Johnson. D. O., Kang, O., & Ghanem, R. (2016). Language proficiency ratings: Human vs. machine. In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds). *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Dallas, TX, October 2015 (pp. 119-129). Ames, IA: Iowa State University.

LANGUAGE PROFICIENCY RATINGS: HUMAN VS. MACHINE

David O. Johnson, Northern Arizona University
Okim Kang, Northern Arizona University
Romy Ghanem, Northern Arizona University

This paper explains a computer model that mechanically assesses the verbal proficiency of audio recordings of unconstrained non-native English speech. The computer model utilizes machine learning and eleven suprasegmental measures split into four categories (stress, pitch, pause, and temporal) to compute the proficiency levels. In an experiment with 120 non-native English speaker's monologs from the speaking section of the Cambridge ESOL General English Examinations, the Pearson's correlation comparing the certified Cambridge English Language Assessment proficiency scores and the computer's computed proficiency scores was 0.718. This human-computer correlation is greater than that of other related computer programs (0.55-0.66) and is nearing that of human examiners (0.70-0.77) with regards to inter-rater reliability.

INTRODUCTION

Language proficiency assessments are intended to measure reading, writing, listening, and speaking abilities. Humans can score proficiency assessments; but they are costly to employ, train, and compensate; they take a long time to score assessments which produces postponements in providing the results to the candidates; and still with multiple raters, rubrics, and frequent inter-rater reliability testing, humans lack consistency and objectivity. For example, Kang and Rubin (2009) found that listener's attitudinal and background factors accounted for 18-23 % of the variance in human assessment. Innovations in artificial intelligence and natural language processing have resulted in computer programs that can automatically rate language proficiency. Automated scoring systems generate assessments quicker and more economically than human scoring and they are more consistent and equitable in scoring than humans. This is especially true with automated delivery and rating of reading, writing, and listening skills (Attali & Burstein, 2006; Burstein et al., 1998; Chodorow & Burstein, 2004; Landauer & Dumais, 1997; Rudner, Garci, & Welch, 2006; Zechner, Higgins, Xi, & Williamson, 2009). Automated reading and listening assessments are characteristically multiple-choice. They are simple to create and manage, comparatively uncomplicated to grade mechanically, and substantiated by a robust foundation of assessment philosophy and statistical practices. Automated writing tests are usually delivered online and scored automatically. They are written constructed response items where the examinees write a succession of compositions on designated subjects.

Automatic Speaking Proficiency Assessment

Speaking skill assessment is more difficult than other assessments. There are two categories for automated scoring systems in speech: constrained and unconstrained (spontaneous). Constrained speech assessment is the easier of the two to automate. Typically test-takers are requested to respond orally to constructed response items like reading aloud, repeating sentences, building sentences, giving short answers to questions, or retelling brief stories. For some tasks, one correct word sequence is expected for each response. In other tasks, items can have multiple correct answers. The computer recognizes the words spoken with an automatic speech recognizer (ASR) and compares them to the hypothesized response (content). It locates linguistic units (segments, syllables, and words) and measures the pace, fluency, and pronunciation of those words in phrases and sentences (prosody). Then, the computer combines the content and prosodic measures using statistical modeling techniques and calculates an overall score as a weighted combination of the sub-scores. Their use in evaluating constrained speech proficiency has been confirmed by establishing that the automated scores were substantially correlated with those that human raters ascertained from speaking proficiency examinations (Bernstein, Van Moere, & Cheng, 2010).

Existing Automatic Unconstrained Speaking Proficiency Assessment

Unlike constrained speech, unconstrained speech is irregular and variable making automatic proficiency scoring of it more challenging. Asking candidates to converse on a subject for one or two minutes (e.g., what is happening in a picture) is the normal means of obtaining unconstrained speech samples to assess. SpeechRater[SM] is an instance of an operational computerized unconstrained English speech proficiency assessment tool (Zechner et al., 2009). As illustrated in Figure 1, SpeechRater[SM] detects the words in the candidate's speech with an ASR.

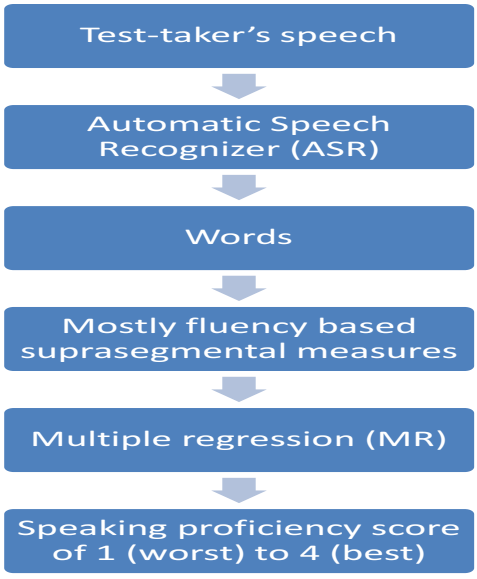

Figure 1. SpeechRater[SM].

It then uses the output from the ASR to compute eleven prosodic measures: average chunk length (in words), where a chunk is segment of contiguous words, articulation rate, mean deviation of chunks (in words), total duration of silent pauses divided by number of words, average silent pause duration (in seconds), average of long silent pause (greater than or equal to 500 ms) duration, frequency of long silent pauses divided by number of words, types of unique words per second, number of types divided by duration of entire transcribed segment exclusive of inter-utterance pauses, normalized global HMM acoustic model score, and normalized global language model score. The eleven measures are then combined with multiple-regression to estimate a speaking proficiency rank of one (lowest) to four (highest). The Pearson's correlation between the ranks assessed by a human and those estimated by SpeechRater^{SM} was 0.55. A Classification and Regression Tree (CART) machine learning version, which was not deployed, had a stronger correlation of 0.62.

New Computer Model for Automatically Scoring Unconstrained Speech Proficiency

We developed a computer model that automatically scores unconstrained English speech proficiency from suprasegmental measures derived from Brazil's (1997) prosody model. The computer programs calculate the suprasegmental measures from the output of an ASR that recognizes phones instead of words. As depicted in Figure 2, in contrast to the method that SpeechRater^{SM} employed, our method has three benefits.

Figure 2. Comparison of SpeechRater^{SM} and our method of automatic proficiency scoring of unconstrained speech.

The first benefit is a consequence of the ASR recognizing phones instead of words. This is because the ASR only has to recognize the relatively tiny number of phones that are used in English words as opposed to recognizing the hundreds of thousands of words that could appear

in unconstrained speech. Since there is a lesser quantity of phones to recognize than words by several orders of magnitude, the phone error rate (PER) of an ASR is predictably less than the word error rate (WER). This lower PER can lead to more correct proficiency scores. The second benefit of our tactic is making use of, along with fluency features, intonational measures drawn from a larger set of suprasegmental measures which were found to explain more than half of the variance in speaking proficiency scores (Kang, Rubin, & Pickering, 2010). Utilizing machine learning, which is a sub-domain of artificial intelligence, results in the third benefit of our approach. Machine learning normally achieves better results than multiple-regression. The fact that Zechner et al. (2009) built a machine learning version (i.e., CART) of SpeechRater^{SM} which had a higher correlation between machine and human proficiency scores than their multiple-regression version is evidence of this.

We begin this paper with an overview of Brazil's (1997) prosody model and a description of the corpus and experimental methods we used to test the computer model we developed to automatically score the English proficiency of unconstrained speech. Then, we report the results and discuss them. We finish with a conclusion and some areas for further study.

METHODS

Brazil's Prosody Model

One of the earliest to put forward the notion of discourse intonation was Brazil (1997). He defined intonation as the linguistically deliberate variation of oral pitch intensity and duration throughout a discourse to relay information beyond that conveyed by the words and grammar. He held forth that the communication purpose of a discourse was realized by the recurring and purposeful selection of one pattern of intonation from an array of patterns. Brazil's model did not require additional phonological or acoustic classifications of the pitch attribute of speech which earlier intonation models had required. Nevertheless, his model assigned fresh inferences and connotations to orthodox intonation components (Chun, 2002). His model is regularly made use of in learning and teaching a language for the reason that it is founded on the use of intonation in a discourse to accomplish linguistic objectives that reach beyond the sentence level. He maintained that the four principal features of his model, i.e., tone unit, prominence, tone choice, and relative pitch, offered a practical structure for examining and studying the use of intonation that speakers exercised in a discourse. The main features of his model remain true for every facet of discourse; whether it is a dialog or a monolog consisting of either unconstrained or constrained speech.

Brazil characterized a tone unit as a fragment of a speech that a listener can perceive has an arrangement of falling and rising tones which is not the same as the arrangement of another fragment of the speech (Brazil, 1997). Then he stated that all tone units include a minimum of one prominent syllable. Chun (2002) added that syllables become prominent by being accentuated with extra pitch (fundamental frequency in Hz), intensity (amplitude in dB), duration (length in seconds), or a mixture of the three. Brazil insisted that prominence was ascribed to the syllable, and not the word. Brazil differentiated prominence from lexical stress. Lexical stress is the normal, or dictionary defined, stress applied to syllables within a word. In opposition, prominence is the application of supplementary pitch, intensity, or duration on a syllable, even if it is lexically stressed, to call attention to a word's importance or to recognize its difference. The

initial prominent syllable is called the key and the last is called the termination. A solitary prominent syllable in a tone unit is considered both the key and termination. The arrangement of falling and rising intonation of a tone unit is characterized by the relative pitch of the key and termination syllables and the tone choice of the termination syllable. Brazil divided the pitch range of an utterance into three uniform dissections: low, mid, and high. The relative pitch of a prominent syllable was defined as the dissection in which its pitch resided. The tone choice of the termination syllable was specified by whether its pitch contour was rising, falling, level, rising then falling (rise-fall), or falling then rising (fall-rise).

Cambridge English Language Assessment (CELA) Corpus

The CELA corpus consists of 120 speech files of non-native English speaker's monologs from the speaking part of the Cambridge ESOL General English Examinations, which was previously used in Kang (2013). The speakers represented 21 first languages: 16 Spanish/Mexican, 11 Korean, eight Italian, seven Dutch, six French, five each of Chinese and Russian, four each of Greek, Portuguese, and Swedish, three German, two each of Swiss and Japanese, and one each of Arabic, Austrian, Bolivian, Brazilian, Bulgarian, Colombian, Estonian, and Turkish. Table 1 describes the Common European Framework of Reference for Languages (CEFR) proficiency level each of the speakers had attained, the equivalent Cambridge proficiency level, the number and gender of the speakers, and a description of the monologs they spoke.

Table 1

Cambridge English Language Assessment (CELA) Corpus

CEFR Proficiency Level	Cambridge Proficiency Level	Males	Females	Subject Of Monologues
B1	Preliminary English Test (PET)	16	16	The speaker is given a color photograph to discuss for one minute.
B2	First Certificate in English (FCE)	11	21	The speaker is provided with two photographs to talk about for one minute.
C1	Certificate in Advanced English (CAE)	11	23	The speaker selects two of three pictures and explains what is happening in the pictures for one minute.
C2	Certificate of Proficiency in English (CPE)	5	17	The speaker converses about a question from a card with various ideas on it for two minutes.

Automatic Scoring of English Speaking Proficiency of Unconstrained Speech

The English proficiency for a speaker is scored by the computer in three stages: (1) process the speech file to ascertain silent pauses, filled pauses, syllables, and the elements of Brazil's (1997) prosody model (i.e., tone units, prominent syllables, tone choices, and relative pitches); (2) compute 35 suprasegmental measures from the amounts and intervals of silent pauses, filled pauses, syllables, and the elements of Brazil's (1997) prosody model; and (3) utilize machine learning to analyze the suprasegmental measures and determine a proficiency score: B1, B2, C1, and C2. The following sections specify each of these stages.

Stage 1: Ascertain the Underlying Variables of the Suprasegmental Measures

A comprehensive discussion about ascertaining the underlying variables of the suprasegmental measures can be found in published articles (Johnson & Kang, 2015a; Johnson & Kang, 2015b) and manuscripts (e.g., Kang & Johnson, under review), which are currently under review for publication in other venues.

Stage 2: Compute the Suprasegmental Measures

Thirty-five suprasegmental measures shown in Table 2 are computed for each utterance based on the time intervals and amounts of silent pauses, filled pauses, syllables, and the four elements of Brazil's (1997) prosody model.

Table 2

Suprasegmental Measures

*Articulation rate	High-fall rate
Phonation time ratio	*Low-fall rate
Tone unit average length	*Mid-fall rate
*Syllable rate	*High-fall-rise rate
*Filled pause average duration	Low-fall-rise rate
Filled pause rate	Mid-fall-rise rate
Silent pause average duration	High-level rate
Silent pause rate	*Low-level rate
Prominent syllables per tone unit (i.e., pace)	Mid-level rate
*Percent of tone units with at least one prominent syllable	*High-rise-fall rate
Percent of syllables that are prominent (i.e., space)	Low-rise-fall rate
Overall pitch range	Mid-rise-fall rate
Non-prominent syllable average pitch	High-rise rate
Prominent syllable average pitch	*Low-rise rate
Paratone boundary onset pitch average height	*Mid-rise rate
Paratone boundary rate	Given lexical item mean pitch
Paratone boundary average pause duration	New lexical item mean pitch
Paratone boundary average termination pitch height	

The 35 suprasegmental measures were established from ones made use of in prior research (Brazil, 1997; Derwing, 1990; Derwing & Munro, 2001; Hincks, 2005; Kang et al., 2010; Kormos & Denes, 2004; Levis & Pickering, 2004; Pickering, 2004; Wennerstrom, 2001; Wichmann, 2000).

Stage 3: Utilize Machine Learning to Determine a Proficiency Score

In the final stage, a boosting ensemble of decision trees receives a subgroup of the suprasegmental measures (designated with an asterisk in Table 2) as input and outputs a proficiency score of B1, B2, C1, or C2. The boosting ensemble of decision trees was tested and trained using three-fold cross-validation of the 120 speech files. Each fold included 40 randomly allocated speakers, divided evenly by gender and proficiency.

The boosting ensemble did not utilize every one of the 35 suprasegmental measures to calculate a speaking proficiency score. The explanations for this are: (1) several of the original variables (i.e., quantities and time spans of silent pauses, filled pauses, syllables, and the four elements of the prosody model) that are utilized to compute the measures could be well correlated, and hence just one of them needs to be taken into account; (2) the measure might possibly not differ sufficiently across proficiency levels to be a suitable predictor; and (3) the original variables might contain inaccuracies, stemming from the intrinsic error rates of the equipment, procedures, and machine learning methods employed to ascertain them, which would make the suprasegmental measure an undependable proficiency prognosticator.

An exhaustive search for the best set of suprasegmental measures would necessitate an unfeasible assessment of 2.81×10^{40} permutations of the suprasegmental measures. To resolve this challenge, a genetic algorithm was utilized. A comprehensive discussion about the genetic algorithm can be found in manuscripts, which are currently under review for publication in other venues.

RESULTS

The objective of this research was to employ a collection of computer programs to automatically rate the oral proficiency of 120 speech files of non-native English examinee monologs from the speaking part of the Cambridge ESOL General English Examinations and to contrast the computer's ratings with the CELA examiners' ratings. The computer produced proficiency ratings of B1, B2, C1, and C2 utilizing the eleven suprasegmental measures shown in Table 3. The computer's proficiency ratings had a Pearson's correlation of 0.718 ($p < 0.01$) with the CELA examiner assigned proficiency ratings.

Table 3

Suprasegmental measures used by computer to rate unconstrained English speaking proficiency

Type	Suprasegmental Measure
Stress	Percent of tone units containing at least one prominent syllable
Pitch	Low-rise rate Mid-rise rate Low-level rate Low-fall rate Mid-fall rate High-rise-fall rate High-fall-rise rate
Pause	Filled pause average duration
Temporal	Syllable rate Articulation rate

DISCUSSION

Although not exactly the same as the CELA corpus, the English proficiency of speakers using unconstrained speech was scored automatically in four levels in two other studies. One of those resulted in SpeechRater^{TM} which is described above (Zechner et al., 2009). In another study, Evanini and Wang (2013) used linear regression of ten features extracted from the output of an ASR configured to recognize the words to automatically score the spoken English responses given by non-native children in an English proficiency assessment of middle school students. The assessment included three different task types intended to measure a student's ability to converse in English. One of these, the Picture Narration task, is similar to the Cambridge 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. The Pearson's correlation between the scores assessed by the humans and those automatically scored was 0.62. This illustrates that the computer correlation of our method exceeds those of other similar computer programs (0.55-0.62). More importantly though, Zechner et al. (2009) reported human inter-rater reliability of 0.77 and Evanini and Wang (2013) reported 0.70 which also shows that the computer model for automatic scoring of unconstrained speech explained herein is nearing that of human raters with respect to inter-rater reliability.

CONCLUSION

In this paper, we presented a computer model for automatically scoring the English proficiency of unconstrained speech. In a test with the CELA corpus, the Pearson's correlation between the automatic scores from the computer model and the scores assigned by two human CELA examiners was 0.718. This correlation is greater than similar computer programs for automatically scoring the proficiency of unconstrained speech and is on the verge of inter-rater reliability of human scoring. The results also imply that stress, pitch, pause, and temporal suprasegmental measures might be the most important with regard to automated English

proficiency scoring systems for unconstrained speech. This has also been shown to be true for human judgement (Kang et al., 2010).

Follow-on research that shows potential is expanding the computer model to automatically score the interactive aspects of English speaking proficiency. This bodes well for the reason that Brazil's (1997) model is markedly strong in elucidating the prosody of dialogs. Besides adding interactive measures to the computer model, augmenting the model with lexical and grammatical measures shows promise, too. A final area for further study is creating L1-specific models.

ABOUT THE AUTHORS

David O. Johnson is a post-doctoral researcher in the Applied Linguistics Speech Laboratory at Northern Arizona University, Flagstaff, AZ, USA. He is currently developing software and computer models to automatically rate English language proficiency. He received his BSEE and MSEE from Kansas State University and his PhD in Computer Science from the University of Kansas. Prior to a post-doctoral research appointment at the Eindhoven University of Technology in the Netherlands, he was an Adjunct Professor in the Computer Science Electrical Engineering department at the University of Missouri – Kansas City. He is interested in natural language processing and human-robot interaction.

Okim Kang is an Associate Professor in the Applied Linguistics Program at Northern Arizona University. Her research concerns aspects of L2 pronunciation, speech perception and production, automated speech scoring, oral language proficiency assessment, language attitudes, and World Englishes.

Romy Ghanem is a 3rd year PhD student in the Applied Linguistics program at Northern Arizona University, Flagstaff, AZ, USA. Her main research interests include speech perception (including nonnative speaker stereotyping and reverse stereotyping) and production, particularly accommodating different interlocutors' linguistic features. She has been working as a coder and trainer on a Speech Automated Recognition project for the past two years under the supervision of Okim Kang and David Johnson.

REFERENCES

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. The Journal of Technology, *Learning and Assessment*, 4(3).

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.

Burstein, J., Kukich, K., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays. *ETS Research Report Series*, 1998(1), i-67.

Cambridge English Language Assessment (2015). www.cambridgeenglish.org, Retrieved March 29, 2015.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1), i-38.

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

Derwing, T. M. (1990). Speech rate is no simple matter. *Studies in Second Language Acquisition*, 12, 303– 313.

Derwing, T. M., & Munro, M. J. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics*, 22, 324–227.

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

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, 33, 575–591.

Johnson, D. O., & Kang, O. (2015a). Automatic prominent syllable detection with machine learning classifiers. *International Journal of Speech Technology*, 18(4), 583-592.

Johnson, D. O., & Kang, O. (2015b). Automatic prosodic tone choice classification of Brazil's intonation model. *International Journal of Speech Technology*, DOI: 10.1007/s10772-015-9327-z.

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, 94(4), 554-566.

Kang, O. (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels B1 to C2 and examinee L1 backgrounds. *Research Notes, 52, 40-48*. http://www.cambridgeenglish.org/images/139525-research-notes-52-document.pdf

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32, 145–164.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, 32, 505–524.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purp*oses, 23, 19–43.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Wennerstrom, A. (2001). *The music of everyday speech*. New York: Oxford University Press.

Wichmann, A. (2000). *Intonation in text and discourse*. London: Longman.

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.