similar computer programs for automatically scoring the proficiency of unconstrained speech.

**Keywords:** automated proficiency scoring; Brazil’s prosody model; simulated annealing; genetic algorithm; suprasegmental measures

**I Introduction**

Automated scoring systems for speech can be partitioned into two categories: those that are intended for scoring constrained speech, such as constructed response items, and those intended to score unconstrained speech. The Versant Spanish Test (Pearson, 2015) and the speaking tasks within the Pearson Test of English (Longman, 2013) are examples of successful automated scoring systems for constrained speech. In these tests, the test-taker responds verbally to a sequence of recorded prompts (e.g., sentence repeats, short questions, sentence builds, passage retells). The computer evaluates their responses to ascertain the contents of the speech and the fashion in which it was articulated. The computer calculates scores on the facets of sentence mastery, vocabulary, fluency and pronunciation. Bernstein, Van Moere, and Cheng (2010) offered proof that the automated scores from the Versant Spanish Test and the Pearson Test of English were highly correlated with scores from oral proficiency interviews, which substantiates using them to assess a person’s proficiency in constrained spoken communication.

Unconstrained speech is unpredictable, and thus, is more difficult to score automatically. In proficiency tests, unconstrained speech is elicited by asking the test-taker to speak about a general topic for a minute or more. For example, the examiner might provide the speaker with a photograph and ask the speaker to talk about it for one minute. SpeechRaterSM is an example of a successful unconstrained English speech automated scoring system (Zechner, Higgins, Xi, & Williamson, 2009). SpeechRaterSM processes the test-taker’s speech with an automatic speech recognizer (ASR) configured to recognize the words in the speech. Then, it derives a set of mostly fluency based (i.e., segmental) measures from the ASR output, which are then analyzed with multiple regression (MR) to predict a speaking proficiency score of one to four, four being the best. In laboratory tests of SpeechRaterSM, the correlation between its scores and those of a human were 0.37, but in actual field tests the correlation was higher at 0.55. A version of SpeechRaterSM, which was not deployed, that used classification and regression trees (CART) had even higher correlations between it and human raters: 0.44 (laboratory test) and 0.62 (field test).

This paper examines an alternative method of automatically scoring unconstrained English speech with an ASR that recognizes phones instead of words and a set of segmental and suprasegmental measures calculated from the output of the ASR and other software designed to detect the elements of Brazil’s (1997) prosody model. This approach has two advantages over the one employed for SpeechRaterSM. First, the ASR recognizes phones instead of words. This approach offers the following two benefits: 1) the ASR only has to be trained on the 60 phones that make up all English words vs. being trained on thousands of words that have to be determined *a* *priori* (which violates the principle of the speech being truly unconstrained) and 2) the phone error rate (PER) of an ASR (e.g., 16%, see Section II.1) is usually lower than the word error rate (WER) of an ASR (e.g., 50%; Zechner et al., 2009) potentially leading to more accurate proficiency scores. The second advantage of this approach is in the utilization of suprasegmental measures derived from the elements of Brazil’s prosody model. This is important because although segmental measures (i.e., those used by SpeechRaterSM) can predict proficiency, suprasegmental measures have been found to account for 50% of the variance in oral proficiency ratings (Author and colleagues, 2010).

The purpose of the research presented in this paper is to ascertain the optimum machine learning classifier and collection of segmental and suprasegmental measures for automating English proficiency scoring of unconstrained speech. The paper begins with a description of how the segmental and suprasegmental measures were calculated. Next, the classification features, classifiers, and experimental methods applied in this research are explained, after that is a presentation of the results of the experiments followed by a discussion of the results and a conclusion.

**II Automatic calculation of segmental and suprasegmental measures**

As part of this research, computer programs were developed to automatically extract from audio speech files, segmental and suprasegmental prosody measures built on Brazil’s (1997) model. The foundation of Brazil’s model is the tone unit. Brazil defines a tone unit as a segment of a monologue or dialog that a listener can recognize as having an intonation pattern that is distinct from those of otherwise similar tone units having other patterns of intonation. Each tone unit has at least one prominent syllable, which can be perceived from three aspects: pitch (fundamental frequency of a syllable in Hz), duration (length of the syllable in seconds), and intensity (amplitude of the syllable in dB) (Chun, 2002). Brazil states that the significance of prominence is on the syllable, and not the word, because there are words whose prominent syllable change depending on the intonational meaning that the speaker is conveying. Brazil further contrasts prominence and lexical stress. Lexical stress specifies the syllable within content words that is stressed, whereas prominence is the application of stress to differentiate those words in an utterance which convey more meaning, emphasis, or contrast. Thus, a syllable within a word that is typically given lexical stress may be given additional pitch, duration, or intensity to emphasize its meaning.

On the other hand, a syllable that is not normally stressed (such as a function word) may be stressed to distinguish it. Each tone unit includes a key (first) and termination (last) prominent syllable. The termination syllable is also called the tonic syllable. (If the tone unit contains only one prominent syllable, then it is both the key and termination prominent syllable). The intonation pattern of a tone unit is defined by the relative pitch of the key and termination prominent syllables and the tone choice of the termination prominent syllable. Brazil specified three equal ranges of relative pitch: low, mid, and high, and five tone choices: falling, rising, rising-falling, falling-rising, and neutral.

The computer programs developed for this project automatically determine the elements of Brazil’s model from a raw audio file in several steps: 1) recognize the phones and pauses that make up the utterance, 2) divide the utterance into tone units, 3) group the phones into syllables, 4) identify the filled pauses, 5) detect the prominent syllables, 6) classify the tone choice (falling, rising, rising-falling, falling-rising, or neutral) of the tonic syllables (last prominent syllable in a tone unit), 7) compute the relative pitch (low, mid, or high) of the tonic syllables, and 8) calculate segmental and suprasegmental measures derived from counts of pauses, filled pauses, tone units, syllables, prominent syllables, tone choices, and relative pitch. Each step is described with more detail in the following sections.

In this research, these eight steps were applied to a corpus consisting of 120 speech files of non-native English speaker’s monologues from the Cambridge English Language Assessment (CELA) corpus (see Section III.2.c).

*1 Phone and silent pause recognition*

First, the process begins with converting audio files of human speech into the 60 phones defined by the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo, Lamel, Fisher, Fiscus, & Pallett, 1993) as illustrated in Figure 1.



**Figure 1.** Phone and silent pause recognition

A large vocabulary spontaneous speech recognition (LVCSR) program generates time-aligned phone transcriptions of the audio speech files. The LVCSR is derived from the KALDI speech recognition engine (Povey, Ghoshal, Boulianne, Burget, Glembek, Goel, Hannemann, Motlíček, Qian, Schwarz, Silovsky, Stemmer, & Vesel, 2011). The LVCSR was trained on the 4,620 sentences in the TIMIT training set. The trained LVSCR is capable of recognizing the phones of the 1,680 sentences in the TIMIT test set with a phone error rate (PER) of only 16%. The TIMIT-trained LVSCR identified 70,498 phones in the CELA audio files.

The TIMIT phone set includes the *pau*, which represents a silent pause in the audio stream. The LVCSR output was filtered as follows to improve silent pause recognition: 1) convert TIMIT phones *f*, *k*, *t*, *n*, and *epi* occurring before a short *pau* phone (i.e., a *pau* less than 100 msec in duration) to a *pau*, 2) convert single consonant phones between two long *pau* phones (i.e., a *pau* more than 100 msec in duration) to a single *pau*, 3) combine adjacent *pau* phones, and 4) replace high intensity *pau* phones or ones which have a pitch contour with a pseudo phone, *?*, representing a non-pause phone of unknown type. These filters were developed by experimentation with a World Englishes corpus (see Section III.1.b). These filters improved the correlation between the KALDI detected pauses and those identified using the Multi-Speech and Computerized Speech Laboratory (CSL) Software (KayPENTAX, 2008) from 0.508 to 0.935.

For the CELA audio files, 1,693 *f*, *k*, *t*, *n*, and *epi* phones occurring before a short *pau* phone were converted to a *pau*, 947 single consonant phones between two long *pau* phones were converted to a single *pau*, and the combination of adjacent *pau* phones resulted in reducing the number phones from 70,498 to 66,098. The software replaced 738 *pau* phones which had high intensity or a pitch contour with a *?* pseudo phone. The final number of silent pauses (*pau*) detected was 4,173.

*2 Tone unit division*

Second, the utterance is divided into tone units by analyzing the silent pauses (*pau*) as depicted in Figure 2.



**Figure 2.** Tone unit division

Silent pauses which have a duration longer than 200 ms or a duration between 150 ms and 200 ms, followed by either a pitch reset or a slow pace, delimit a tone unit. Pitch reset means the relative pitch of the syllable before the *pau* is high and the relative pitch of the syllable after it is low, or just the opposite (i.e., low before and high after). Slow pace is defined as the duration of the syllable after the *pau* being greater than the normal duration of a syllable. The normal duration of each syllable is calculated by summing the average duration of the phones in the syllable. In these calculations a syllable is defined as three phones. The number three was derived as follows:

*s* = Average number of syllables per word = 1.5 (Sakti, 2009) (1)

*t* = average number of TIMIT phones per word = 3.9 (Garofolo et al., 1993) (2)

*l* = ┌ *t*/*s* ┐ = 3 (3)

In tests with the World Englishes corpus (Author 2010; Author & colleagues, 2014; Author & colleagues, 2015), the correlation between the number of tone units recognized by a trained analyst employing Computerized Speech Laboratory (CSL) Software (KayPENTAX, 2008) equipment and a computer applying these algorithms was 0.959. In the CELA data, of the 4,173 silent pauses detected, the computer concluded that 120 were leading silences, 2,845 delimited a tone unit, 1,088 did not delimit a tone unit, and 120 were trailing silences.

*3 Syllabification*

Third, the phones are combined into syllables as portrayed in Figure 3.



**Figure 3.** Syllabification

Syllables are formed by grouping consonant phones with the vowel phone closest to them in time. The vowel phones are: *aa*, *ae*, *ah*, *ao*, *aw*, *ax*, *ax-h*, *axr*, *ay*, *eh*, *el*, *em*, *en*, *er*, *ey*, *ih*, *iy*, *ix*, *ow*, *oy*, *uh*, *uw*, and *ux*. Syllabic consonants are considered vowels in the syllabification process. The closet vowel is determined by measuring the duration from the consonant phone to the vowel before and after it. The before-duration is biased by multiplying it by a factor, *b*, where *b* > 0. The bias has a tendency to group a consonant more often with the next vowel. The consonant is grouped with the previous vowel, if the biased before-duration is less than the after-duration. The factor *b* was calculated as 1.643 by an exhaustive search of possible values from 0.001 to 2.000 in 0.001 increments as illustrated in Figure 4.



**Figure 4**. Syllabification alignment error vs. factor *b*

The exhaustive search was conducted by minimizing the syllable alignment error resulting from syllabifying the 1,680 TIMIT test utterances. Syllable alignment error (*erralign*) is the percent of the utterance duration where the computer generated syllables do not match with the “gold standard” syllables defined in the TIMIT corpus. With *b* set to the optimum value of 1.643, the alignment error for the 1,680 TIMIT test utterances was 19.761%. Syllabifying the CELA corpus with *b* set to 1.643 yielded 21,776 syllables.

*4 Filled pause identification*

Fourth, the filled pauses in the utterance are located as represented in Figure 5.



**Figure 5.** Filled pause identification

A filled pause is defined as any vocalized word or sound used to fill breaks in speech (e.g., erm, huh, like, uh, um, well, you know), but it excludes repetitions, restarts, or repairs. They are segments of an utterance which are not commonly understood to express formal meaning or to be purposeful. They are called filled pauses because they are similar to silent pauses. For this research, a tone unit containing only syllables from the set of {ah, axr, er, em, m, w\_el, ax, aa\_hh} was considered a filled pause. (Note: A syllable is represented as phones concatenated with the underscore character, e.g., w\_el is the syllable for “well”). The software detected 20 filled pauses in the CELA corpus.

*5 Prominent syllable detection*

Fifth as characterized in Figure 6, a bagging ensemble of 100 decision tree learners analyzed the pitch, duration, and intensity features extracted from the original waveform to detect prominent syllables (Author 2 & Author1, accepted).

**Figure 6.** Prominent syllable detection

In 5-fold cross-validation experiments with a subset of the TIMIT corpus, this bagging ensemble performed with an Accuracy of 95.9% ±0.2%, an F-measure of 93.7 ±0.4, and a *κ* of 0.907 ±0.005 (Authors, accepted). The ensemble utilized to detect prominent syllables in the CELA corpus was trained on the prominent syllables in 839 utterances from the TIMIT corpus identified by a trained analyst (Authors,, accepted). Of the 21,776 syllables in the CELA corpus, the computer detected 6,508 prominent syllables

*6 Tone choice classification*

Figure 7 illustrates the sixth step which is a rule-based classifier ascertaining the tone choice of the tonic syllable (i.e., last prominent syllable of the tone unit) utilizing a set of extracted pitch features (Authors,, under review).



**Figure 7.** Tone choice classification

The pitch features are calculated using a 4-point model developed by Authors, (under review). The rule-based classifier achieved an accuracy of 75.1% and a Cohen’s kappa coefficient of 0.73 within 5-fold cross-validation experiments of a portion of the TIMIT corpus (Authors,, under review). The rule-based classifier, trained on the tone choices of 839 utterances from the TIMIT corpus (Authors,, under review), was employed to classify the tone choice of the CELA corpus. The CELA corpus contained 2,945 tone units that were not filled pauses. Of those, 2,898 included at least one prominent syllable. The computer classified the following tone choices for the tonic syllables of those tone units: 416 rising, 592 neutral, 1,257 falling, 261 fall-rising, and 372 rise-falling.

*7 Relative pitch calculation*

Figure 8 explains the relative pitch (low, mid, and high) calculation of the key (i.e., first prominent syllable in the tone unit) and tonic syllables, which is the seventh step.



**Figure 8.** Relative pitch calculation

The relative pitch is computed with a straightforward algorithm. The pitch range of the entire utterance is divided into three equal bands: low, mid, and high. The relative pitch of the key or tonic syllable is the band within which the majority of pitch contour points fall. There were 4,522 key and tonic syllables detected in the CELA corpus of which 3,108 had a low-relative pitch, 1,349 had a mid-relative pitch, and 65 had a high-relative pitch.

*8 Computation of segmental and suprasegmental measures*

The final step of computing the 35 segmental and suprasegmental measures defined in Author et al. (2010) is summarized in Figure 9.



**Figure 9.** Computation of segmental and suprasegmental measures

These measures are computed from the counts, pitch ranges, and durations of tone units, pauses, syllables, filled pauses, prominent syllables, tone choices, and relative pitches as specified in Table 1. These 35 measures comprise the superset of features with which we tested the classifiers described in the next section.

**Table 1.** Segmental and suprasegmental measures.

|  |  |  |
| --- | --- | --- |
| Abbreviation | Category | Description |
| SMARTI | Rate | Articulation rate (syllables per second excluding silent pause time) |
| SMAVNP | Pitch | Non-prominent syllable mean pitch |
| SMAVPP | Pitch | Prominent syllable mean pitch |
| SMFALH | Pitch | Falling-high rate (tonic syllables with falling tone choice and high-relative pitch per second) |
| SMFALL | Pitch | Falling-low rate (tonic syllables with falling tone choice and low-relative pitch per second) |
| SMFALM | Pitch | Falling-mid rate (tonic syllables with falling tone choice and mid-relative pitch per second) |
| SMFPLN | Pause | Filled pause mean length |
| SMFPRT | Pause | Filled pauses per second |
| SMFRSH | Pitch | Fall-rise-high rate (tonic syllables with fall-rise tone choice and high-relative pitch per second) |
| SMFRSL | Pitch | Fall-rise-low rate (tonic syllables with fall-rise tone choice and low-relative pitch per second) |
| SMFRSM | Pitch | Fall-rise-mid rate (tonic syllables with fall-rise tone choice and mid-relative pitch per second) |
| SMGIVP | Pitch | Given lexical item (i.e., prominent syllables) mean pitch |
| SMNEUH | Pitch | Neutral-high rate (tonic syllables with neutral tone choice and high-relative pitch per second) |
| SMNEUL | Pitch | Neutral-low rate (tonic syllables with neutral tone choice and low-relative pitch per second) |
| SMNEUM | Pitch | Neutral-mid rate (tonic syllables with neutral tone choice and mid-relative pitch per second) |
| SMNEWP | Pitch | New lexical item (i.e., prominent syllables) mean pitch |
| SMOPTH | Paratone | Paratone boundary onset pitch mean height |
| SMPACE | Stress | Pace (prominent syllables per tone unit) |
| SMPARA | Paratone | Paratone boundaries (low-relative pitch termination syllable followed by high-relative pitch key syllable onsets) per second |
| SMPCHR | Stress | Percent of tone units containing at least one prominent syllable |
| SMPHTR | Rate | Phonation time ratio (percent of time spent speaking, including filled pauses) |
| SMPPLN | Paratone | Paratone boundary mean pause length |
| SMPRAN | Pitch | Overall pitch range |
| SMRFAH | Pitch | Rise-fall-high rate (tonic syllables with rise-fall tone choice and high-relative pitch per second) |
| SMRFAL | Pitch | Rise-fall-low rate (tonic syllables with rise-fall tone choice and low-relative pitch per second) |
| SMRFAM | Pitch | Rise-fall-mid rate (tonic syllables with rise-fall tone choice and mid-relative pitch per second) |
| SMRISH | Pitch | Rising-high rate (tonic syllables with rising tone choice and high-relative pitch per second) |
| SMRISL | Pitch | Rising-low rate (tonic syllables with rising tone choice and low-relative pitch per second) |
| SMRISM | Pitch | Rising-mid rate (tonic syllables with rising tone choice and mid-relative pitch per second) |
| SMRNLN | Rate | Tone unit mean length |
| SMSPAC | Stress | Space (percent of syllables that are prominent) |
| SMSPLN | Pause | Silent pause mean length |
| SMSPRT | Pause | Silent pauses per second |
| SMSYPS | Rate | Syllables per second |
| SMTPTH | Paratone | Paratone boundary mean termination pitch height |

Various combinations of these measures have been tested with the classifiers explained in the next section to ascertain the best classifier and set of measures for automating English proficiency scoring of unconstrained speech.

**III Methods**

*1 Corpora*

This research made use of three speech corpora, TIMIT, World Englishes, and CELA, which are described below.

*a TIMIT corpus*

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) comprises ten sentences spoken by each of 630 speakers from eight major dialect regions of the United States for a total of 6300 sentences (Garofolo et al., 1993). The speakers read text made up of two dialect sentences, 450 phonetically-compact sentences, and 1890 phonetically-diverse sentences. All 630 speakers read the dialect sentences which were designed to expose the speaker’s dialect. The phonetically-compact sentences included a rich selection of typical phone pairs and additional incidences of phonetic contexts believed to be either of specific significance or challenging to speak. Seven different speakers read each of these sentences and five of these sentences were spoken by each speaker. The phonetically-diverse sentences were intended to use the full range of allophonic contexts in the texts. Only one speaker read each of these sentences and each speaker read three of them. Hand corrected start and end times for the phones, phonemes, pauses, syllables, and words are provided with the corpus. The TIMIT corpus defines 60 phones, which have become a de facto standard for other corpora.

*b World Englishes corpus*

The World Englishes corpus is a set of speech samples and orthographic transcriptions in another study of the authors (Author1 and colleagues, XXX). It contains speech files of 3-5 minute academic lectures from 18 English speakers, three (one female and two male) from each of six separate categories of English. The speakers’ first languages characterized all of Kachru’s (1992) three concentric circles of World Englishes: the inner circle Englishes (North American and British), the outer circle Englishes (Indian and non-Anglophone South African), and the expanding circle Englishes (Chinese and Spanish). Inner circle speakers had Standard American and British accents, specifically California and Southern England. Indian and South African speakers were selected for the outer circle speakers. To maintain homogeneity all of the Indian speakers were proficient in Hindi and all of the South African speakers were proficient in Afrikaans. The expanding circle of speakers were Chinese and Spanish speakers because their first languages were included in the most frequent native languages of TOEFL takers (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002), and they characterized unalike language families. All of the Chinese English speakers spoke Mandarin as a first language. Each of the Spanish English speakers was from the state of Sonora in Mexico. The South African, Indian, Chinese, and Mexican speakers were very proficient in English, but still had an obvious foreign accent.

All of the speakers were selected based on qualities proposed by Major et al. (2002). First, they sounded like authentic speakers of a given accent and conversational as a lecturer. Second, they could fluently manage the terminology of the lectures they read and consequently acted like experts in the field covered in the lecture. And lastly, they had the seasoned voice quality and pitch of a practiced scholarly speaker.

*c Cambridge English Language Assessment (CELA)*

The CELA corpus consists of the speaking part of the Cambridge ESOL General English Examinations (Cambridge, 2015). The subject of the monologues, number of speech files, and proficiency levels, using the Common European Framework of Reference for Languages (CEFR) from B1-C2, are as follows: Preliminary English Test (PET, B1) is an intermediate level qualification (32 files) – the examiner provides the speaker with a color photograph and asks the speaker to talk about it for one minute. First Certificate in English (FCE, B2) is an upper-intermediate level qualification (32 files) – the examiner shows the speaker two photographs and the speaker discusses them for one minute. Certificate in Advanced English (CAE, C1) tests the ability of the speaker to use English in work or studies (34 files) – the examiner presents the speaker with three pictures and asks the speaker to pick two of them and discuss what he or she sees in the picture for one minute. Certificate of Proficiency in English (CPE, C2) is the most advanced qualification (22 files) – the examiner gives the speaker a card with a question and some ideas on it. The speaker has to speak about the question for two minutes. The speakers have 21 different first languages (L1s). There are 23 females and 11 males in the CAE group; 17 females and 5 males in the CPE group; 21 females and 11 males in the FCE group; and 16 females and 16 males in the PET group. There were 21 first languages included: 16 Spanish (including Mexican and Spanish), 11 Korean, 8 Italian, 7 Dutch, 6 French, 5 Chinese, 5 Russian, 4 Greek, 4 Portuguese, 4 Swedish, 3 German, 2 Swiss, 2 Japanese, 1 Brazilian, 1 Bulgarian, 1 Bolivian, 1 Austrian, 1 Turkish, 1 Arabic, 1 Colombian, and 1 Estonian. Each speaker was evaluated by two examiners and received a passing grade greater than or equal to 75 on a scale of 100.

*2 Classification features*

The machine learning classification features are the 35 segmental and suprasegmental measures specified in Table 1 above.

*3 Classifiers*

Decision tree learning utilizes a decision tree as a predictive model to map observations (e.g., suprasegmental and segmental measures) about an item (e.g., speech file) to conclusions about the item's target value (e.g., English proficiency level). It is a predictive modeling approach frequently found in statistics, data mining and machine learning. Ensemble methods are techniques for improving the performance of machine learning classifiers by constructing more than one classification, or decision, tree. We tested the performance of 17 machine learning ensemble classifiers, as depicted in Table 2.

**Table 2.** Machine learning ensemble classifiers evaluated in the current study.

|  |  |
| --- | --- |
| **Ensemble** | **Configuration** |
| Bagging | Multi-class |
| AdaBoostM2 | Multi-class |
| LPBoost | Multi-class |
| TotalBoost | Multi-class |
| RUSBoost | Multi-class |
| Subspace Discriminant | Multi-class |
| Subspace kNN | Multi-class |
| Bagging | Pairwise coupling |
| AdaBoostM1 | Pairwise coupling |
| LogitBoost | Pairwise coupling |
| GentleBoost | Pairwise coupling |
| RobustBoost | Pairwise coupling |
| LPBoost | Pairwise coupling |
| TotalBoost | Pairwise coupling |
| RUSBoost | Pairwise coupling |
| Subspace Discriminant | Pairwise coupling |
| Subspace kNN | Pairwise coupling |

We utilized two classification configurations: multi-class and pairwise coupling. Multi-class classifiers make a 1-of-n choice. Binary classifiers (i.e., 1-of-2 choices) generally perform better than multi-class ones. González-Ferreras, Escudero-Mancebo, Vivaracho-Pascual, and Cardeñoso-Payo (2012) proposed a method of breaking a multi-class problem into a number of more accurate binary classification problems which they named pairwise coupling.

The Bagging ensembles were implemented with the Matlab *TreeBagger* function (MathWorks, Inc., 2013). The multi-class configuration contained 1,000 decision tree learners while the pairwise coupling configuration contained 100.

The Matlab *fitensemble* function with 100 weak learners was utilized for all of the boosting ensembles (i.e., AdaBoost, LogitBoost, GentleBoost, RobustBoost, LPBoost, TotalBoost, and RUSBoost). Matlab does not have a built-in function for the multi-class configuration of LogitBoost, GentleBoost, or RobustBoost.

*4 Three-fold cross-validation*

We executed a number of experiments to determine the best machine learning classifier and set of segmental and suprasegmental measures for classifying the English proficiency of the CELA corpus. In all the experiments, we applied three-fold cross-validation. Speakers were randomly assigned to folds of 40, divided evenly by gender and proficiency as portrayed in Table 3. More than three folds would have been too many because of the small number of male CPE speakers.

**Table 3.** Distribution of gender and proficiency across cross-validation folds.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Male | | | | Female | | | |
| Fold | CPE | CAE | FCE | PET | CPE | CAE | FCE | PET |
| 1 | 1 | 4 | 4 | 5 | 6 | 8 | 7 | 5 |
| 2 | 2 | 3 | 4 | 5 | 5 | 8 | 7 | 6 |
| 3 | 2 | 4 | 3 | 6 | 6 | 7 | 7 | 5 |

*5 Feature selection*

We conducted one experiment for each combination of the 17 ensemble classifiers and feature sets composed of selected segmental and suprasegmental measures, i.e., features. The number of feature combinations taken one-at-a-time, two-at-a-time, etc. is 3.44 x 1010. Some of the classifiers are sensitive to the order in which the features are presented to them, increasing the number of feature sets to explore to 2.81 x 1040. Thus, an exhaustive search of the feature space is impractical. Therefore, we employed three techniques of feature selection: 1) take-away/add-in, 2) simulated annealing (Kirkpatrick & Vecchi, 1983; Černý, 1985), and 3) a genetic algorithm (GA). Each of these feature selection techniques is described in the following sections.

*a Take-away/add-in feature selection*

The take-away/add-in technique is a greedy algorithm that begins with an initial seed set of features (i.e., segmental and suprasegmental measures). The first step is to calculate a baseline performance number in terms of the Pearson’s correlation (*r*) between the computer classifier proficiency ratings using all the features and the human ratings. The next step is to take each of the features away from the seed set one-at-a-time. If the resulting feature set improves the performance of the classifier (in terms of Pearson’s correlation), then the feature is left out; otherwise it is added back into the set. Next, the remaining features (i.e., those not in the seed set) are added to the seed set one-at-a-time. If the resulting set improves the classifier performance, it is left in the set; if not, it is taken back out. Thus, for each classifier, 35 three-fold cross-validations were performed per seed. At the end of this process, if the performance for the final resulting set is better than the baseline performance, or the performance is the same and the final set of features is different, then the entire process is repeated with the final set replacing the seed set and the final performance number becoming the baseline.

This technique is highly dependent on the initial seed set. We defined three different initial seeds. The first initial seed is the measures in Figure 10, which have an Out-Of-Bag Feature Importance greater than 0.2: syllables per second (SMSYPS), articulation (SMARTI), run length (SMRNLN), and prominent syllables per second (SMPACE). We selected the measures that have a linear relationship with the proficiency ratings for the second initial seed: syllables per second (SMSYPS), articulation (SMARTI), phonation time ratio (SMPHTR), prominent syllables per second (SMPACE), run length (SMRNLN), silent pause length (SMSPLN), proportion of prominent syllables (SMSPAC), and prominence characteristics (SMPCHR). The strategy for determining whether or not a relationship is linear was based on the 2-tailed significance test for the Pearson *r* correlation coefficient (p < 0.01). Table 4 lists the Pearson correlation coefficient (*r*) and 2-tailed significance test (*p*) for each measure.



**Figure 10.** Out-of-Bag feature importance of segmental and suprasegmental measures

**Table 4.** Pearson correlation coefficient (*r*) between each measure and the accent ratings and 2-tailed significance test (*p*) sorted by *p*.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure** | ***r*** | ***p*** |  | **Measure** | ***r*** | ***p*** |  | **Measure** | ***r*** | ***p*** |
| SMSYPS | -0.590 | 0.00 |  | SMGIVP | -0.146 | 0.11 |  | SMRISM | 0.074 | 0.42 |
| SMARTI | -0.571 | 0.00 |  | SMNEWP | -0.138 | 0.13 |  | SMFRSL | -0.060 | 0.51 |
| SMPHTR | -0.406 | 0.00 |  | SMFPLN | 0.131 | 0.16 |  | SMRFAH | -0.055 | 0.55 |
| SMPACE | -0.399 | 0.00 |  | SMRFAL | -0.129 | 0.16 |  | SMNEUH | 0.038 | 0.68 |
| SMRNLN | -0.378 | 0.00 |  | SMAVPP | -0.126 | 0.17 |  | SMFRSM | -0.038 | 0.68 |
| SMSPLN | 0.370 | 0.00 |  | SMPPLN | 0.117 | 0.20 |  | SMPARA | 0.038 | 0.68 |
| SMSPAC | 0.345 | 0.00 |  | SMNEUM | 0.115 | 0.21 |  | SMTPTH | 0.037 | 0.69 |
| SMPCHR | -0.303 | 0.00 |  | SMFPRT | 0.105 | 0.26 |  | SMNEUL | 0.037 | 0.69 |
| SMSPRT | 0.206 | 0.02 |  | SMAVNP | -0.096 | 0.29 |  | SMRFAM | 0.029 | 0.76 |
| SMPRAN | -0.200 | 0.03 |  | SMRISL | 0.084 | 0.36 |  | SMFALH | -0.026 | 0.78 |
| SMFALM | 0.160 | 0.08 |  | SMFALL | -0.082 | 0.37 |  | SMOPTH | 0.018 | 0.84 |
| SMRISH | 0.155 | 0.09 |  | SMFRSH | -0.075 | 0.41 |  |  |  |  |

For the third initial seed, we chose the most heavily weighted measures in the first principle component from Principle Component Analysis (PCA): mean height of onset pitch (SMOPTH), mean height of terminating pitch (SMTPTH), overall pitch range (SMPRAN), mean pitch of new lexical items i.e., prominent syllables (SMNEWP), mean pitch of given lexical items i.e., prominent syllables (SMGIVP), mean prominent syllable pitch (SMAVPP), and mean non-prominent syllable pitch (SMAVNP). Since the second and third principal components contained the same measures as the first principle component, no additional initial seeds could be derived. The first three principal components explained 93% of the variance and so none of the other principle components were analyzed

*b Simulated annealing feature selection*

The pseudo code for the simulated annealing feature selection algorithm is:

*m* ← *m0*

For *i* = [0, …, *imax*-1]

*t* ← *ft*(*i* ∕ *imax*)

*mnew* ← *fn*(*m*)

If *fp*(*fe*(*m*), *fe*(*mnew*), *t*) > *rrandom*

*m* ← *mnew*

Output: *m*

where,

*m* = {set of measures from Table 1}

*m0* = {initial set of measures from Table 1}

*imax* = maximum number of iterations (1000)

*ft*(*x*) = 1 – *x*

*fn*(*x*) = transformation of *x* by randomly adding, deleting, or changing one measure

*fe*(*x*) = 1 – (Pearson’s *r* between proficiency ratings determined by classifier and by humans)

*fp*(*x*, *y*, *z*) =

*rrandom* ∈ *R*

*R* ~ *U*(0,1)

Three initial sets of measures (*m0*) were tested for the simulated annealing feature selection algorithm. The first *m0* was the best set of measures chosen during the take-out/add-in feature selection process: prominence characteristics (SMPCHR), neutral-mid rate (SMNEUM), fall-rise-low rate (SMFRSL), fall-rise-mid rate (SMFRSM), rise-fall-high rate (SMRFAH), average non-prominent syllable pitch (SMAVNP), articulation (SMARTI), and filled pause rate (SMFPRT). A random set of 18 measures from Table 1 was the second *m0*: space (SMSPAC), pace (SMPACE), rise-mid rate (SMRISM), neutral-low rate (SMNEUL), neutral-high rate (SMNEUH), fall-mid rate (SMFALM), fall-rise-low rate (SMFRSL), fall-rise-high rate (SMFRSH), rise-fall-mid rate (SMRFAM), overall pitch range (SMPRAN), average prominent syllable pitch (SMAVPP), syllables per second (SMSYPS), mean run length (SMRNLN), silent pause rate (SMSPRT), mean silent pause length (SMSPLN), mean onset pitch height (SMOPTH), mean paratone pause length (SMPPLN), and mean pitch of given lexical items (SMGIVP). The 17 measures in Table 1 that were not in the second *m0* comprised the third *m0*: prominence characteristics (SMPCHR), rise-low rate (SMRISL), rise-high late (SMRISH), neutral-mid rate (SMNEUM), fall-low rate (SMFALL), fall-high rate (SMFALH), fall-rise-mid rate (SMFRSM), rise-fall-low rate (SMRFAL), rise-fall-high rate (SMRFAH), average non-prominent syllable pitch (SMAVNP), average prominent syllable pitch (SMPHTR), articulation rate (SMARTI), filled pause rate (SMFPRT), mean filled pause length (SMFPLN), paratone rate (SMPARA), mean terminating pitch height (SMTPTH), and mean pitch of new lexical items (SMNEWP).

We set the maximum number of iterations (*imax*) to 1,000 after an initial test indicated that it produced significantly better results than 100 iterations. We judged its superiority by comparing the mean correlation (Pearson’s *r*) between the machines proficiency ratings and those of humans of all 17 classifiers. For 100 iterations, the mean correlation was 0.539 while it was 0.588 for 1,000 iterations.

*c Genetic algorithm feature selection*

A genetic algorithm (GA) is a search heuristic which solves optimization problems utilizing methods motivated by natural evolution, such as inheritance, mutation, selection, and crossover. In these experiments, each generation consisted of an initial population of five sets of segmental and suprasegmental measures, the recombination of those sets, and the mutation of those sets. The recombination was performed by taking the union and intersection of each pair of sets, resulting in at most 20 sets. (Null sets and duplicate sets were not evaluated and thus would reduce the number of sets from 20). The mutation was carried out by randomly changing, deleting, or adding a measure in the original five sets, ten union sets, and the ten intersection sets, leading to potentially 25 mutated sets. The fitness of each generation (possibly 50 sets) was assessed with the Pearson correlation between the classifier’s proficiency ratings and the human judged ones. The five with the highest correlation became the initial population for the next generation.

We began the first generation with five different initial populations. The first one was created by randomly dividing the 35 measures into five sets (Table 6). The theory behind the second one (Table 7) was to use the five best sets from the take-away/add-in and simulated annealing selection processes, but the simulated annealing algorithm did not produce any of the top five. The third, fourth, and fifth sets were the best five from the experiments with the second, third, and fourth sets. The final two sets produced the same results, so we stopped the experiments. We ran the genetic algorithm for 50 generations each time.

**Table 6.** First generation initial populations for genetic algorithm (five random sets)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SMPPLN  SMSPAC  SMFALM  SMSYPS  SMFALL  SMFRSL  SMPHTR | SMFRSH  SMAVNP  SMNEUM  SMAVPP  SMARTI  SMRFAL  SMNEUH | SMFRSM  SMFALH  SMRISL  SMPRAN  SMRISH  SMPACE  SMFPLN | SMFPRT  SMSPLN  SMPARA  SMGIVP  SMRISM  SMRFAH  SMNEUL | SMTPTH  SMRFAM  SMRNLN  SMSPRT  SMPCHR  SMOPTH  SMNEWP |

**Table 7.** First generation initial populations for genetic algorithm (best sets from take-out/add-in)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SMPCHR SMNEUM SMFRSL SMFRSM SMRFAH SMAVNP SMARTI SMFPRT | SMRISH SMNEUL SMFRSH SMRFAH SMPRAN SMAVNP SMAVPP SMSYPS SMARTI SMFPRT SMPARA SMPPLN SMGIVP SMOPTH | SMRISH SMNEUL SMFRSH SMPRAN SMAVNP SMAVPP SMSYPS SMARTI SMFPRT SMPARA SMOPTH SMPPLN SMGIVP | SMSPAC SMPCHR SMPACE SMRISM SMRISH SMNEUM SMFALL SMFRSM SMPRAN SMARTI SMFPRT SMPPLN SMNEWP | SMPCHR SMPACE SMNEUM SMFRSL SMFRSM SMRFAH SMAVNP SMARTI SMFPRT |

**IV Results**

Determining the most appropriate machine learning classifier and group of segmental and suprasegmental measures for automating English proficiency scoring of unconstrained speech is the objective of the study described here. To that end, we assessed the operation of 17 machine learning classifiers. For each classifier, we examined a number of combinations of features taken from a super set of segmental and suprasegmental measures grounded on Brazil’s (1997) prosody model. Take-out/add-in, simulated annealing, and a genetic algorithm feature selection procedure were employed to choose the feature sets. The correlation between the computer’s calculated proficiency ratings and those scored by humans determined which machine learning classifier and group of segmental and suprasegmental measures was the most effective. In total, we performed 493 experiments with thousands of combinations of measures. Table 8 shows the results of the top 10 experiments in terms of correlation. The feature selection method for the top 10 experiments was the genetic algorithm

**Table 8.** Results of the top 10 experiments sorted by correlation between human and computer

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Best Measures** | **Best *r***  ***p* < 0.01** | **Run** | **Seed** |
| Pairwise GentleBoost | SMPCHR SMNEUL SMNEUH SMFRSM SMPRAN SMSYPS SMARTI SMFPLN SMOPTH SMGIVP SMFALH | 0.677 | 4 | Note 1 |
| Pairwise TreeBagger | SMPCHR SMPACE SMNEUM SMFRSL SMFRSM SMAVNP SMARTI SMFPRT SMGIVP | 0.675 | 3 | Note 1 |
| Multi-class LPBoost | SMSPAC SMPCHR SMPACE SMNEUL SMNEUM SMNEUH SMFALL SMFALH SMFRSL SMFRSM SMFRSH SMRFAH SMPRAN SMAVNP SMARTI SMFPRT SMOPTH SMPPLN | 0.675 | 2 | Note 2 |
| Pairwise GentleBoost | SMPCHR SMNEUL SMFRSM SMRFAH SMPRAN SMNEUH SMSYPS SMARTI SMFPLN SMGIVP | 0.670 | 3 | Note 1 |
| Pairwise AdaBoostM1 | SMPCHR SMRISM SMRISH SMNEUL SMFALL SMFALH SMFRSM SMRFAH SMPRAN SMAVNP SMSYPS SMARTI SMFPRT SMPARA | 0.670 | 5 | Note 1 |
| Pairwise LogitBoost | SMPCHR SMRISM SMRISH SMNEUL SMNEUH SMFALL SMFALH SMFRSM SMFRSH SMPRAN SMAVNP SMSYPS SMARTI SMFPRT SMPARA SMGIVP | 0.670 | 5 | Note 1 |
| Pairwise LogitBoost | SMPCHR SMRISM SMRISH SMNEUL SMNEUH SMFALL SMFALH SMFRSM SMFRSH SMRFAH SMPRAN SMAVNP SMSYPS SMARTI SMFPRT SMPARA SMGIVP | 0.669 | 4 | Note 1 |
| Pairwise AdaBoostM1 | SMPCHR SMRISM SMRISH SMNEUL SMFALL SMFALH SMFRSM SMFRSH SMPRAN SMAVNP SMSYPS SMARTI SMFPRT SMOPTH SMRFAH | 0.669 | 3 | Note 1 |
| Multi-class Subspace Disc | SMSPAC SMPCHR SMRISH SMNEUL SMFALH SMFRSM SMFRSH SMPRAN SMSYPS SMARTI SMFPRT SMFPLN SMAVPP | 0.665 | 3 | Note 1 |
| Pairwise RobustBoost | SMPCHR SMRISM SMNEUL SMNEUH SMFALL SMFALM SMFALH SMFRSM SMFRSH SMRFAH SMPRAN SMSYPS SMARTI SMFPRT SMOPTH SMGIVP | 0.665 | 5 | Note 1 |

The findings in Table 8 indicate that the optimum machine learning classifier for English proficiency scoring of unconstrained speech is Pairwise GentleBoost utilizing the following segmental and suprasegmental measures: percent of tone units containing at least one prominent syllable (SMPCHR), neutral-low rate (SMNEUL), neutral-high rate (SMNEUH), fall-rise-mid rate (SMFRSM), overall pitch range (SMPRAN), syllables per second (SMSYPS), articulation rate (SMARTI), filled pause mean length (SMFPLN), paratone boundary onset pitch mean height (SMOPTH), given lexical item mean pitch (SMGIVP), and falling-high rate (SMFALH).

From an algorithmic perspective, of the 17 classifiers we tested, the multi-class Multi-class LPBoost, Subspace Disc, and TreeBagger and the pairwise AdaBoostM1, GentleBoost, LogitBoost, RobustBoost, Subspace Disc, and TreeBagger performed the best. The genetic algorithm was clearly the superior feature selection algorithm. Additionally, the boot-strap technique of initially populating the first generation of a genetic algorithm run with the five best sets from the previous run seems to be a successful policy.

**V Discussion**

In this paper, we evaluated the performance of 17 machine learning classifiers in automatically scoring the English proficiency of unconstrained speech. For each of the classifiers, we considered a number of sets of features drawn from a master set of segmental and suprasegmental measures which were derived from elements of Brazil’s (1997) prosody model. The sets of features were chosen by means of three different feature selection algorithms: take-out/add-in, simulated annealing, and genetic algorithm. We assessed the performance of the classifiers in terms of the correlation between the computer’s calculated proficiency ratings and those scored by humans.

The outcomes of our study provide evidence that a computer can automatically score the English proficiency of unconstrained speech with a Pearson’s correlation (r) of 0.677 (p < 0.01) when compared with a human expert.

There have been no other studies of automatically scoring the English proficiency of unconstrained speech utilizing the CELA rating system (i.e., PET, FCE, CAE, and CPE). Even though, cross-corpus comparisons are not always reliable, there are two similar studies where the English proficiency of unconstrained speech was tested in a similar manner and rated on a scale from one to four: (1) SpeechRaterSM (Zechner et al., 2009) and (2) a study of automated scoring of spoken English responses given by non-native children in an English proficiency assessment of middle school students (Evanini & Wang, 2013). The assessment included three different task types intended to measure a student’s ability to converse in English. One of these tasks, Picture Narration, is similar to the CELA test tasks. In the Picture Narration task, the child is presented with six pictures that portray a series of events and is asked to describe what is transpiring in the images. For an automated scoring instrument, they applied linear regression of ten features extracted from the output of an ASR configured to recognize the words spoken by the children. Table 9 compares these two studies with ours in terms of mean (*μc*, *μh*), standard deviation (*σc*, *σh*), standard deviation ratio (*σh*/*σc*), standardized mean difference (*μh - μc*), proportion of agreement (exact - *ae*, adjacent - *aa*, and non-adjacent - *an*), linear-weighted kappa (*κl*), quadratic-weighted kappa (*κq*), and Pearson’s correlation (*r*).

**Table 9.** Comparison of current study with two other similar studies

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Current Study | Zechner et al. (2009)  TPO sm-eval + rec-train sets MR1 | Zechner et al. (2009)  iBT Field Study  MR1 | Zechner et al. (2009)  iBT Field Study  CART1 | Evanini and Wang (2013) |
| *μc* | 2.44 | 2.79 | 2.45 | 2.47 | 2.283 |
| *μh* | 2.62 | n/a | n/a | n/a | 2.253 |
| *σc* | 1.03 | 0.712 | 0.61 | 0.92 | 0.543 |
| *σh* | 1.07 | n/a | n/a | n/a | 0.873 |
| *σh*/*σc* | 1.03 | n/a | n/a | n/a | 1.61 |
| *μh - μc* | 0.18 | n/a | n/a | n/a | -0.03 |
| *ae* | 52.5% | 57.8% | 44.2% | 50.6% | n/a |
| *aa* | 92.5% | 98.4% | 95.1% | 93.3% | n/a |
| *an* | 7.5% | 1.6% | 4.9% | 6.7% | n/a |
| *κl* | 0.53 | 0.33 | 0.51 | 0.59 | n/a |
| *κq* | 0.67 | n/a | n/a | n/a | n/a |
| *r* | 0.68 | 0.37 | 0.55 | 0.62 | 0.624 |
| 1Table 9 (Zechner et al., 2009), except as noted below  2estimated from Table 1 (Zechner et al., 2009)  3Table 6 (Evanini & Wang, 2013)  4Table 5 (Evanini & Wang, 2013) | | | | | |

The metrics in Table 9 are calculated as follows:

*μc* = mean of computer generated ratings (4)

*μh* = mean of human generated ratings (5)

*σc* = standard deviation of computer generated ratings (6)

*σh* = standard deviation of human generated ratings (7)

*ae* = % of ratings where computer and human matched exactly (8)

*aa* = *ae* + % of ratings where computer and human matched ± 1 of each other (9)

*an* = 1 - *aa* (10)

*sc* = computer score (11)

*sh* = human score (12)

*κl* = Cohen’s (1968) weighted kappa, where *wij* = | *sc* – *sh* | (13)

*κq* = Cohen’s (1968) weighted kappa, where *wij* = (*sc* – *sh*)2 (14)

*r* = Pearson’s correlation (15)

Comparing Pearson’s correlation leads to the conclusion that the classifier and features utilized in the current study are significantly better at scoring the English proficiency of unconstrained speech than those tested in either of the other two studies. The better standard deviation ratio (*σh*/*σ*) of the current study implies that the classifier and features employed in the current study are better than those Evanini and Wang (2013) made use of in their study. Examining accuracy (i.e., *ae*, *aa*, and *an*) indicates the results of Zechner’s et al. (2009) laboratory tests (i.e., TPO sm-eval + rec-train sets MR) are better than the results of the current study in all accuracy metrics. However, the current study is better than their field tests in exact accuracy (*ae*), but not adjacent (*aa*) nor non-adjacent (*an*) accuracy. The linear-weighted kappa (*κl*) reveals the current study is better than either the laboratory or field tested MR classifier, but not the field tested CART classifier.

One reason for the better performance of the current study’s classifier may be related to the inclusion of suprasegmental measures in the features set. Both Zechner (2009) and Evanini and Wang (2013) calculated the automated scores primarily from segmental features, whereas the best feature set report for the current study contained nine suprasegmental measures out of eleven measures. This is consistent with the findings of Author et al. (2010) that suprasegmental measures explained 50% of the variance in oral proficiency ratings.

**VI Conclusion**

The findings reported here offer empirical evidence that a Pairwise GentleBoost classifier and a set of features rich in suprasegmental measures can automatically score the English proficiency of unconstrained speech better than other methods of automatic scoring. Possible next steps in this research include improving the algorithms that produce the underlying numbers that are drawn on to calculate the segmental and suprasegmental measures, specifically: silent pause detection, filled pause detection, tone unit detection, syllabification, prominent syllable detection, and tone choice classification.

Another promising area to investigate is the application of Brazil’s (1997) model to automatically scoring the dialogic aspects of English proficiency. This is an interesting area because Brazil’s framework is especially strong in modelling dialogs. Specifically, in an interactive dialog between two persons, Brazil’s model includes pitch concord, which is matching the relative pitch of the key and termination prominent syllables between two speakers. For example, high pitch on the termination of one speaker’s utterance anticipates a high pitch on the key of the next speaker’s utterance, while mid termination anticipates a mid key. There are no expectations for low termination. Pitch concord is a strong indicator of language proficiency in non-native speakers (Pickering, 1999). In this line of research, the computer would calculate English proficiency from interactive suprasegmental measures. Interactive suprasegmental measures are a superset of the suprasegmental measures employed in the current research with additional measures specific to dialogs. Author and colleague (2014) found these to be highly correlated with human-judged accentedness and comprehensibility measures.

The results in this paper affirm the potential of using an alternative method of automatically scoring unconstrained speech using a phone ASR and set of segmental and suprasegmental measures based on Brazil’s model.

**References**

Authors (accepted)

Authors ( under review)

Author (2010)

Author and colleagues (2014)

Author and colleagues (2010)

Author and colleagues (2015)

Author and colleagues (2014)

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*.

Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge, England: Cambridge University Press.

Brown, A. (2000) Tongue slips and Singaporean English pronunciation. *English Today*, 16 (3), 31–6

Cambridge English Language Assessment (2015). [www.cambridgeenglish.org](http://www.cambridgeenglish.org), Retrieved March 29, 2015.

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of optimization theory and applications*, 45(1), 41-51.

Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice* (Vol. 1). John Benjamins Publishing.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4), 213.

Crystal, D. (2003). *English as a global language*. Cambridge: Cambridge University Press.

Evanini, K., & Wang, X. (2013). Automated speech scoring for non-native middle school students with multiple task types. In *INTERSPEECH* (pp. 2435-2439).

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93, 27403.

González-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., & Cardeñoso-Payo, V. (2012). Improving automatic classification of prosodic events by pairwise coupling. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(7), 2045-2058.

Hamp-Lyons, L., & Davies, A. (2008). The English of English tests: Bias revisited. *World Englishes*, 27 (1), 27-39.

Jenkins, J. (2003) *World Englishes: A Reference Book for Students*. New York: Routledge.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *English Language Teaching Journal*, 60 (1), 42-50.

Kachru, B. B. (1986). *The alchemy of English: The spread, functions, and models of non-native Englishes*. University of Illinois Press.

KayPENTAX (2008). *Multi-Speech and CSL Software*. Lincoln Park, NJ: KayPENTAX.

Kirkpatrick, S., & Vecchi, M. P. (1983). Optimization by simmulated annealing. *science*, 220(4598), 671-680.

Longman, P. (2013). Official guide to Parson test of English academic.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, 36(2), 173–190.

MathWorks, Inc. (2013). MATLAB Release 2013a. [Computer program]. Retrieved February 15, 2013.

Modiano, Marko (2001) Linguistic imperialism, cultural integrity, and EIL. *ELT Journal*, 55, 339 – 346.

Pearson Education, Inc. (2015). *Versant Spanish Test*. Retrieved from <http://www.versanttest.com/products/spanish.jsp>.

Pickering, L. (1999). *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants* (Doctoral dissertation, University of Florida).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesel, K. (2011). The Kaldi speech recognition toolkit.

Sakti, S., Markov, K., Nakamura, S., & Minker, W. (2009). *Incorporating Knowledge Sources into Statistical Speech Recognition* (Vol. 42). Springer.

Yano, Y. (2001). World Englishes in 2000 and beyond. *World English*, 20, 119-131.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10), 883-895.