Johnson, D.O., Kang, O. (2017). Measures of intelligibility in different varieties of English: Human vs. machine. In M. O'Brien & J. Levis (Eds). *Proceedings of the 8th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Calgary, AB, August 2016 (pp. 58-72). Ames, IA: Iowa State University.

MEASURES OF INTELLIGIBILITY IN DIFFERENT VARIETIES OF ENGLISH: HUMAN VS. MACHINE

David O. Johnson, University of Kansas
Okim Kang, Northern Arizona University

This paper demonstrates the feasibility of a tool for measuring the intelligibility of English speech utilizing an automated speech system. The system was tested with eighteen speakers from countries representing six Englishes (American, British, Indian, South African, Chinese, and Spanish) who were carefully selected to represent a range of intelligibility. Intelligibility was measured via two different methods: transcription and nonsense. A computer model developed for automated oral proficiency scoring based on suprasegmental measures was adapted to predict intelligibility scores. The Pearson's correlation between the human assessed and computer predicted scores was 0.819 for the nonsense construct and 0.760 for the transcription construct. The inter-rater reliability Cronbach's alpha for the nonsense intelligibility scores was 0.956 and 0.932 for the transcription scores. Depending on the type of intelligibility measure, the computer utilized different suprasegmental measures to predict the score. The computer employed 11 measures for the nonsense intelligibility score and eight for the transcription score. Only two features were common to both constructs. These results can lead L2 researchers to different perspectives of measuring intelligibility in future research.

INTRODUCTION

In L2 pronunciation, the importance of intelligible speech has been emphasized both in classroom and assessment contexts. Researchers have aimed to determine the specific features that affect intelligibility (Field, 2005; Hahn, 2004) and assessment scores for listeners (Iwashita, Brown, McNamara, & O'Hagan, 2008; Kang, 2013). However, there has been no universally accepted method of measuring intelligibility (Munro & Derwing, 1999). Intelligibility has been ascertained most commonly using transcription or other methods (e.g., true/false statements in Munro and Derwing, (1995) or nonsense, or filtered speech (Kang, Thomson, & Moran, 2015). Due to the technical aspects of such measures, their operationalization has often been supplemented by other constructs, i.e., comprehensibility or accentedness. Currently, advances in computing technology and artificial intelligence have produced automated systems (e.g., SpeechRater^{SM}) that can assess oral proficiency of accented speech, but not intelligibility. The advantages of automated systems are that they can be faster, less expensive, more consistent, and equitable scoring.

This paper introduces an exploratory method of providing an alternative and complementary tool for measuring the intelligibility of different varieties of World Englishes using an automated speech system. The system was tested with 90 sentences from a corpus representing the three circles of World Englishes (Kachru, 1992). Eighteen speakers from countries representing six English varieties (American, British, Indian, South African, Chinese, and Spanish) are included. Sixty listeners from the respective countries listened to speech stimuli to determine the intelligibility of the speech using transcriptions and scalar judgments. The computer system used

a machine learning model with suprasegmental measures being the input and the output being an intelligibility measure. The model was trained using the annotated World Englishes corpus and tested on the World Englishes corpus with k-fold cross-validation. Correlations were conducted to compare the computer's calculated intelligibility scores with those of humans. We will discuss how different sets of phonological features influenced the computer's determination of the intelligibility of different varieties of English. These results suggest different perspectives for measuring intelligibility that can be applied in future research.

Intelligibility Measures

There is a growing recognition that L2 speech should aim for intelligibility rather than nativeness (Levis, 2005). In light of this trend, various methods have been utilized in measuring intelligibility in the field of L2 pronunciation. Currently the most commonly used method is a transcription test, which requires a participant to listen to a sound file and transcribe it. Intelligibility scores are based on the percentage of an utterance or word that is transcribed correctly by listeners (Derwing & Munro, 1997). A less-frequently-used method is a cloze test that asks listeners to fill in blanks from a transcript of speech (Smith & Nelson, 1985). The number of words correctly identified determines intelligibility scores. Munro and Derwing (1995) also used True/False judgments in which listeners are asked to make true/false decision about a short sentence they hear. This approach assumes that more intelligible speech will allow listeners to correctly understand the intended message and to correctly evaluate the truth or falsity of sentences. More recently, Kang et al. (2015) introduced exploratory methods of measuring intelligibility, i.e., nonsense statement and filtered speech methods. The nonsense statement task involves decontextualized sentences, which do not make sense semantically (e.g., "Our deaf ads traced my ants."). Listeners were asked to type missing content words into blank boxes provided. Each nonsense sentence receives a score based on the number of correct content words.

Overall, even though varied techniques are available for assessing L2 speech intelligibility, how best to measure intelligibility is still not well understood nor are some intelligibility measures necessarily easy to implement. In the current study, we attempted to explore a new way of measuring intelligibility by adopting a computer model developed for automated oral proficiency scoring to predict intelligibility scores. The findings of the study are exploratory and should not be over-generalized to other contexts of speech corpora or language assessment.

METHODS

World Englishes Speech Corpus

The data set we used is from a project investigating intelligibility of different varieties of World Englishes. Because we wanted to do an exploratory experimental study that could simply compare human vs. machine's intelligibility ratings, for the sake of convenience, we used the World Englishes corpus.

The World Englishes speech files were developed as part of a TOEFL listening test project supported by Educational Testing Service. Eighteen speakers (ages 30-50) were carefully chosen, one female and two males from each of six countries: United States and England (Inner

Circle), India and South Africa (Outer Circle), and Mexico and China (Expanding Circle). All of the Inner Circle and Outer Circle speakers were highly proficient in English but retained a noticeable foreign accent. The speakers were selected to represent a range of intelligibility, as determined by nine trained raters' scalar judgments. The speakers were either currently teaching in English as professors or graduate students.

The speakers were asked to record in a quiet location 72 nonsense sentences, 72 true/false sentences, and 18 iBT listening passages (4-5 minutes) to be utilized for the intelligibility test and computer model training. Using Audacity, a research assistant acoustically edited noises and sound quality for practical uniformity and added three seconds of pause time before and after each passage.

The intelligibility of the speakers was scored by 60 listeners, consisting of ten listeners representing each of the six World English varieties. The listeners were recruited both nationally and internationally. Listeners of non-inner circle English varieties were highly proficient with TOEFL iBT scores greater than 100. They were undergraduate and graduate students (43% males and 57% females). The intelligibility scoring was administered via computer, using headphones, and in a highly controlled laboratory setting under careful supervision. The speech files were randomly presented to the listeners. Two measures of intelligibility were used: *transcription* (Derwing & Munro, 1997) and *nonsense* (Kang et al., 2015).

For the transcription measure, listeners heard each of the 18 speakers read four sentences (72 sentences total), four to eight words in length, that were syntactically correct, but semantically incorrect. An example of the sentences is '*gasoline is an excellent drink'*. They listened to each sentence one time only and then were asked to transcribe what was said. Each speaker received an intelligibility score of 0-100% based on the number of words the listeners transcribed correctly in all four sentences. The actual transcription intelligibility scores ranged from 88.02 to 99.14%.

The nonsense intelligibility score was determined by listening to four nonsense sentences recited by each of the 18 speakers. The nonsense sentences were semantically meaningless, though syntactically normal, containing frequently used monosyllabic words. The sentences were adopted from studies on native language (L1) intelligibility (Nye & Gaitenby, 1974; Picheny, Durlach, & Braida, 1985). To score intelligibility, the listeners were asked to type missing words from the nonsense sentences into each of four text boxes. An example nonsense sentence with the missing words underlined is '*The tall kiss can draw with an oak'*. The speaker's nonsense intelligibility score was then calculated as the total number of blanks out of 16 correctly filled in by the listeners. 8.62 to 13.62 was the actual range of the nonsense intelligibility scores. Scores from the both methods were normalized for comparison. Table 1 shows the raw and normalized values for both measures of intelligibility assessed by the humans: nonsense and transcription.

Table 1

Raw and normalized intelligibility scores

Speaker	Gender	World English	Nonsense Raw	Nonsense Normalized	Transcription Raw	Transcription Normalized
1	M	Inner	13.62	6	98.31	6
2	M	Inner	13.17	6	98.65	6
3	F	Inner	12.62	5	98.25	6
4	F	Inner	12.52	5	99.14	6
5	M	Inner	11.47	4	98.93	6
6	M	Inner	11.35	4	99.13	6
7	M	Outer	13.28	6	98.82	6
8	M	Outer	11.70	4	95.91	5
9	M	Outer	10.90	3	98.46	6
10	F	Outer	9.88	2	93.37	3
11	M	Outer	8.75	1	88.02	1
12	F	Outer	8.62	1	92.27	3
13	M	Emerging	13.13	6	94.30	4
14	F	Emerging	12.92	5	97.91	5
15	M	Emerging	10.45	3	95.80	4
16	F	Emerging	10.33	3	96.85	5
17	M	Emerging	10.12	3	92.62	3
18	M	Emerging	9.10	1	96.71	5

Intelligibility Score: Computer Prediction

Figure 1 illustrates the computer model employed to predict the intelligibility scores.

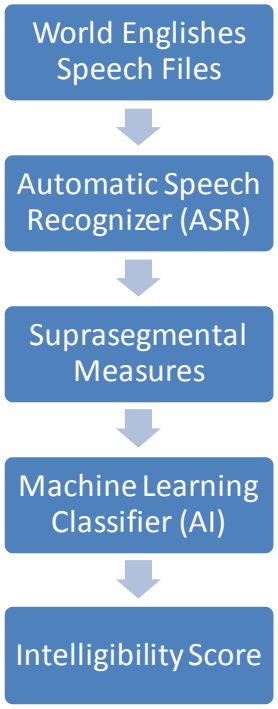

Figure 1. Computer model for predicting intelligibility scores

To predict the intelligibility scores, the computer first analyzed the 144 speech files with an automatic speech recognizer (ASR). From the output of the ASR, the computer calculated 35 suprasegmental measures of rate (e.g., syllables per second, articulation, phonation ratio), pause (e.g., silent and filled pauses per minute and mean length), stress (e.g., pace, space, percent tone units with termination), pitch (e.g., pitch range, mean prominent syllable pitch), paratone (e.g., low terminations followed by high key resets), and intonation (e.g., percent of tone choice and relative pitch). Then, the computer utilized AI (artificial intelligence) machine learning techniques to predict normalized intelligibility scores from one to six. The computer utilized a genetic algorithm to select the most salient suprasegmental measures and then built a decision tree classifier to predict the intelligibility scores from salient suprasegmental measures. The classifier was trained to achieve the best human-computer correlation by 3-fold cross-validation of the speech files. Each set of the 72 speech files was used to train a separate computer model for the two different intelligibility measures: transcription and nonsense. The computer model was developed for automated oral proficiency scoring (Johnson, Kang, & Ghanem, 2015), but in this case the output was the intelligibility score instead of the proficiency score.

RESULTS

First, we examined the Pearson's correlation between the intelligibility scores assessed by the humans and those predicted by the computer model. Figure 2 gives the Pearson's correlation between the human assessed intelligibility scores and the computer predicted scores. The

correlation between the human and computer nonsense intelligibility scores is 0.819; the correlation between the two transcription intelligibility scores is 0.760.

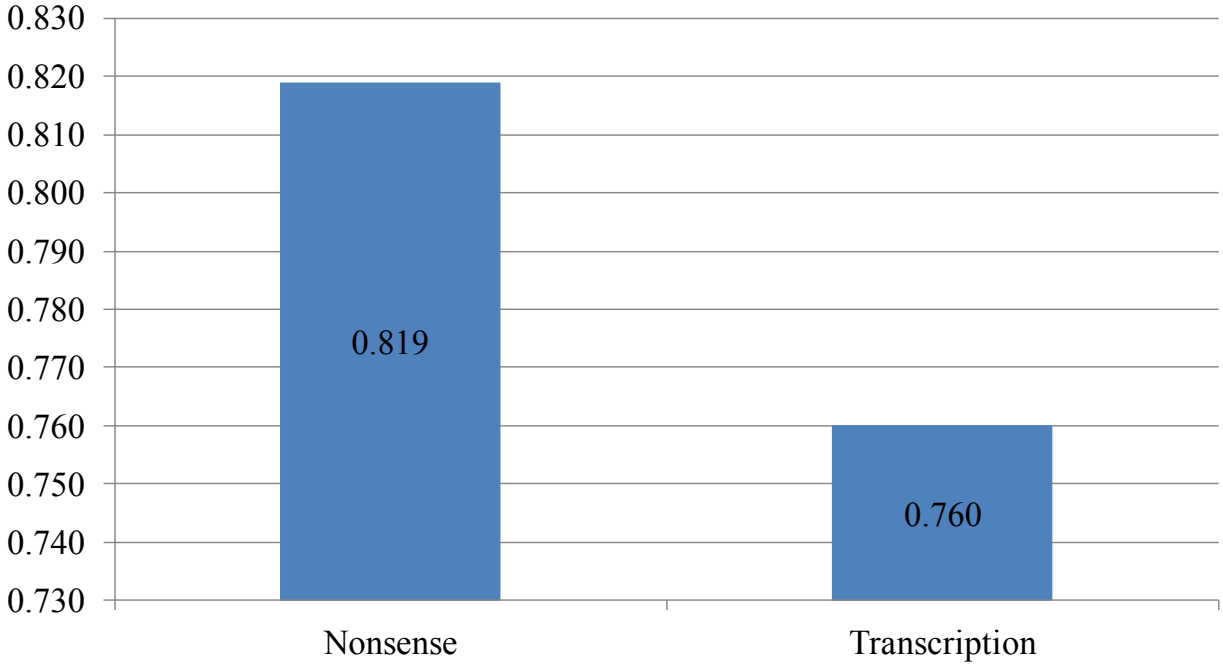

Figure 2. Human-computer correlation (*r*)

Next, we analyzed the individual correlations between humans for the nonsense intelligibility score as depicted in Figure 3. The pair-wise correlations between the 60 humans ranged from 0.989 to -0.581.

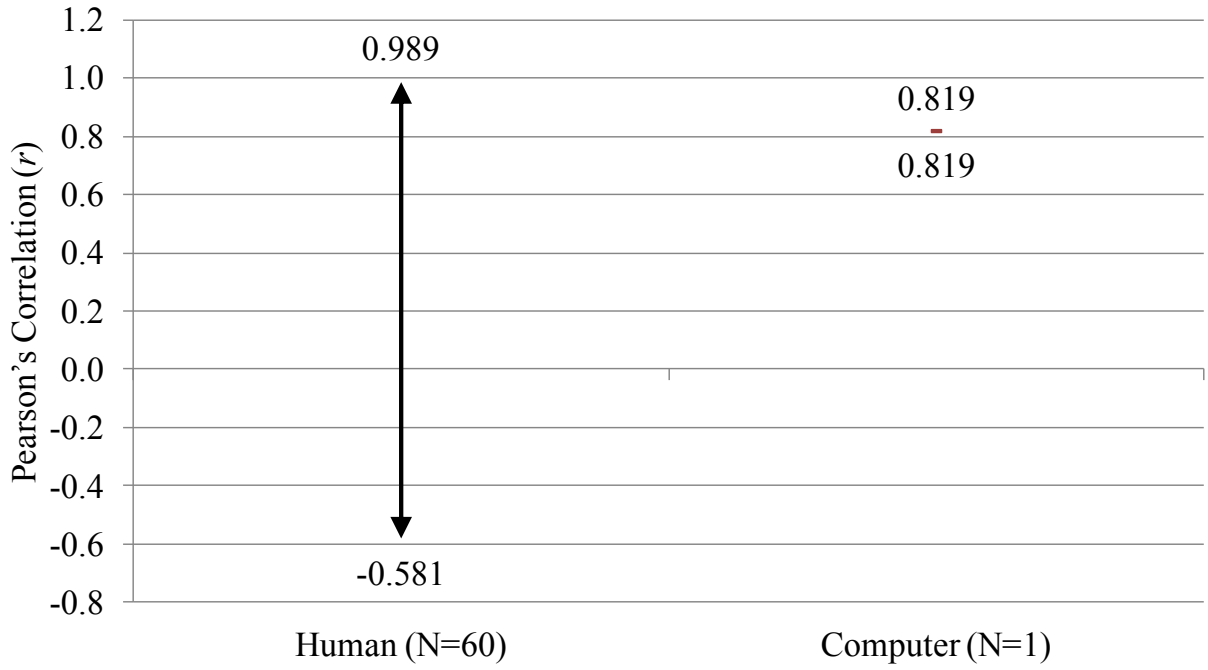

Figure 3. Human-human correlation vs. human-computer correlation (*r*) for nonsense scores

Next, we investigated interrater reliability with Cronbach's alpha. Figure 4 shows that the Cronbach's alpha (α=0.85) for the nonsense intelligibility scores was 0.956 (N=60) and 0.710 (N=5).

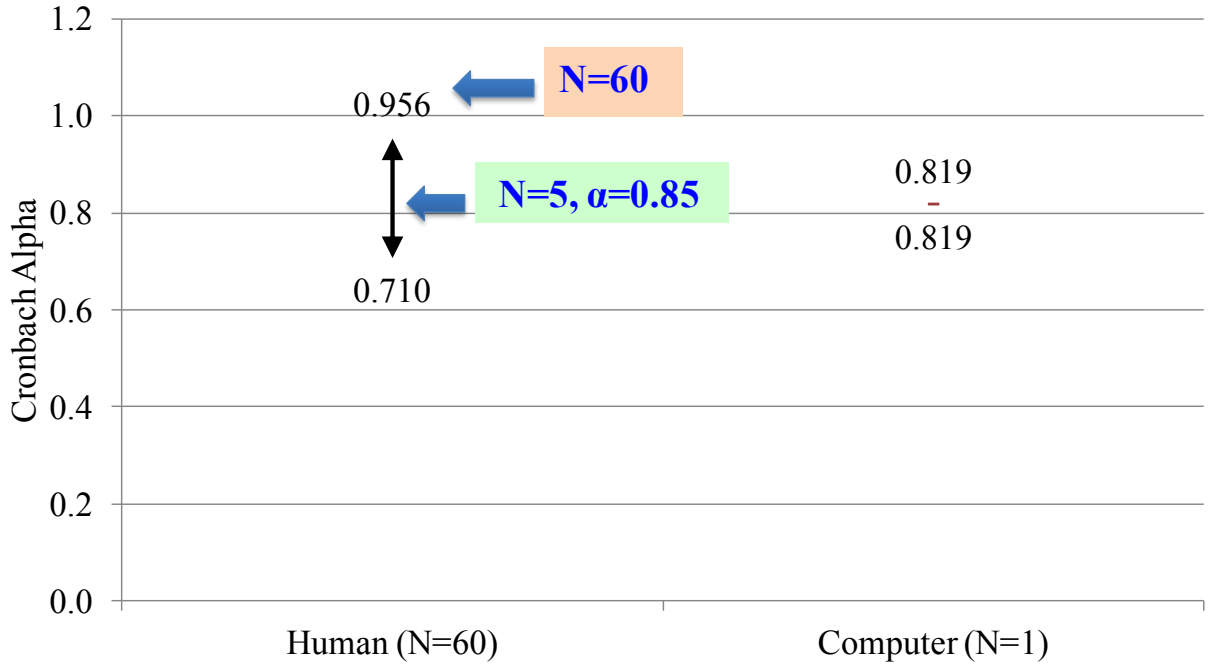

Figure 4. Interrater reliability for nonsense intelligibility scores

As described in Figure 5, the Cronbach's alpha for the transcription intelligibility scores was 0.932.

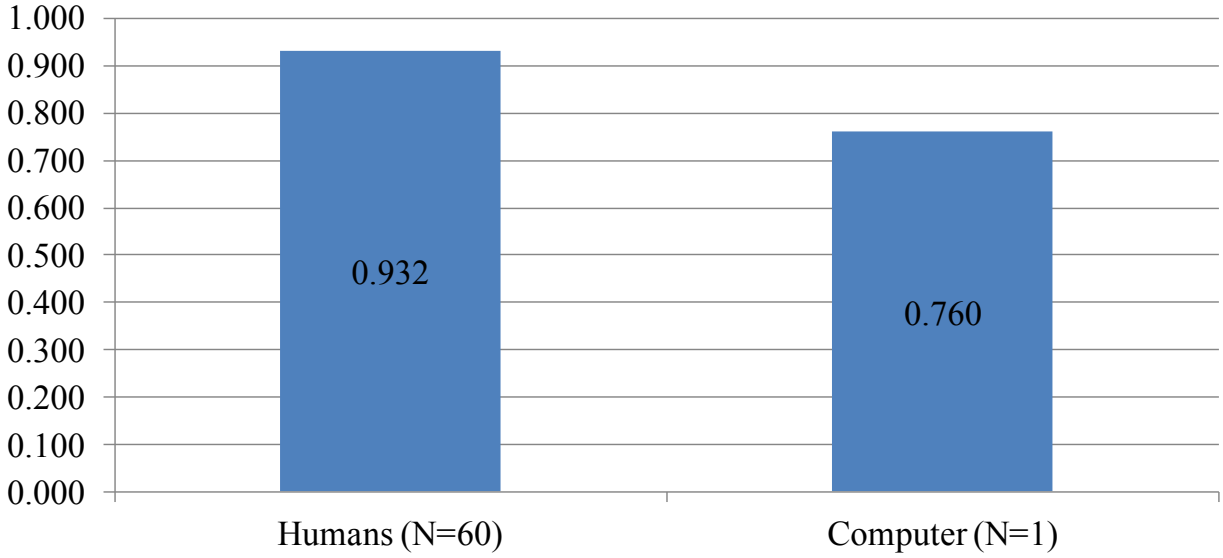

Figure 5. Inter-rater reliability for transcription intelligibility scores

After comparing the intelligibility scores appraised by the humans and those calculated by the computer model, we looked at what salient suprasegmental features were selected by the computer model. Table 2 gives the salient suprasegmental features selected by the genetic algorithm of the computer model to forecast each of the two intelligibility scores (marked with an X).

Table 2

Salient suprasegmentals in the computer model

Type	Suprasegmental	Nonsense	Transcription
Rate	phonation time ratio		X
	syllables per second		X
Pause	mean length of filled pauses	X	
	mean length of silent pauses	X	
	number of silent pauses per minute	X	
Stress	number of prominent syllables per run (pace)	X	
	prominence characteristics		X
	proportion of prominent syllables (space)		X

Pitch	average pitch of new prominent syllables	X	
	overall pitch range	X	X
	average height of terminating pitch	X	
Tone choice	% mid falling tone choices	X	
	% high falling tone choices	X	
	% mid fall-rise tone choices		X
	% high fall-rise tone choices	X	
	% low rising tone choices	X	X
	% mid rising tone choices		X

We found that, depending on the type of intelligibility measure, the computer picked different features. For the nonsense intelligibility score, the computer employed 11 features: mean length of filled pauses, mean length of silent pauses, number of silent pauses per minute, number of prominent syllables per run (pace), average pitch of new prominent syllables, overall pitch range, average height of terminating pitch, % falling mid tone choices, % falling high tone choices, % fall-rise high tone choices, and % rising low tone choices. On the other hand, the computer utilized only eight features for the transcription intelligibility score: phonation time ratio, syllables per second, prominence characteristics, proportion of prominent syllables (space), overall pitch range, % fall-rise mid tone choices, % rising low tone choices, and % rising mid tone choices. Only two features, overall pitch range and % rising low tone choices, were common to both nonsense and intelligibility predictions.

DISCUSSION

The results show there is a correlation between the human assessed and computer predicted scores for both intelligibility constructs, nonsense (r = 0.819) and transcription (r = 0.760). The computer utilized different suprasegmental measures to predict the score for each construct, 11 measures for nonsense and eight for transcription, with two the same for both constructs (i.e., pitch range and mid rising). In comparing this work with other similar research, Kang et al. (2015) found mean length of run, expected pause ratio, number of prominent syllables per run, word stress errors, % of falling tone choice, and vowel and consonant substitution/deletion errors as the salient variables in human ratings when the nonsense sentence was utilized to gauge intelligibility. Table 3 compares these with the ones selected by the computer model in this study.

Table 3

Salient pronunciation features that predict intelligibility with the nonsense sentence construct

Human Ratings (Kang et al., 2015)	Computer Modeling
mean length of run	
expected pause ratio	mean length of silent and filled pauses number of silent pauses
number of prominent syllables per run word stress errors	number of prominent syllables per run
	overall pitch range
% of falling tone choice	% of falling tone choice % of rising tone choice
vowel & consonant substitution/deletion errors	The computer model is only trained for suprasegmental features at the moment.

The salient variables for the human and computer ratings for the nonsense sentence intelligibility measure differ in four areas: pausing, prominence, pitch, and tone choice. (Note: the computer model was only trained for suprasegmental measures; therefore, any discussion of segmentals is excluded). For pausing, the salient variables in human ratings were mean length of run and expected pause ratio. One could argue that these are very similar to the measures selected by the computer because they gauge how a speaker uses pausing (both filled and silent) to articulate the speech. Thus, from this perspective both humans and the computer model found pausing to be a salient predictor of intelligibility. Both discourse and intonation units are delineated by pauses (Brazil, 1997; Wagner & Watson, 2010). Previous studies have recognized that non-native speakers pause silently more frequently, longer, and more unevenly than native speakers (Anderson-Hsieh & Venkatagiri, 1994; Riggenbach, 1991; Rounds, 1987).

With regard to prominence, pace (number of prominent syllables per run) was a salient variable for humans and the computer model. The computer did not measure word stress errors; accordingly, this variable could not be compared. It appears that both humans and the computer model agree that prominence significantly predicts the intelligibility of accented speech. The prominent syllable is a basic aspect of Brazil's (1997) model of English prosody. The proper use of prominent syllables by speakers should then be an important suprasegmental measure of intelligibility. Thus, it is consistent that the computer model employed number of prominent syllables per run as a predictor of intelligibility.

The computer model also used average pitch of new prominent syllables, overall pitch range, and average height of terminating pitch as predictors of intelligibility. The saliency of average pitch

of new prominent syllables and average height of terminating pitch is suspect because these are paratone measures and the nonsense test did not really include paratones. Thus, more research is necessary to determine if these are salient or an anomaly. Likewise, overall pitch range is more likely an indication of the variation in voice pitch between speakers, rather than an indication of an individual speaker utilizing intonation as a discourse signal to begin, maintain, and end a thought group or as a means of differentiating the information content of specific lexical items (Cutler, Dahan, & Donselaar, 1997; Kang, Rubin, & Pickering, 2010).

In the category of tone choice, the computer model and the human study mutually found falling tone choices to be indicative of relative intelligibility. This is in harmony with earlier research which puts forward that lower proficiency speakers tend to overuse falling tones until they learn that it has a negative impact on intelligibility (Kang, 2012). Additionally, the computer model found the use of rising tone to be a significant predictor of NNS intelligibility. That is, intelligible speech tends to contain more use of rising tone but less use of falling tone. Overall, when employing the nonsense intelligibility test, the human raters and computer model agreed that pausing, prominence, and tone choice are salient features of intelligibility.

Mean length of filled pauses is the only one of the computer model's predictors not supported by other research. According to Goldman-Eisler (1968), filled pauses may imply more regarding a speaker's style of articulation and cognitive load and less about speaking proficiency. Similarly, Fulcher (1996) said that more capable speakers make an impression on listeners because they pause for distinct reasons, not because they pause at a different rate than less capable speakers.

Kang et al. (2015) also found mean length of run, syllable per second, proportion of prominent syllables (space), word stress error, % of rising tone choice, vowel consonant substitution/deletion errors, consonant deletions, syllable reductions, and consonant cluster errors to be the salient variables when assessing intelligibility with the transcription construct. These are contrasted with those utilized by our computer model in Table 4.

Table 4

Salient pronunciation features that predict intelligibility with the transcription construct

Human Ratings (Kang et al., 2015)	Computer Modeling
mean length of run syllable per second	phonation time ratio syllables per second
proportion of prominent syllables (space) word stress error	proportion of prominent syllables (space) prominence characteristics
	overall pitch range
% of rising tone choice	% of rising tone choice
vowel consonant substitution/deletion errors & consonant deletions, syllable reductions, consonant cluster errors	The computer model is only trained for suprasegmental features at the moment.

Like the nonsense construct, except for overall pitch range, the human study and the computer model seem to agree on the salient features for measuring intelligibility by the transcription method. They both found prominence (prominence characteristics and space), falling tone choice (% fall-rise mid tone choices), and rising tone choice (% rising low and mid tone choices) to be prognosticators of intelligibility, all of which are consistent with prior research as discussed above.

In the area of speech rate, both concur on syllables per second as salient. This is in line with the conjecture of Kormos and Dénes (2004) that proficiency is a speech rate phenomenon in addition to an intonational one. Ginther, Dimova, and Yang (2010) noted strong to moderate correlations linking oral English proficiency scores and speech rate (i.e., syllables per second or syllable rate), articulation rate, and mean length of run (i.e., the average number of syllables per run). The computer also found that speech rate and phonation time ratio, which is speech rate divided by articulation, were predictive of intelligibility. Mean length of run was a leading measure in the human study. However, the computer did not find mean length of run to be an indicator of intelligibility. This may be because of the study's use of sentences rather than extended discourse. Even though the computer did not end up using it in its scoring, one could argue that speech rate, articulation rate, and mean length of run are inter-related and it is not necessary to consider all three of them.

With regard to prominence, the computer model is consistent with the human study in the salience of space as a predictor of intelligibility. The human study also found word stress error to

be salient and the computer model found prominence characteristics to be. The human study did not measure prominence characteristics, nor did the computer model consider word stress errors, so it is impossible to compare these two aspects. Hence, rate, stress, and tone choice emerge as salient features in both the computer model and human study when the measuring intelligibility via the transcript construct.

CONCLUSION

This paper presents an exploratory approach to automating intelligibility scoring using a computer model that predicts scores based on suprasegmental measures derived from an automated speech system. The results suggest the importance of suprasegmentals in human judgments as well as machine scoring. We also found that the salient suprasegmental measures used by the computer model depend on which intelligibility measurement method is employed, either nonsense or transcription. Further research is needed on this topic, however. First, additional work is necessary to improve the accuracy of our current computer modeling. The accuracy of the underlying components of the suprasegmental measures (e.g., prominent syllables) obtained from the output of the ASR varies due to the inherent error rates of the instrumentation, algorithms, and machine learning techniques applied. Second, the current computer model only made use of suprasegmental features. Incorporating segmental features along with other linguistic properties (e.g., grammatical and lexical features) into the computer model could improve its prognostic capabilities.

Currently, the computer model predicts a normalized intelligibility score ranging from one to six. This could be expanded to the full range of the intelligibility measures which is 0-100 for the transcription score and 0-16 for the nonsense score. The model could also be enhanced to predict intelligibility based on other assessment constructs such as accentedness or comprehensibility. A larger corpus of speakers is also recommended to validate the computer model. Although the corpus was carefully created to represent a wide range of World Englishes speakers, it only provided a training set of 12 speakers and a testing set of 6 speakers.

ABOUT THE AUTHORS

Okim Kang is an Associate Professor in the Applied Linguistics Program at Northern Arizona University, Flagstaff, AZ, USA. Her research interests are speech production and perception, L2 pronunciation and intelligibility, L2 oral assessment and testing, automated scoring and speech recognition, World Englishes, and language attitude.
Author's contact information: okim.kang@nau.edu

David O. Johnson is a Lecturer in the Electrical Engineering and Computer Science department at the University of Kansas in Lawrence, KS, USA. At the time of this research, he was a post-doctoral researcher in the Applied Linguistics Speech Laboratory at Northern Arizona University, Flagstaff, AZ, USA developing software and computer models to automatically score English language proficiency and intelligibility. 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.
Author's contact information: Email: davidojohnson@aol.com

REFERENCES

Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of Chinese ESL speakers. *TESOL Quarterly*, 28, 807–812.

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

Cutler, A., Dahan, D., & Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40, 141–201.

Derwing, T. M., Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition 19*, 1–16.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39(3), 399-423.

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

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379-399.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Hahn, L.D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201–223. doi: 10.2307/3588378.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24-49.

Johnson, D. O., Kang, O., & Ghanem, R. (2015). 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*, (pp. 119-129), ISSN 2380-9566, Dallas, TX, October 2015. Ames, IA: Iowa State University.

Kachru, B. B. (1992). *The other tongue: English across cultures*. University of Illinois Press.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, 9(3), 249-269.

Kang, O. (2013). Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (pp. 10-15). Ames, IA: Iowa State University.

Kang, O., Thomson, R., & Moran, M. (2015), 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.

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.

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

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369-377.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73-97.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 49(s1), 285-310.

Nye, P. W., & Gaitenby, J. H. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research*, 37(38), 169-190.

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing intelligibility differences between clear and conversational speech. *Journal of Speech, Language, and Hearing Research*, 28(1), 96-103.

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes*, 14, 423–441.

Rounds, P. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly*, 21, 643–672.

Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes*, 4(3), 333-342.

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, 25(7-9), 905-945.