BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

LINGUISTIC FEATURES AND AUTOMATIC SCORING OF APTIS SPEAKING PERFORMANCES

AR-G/2021/3

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

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITORS: RICHARD SPIBY AND CAROLYN WESTBROOK
ISSN 2057-5203

ABSTRACT

The purpose of this study is to investigate linguistic features of spoken performances on Aptis, which can be automatically extracted and further utilised for automated speaking assessment, and to actually predict Aptis English proficiency levels through an automated scoring system. The project aims to adapt automatic speaking proficiency rating software, initially developed by the research team, which automatically scores unconstrained English speech proficiency (e.g., Johnson, Kang & Ghanem, 2016; Johnson & Kang, 2016; Kang & Johnson, 2015; 2018). The current project further plans to augment the existing computer model to employ lexico-grammatical features along with suprasegmental/fluency measures to predict speaking proficiency scores in the Aptis speaking test.

The project provides validity evidence for the Aptis speaking test by demonstrating automatically extracted speaking features aligned with speaking criteria in Aptis. Findings of the study inform the development of automated scoring systems in the Aptis speaking exam. The project improves the computer model's capability to automatically score unconstrained English speech in second language testing and assessment with increased reliability and accuracy. It further helps better understand test-takers' speech properties in testing and assessment contexts.

Authors

Okim Kang is Professor of Applied Linguistics and Director of the Applied Linguistics Speech Lab at Northern Arizona University, Flagstaff, AZ. 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. She has recently published three edited books: *The Routledge Handbook of English Pronunciation*; *Assessment in Second Language Pronunciation*; and *The Encyclopedia of Educational Innovations*, as well as a set conference proceedings (PSLLT 2019). She is currently working on two new books: *Linguistic Analysis of Spoken Discourse* and *Discourse Prosody and Computer Modeling*.

David Johnson is an Associate Teaching Professor in the Electrical Engineering and Computer Science Department at the University of Kansas in Lawrence, KS, USA. He received his BSEE and MSEE from Kansas State University and his PhD in Computer Science from the University of Kansas. Prior to two post-doctoral research appointments at the Eindhoven University of Technology in the Netherlands and in the Applied Linguistics Speech Laboratory at Northern Arizona University, Flagstaff, AZ, USA, he was an Adjunct Professor in the Computer Science Electrical Engineering Department at the University of Missouri – Kansas City. Before beginning his academic career, he spent three decades in industry as a manager, software developer, and sales consultant.

CONTENTS

List of tables

List of figures

1. BACKGROUND AND THEORETICAL FRAMEWORK

1.1 Theoretical framework

Much debate over the construct validity of a speaking test tends to reflect the transformation in traditional views of validity from multiple forms to a contemporary notion of validity as a unified concept (Messick, 1975). **I**n practice, speaking proficiency or ability has been used more arbitrarily to refer to knowledge or competence in the use of a language (Bachman, 1990). Due to the difficulty and variability in defining the construct, it is still unclear as to what proficiency entails exactly in speaking performance **(**Iwashita, Brown, McNamara & O'Hagan, 2008**;** Jin & Mak, 2012). One could argue that a defining characteristic of language proficiency may lie in its capacity to assess the test-taker's ability to process linguistic information to construct meaning.

Research findings on linguistic criteria of speaking proficiency have been mixed (Brown, 2006; Iwashita et al., 2008). In general, L2 learners' speaking proficiency has been described in high-stakes tests as different *levels* or *bands* by certain representative components of language. Linguistic measures often include various aspects of speech properties such as pronunciation, vocabulary or grammar. However, how proficiency can be characterised by these features or components may vary across proficiency levels and assessment contexts (De Jong, Steinel, Florijn, Schoonen & Hulstijn, 2012).

The current research project is positioned within the socio-cognitive framework for test development and validation (O'Sullivan & Weir, 2011; Weir, 2005). It focuses on scoring validity evidence for the Aptis speaking test, by investigating the linguistic features of Aptis speaking performances, which can be automatically extracted and utilised in automatic assessment and further predict English proficiency levels via automated scoring systems. The socio-cognitive framework is claimed to offer concrete explanations about how criterial distinctions at different proficiency levels are made between tests as it provides a practical and achievable methodology for generating evidence-based arguments about the test use in the real-world contexts (Weir, 2005). The model consists of various components (e.g., *test-takers, cognitive validity, scoring validity, context validity, criterion-related validity* and *consequential validity).* Although these multiple components are considered as being independent of one another for purposes of transparency and focus, they offer a comprehensive and coherent perspective on the process of test development and validation activity (Shaw & Weir, 2007).

1.2 Background in automated scoring systems

Automated scoring systems generate test scores more quickly and more economically than human scoring and they are more consistent and equitable in scoring than humans (Attali & Burstein, 2006; Zechner, Xi, Higgins & Williamson, 2009). There are two categories for automated scoring systems in speaking assessment: constrained and unconstrained (spontaneous). In constrained speech assessment, test-takers are requested to respond orally to constructed response items. The computer recognises the words spoken with an automatic speech recogniser (ASR) and compares them to the hypothesised response. The use in evaluating constrained speech proficiency has been confirmed by various research studies (e.g., Bernstein, Van Moere & Cheng, 2010; Van Moere & Suzuki, 2017). On the other hand, unconstrained speech is irregular and variable, making automatic proficiency scoring of it more challenging. SpeechRater^{SM} is an example of an operational computerised unconstrained English speech proficiency assessment tool (Loukina, Davis & Xi, 2017; Zechner et al., 2009). The Pearson correlation between the ranks assessed by a human and those estimated by SpeechRater^{SM} was 0.55 (Zechner et al., 2009) or 0.62 (Loukina et al., 2017). However, these models are largely based on fluency and segmental features.

The research team has created a set of computer programs that automatically score the proficiency of unconstrained English speech (e.g., Johnson et al., 2016; Johnson & Kang, 2016; Kang & Johnson, 2018), using an advanced prosodic model. The Pearson correlation between the official Cambridge Language Assessment proficiency levels and those calculated by the computer was 0.718, which is higher than that of other similar computer programs (Zechner et al., 2009; Evanini & Wang, 2013). The advanced suprasegmental measures calculated by the computer model to score proficiency are available for more in-depth analysis. Adapting Kang & Johnson's (2018) current computer models, the current study demonstrates how the Aptis speaking test can be automatically scored and ultimately improve the computer's ability to automatically score English speaking proficiency.

1.3 Aptis spoken responses

The Aptis speaking test is divided into four parts or task types and takes about 12 minutes to complete. It involves multiple tasks, such as personal information about test-takers themselves and their interests, description of opinions, or discussion of personal experience on an abstract topic. (See Table 1 for detail.) The current study focused on the long, uninterrupted, monologic portions of speech in Tasks 3 and 4. The speaking tasks of the test, though varying with topics, were of the similar task type: either *individual picture-based description and discussion,* in which candidates were asked to describe pictures without interruption, or description of personal experience and opinion. As for Task 3, responses to the first question in particular have been used in this project, and there were three different forms (different topics) of the Task 3 (a, b, and c). Tasks 3 and 4 also contained increased complexity in responses. Previous research (e.g., Sun, 2011) demonstrated that different topics did not affect test-takers' speaking performances significantly; accordingly, the current study selected Tasks 3 and 4 for analysis, but not the other two tasks.

Table 1: Four different tasks in the Aptis speaking test

Task	Types	Description
Task 1	Personal information	Test-takers are asked to answer three questions on personal topics. They are expected to talk for 30 seconds per question.
Task 2	Describe, express your opinion, provide reasons and explanations	Test-takers are asked to describe a photograph and then answer two questions related to the topic illustrated in the photo. The three questions increase in complexity (from description to opinion). They are expected to talk for 45 seconds per question.
Task 3	Describe, compare and provide reasons and explanations	Test-takers are asked to compare two pictures and then answer two questions related to the topic. The three questions increase in complexity (from description to speculation). They are expected to talk for 45 seconds for each question.
Task 4	Discuss personal experience and opinion on an abstract topic	Test-takers will see a picture and be asked three questions about an abstract topic. They are given one minute to prepare an answer and can take notes. They are expected to talk for two minutes.

(O'Sullivan et al., 2020)

Speaking scores are marked by a team of trained examiners. A CEFR level is assigned according to the score obtained. Separate task based holistic scales are used for each task. The following aspects of performance are addressed: 1) grammatical range and accuracy; 2) lexical range and accuracy; 3) pronunciation; 4) fluency; and 5) cohesion and coherence. Although a range of analytical aspects (e.g., part-scoring or task-specific rating scales) are considered, raters are required to give only holistic scores (O'Sullivan et al., 2020).

2. RESEARCH DESIGN AND METHODS

2.1 Research questions

The proposed project is guided by the following research questions.

1. What are the salient linguistic features in Aptis speaking performances that can contribute to automated scoring systems?
2. Can the speaking features utilised by the computer to assess oral proficiency in the Aptis speaking test be validated by other research findings in the field?

2.2 Research design

To evaluate the research questions, we used an automated scoring system to predict the Aptis oral proficiency rating of speakers from audio data files of examinees' responses to the Aptis speaking test for two different tasks (Task 3 and Task 4). Each speech sample was transcribed in a computer readable format by a trained transcriptionist. Then, two trained coders/linguists identified grammatical and lexical features, which were combined into 36 lexico-grammatical measures. Note that in this study, coders and linguists are used interchangeably as two coders who analysed the speech files are trained applied linguists in the North American doctoral programs.

2.3 Materials received

The British Council provided audio data files of the examinees' responses to the Aptis speaking test for two different tasks (Task 3 and Task 4). They were 105 speech files for 83 speakers. Each speech sample was transcribed in a computer readable format by trained coders/linguists. The length of the speech samples was adjusted for linguistic analysis, depending upon the availability and quality of the speech samples. The total length of responses per question ranged from 45 seconds to 2 minutes; accordingly, normalisation took place for linguistic analysis.

2.4 Materials used in the study

The 105 audible speech files received are described in Table 2 below. The average speech time per file was 1.32 minutes. The Phase(s) column in Table 1 indicates which Phase(s) of the experiment, as described above, the speaker was included in. The Speaker column is the number assigned to the audio file for analysis. The Region and Country columns provide the L1 background of the speaker. (Note: A blank in these columns indicates no L1 background was provided by the British Council.) The Tasks column identifies which Aptis tasks, (i.e., 3a, 3b, 3c, or 4) the speaker was recorded performing. The Aptis column gives the speech proficiency level of the speaker assigned by Aptis.

Table 2: Speech files used for the study

Phase(s)	Speaker	Region	Country	Tasks	APTIS
1, 2	1	EU	SPA	3a, 4	A2
1, 2	2	AM	MEX	3a, 4	A2
1, 2	3	AM	MEX	3a, 4	A2
1, 2	4	EU	SPA	3a, 4	A2
1, 2	5	SA	PAK	3a, 4	B1
1, 2	6	ME	MOR	3a, 4	B1
1, 2	7	SA	PAK	3a, 4	B1
1, 2	8	EU	SPA	3a, 4	B1
1, 2	9	AM	BRA	3a, 4	B1
1, 2	10	ME	MOR	3a, 4	B2
1, 2	11	EU	HUN	3a, 4	B2
1, 2	12	WE	BOH	3a, 4	B2
1, 2	13	AM	BRA	3a, 4	B2
1, 2	14	SA	PAK	3a, 4	B2
1, 2	15	EA	VIE	3a, 4	B2
1, 2	16	WE	TUR	3a, 4	C
1, 2	17	EU	SLO	3a, 4	C
1, 2	18	ME	MOR	3a, 4	C
1, 2	19	EU	HUN	3a, 4	C
1, 2	20	WE	BOH	3a, 4	C
1, 2	21	EU	SLO	3a, 4	C
2	22	AM	MEX	3a, 4	A1
2	23	EA	VIE	4	A1
2	24			4	A1
2	25			4	A1
2	26			4	A1
2	27			4	A1
2	28			4	A1
2	29			4	A1
2	30			4	A1
2	31			4	A1
2	32			4	A1
2	33			4	A2
2	34			4	A2
2	35			4	A2
2	36			4	A2
2	37			4	A2
2	38			4	A2
2	39			4	A2
2	40			4	A2
2	41			4	A2
2	42			3c	A2
2	43			4	B1
2	44			4	B1
2	45			4	B1
2	46			4	B1
2	47			4	B1
2	48			4	B1
2	49			4	B1
2	50			4	B1
2	51			4	B1
2	52			4	B1
2	53			4	B1
2	54			4	B1
2	55			4	B2
2	56			4	B2
2	57			4	B2
2	58			4	B2

Phase(s)	Speaker	Region	Country	Tasks	APTIS
2	59			4	B2
2	60			4	B2
2	61			4	B2
2	62			4	B2
2	63			4	B2
2	64			4	B2
2	65			4	B2
2	66			4	B2
2	67			4	B2
2	68			4	B2
2	69			4	B2
2	70			4	C
2	71			4	C
2	72			4	C
2	73			4	C
2	74			4	C
2	75			4	C
2	76			4	C
2	77			4	C
2	78			4	C
2	79			4	C
2	80			4	C
2	81			4	C
2	82			4	C
2	83			3a	C

2.5 Preparation of the materials for analysis

The spoken responses were coded for linguistic features for the criteria of lexico-grammatical features by two trained human coders. The lexical and grammatical features selected by the computer model were verified by those coded by human analysts. Suprasegmental features were automatically extracted from the current prosody model directly (Kang & Johnson, 2018; Johnson et al., 2016).

2.5.1 Transcription

The speech files were manually transcribed. The following is an excerpt of a typical transcription:

my life (xxx) was normal plus some
I went to I went to my family to southern Morocco
to (xxx) and the marriage it was uh it was lovely
it was a lovely journey uhh we we discovered new people
new beaches new people we meet new people but
it was so much to enjoy the summer the summer sum
the summer (xxx) of uhh of uhh the beaches

The (xxx) is where the transcriptionist could not hear what the speaker was saying. This was quite frequent in the speech samples.

2.6 Phases of the experiment

The experiment was conducted in two phases.

Phase 1: The first phase of the project was carried out using speakers who had speech files for both Task 3a and Task 4. Task 3 has three forms (a, b and c) depending on the topic of the task. They were all the same task with a slightly different topic. Because we had speech files for both Task 3a and Task 4, we employed 72 measures, each of the 36 lexico-grammatical measures for each task. There was only one example of an A1 speaker with both Task 3a and Task 4. It is impossible to train a machine learning computer model with only a single example. The Phase 1 experiments were conducted by leaving the single A1 speaker out. Thus, we analysed 21 speakers with 72 measures per speaker.

Phase 2: The second phase of the project was executed using the 21 speakers from the first phase, plus the single A1 speaker with both Task 3a and Task 4, plus the 61 speakers with only one task. Because we did not have the same tasks for each speaker, we only used the 36 lexico-grammatical measures for each speaker. Accordingly, we analysed 105 speech samples with 36 measures per speech sample.

2.7 How the data was analysed

2.7.1 Lexico-grammatical feature coding protocol

The transcribed scripts above were analysed by two trained coders. Both linguists coded the 22 speakers for which we had both Task 3a and Task 4. The linguists calibrated their coding techniques on both tasks for three speakers: 51848 (Aptis = C), 55097 (A1), and 57622 (B1). Then they coded the remaining speech files independently.

After all 44 speech files were coded, the Matlab computer program compared the coding from each linguist/coder. The agreement rate of the two human coders' lexico-grammatical analyses was 81% after any obvious coding errors were corrected. In some cases, the differences could not be resolved easily. This was for two reasons: 1) the grammatical mistakes could have been interpreted in several ways; and 2) the quality of the audio files created challenges for the interpretation of content or missing words. For those cases, the mean value of the two linguists coding was used to produce 39 lexico-grammatical features for each of the 44 speech files.

Once the desirability rate of the inter-rater reliability was achieved (80% or higher), one of the trained linguists coded the remainder of the speech files independently. Table 3 provides 39 lexico-grammatical features coded for the project. These features were selected based on empirical findings of a number of previous studies, i.e., lexical richness and type/token measures (Brown et al., 2005); vocabulary range (Iwashita et al., 2008); grammatical accuracy (Brown et al., 2005); specific types of errors (Iwashita et al., 2008); and grammatical complexity (Wolfe-Quintero et al., 1998).

Table 3: Lexico-grammatical features

Number	Features
1	Total number of different words (types)
2	Total number of words spoken (tokens)
3	Percent of K1 tokens (the most frequent 1,000 words of English) (Cobb, 2002)
4	Percent of K2 tokens (the second most frequent thousand words of English i.e. 1,001–2,000) (Cobb, 2002)
5	Percent of academic word list (AWL) tokens (Cobb, 2002)
6	TTR (a ratio of the total number of types to the total number of tokens)
7	Lexical density (the number of content words divided by total number of words)
8	Total number of word families
9	Average word length (number of characters)
10	Total number of T-units
11	Total number of error-free T-units
12	T-unit complexity (number of clauses divided by the total number of T-units; Hunt, 1965; Wolfe-Quintero et al., 1998)
13	Total number of clauses
14	Total number of independent clauses
15	Total number of dependent clauses
16	Tense errors (TENSE)
17	Singular/plural errors (SING/PL)
18	Preposition errors (PREP)
19	Article errors (ART)
20	Adverb errors (ADV)
21	Pronoun errors (PRO)
22	Adjective errors (ADJ)
23	Verb errors (VERB)
24	Determiner errors (DET)
25	Coordinator errors (CO)
26	Subject errors (SUB)
27	Object errors (OBJ)
28	Negation errors (NEG)
29	Comparative/superlative errors (COMP/SUP)
30	Copula errors (COPU)
31	Modal errors (MOD)
32	Nominalisation errors (NOMIN)
33	Relative clause errors (RELCLS)
34	Complement clause errors (COMPCLS)
35	Non-finite clause errors (NFCLS)
36	Subject-verb agreement errors (AGREE)
37	Formation of subjunctive structure errors (SUBJ)
38	Formation of conditional structure errors (COND)
39	Passive errors (PASSIVE)

In the analysis, a T-unit is defined as consisting of “one main clause plus all subordinate clauses and non-clausal structures that are attached to or embedded in it” (Hunt, 1965, p. 49). The linguists counted independent clauses (main clause) with a dependent clause attached to the main clause as one T-unit. The following two examples have one T-unit each: (a) *Because I was tired, I didn't go to school*. (b) *I thought that she didn't come to schoo*l. However, the following sentences were counted as having two T-units: (a) *I was tired, and I didn't go to school*. (b) *I was tired, but I still went to school*. In other words, clauses combined by the conjunctions “and, but, or” were counted as a separate T-unit. “Error free T-units are T-units free from any grammatical errors including both the specific errors defined above as well as other grammatical errors (e.g. word-order, omission of pronouns)” (Iwashita et al., 2008, p. 8).

Researchers often use errors per T-unit as measure of accuracy in spoken language. The T-unit was originally developed for analysis of written language, it has counterparts for spoken language in the C-unit and the AS-unit. All of these units (i.e., T-unit, C-unit, or AS-unit) are syntactic measures that allow the analyst to give credit to performers who can embed clauses and construct chunks of speech which can reflect sophisticated planning processes (Foster, Tonkyn & Wigglesworth, 2000). In this study, we adopted Brown et al.'s (2005) method because their study showed that all measures of grammatical accuracy/error within T-units had significant effects on proficiency level of the speakers. In the study, the percentage of error free T-units was measured for global accuracy. That is, specific errors were counted for tense, third person singular verbs/copula, plural nouns, article use, and prepositions.

2.8 Lexico-grammatical measures

The Matlab computer program was utilised to normalise the 39 lexico-grammatical features to produce 36 lexico-grammatical measures shown in Table 4. Seven of the features (Percent of K1, K2, and AWL tokens, TTR, lexical density, average word length, and T-unit complexity) were already normalised producing seven measures (Measures 1–7 in Table 4). The number of error-free T-units was normalised by the total number of T-units producing one measure (Measure 8 in Table 4). The total number of tokens was employed to normalise the number of word families, independent clauses, and dependent clauses producing three measures (Measures 9–11 in Table 4). Normalising the 24 grammatical errors (Numbers 16 through 39 in Table 3 above) was accomplished by dividing them by the Total number of tokens also, which created 24 measures (Measures 12–35 in Table 4). An additional measure was calculated by dividing the total number of grammatical errors by the Total number of tokens (Measure 36 in Table 4). Thus, the 39 lexico-grammatical features were utilised to produce 36 lexico-grammatical measures (7 + 1 + 3 + 24 + 1 = 36).

Table 4: Lexico-grammatical measures

		Salient measures	
Number	**Measures**	**Phase 1**	**Phase 2**
1	Percent of K1 tokens (the most frequent 1,000 words of English) (Cobb, 2002)		X
2	Percent of K2 tokens (the second most frequent thousand words of English i.e. 1,001–2,000) (Cobb, 2002)	(3a)	
3	Percent of academic word list (AWL) tokens (Cobb, 2002)		
4	TTR (a ratio of the total number of different types to the total number of tokens)		
5	Lexical density (the number of content words divided by total tokens)		
6	Average word length (number of characters)	(3a)	X
7	T-unit complexity (number of clauses divided by the total number of T-units; Hunt, 1965; Wolfe-Quintero et al., 1998)		
8	Total number of error-free T-units/Total number of T-units	(3a)	X
9	Total number of word families/Total number of tokens		X
10	Total number of independent clauses/Total number of tokens		
11	Total number of dependent clauses/Total number of tokens	(3a)	
12	Tense errors (TENSE) /Total number of tokens		
13	Singular/plural errors (SING/PL) /Total number of tokens	(3a)	X
14	Preposition errors (PREP) /Total number of tokens		

Number	Measures	Phase 1	Phase 2
15	Article errors (ART) /Total number of tokens	(4)	X
16	Adverb errors (ADV) /Total number of tokens		
17	Pronoun errors (PRO) /Total number of tokens	(4)	
18	Adjective errors (ADJ) /Total number of tokens	(3a), (4)	
19	Verb errors (VERB) /Total number of tokens	(3a)	
20	Determiner errors (DET) /Total number of tokens		X
21	Coordinator errors (CO) /Total number of tokens		X
22	Subject errors (SUB) /Total number of tokens		
23	Object errors (OBJ) /Total number of tokens	(4)	
24	Negation errors (NEG) /Total number of tokens		X
25	Comparative/superlative errors (COMP/SUP) /Total number of tokens		
26	Copula errors (COPU) /Total number of tokens		X
27	Modal errors (MOD) /Total number of tokens	(3a)	
28	Nominalisation errors (NOMIN) /Total number of tokens	(4)	
29	Relative clause errors (RELCLS) /Total number of tokens		X
30	Complement clause errors (COMPCLS) /Total number of tokens	(3a)	X
31	Non-finite clause errors (NFCLS) /Total number of tokens		X
32	Subject-verb agreement errors (AGREE) /Total number of tokens	(3a)	X
33	Formation of subjunctive structure errors (SUBJ) /Total number of tokens		X
34	Formation of conditional structure errors (COND) /Total number of tokens		
35	Passive errors (PASSIVE) /Total number of tokens		
36	Total Grammatical Errors (16-39) /Total number of tokens		

2.9 Suprasegmental features

Thirty-five suprasegmental measures (Kang et al., 2010) shown in Table 5 have been computed for each utterance based on the time intervals and amounts of silent pauses, filled pauses, syllables, and the elements of Brazil's (1997) prosody model. Previous experiments showed that suprasegmental measures could be utilised to predict the CEFR proficiency ratings (Johnson, et al., 2016; Johnson & Kang, 2016; Kang & Johnson, 2015; 2018).

Table 5: Suprasegmental measures

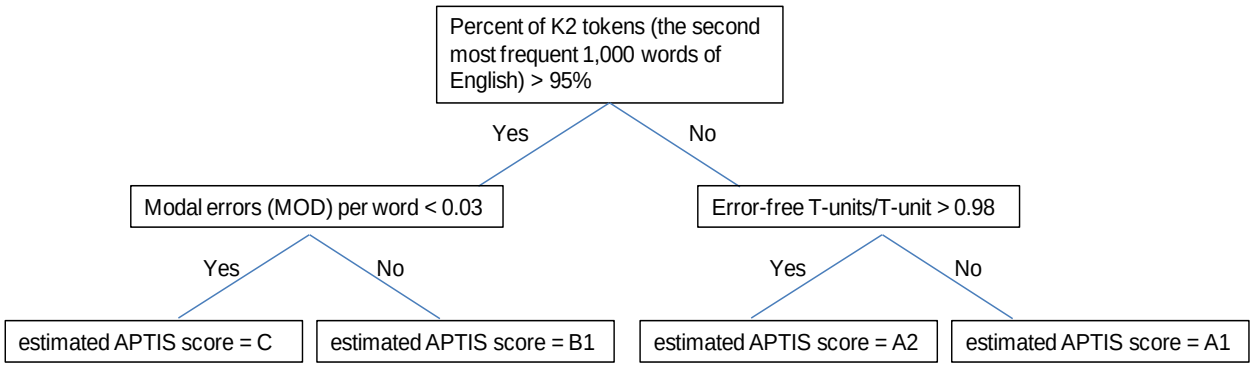
1. Articulation rate	19. High-fall rate
2. Phonation time ratio	20. Low-fall rate
3. Tone unit average length	21. Mid-fall rate
4. Syllable rate	22. High-fall-rise rate
5. Filled pause average duration	23. Low-fall-rise rate
6. Filled pause rate	24. Mid-fall-rise rate
7. Silent pause average duration	25. High-level rate
8. Silent pause rate	26. Low-level rate
9. Prominent syllables per tone unit (i.e., pace)	27. Mid-level rate
10. Percent of tone units with at least one prominent syllable	28. High-rise-fall rate
11. Percent of syllables that are prominent (i.e., space)	29. Low-rise-fall rate
12. Overall pitch range	30. Mid-rise-fall rate
13. Non-prominent syllable average pitch	31. High-rise rate
14. Prominent syllable average pitch	32. Low-rise rate
15. Paratone boundary onset pitch average height	33. Mid-rise rate
16. Paratone boundary rate	34. Given lexical item mean pitch
17. Paratone boundary average pause duration	35. New lexical item mean pitch
18. Paratone boundary average termination pitch height	

2.10 Computer estimation of Aptis level

The computer utilised a genetic algorithm to select the most salient lexico-grammatical measures and then built a decision tree classifier to predict the Aptis level from the salient measures. The classifier was trained to achieve the best human-computer correlation by three-fold cross-validation of the speech files. Each iteration of the genetic algorithm tries up to 50 different subsets of measures. The best five subsets are then used to generate another 50 different subsets by recombining and mutating the best ones. In the first iteration, the computer recombined and mutated five “seed” subsets determined by dividing all measures into five groups based on the correlation between the measure and the Aptis scores. The ones with the best correlation went into the first group; the ones with the next best correlations went into the second group; and so forth. The genetic algorithm was iterated 50 times, potentially trying up to 2500 different subsets of measures to find the subset that best predicts the highest human-computer correlation. The genetic algorithm was iterated 50 times for each of 17 different decision tree models. The end result of each computer run is then a decision tree model (1 of 17) and a subset of lexico-grammatical features (1 of 2500) that best predicts the speaker’s Aptis score. To further improve the results, the five best subsets of measures across all 17 decision tree models became the seed subsets for another run of the entire process again.

The construction of the decision tree is fully explained in other documentation (Johnson, et al., 2016; Johnson & Kang, 2016; Kang & Johnson, 2015; 2018). Briefly, a decision tree is a simple representation for classifying examples, in this case, English oral proficiency. It is a supervised machine learning method where the data is continuously split according to a certain parameter. A decision tree consists of: 1) nodes which test for the value of a certain parameter (e.g., lexico-grammatical measure); 2) edges/branches which correspond to the outcome of a test and connect to the next node or leaf; and 3) leaf nodes which are terminal nodes that predict the outcome (e.g., oral proficiency level). Figure 1 is an illustrative example of part of a decision tree for predicting oral proficiency.

Figure 1: Illustrative example of oral proficiency decision tree



2.11 Analysis

The proposed study applied a quantitative approach and correlational research method to the analysis of linguistic features and speaking scores in automated systems. It utilised a set of machine learning computer programs to automatically predict the unconstrained English speaking proficiency levels of 105 non-native speakers in the Aptis speaking test and to compare the computer's results with those from the Aptis exam. Selected features through the computer model were discussed in terms of prediction of English-speaking proficiency (CEFR, A1-C).

3. RESULTS

3.1 Lexico-grammatical features: Phase 1 (21 speakers – 72 measures per speaker – 4 Aptis levels)

This phase of the experiment produced a correlation between four computer estimated Aptis scores and the Aptis assigned scores (C, B2, B1 and A2) of 0.90 based on the following 15 lexico-grammatical measures:

- (3a) Percent of K2 tokens (the second most frequent 1,000 words of English)
- (3a) Modal errors (MOD)
- (3a) Error-free T-units
- (3a) Dependent clauses
- (3a) Average word length (number of characters)
- (3a) Adjective errors (ADJ)
- (3a) Subject-verb agreement errors (AGREE)
- (3a) Complement clause errors (COMPCLS)
- (3a) Singular/plural errors (SING/PL)
- (3a) Verb errors (VERB)
- (4) Adjective errors (ADJ)
- (4) Article errors (ART)
- (4) Nominalisation errors (NOMIN)
- (4) Object errors (OBJ)
- (4) Pronoun errors (PRO).

We let the computer repeat runs with the five best subsets of measures across all 17 decision tree models as the seed for the next run. Each run takes approximately 20 hours of computer time. We repeated the runs 34 times. The confusion matrix for the computer estimated Aptis scores versus the Aptis assigned scores is shown in Table 6.

Table 6: Confusion matrix for Phase 1

		Estimated Aptis score			
		C	**B2**	**B1**	**A2**
Aptis assigned score	**C**	6	0	0	0
	B2	4	2	0	0
	B1	0	0	5	0
	A2	0	0	3	1

3.2 Lexico-grammatical features: Phase 2 (105 speech files – 36 measures per speech file – 5 Aptis levels)

This phase of the experiment produced a correlation between five computer estimated Aptis scores and the Aptis assigned scores (C, B2, B1, A2 and A1) of 0.76 based on the following 15 lexico-grammatical measures:

- Percent of K1 tokens (the most frequent 1,000 words of English)
- Word families
- Error-free T-units
- Average word length (number of characters)
- Subject-verb agreement errors (AGREE)
- Article errors (ART)
- Coordinator errors (CO)
- Complement clause errors (COMPCLS)
- Determiner errors (DET)
- Negation errors (NEG)
- Non-finite clause errors (NFCLS)
- Relative clause errors (RELCLS)
- Singular/plural errors (SING/PL)
- Formation of subjunctive structure errors (SUBJ)
- Copula errors (COPU).

As with Phase 1, we let the computer repeat runs 50 times. Each run takes approximately 32 hours of computer time. The confusion matrix for the computer estimated Aptis scores versus the Aptis assigned scores is shown in Table 7.

Table 7: Confusion matrix for Phase 2

		Estimated Aptis score				
		C	**B2**	**B1**	**A2**	**A1**
Aptis assigned score	**C**	21	5	0	0	0
	B2	12	9	6	0	0
	B1	2	11	6	3	0
	A2	0	4	8	3	3
	A1	0	0	5	6	1

3.3 Suprasegmental features: Phase 1 and Phase 2

The objective of this research was to employ a collection of computer programs to automatically rate the oral proficiency of 105 speech files of non-native English examinee responses from the Aptis test. The computer produced proficiency ratings (.73) by utilising the 12 suprasegmental measures shown in Table 8, which have been supported by our previous research (Kang & Johnson, 2018; Kang et al., 2010).

Table 8: Suprasegmentals used by computer to predict proficiency

Type	Suprasegmental measure
Stress	Prominent syllables per tone unit (i.e., pace)
Pitch	Low-rise rate
	Mid-rise rate
	Low-fall rate
	High-rise-fall rate
	High-fall-rise rate
	Overall pitch range
Pause & Length	Silent pause average duration
	Tone unit average length
	Filled pause average duration
Speech rate	Syllable rate (syllable per second)
	Articulation rate

The second research question of the current project was to validate the speaking features utilised by the computer to assess oral proficiency in the Aptis speaking test through other research findings in the field. Table 9 provides a three-way summary of salient linguistic features to demonstrate how the features selected in the study link to previous research findings and the Aptis rating descriptors. The first column shows the features used by the current project to predict oral proficiency scores in the Aptis speaking test in Tasks 3 and 4. The second column lists the previous studies that supported the relationships between testing scores and linguistic features, and the third column demonstrates the features currently included in the Aptis scoring rubric descriptors. The lexico-grammatical features from Phases 1 and 2 have been combined.

Table 9: Comparisons between linguistic features utilised by the computer model and Aptis descriptors

Linguistic measures	Previous research	Current Aptis descriptor
Grammatical complexity & accuracy Modal errors Error-free T-units Dependent clauses Adjective errors Subject-verb agreement errors Complement clause errors Singular/plural errors Verb errors Adjective errors Article errors Nominalisation errors Object errors Pronoun errors Coordinator errors Determiner errors Negation errors Non-finite clause errors Relative clause errors Formation of subjunctive structure errors Copula errors	Biber et al., 1999; Brown et al., 2005; Brown, 2006; Espada-Gustilo, 2011; Foster & Skehan, 1996; Grant & Ginther, 2000; Hinkel, 2003; Iwashita et al. 2008; Kang & Yan, 2018; Lennon, 1990; Norrby & Hakansson, 2007; Skehan & Foster, 1999	• Uses a range of complex grammar constructions accurately. • Control of grammatical structures • A range of cohesive devices used to clearly indicate the links between ideas.
Lexical diversity Percent of K2 tokens Average word length Percent of K1 tokens Word families	Biber et al., 1999; Lu, 2012; Iwashita et al., 2008; Malvern & Richards, 2002; Jamieson & Poonpon, 2013; Yu, 2010	• Uses a range of vocabulary to discuss the topics required by the task. • Some awkward usage or slightly inappropriate lexical choices.
Fluency Silent pause average duration Tone unit average length Filled pause average duration Syllable rate (syllable per second) Articulation rate	Brown et al., 2005; Brown, 2006; Ginther et al., 2010; Iwashita et al., 2008; Jin & Mak, 2012; Kang, 2010; Kang et al., 2010; Kormos & Dénes, 2004; Munro & Derwing, 2001	• Noticeable pausing, false starts, reformulations and repetition. • Backtracking and reformulations not fully interrupting the flow of speech.
Prosody and pronunciation Prominent syllables per tone unit Mid-rise, low-rise, low-fall, high-rise-fall, and high-fall-rise tone choice Overall pitch range	Hewings, 1995; Kang et al., 2010; Pickering, 2004, 2018; Wennerstrom, 1994; 2000	•Inappropriate mispronunciations put an occasional strain on the listener. •Pronunciation is intelligible.

4. DISCUSSION

The English proficiency of speakers using unconstrained speech was scored automatically in four levels in two other studies. In this paper, we presented a computer model for automatically scoring the English proficiency of unconstrained speech. In a test with the Aptis corpus, the Pearson correlation between the automatic scores from the computer model and the scores assigned by the British Council examiners was 0.90 in Phase 1 and 0.76 in Phase 2. (It should be noted that the results are subject to the accuracy of transcripts provided by human transcribers and do not take into account errors that could occur at the speech recognition stage). 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. This is even higher than Kang & Johnson's (2018) study, which has a correlation of 0.71. The results also imply that suprasegmental measures, along with lexical and grammatical features, are important with regard to automated English proficiency scoring systems for unconstrained speech. This has also been shown to be true for human judgments (Kang et al., 2010; Kang & Johnson, 2018).

Evanini and Wang (2013) used linear regression of 10 features extracted from the output of an ASR configured to recognise the words to automatically score the spoken English responses given by non-native children in an English proficiency assessment of middle school students. The Pearson 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.

Phase 1 (21 speakers – 72 measures per speaker – 4 Aptis levels) of the experiment showed a very strong correlation of 0.90 between lexico-grammatical features and APTIS scores. Phase 2 (105 speech files – 36 measures per file– 5 Aptis levels) indicated a strong correlation of 0.76 between the features and APTIS scores.

The difference between the correlations of the two phases of the experiment could be due to a number of factors. Phase 1 involved more measures (72 vs. 36), fewer Aptis scores (A2-C in Phase 1 and A1-C in Phase 2), and fewer samples (21 vs. 105). More measures typically leads to better machine learning results, which might be why the Phase 1 correlation was stronger. However, the genetic algorithm indicated that only 15 of the 72 Phase 1 measures were salient versus 15 of the 36 Phase 2 measures. (Note: The 15 measures from Phase 1 were completely different from those in Phase 2.) Machine learning tools also perform dramatically better with fewer target classes (Aptis scores in these experiments). The fewer Aptis scores in Phase 1 (4 vs. 5) is a more likely cause of the stronger Phase 1 correlation than the number of measures. Equally important to the success of machine learning tools is the number of samples. A small number of samples tends to over-train the computer model which results in better performance than would be expected over a larger population. Consequently, the relatively small number of samples for Phase 1 (21 vs. 105) was also a more probable cause of the stronger correlation than the number of measures.

The expected gradient of increasing complexity per level was found for most of the measures (e.g., number of error free T-units, world families, word length, subject-verb agreement errors, etc.). These findings concur with previous literature (e.g., Iwashita et al. 2008), using the TOEFL iBT spoken data. The complexity of these features was significantly different for each level in this study. The computer also was able to choose these lexical grammatical features to distinguish proficiency levels. Moreover, in terms of grammatical accuracy, the frequency of errors decreased as proficiency increased. Especially, variables such as *article, coordinator, complement clause, determiner, negation, non-finite clause, relative clause, singular/plural, subjunctive structure, and copular errors* were particularly outstanding in the computer selection process. This is, in fact, a common pattern found in various studies manually coded by human researchers (Iwashita et al., 2008; Skehan & Foster, 1999).

In two phases of computer experiments for the lexico-grammatical analysis, some features emerged significantly, which included the following: error-free t-units, average word length, complement clause

errors, and singular/plural errors. It means that, as proficiency increased, Aptis test-takers produced fewer error-free t-units and agreement/clausal errors and their word length become longer. The findings from grammatical measures reflected the complexity of utterances at the level of clause relations and within-sentence sophistication. These results concur with the findings of previous studies (e.g., Brown et al., 2005; Brown, 2006) concluding that advanced learners used more features such as be-copula as the main verb (Hinkel, 2003) or subordinators, verbs, or pronouns (Espada-Gustilo, 2011, Grant & Ginther, 2000).

In Phase 1, dependent/subordinate clause error was selected by the computer model, which is in line with Norrby and Hakansson's (2007) study in which subordination proved a significant indicator of complexity. This finding supports Foster and Skehan's (1996) study on L2 oral performance, illustrating that amount of subordination can be used to measure L2 learners' oral complexity. In Phase 2, another salient feature was relative class errors, which was found to be a significant variable in Lennon's (1990) longitudinal study of the development of NNSs' oral performance over 23 weeks where the use of relative clauses increased from 18–118% between weeks 1–17 and 19–23.

In Phase 2, examining 105 responses, error rates of some features (e.g., number of error free T-units, articles, singular/plural, and subject-verb agreement) dropped substantially as proficiency increased; therefore, the computer program was able to extract these features. This finding is parallel with previous findings seen from Kang &Yan's (2018) study which investigated linguistic features that distinguish proficiency levels in Cambridge English Language Assessment. In addition, compared to high-level respondents, low-proficient speakers used more clausal coordinators (e.g., *and* or *but*), but high-level candidates used grammatically more complex and more structured expressions and phrases such as emphatics, *be* as a main verb, subordinate clauses, perfect aspects, time adverbs, modals, and conjunctions. Additionally, they used more complicated adverbial expressions such as causal adverbs, *to*-infinitives, and adverbial subordinators for condition. Overall, the occurrence of grammatical features which involve more complicated forms of structure or arrangement seem to increase with higher proficiency levels, but simplified forms or content-word based formations, such as nouns, were used among low-proficiency speakers.

The results of lexical analysis also revealed noticeable patterns in the computer extraction process. Increase in proficiency resulted in an increase in the number of words produced (tokens) and a wider range of words (types). This finding is in agreement with many other previous studies (e.g., Iwashita et al., 2008; Jamieson & Poonpon, 2013; Malvern & Richards, 2002; Yu, 2013). Previous research findings suggested that the increase in proficiency is associated with the increase in the amount of vocabulary produced and the range of words used **(**e.g., Biber, Johansson, Leech, Conrad & Finnegan, 1999; Lu, 2012). There was also a significant increase in the frequencies of the 1,000-word usage and the choices of various word families as proficiency levels improved. In addition, words chosen by high-proficiency candidates were longer than those by low-proficiency candidates especially in the level changes from lower to higher proficiency levels.

With regard to fluency features, the computer selected the silent pauses, articulation rate, and syllable per second. The silent and filled pause frequency was negatively associated with proficiency level while the tone unit average length was positively related. In other words, proficient candidates produced more syllables per second and used less pauses as seen from other studies (Ginther, Slobodanka, & Yang, 2010; Jin & Mak, 2012; Kang, 2010; Kang et al., 2010; Kormos & Dénes, 2004), but they produced a longer unit of speech. This suggests that examinees talked faster with a shorter duration of pauses and hesitation markers in their monologues as their proficiency increased.

In terms of prosody features, low-proficiency speakers emphasised words with stress more frequently than high-proficiency speakers. Typically, low-proficiency NNSs use primary stress on every lexical item, regardless of its function or semantic importance (Kang, 2010; Wennerstrom, 2000). Given that the importance of intonation and tone use is well recognised (Kang et al., 2010), the computer model in the current study also included tone choices (e.g., low rising, mid rising, low falling, high-rise falling, high-fall rising, and overall pitch range) for prosody measures.

Among the 15 tone choices, the frequencies of high-rising, mid-rising, and low-falling tones were selected to distinguish across proficiency levels. The findings were parallel with Kang et al.'s study (2010). That is, while mid-rising and high-rising tones were positively associated with proficiency, mid-level and low-falling tones were negatively associated with proficiency. The use of rising tone has been particularly emphasised in the native speaker's discourse context (Brazil, 1997) as it can signal solidarity with speakers or common group or shared background. In the situation of discourse production, for example, it has been known that non-native, low-proficiency speakers tend to use low-falling tones between related propositions, whereas rising and mid-level tones would be anticipated by NS listeners (Hewings, 1995; Pickering, 2001, 2018; Wennerstrom, 2000). Furthermore, in line with previous research findings (Kang, 2010; Wennerstrom, 1994), the number of prominent syllables per tone unit decreased as proficiency increased. Overall, these tone choice and prominence variables appeared to be good indicators of distinguishing candidates' speaking performance across Aptis levels for the criterion of pronunciation

Overall, the speaking features used by the computer to assess oral proficiency in the Aptis speaking test are well supported by previous research findings. As for the criterion of the grammatical complexity and accuracy, the current study has selected various error types and complexity (e.g., dependent clause and formation of subjunctive), which are closely linked to learners' proficiency (e.g., Iwashita et al. 2008; Kang & Yan, 2018). Lexical features selected (i.e., percent of K2 tokens, average word length, percent of K1 tokens, and word families) are also well in line with previous research **(**e.g., Biber et al., 1999; Iwashita et al., 2008; Jamieson & Poonpon, 2013). The computer program has further selected most of the fluency features (i.e., speech rates and pauses) which are often recommended as strong indicators of oral proficiency by various research studies (e.g., Jin & Mak, 2012; Kang, 2010; Kang et al., 2010; Kormos & Dénes, 2004). Finally, research in the field of L2 pronunciation can sufficiently approve of the tone choice, prominence, and pitch range features selected for the current study as their association with oral performances has been well documented (e.g., Kang et al., 2010; Kang & Yan, 2018).

5. CONCLUSION

The project identified computer-selected salient linguistic features that characterise Aptis speaking performances, which can be used for the development of automated speaking assessment. It offers validity evidence for the Aptis speaking test by describing relationships between linguistic features of key criteria determined by human raters and those by computer. Even though the project could benefit from more controlled speech tasks with a bigger sample size, the findings demonstrated that various linguistic features could be used to improve the computer model's prediction of proficiency beyond what could be predicted by fluency features. Future research could involve one task (e.g., Task 4 only) with the same topic to see if the same linguistic features could emerge. In addition, the current project did not include any linguistic features related to cohesive devices, which could be added to the future studies.

Overall, the project not only introduces a new computer-modeling approach to the field of second language assessment, but also applies this comprehensive method to the Aptis speaking test. It further helps better understand non-native speakers' speech properties in testing and assessment contexts.

REFERENCES

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finnegan, E. (1999). *Longman Grammar of Spoken and Written English*. Essex: Pearson Education Limited. Educational Testing Service.

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

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

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. *IELTS Research Reports Vol 6,* (pp. 71–89). Canberra & London: IELTS Australia and British Council.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purpose speaking tasks* (TOEFL Monograph No. 29). Princeton, NJ: Educational Testing Service.

Cobb, T. (2002). *The Web Vocabulary Profile*. http://www.er.uqam.ca/nobel/r21270/textools/web_vp.html

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R. & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, pp. 5–34.

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

Espada-Gustilo., L. (2011). Linguistic features that impact essay scores: A corpus linguistic analysis of ESL writing in three proficiency levels. *The Southeast Asian Journal of English Language Studies*, *17* (1), pp. 55–64.

Foster, P., & Skehan, P. (1996). The influence of planning and performance in task based learning. *Studies in Second Language Acquisition, 18,* pp. 299–324.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21,* pp. 354–375.

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), pp. 379–399.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9* (2), pp. 123–145.

Hewings, M. (1995). Tone choice in the English intonation of nonnative speakers. *International Review of Applied Linguistics, 33*(3)*,* pp. 251–266.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, *37*(2), pp. 275–301.

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. *NCTE Research Report No. 3*.

Iwashita, N, Brown, A, McNamara, T, and O'Hagan, S (2008) Assessed levels of second language speaking proficiency: How difficult? *Applied Linguistics* 29, pp. 24–49.

Jamieson, J., & Poonpon, K. (2013). *Developing analytic scoring guides for TOEFL iBT's* Speaking Measure. TOEFL monograph Series. Retrieved from: https://www.ets.org/Media/Research/pdf/RR-13-13.pdf

Jin, T., & Mak, B. (2012). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing, 30*(1), pp. 23–47.

Johnson, D., Kang, O., & Ghanem, R. (2016). Improved automatic English proficiency rating of unconstrained speech with multiple corpora. *International Journal of Speech Technology, 19* (4), pp. 755–768. DOI: 10.1007/s10772-016-9366-0

Johnson, D., & Kang, O. (2016). Automatic detection of Brazil's prosodic tone unit. *Speech Prosody 2016*. https://drive.google.com/file/d/0B5tX0NnTCf-IV0wwbUw3a2FGWVE/view

Kang, O. (2010). Relative salience of suprasegmental features on judgements of L2 comprehensibility and accentedness, *System* 38, pp. 301–315.

Kang, O. and Johnson, D. O. (2018). Phone-based automated English proficiency scoring of unconstrained speech using prosodic features. *Language Assessment Quarterly*.

Kang, O., & Johnson, D. O. (2015). Comparison of Inter-rater Reliability of Human and Computer Prosodic Annotation using Brazil's Prosody Model. *English Linguistics Research*, 4(4), pp. 58–68.

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), pp. 554–566.

Kang, O., & Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing and Assessment, 1,* pp. 24–39,

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

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40,* pp. 387–417.

Loukina, A., Davis, L., & Xi, X. (2017). Automated assessment of pronunciation in spontaneous speech. In O. Kang & A. Ginther (Eds.) *Assessment in second language pronunciation* (pp. 153–171). New York: Routledge.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96* (2), pp. 190–208.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), pp. 85–104.

Messick, S. (1975). Meaning and values in measurement and evaluation. *American Psychologist*, *30*, pp. 955–966.

Munro, M. J., & Derwing, T. M. (2001). Modelling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies of Second Language Acquisition, 23*, pp. 451–468.

Norrby, C., & Hakansson, G. (2007). The interaction of complexity and grammatical processability: The case of Swedish as a foreign language. *International Review of Applied Linguistics, 45,* pp. 45–68.

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., Dunn, K. (2020). *Aptis General Technical Manual,* version 2.2. London: British Council.

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In O'Sullivan, B. (Ed.) *Language Testing: Theories and Practices* (pp. 13–32), Basingstoke: Palgrave Macmillan.

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*(2), pp. 233–255.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes, 23*(1), 19.

Pickering, L. (2018). Discourse Intonation: A discourse-pragmatic approach to teaching the pronunciation of English. Ann Arbor: The University of Michigan Press.

Shaw, S. D. and Weir, C. J. (2007) Examining Writing: Research and practice in assessing second language writing. Studies. *Studies in Language Testing, 26*, Cambridge: UCLES/ CUP.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49,* pp. 93–120.

Sun, Y. (2011). The influence of the social interactional context on test performance: A sociocultural view. *The Canadian Journal of Applied Linguistics*, *14* (1), pp. 194–221.

Van Moere, A., & Suzuki, M. (2017). Using speech processing technology in assessing pronunciation. In O. Kang & A. Ginther (Eds.) *Assessment in Second Language Pronunciation* (pp. 137–152). New York: Routledge.

Weir, C. J. (2005). *Language Testing and Validation*. Hampshire: Palgrave McMillan.

Wennerstrom, A. (1994). Intonational meaning in English discourse: A study of nonnative speakers. *Applied Linguistics*, *15*, pp. 399–421.

Wennerstrom, A (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.) *Perspectives on fluency* (pp. 102–127) Ann Arbor, MI: University of Michigan.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (No. 17). University of Hawaii Press.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), pp. 236–259.

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), pp. 883–895.

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

LINGUISTIC FEATURES AND AUTOMATIC SCORING OF APTIS SPEAKING PERFORMANCES

Okim Kang
Northern Arizona University
David Johnson
University of Kansas

AR-G/2021/3

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203