9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Automated English Proficiency Scoring of Unconstrained Speech Using Prosodic Features

Okim Kang[1], *David O. Johnson*[2]

[1]Northern Arizona University, USA
[2]University of Kansas, USA
okim.kang@nau.edu, davidojohnson@ku.edu

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 groups of features drawn from a master set of prosodic measures founded in Brazil's model [3]. The prosodic measures were calculated from the output of an ASR that recognizes phones instead of words and other software designed to detect the elements of Brazil's prosody model. The performance of the best classifier was 0.68 ($p < 0.01$) in terms of the correlation between the computer's calculated proficiency ratings and those scored by humans. Using only prosodic features, this correlation is in the range of other similar computer programs for automatically scoring the proficiency of unconstrained speech.

Index Terms: Brazil's prosody model, automatic speech recognition (ASR), World Englishes, large vocabulary spontaneous speech recognition (LVCSR)

1. Introduction

The growth in the use of automated scoring is because of the capability of computer systems to generate scores more swiftly and less expensively than human scoring. Automated scoring systems are also more consistent and equitable in scoring than humans. Automated scoring systems for speech can be partitioned into two categories: those that are intended for scoring constrained speech and those intended to score unconstrained speech. The Versant Spanish Test [24] and the speaking tasks within the Pearson Test of English [20] are examples of successful automated scoring systems for constrained speech. Bernstein, Van Moere, and Cheng offered proof that the automated scores from the Versant Spanish Test and the Pearson Test of English were highly correlated with scores from trained human examiners, which substantiates using them to assess a person's proficiency in constrained spoken communication [1].

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 [26]. 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 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 tests of SpeechRaterSM, the correlation between its scores and those of a human was higher at 0.55. A version of SpeechRaterSM, which was not deployed, that used classification and regression trees (CART) had even higher correlations: 0.62 (field test). In more recent work, Loukina, Zechner, Chen, and Heilman [21], reported higher correlations ($r = 0.649$-0.667) using a linear regression scoring model.

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 prosodic measures calculated from the output of the ASR and other software designed to detect the elements of Brazil's prosody model [3]. This approach is important because prosodic properties have been found to account for 50% of the variance in oral proficiency ratings [16]. The purpose of the research is to ascertain the optimum machine learning classifier and collection of prosodic measures for automating English proficiency scoring of unconstrained speech.

2. Method

2.1. Corpora

This research made use of three speech corpora: TIMIT, World Englishes, and CELA.

2.1.1. TIMIT

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 [8]. The speakers read text made up of two dialect sentences, 450 phonetically-compact sentences, and 1890 phonetically-diverse sentences. We labeled the prominent syllables and tone choices in 839 utterances using the syllable boundaries. Two trained linguists labeled the prominent syllables and tone choices by utilizing Praat [2] with inter-rater reliability of 87%.

2.1.2. World Englishes corpus

The World Englishes corpus 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 (British, American, Indian, South African, Mexican, and Chinese) [15] [17]. The World English orthographic transcriptions included silent pause and tone unit labeling. The silent pause and tone unit labels were utilized to developed the silent pause recognition and tone unit division algorithms [13].

10.21437/SpeechProsody.2018-125

2.1.3. Cambridge English Language Assessment (CELA)

The CELA corpus consists of audio recordings of 120 non-native English speakers' monologues from the oral section of the CELA [4]. We utilized this corpus because the Cambridge ESOL General English Examinations is a set of criteria and assessments for evaluating English proficiency that is accepted worldwide. The CELA use a four level proficiency scoring system of PET (lowest, 32 files), FCE (32 files), CAE (34 files), and CPE (highest, 22 files) which is equivalent to the Common European Framework of Reference for Languages (CEFR) B1, B2, C1, and C2, respectively. 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.

2.2. Classification features and outputs

The machine learning classification features (i.e., inputs) are a subset of the 35 prosodic measures specified in categories of speech rates, pauses, stress, pauses, tone, pitch range, and paratone (see Table 1). The output of the machine learning classifiers is one, two, three, or four corresponding to the proficiency levels of PET (lowest), FCE, CAE, or CPE (highest). The 35 suprasegmental measures were drawn from a variety of prosodic studies [3] [6] [12] [16] [19]. 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.

2.3. Classifiers and three-fold cross-validation

For this study, we only examined decision tree ensembles because decision tree learning is one of the more successful machine learning techniques. Leaves of the tree represent class labels (e.g., CPE, CAE, FCE, and PET) and branches are combinations of features that lead to those class labels (e.g., prosodic measures). We executed a number of experiments to determine the best machine learning classifier and set of prosodic measures for classifying the English proficiency of the CELA corpus. We applied three-fold cross-validation. Speakers were randomly assigned to folds of 40 divided evenly by gender.

2.4. Feature selection

The feature selection was done separately from the machine learning approach because the MATLAB machine learning tools did not select the features used automatically [22]. We conducted one experiment for each combination of the 17 ensemble classifiers and feature sets composed of selected 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 [18] [5], and 3) a genetic algorithm (GA).

3. Results

Determining the most appropriate machine learning classifier and group of prosodic 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 suprasegmental measures grounded on Brazil's prosody model [3]. 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 suprasegmental measures was the most effective. In total, we performed 493 experiments with thousands of combinations of measures.

Table 1: *35 Prosodic features.*

Description (Abbreviation)
Articulation rate (SMARTI)
Non-prominent syllable mean pitch (SMAVNP)
Prominent syllable mean pitch (SMAVPP)
Falling-high rate (SMFALH)
Falling-low rate (SMFALL)
Falling-mid rate (SMFALM)
Filled pause mean length (SMFPLN)
Filled pauses per second (SMFPRT)
Fall-rise-high rate (SMFRSH)
Fall-rise-low rate (SMFRSL)
Fall-rise-mid rate (SMFRSM)
Given lexical item mean pitch (SMGIVP)
Pitch neutral-high rate (SMNEUH)
Pitch neutral-low rate (SMNEUL)
Pitch neutral-mid rate (SMNEUM)
New lexical item mean pitch (SMNEWP)
Paratone boundary onset pitch mean height (SMOPTH)
Pace (SMPACE)
Paratone boundaries (SMPARA)
% of tone units with at least one prominent syllable (SMPCHR)
Phonation time ratio (SMPHTR)
Paratone boundary mean pause length (SMPPLN)
Overall pitch range (SMPRAN)
Rise-fall-high rate (SMRFAH)
Rise-fall-low rate (SMRFAL)
Rise-fall-mid rate (SMRFAM)
Rising-high rate (SMRISH)
Rising-low rate (SMRISL)
Rising-mid rate (SMRISM)
Tone unit mean length (SMRNLN)
Space (SMSPAC)
Silent pause mean length (SMSPLN)
Silent pauses per second (SMSPRT)
Syllables per second (SMSYPS)
Paratone boundary mean termination pitch height (SMTPTH)

Table 2: *The top 10 experiments sorted by correlation ($p < 0.01$) between human and computer.*

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

Table 2 reports the results in terms of correlation ($p < 0.01$) for the top ten decision tree ensembles (Column 1) using the features selected (Column 2) by the genetic algorithm.

The findings indicate that the optimum machine learning classifier for English proficiency scoring of the CELA corpus is Pairwise GentleBoost utilizing the following 11 prosodic 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).

4. 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 suprasegmental measures which were derived from elements of Brazil's prosody model [3]. 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. Note that 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).

One main reason for the better performance of the current study's classifier may be related to the inclusion of prosodic measures in the features set. The best feature set reported for the current study contained nine suprasegmental measures out of eleven measures. This is consistent with the findings of Kang et al. that suprasegmental measures explained 50% of the variance in oral proficiency ratings [16].

The prominent syllable is a central component of Brazil's prosody model [3]. The computer observed that the prominence characteristics (SMPCHR) measure was predictive of English speaking proficiency. Moreover, low proficient non-native speakers typically stressed every word in a sentence [14]. Thus, the utilization of prominence characteristics (SMPCHR) as a predictor of proficiency by the computer is in harmony with the previous research [14].

Four of the eleven suprasegmental measures, neutral-low rate (SMNEUL), neutral-high rate (SMNEUH), fall-rise-mid rate (SMFRSM), and falling-high rate (SMFALH) assess the use of intonation in the speech. It should be mentioned that not all linguists differentiate among rising and fall-rising or falling and rise-falling tone choices as in Brazil [3]. Often times, linguists consider three tone choices: rising, neutral, and falling. Hence, from those linguists' viewpoint, the computer's set of suprasegmental measures includes measures of all three tone choices.

With regards to setting up the framework of a dialog, a native English speaker tends to invoke high-pitch levels to begin a new subject, mid-pitch levels to continue a subject, and low-pitch levels to end a subject [23]. This is in concert with the computer employing prosody that gauge the use of one low relative pitch level (SMNEUL), one mid relative pitch level (SMFRSM), and two high relative pitch levels (SMNEUH and SMFALH).

Non-native speakers are prone to employ low-falling pitches to connect associated topics in a dialog whereas native speakers anticipate mid-rising and mid-neutral pitches [16] [11] [25]. Although the computer found the mid-rising rate

(SMFRSM) to be an indicator of proficiency levels, it did not find the mid-neutral (SMNEUM) or low-falling rate (SMFALL) as indicative. The reason for this might be that the scarcity of mid-rising tone usage (SMFRSM) by less proficient speakers is more indicative of high proficiency than an abundance of mid-level tones (SMNEUM) and a scarcity of low-falling tones (SMFALL). The computer's consideration of overall pitch range (SMPRAN) in evaluating proficiency is supported by other research.

The computer evaluated the speaker's pitch height with two of the eleven prosodic measures: paratone boundary onset pitch mean height (SMOPTH) and given lexical item mean pitch (SMGIVP). Pitch height has been found to influence a hearer's opinions of a speaker even though it is not symptomatic of proficiency [9]. The computer also utilized speech rate (SMSYPS) and articulation rate (SMARTI) to score speaking proficiency. Nonetheless, tone unit mean length (SMRNLN) was not one of the measures that the computer included in rating speaking proficiency.

Filled pause mean length (SMFPLN) is the only computer measure which is not backed by other studies. Filled pauses may not imply as much about speaking proficiency as they point toward the cognitive load and delivery style of the speaker [10]. Also, a more proficient speaker makes a better impact on listeners than a less proficient speaker because of where they pause as opposed to how often or how long they pause [7].

5. Conclusions

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 prosodic 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 model to automatically scoring the dialogic aspects of English proficiency [3]. 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.

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 prosodic measures based on Brazil's model.

6. References

[1] J. Bernstein, A. Van Moere, and J. Cheng, J, "Validating automated speaking tests," *Language Testing*, 2010.

[2] P. Boersma and D. Weenink, D. "Praat: doing phonetics by computer (version 5.3.83)". [Computer program]. Retrieved August 19, 2014, 2014.

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

[4] *Cambridge English Language Assessment*. www.cambridgeenglish.org, Retrieved March 29, 2015, 2015.

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

[6] C. Cucchiarini, A. Neri, and H. Strik, H. "Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback," *Speech Communication*, 51(10), 853-863, 2009.

[7] G. Fulcher. "Does thick description lead to smart tests? A data-based approach to rating scale construction," *Language Testing*, 13, 208–238, 1996.

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

[9] H. Giles and P.F. Powesland. "Speech style and social evaluation," *London: European Association of Experimental Social Psychology*, 1975.

[10] F. Goldman-Eisler. "Psycholinguistics: Experiments in spontaneous speech," *London: Academic Press*, 1986.

[11] M. Hewings. "The English intonation of native speakers and Indonesian learners: A comparative study," *IRAL*, 3, 251–265, 1995.

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

[13] D. O. Johnson and O. Kang, O. "Automatic detection of Brazil s prosodic tone unit," *Speech Prosody 2016*, 287-291, 2016.

[14] O. Kang and L. Pickering. "Acoustic and Temporal Analysis for Assessing Speaking," *The Companion to Language Assessment, Wiley Online Library*, 2013.

[15] O. Kang, M. Moran, and R. Thomson. "Pronunciation features of intelligible speech among different varieties of world englishes," presentation at *Pronunciation and Second Language Learning and Teaching Conference, Santa Barbara, CA, September 5-6, 2014*.

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

[17] O. Kang, R. Thomson, and M. Moran. "Intelligibility of Different Varieties of English: The Effects of Incorporating "Accented" English into High Stakes Assessment," presentation at *American Association of Applied Linguistics Conference, Toronto, ON, Canada, March 21–24, 2015*.

[18] S. Kirkpatrick and M. P. Vecchi. "Optimization by simulated annealing," *science*, 220(4598), 671-680, 1983.

[19] J. Kormos and M. Denes. "Exploring measures and perceptions of fluency in the speech of second language learners," *System*, 32, 145–164, 2004.

[20] P. Longman. *Official guide to Parson test of English academic*, 2013.

[21] A. Loukina, K. Zechner, L. Chen, and M. Heilman, M. "Feature selection for automated speech scoring," In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 12-19), 2015.

[22] MathWorks, Inc. *MATLAB Release 2013a*. [Computer program]. Retrieved February 15, 2013.

[23] S. Nakajima and J. Allen, "A study on prosody and discourse structure in cooperative dialogues," *Phonetica*, 50, 197–210, 1993.

[24] Pearson Education, Inc. *Versant Spanish Test*. Retrieved from http://www.versanttest.com/products/spanish.jsp, 2015.

[25] A. Wennerstrom. "Intonational meaning in English discourse: A study of nonnative speakers," *Applied Linguistics*, 15, 399–421, 1994.

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