Capturing L2 Oral Proficiency with Composite CAF Measures:

A Focus on Fluency

MIYAMOTO, Mayu (Nagoya University of Foreign Studies)

Summary

Despite an emphasis on oral communication in language courses, the resource-intensive nature of speaking tests hinders regular oral assessments. A possible solution is the development of a (semi-) automated scoring system, as the consistency of computers can complement human raters' comprehensive judgments and increase efficiency in scoring. In search of objective and quantifiable variables, a number of studies have reported that some utterance-fluency variables (e.g., speech rate) are strongly correlated with overall L2 oral proficiency. While these studies focused on finding a single fluency variable as a predictor, given the complex nature of L2 oral proficiency, it is also important to examine a composite variable to predict learner proficiency. Consequently, this study investigated the relationship between complexity, accuracy, and fluency (CAF) variables and L2 oral proficiency. Utilizing audio samples from the Oral Proficiency Interview (OPI), a well-established speaking test by the American Council on Teaching Foreign Languages, this study analyzes spontaneous speech samples collected from 170 L2 Japanese learners with varied proficiency levels. The first part of the study investigated the relationship between CAF variables and learners' oral proficiency. The results revealed that speech speed and complexity variables demonstrated strong correlations to the OPI levels, and moderately strong correlations were found for the variables in the following categories: speech quantity, pause, pause location (silent pause ratio within AS-unit), dysfluency (repeat ratio), and accuracy. The second part investigated an optimal composite measure that could best predict the OPI levels. A series of multiple regression analyses revealed that a combination of five measures (effective articulation rate, silent pause ratio, repeat ratio, syntactic complexity, and error-free AS-unit ratio) can predict 72.3% of the variance in the OPI levels. This regression model includes variables that correspond to three categories of fluency (speed, breakdown, and repair) and variables that represent CAF.

Key words : oral proficiency, assessment, automated scoring, fluency, CAF

1. Introduction

As second language (L2) proficiency is transitional and invisible, assessment plays a crucial role in monitoring the learning progress and achievement of learners. Ideally, the four language skills (listening, reading, writing, and speaking) should be assessed separately as they develop at different speeds. In reality, however, L2 oral proficiency assessments are not administered as frequently as they are for other skills due to the complexity of assessing productive skills. While receptive skills can be assessed easily by multiple-choice or fill-in-the-blank style items, productive skills require learners to provide constructive responses, such as a composition or oral interaction with an examiner. Furthermore, written and spoken responses usually need to be rated by experts in their field. Although it is necessary, it is unfortunately not practical to conduct frequent oral tests, as it is extremely costly in terms of time, human resources, and money. One approach to the cost problem may be (semi-) automated scoring. This article is a basic research study that explores algorithms to support such automation.

The development and adaptation of automated scoring have recently become popular in the language testing industry (e.g., ETS, Pearson Education; Xi, 2010). For example, the TOEFL iBT writing task is now rated in part by an automated scoring system. Enright and Quinlan (2010) claim that the consistency of computers can complement human raters' comprehensive and sophisticated judgments and increase the efficiency of scoring. For writing tasks, the computer looks at surface linguistic features, such as length of essay and vocabulary sophistication, to predict overall writing proficiency. It cannot replace humans since it does not score the essays with the same sophistication as trained human raters; however, when it is used to support human raters, it can increase efficiency. This leads one to ask, could this be applied to speaking tests as well? Unfortunately, when compared to written responses, speech is much more fragmented, repetitive, and unstructured, which makes automated evaluation even more difficult (Xi, 2010). With current voice recognition technology, it is not yet feasible. Researchers have attempted to tackle this problem by investigating the possibility of using fluency variables, quantifiable subcomponents of oral skills, as predictors of overall speaking proficiency.

2. Literature Review

2.1 Complexity, Accuracy, and Fluency (CAF) as Concepts of Oral Proficiency

In second language acquisition (SLA), many researchers hold the view that L2 proficiency is a multidimensional construct rather than a unitary one, and it can be captured by assessing complexity, accuracy, and fluency (CAF; Housen et al., 2012). Skehan (1996) was the first to combine the three components, proposing that complexity (e.g., syntactic complexity, elaboration in speech), accuracy (e.g., grammatical correctness), and fluency (e.g., ease of L2 speech production) are key elements of L2 proficiency. These components were identified as distinct and competing areas of L2

proficiency, suggesting all three components must be considered together, not separately (Larsen-Freeman, 2009).

Various procedures have been employed to capture or evaluate CAF, including holistic/subjective and objective quantitative measures, but the latter seems to be the preferred method in L2 production (Housen et al., 2012). One important issue that has been addressed by some researchers is the validity of these measures and their inconsistent application. Many of the CAF measures that were originally developed for L2 writing studies have been adapted to L2 speaking studies, and therefore, need some adjustments to account for the unique characteristics of speaking. Sakuragi (2011) investigated the construct validity of 10 CAF measures using L2 Japanese narrative speech samples. The results of a factor analysis revealed that the validity of syntactic complexity and accuracy measures were supported, but not fluency measures. The fluency variables, the speed measure, and the dysfluency measure did not share the same factor as one construct of fluency, and the author opined that it might be due to a lack of consensus on the definition of fluency and its measurements. The author encouraged further research with a greater number of fluency measures to investigate the validity of fluency measures and indicated that while several CAF investigations were conducted in Indo-European languages, there are not enough studies done in languages such as Japanese. Encouraged by Sakuragi's (2011) research, this study reviews definitions of fluency and investigates which fluency measures can validly capture L2 proficiency when combined with accuracy and complexity measures. Since the measures for accuracy and complexity were found to be valid in Sakuragi's study, this study adopts one measure from each category and investigates various fluency measures.

2.2 Definitions of Fluency

Many researchers agree that fluency is a fundamental component of L2 oral proficiency; however, no consensus has been reached on its definition (Chambers, 1997; Lennon, 1990). Since the conceptualization of the term remains vague and is difficult to operationalize, Lennon's (1990) narrower sense of fluency and the use of temporal measures have attracted researchers' attention. While some researchers argue that the complex nature of fluency cannot be reduced to a handful of temporal measures, others argue that the term fluency should be restricted to temporal and other fluency-related characteristics of spoken discourse, because they are observable, quantifiable, and therefore reliable (Chambers, 1997; Leclercq et al., 2014).

Although there is some variability, many researchers seem to agree that speed-related aspects of speech, especially *speech rate*, *mean length of run*, and *articulation rate*, are strongly correlated with overall oral proficiency (e.g., Ginther et al., 2010; Lennon, 1990). Chambers (1997) encourages more research into temporal variables in speech production to provide valuable empirical evidence that "can contribute to a more precise definition of fluency" (p. 535). In response to Chambers' suggestion, this study takes an empirical-based approach to capture L2 oral proficiency with CAF

measures and focuses on fluency-related measures.

2.3 Motivation for the Present Study

While previous studies have contributed significantly to understanding L2 fluency, most studies were conducted with a small number of participants with a limited range of proficiency levels. Additionally, much of the audio data used for the analyses are responses to a controlled task (e.g., a recall picture description task), rather than spontaneous speech. Although it is important to control for task variability, it is also necessary to investigate what information can be obtained from unprepared audio samples. Such data can provide information about learners' true ability to carry on a conversation in a natural discourse. Furthermore, while there is a considerable amount of literature discussing fluency of English learners, little is known about other languages. It is worthwhile to investigate whether the findings of studies on English learners are applicable to other languages, especially in Japanese, where empirical data are scarce. While most previous studies have attempted to find a single fluency variable that is able to predict L2 proficiency levels, to the best of the authors' knowledge, no study has attempted to find a composite variable including CAF to predict learner proficiency. Because L2 proficiency is a complex ability to capture, it is necessary to consider multiple variables.

Considering these research gaps, the current study investigates the relationship between CAF variables and L2 oral proficiency. Utilizing audio samples from the American Council on Teaching Foreign Languages' (ACTFL) well-established speaking test, the Oral Proficiency Interview (OPI), this study analyzes spontaneous speech samples collected from 170 L2 Japanese learners at a wide range of proficiency levels. The first part is a correlational study that investigates the relationship between CAF variables and learner oral proficiency assessed by the OPI. The second part of the study created an optimal composite measure for predicting the OPI levels. The study's findings will contribute to the discussion of using a composite CAF measure as a predictor of L2 overall proficiency and the future development of a (semi-) automated scoring system. The research questions (RQs) include:

RQ 1: Which CAF variables correlate with L2 Japanese proficiency levels measured by the ACTFL OPI, and to what extent do they correlate?

RQ 2: Which combination of CAF variables can best predict examinees' L2 proficiency levels?

3. Methodology

3.1 The Database

The speech samples used in this study were obtained from the L2 Japanese learners' conversation online database (https://mmsrv.ninjal.ac.jp/kaiwa/), published by the National Institute for Japanese Language and Linguistics in 2009. The database includes 339 transcriptions of each participant's

30-minute face-to-face interaction at the ACTFL OPI session, and 215 of them have accompanying audio recordings. The database also details the awarded OPI proficiency level of each participant along with their background information. The ACTFL OPI is widely known as a reliable and valid test for assessing language proficiency, and its scale has 10 levels, ranging from novice to superior[1)], with sublevels. Each speech sample is carefully rated by two to three human raters to determine the awarded OPI level. Moreover, the ACTFL OPI provides speech samples in a dialog format rather than a monolog format (e.g., talking to a computer), which has higher face validity for measuring test takers' functionality in daily conversations. Since this database can provide (i) speech samples obtained from a valid and reliable speaking test, (ii) samples from a wide range of proficiency levels, (iii) a large number of samples, (iv) spontaneous speech samples, and (v) speech samples in Japanese, it was selected as a suitable data source for this study.

3.2 Procedure

3.2.1 Retrieving Audio Recordings

For the analyses, 170 out of 215 audio recordings were retrieved from the database. Since the coding process was labor-intensive, the number of audio samples for Intermediate–Mid and Intermediate–High was limited to 30 to ensure the quality of the coding process. Table 1 summarizes the number of samples used in this study.

Table 1. Number of samples at each level

OPI Levels	Database	Current Study
Novice–Low	0	0
Novice–Mid	6	6
Novice–High	12	12
Intermediate–Low	21	21
Intermediate–Mid	58	30
Intermediate–High	47	30
Advanced–Low	27	27
Advanced–Mid	20	20
Advanced–High	19	19
Superior	5	5
Total	215	170

3.2.2 Speech Sample Selection

Due to the adaptive nature of OPI, various speech tasks were conducted during a 30-minute interview session. Although there is a well-established procedure for administering the OPI, the questions asked by testers are not predetermined. Rather, the tester asked questions according to the

examinees' level and natural conversation flow, using level checking and probing approaches. In order to account for task variability, this study focused on responses to descriptive speech tasks. A descriptive speech is defined as a speech segment where an examinee provides a new piece of information by explaining or describing a particular person, object, location, event, or one's thoughts or reasons. Example questions that elicit descriptive speech responses are, "What are the differences between your hometown and your current residence?" or "Tell me about the most famous food from your city." If one examinee provided several descriptive task responses, the longest one with the most information was selected because it represents one's best performance, regardless of the tester variability. After undergoing a careful selection process, speech samples were extracted from the original recordings and noise was reduced to maximize the audio quality.

3.2.3 Data Processing and Coding

The retrieved speech samples were coded by the researcher using *Praat* (Boersma & Weenink, 2017) based on the following eight categories: (i) morae count, (ii) sounding and silent boundaries, (iii) filled pause boundaries, (iv) Analysis of Speech (AS) unit[2] boundaries, (v) AS-unit with or without grammatical errors, (vi) clause counts within an AS-unit, (vii) sound boundaries for dysfluency factors (repetitions, stutters, and self-corrections), and (viii) sentence count. To test for coding consistency, 10 speech samples were randomly selected from the 170 samples and coded by another expert in the field, and a high degree of reliability was found between the two coders.[3]

3.2.4 CAF measures

The coded data were then submitted to a tool called the CAF Calculator (Fukada et al., 2019). It is a computer software that automatically generates 50 objective CAF measures from annotated *Praat* scripts. For the current study, 18 measures, that are not affected by the extracted speech sample lengths, were used for the analysis. This is because the average speech sample ranged from 17.85 to 66.48 seconds depending on the OPI levels, and the removed output measures were computed based on the speech length, making the comparison across the levels difficult. Since the 18 measures are those incomparable measures converted into ratio or rate measures, the selected measures can represent the removed ones. For explanations and calculations, see the Appendix.

The most frequently reported fluency variables that are positively correlated to the proficiency levels in the literature are *speech rate*, *articulation rate*, and *mean length of run*. While *speech rate* and *articulation rate* account for speech speed, *mean length of run* represents speech speed and density. However, the researcher found the variable, *mean length of run*, to be somewhat problematic. *The mean length of run* is typically calculated as (total number of syllables) / (total number of runs in a given speech sample). What is problematic here is that this formula does not consider repairs and other dysfluency phenomena. For example, if the total number of syllables includes some repetitions, those syllables should not be counted because that portion does not add density to the speech. If one wants to capture speech density more accurately, it makes more sense to eliminate

syllables for dysfluency markers such as stutter, self-correction, and repetition. Therefore, in this study, two new speech-density variables were proposed: *effective speech rate* and *effective articulation rate*. *Effective speech rate* represents how fast a speaker can produce effective syllables within the total response time, and *effective articulation rate* represents how fast a speaker can produce effective syllables if they are not interrupted by any pauses or dysfluency markers. Although these two new measures are introduced in this study, the original *speech rate* and *articulation rate* were kept in the analysis to make comparisons between the findings and the previous literature.

3.3 Data Analysis

IBM SPSS Statistics version 25 was used for statistical analyses. The independent variable was the ACTFL OPI level (a total of 10 levels including sublevels), and the dependent variable was the 18 CAF measures. To address RQ1, Spearman's rank-order correlation coefficients were calculated between the CAF measures and the OPI levels, as well as the correlation among the CAF measures. Next, a series of multiple linear regressions was conducted to investigate which combination of the CAF variables can best predict L2 Japanese speakers' proficiency levels (RQ2).

4. Results and Discussion for RQ1

4.1 Results (Relationships between the 18 CAF measures and OPI levels)

Between the 18 CAF measures and OPI levels, strong correlations ($|r| = .60 – .79$, $p < .001$) were observed for the following six measures in three categories (refer to Appendix for categories and measure numbers):

Speed: 10. Speech rate ($r = .74$), 11. Articulation rate ($r = .64$)

Speed/Density: 12. Mean length run ($r = .67$), 13. Effective speech rate ($r = .78$),
14. Effective articulation rate ($r = .67$)

Complexity: 45. Syntactic complexity ($r = .63$)

Moderately strong correlations ($|r| = .40$ - $.59$, $p < .001$) were observed for the following seven measures in the five categories:

Speech Quantity: 9. Phonation time ratio ($r = .57$)

Pause: 19. Silent pause ratio ($r = -.55$), 20. Silent & filled pause ratio ($r = -.56$)

Pause Location: 29. Silent pause ratio within AS ($r = -.44$)

Dysfluency: 40. Repeat ratio ($r = -.41$), 43. Dysfluency ratio ($r = -.42$)

Accuracy: 48. Error-free AS-unit ratio ($r = .46$)

Weak correlations ($|r| = .20 – .39$, $p < .001$) were found for the following three measures in the *Pause Location* category: 30. Silent & filled pause ratio within AS-unit ($r = -.37$), 31. Ratio of silent pause time between AS-unit to total response time ($r = -.24$), and 32. Ratio of silent & filled pause time between AS-unit to total response ($r = -.24$).

4.2 Discussion (Relationships between the 18 CAF measures and OPI levels)

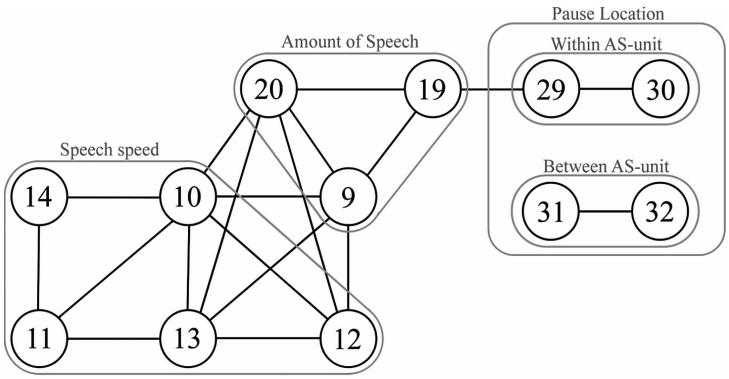
Strong correlations were found with *Speed* (*speed/density*) and *Complexity* measures. Of all six categories, *Speed*-related measures demonstrate the strongest relationship with OPI levels, because all five measures in this category show strong positive correlation coefficients. This finding is consistent with the previous literature, because the top three most frequently reported measures correlating strongly to the OPI levels (*speech rate, mean length of run,* and *articulation rate*) were also found to be strong in this study. Among these three measures, *speech rate* demonstrated the strongest correlation, followed by *mean length of run* and *articulation rate*, in that order. Interestingly, although most of the previous findings are based on English, the results of the current study suggest that they can be extended to Japanese as well. This indicates that these speed-related measures, especially *speech rate*, may be applicable cross-linguistically. Moreover, this study included two new measures that represent *Speed* and speech *Density* (*effective speech rate* and *effective articulation rate*). When correlation coefficients are compared, both new measures show stronger correlations than the measures used earlier (*speech rate, mean length of run,* and *articulation rate*). Notably, the *effective speech rate* showed the highest correlation coefficients ($r = .78$, $p < .001$) of all *Speed*-related measures. This is probably because, while *speech rate* only accounts for speech speed, *effective speech rate* considers speech density (only counting meaningful production). This means that as the OPI level advances, the rate of producing meaningful syllables increases. In addition to *Speed-related* measures, the variable representing *Complexity* (*syntactic complexity*) shows a correlation as high as *Speed*-related variables. Although the complexity measure in this study was very simple, the results indicated that *syntactic complexity* has a strong relationship with OPI levels.

Moderately strong correlations were found in *Speech Quantity, Pauses* (also *Pause-Location*), *Dysfluency*, and *Accuracy* variables. The results suggest that as the OPI rating advances, the amount of time spent on speech increases as the pausing time decreases. This finding supports the previous findings. Furthermore, the results revealed that there is only a weak relationship between *Pause-location* and OPI levels, except for *the silent pause ratio within AS-unit*. The negative correlation coefficient suggests that the more silent pauses within AS-units, the lower their OPI rating is; the pauses outside of AS-units do not matter much. For *Dysfluency* variables, among the three types (repetition, stutter, and self-correction), only the *repeat ratio* showed a moderate relationship with OPI levels. The negative correlation coefficient suggests that as examinees' OPI level advances, the amount of time spent on repetition decreases; however, the occurrence of stutter or self-correction does not have much effect on the OPI levels. The *Accuracy* variable was also moderately correlated. Similar to the *Complexity* variable, the *Accuracy* measure follows a simple method of quantifying the grammatical/vocabulary accuracy by counting the number of AS-units with or without errors. Despite the simple method, it still demonstrated a moderate relationship with OPI levels. The positive correlation coefficient of *the error-free AS Unit ratio* suggests that as oral proficiency improves,

examinees can speak more accurately in terms of grammar and vocabulary.

4.3 Results and Discussions (Relationships among the 18 CAF measures)

Since some of the 18 CAF measures are believed to be closely related to each other, the intercorrelation among the CAF measures was also calculated. As expected, strong to very strong correlations were found, especially among *Speech quantity*, *Speed*, and *Pause-related* measures. This section only focuses on the high correlation ($|r| > .80$), as they might represent the same construct and are highly dependent, and those with weak correlations are considered as unique or relatively independent measures. Figure 1 visualizes the complex relationships of all pairs that are strongly correlated with each other. The relationships can be categorized into three major groups: *Speech speed*, *Amount of speech*, and *Pause-location*. Five variables represent speech speed: *speech rate, articulation rate, mean length of run, effective speech rate*, and *effective articulation rate*. These measures are strongly correlated because the only difference between them is the inclusion or exclusion of pauses and dysfluency phenomena. The *Amount of speech* category consists of *phonation time ratio*, *silent pause ratio*, and *silent & filled pause ratio*. Although the latter two are pause-related measures, they also represent the amount of speech because the opposite of time spent for pausing is speaking time with or without filled pauses. Lastly, variables that represent *Pause-location* also show very strong correlations. Figure 1 shows that the pauses made within an AS-unit have a greater impact on the other fluency variables; the pauses between AS-units do not.



9. Phonation time ratio
10. Speech rate
11. Articulation rate
12. Mean length run
13. Effective speech rate
14. Effective articulation rate
19. Silent pause ratio
20. Silent & Filled pause ratio
29. Silent pause ratio within AS-unit
30. Silent & Filled pause ratio within AS-unit
31. Ratio of silent pause time between AS-unit to total response time
32. Ratio of Silent & Filled pause time between AS-unit to total response time

Figure 1. Visualization Map for Very Strongly Correlated Variables

5. Results and Discussion for RQ2

5.1 Procedure and Testing of Assumptions

A series of multiple linear regression (MLR) analyses[4] using the stepwise approach was conducted to find a parsimonious model that could best predict the OPI levels with the CAF measures currently available for the study. In this study, predictive power was defined as adjusted R^2, and efficiency was defined as the number of predictors in the model. The model fit was evaluated in terms of predictive power and efficiency and carefully examined for theoretical plausibility in light of the literature. The data satisfy all assumptions except for multicollinearity. As explained in the previous section, this was expected as some variables displayed strong correlations with one another. Although the data violate the assumption, since this study is exploratory, it is more meaningful to keep all the variables and keep adjusting them throughout the process of identifying the final model. Once the final model is determined, the model is tested again to ensure that there is no multicollinearity issue.

5.2 Analyses and Results

As a first step, a stepwise multiple regression was used to evaluate whether all 18 CAF variables were necessary to predict the OPI levels, and which variable had the most predictive power. The first variable added to the regression equation was *effective speech rate*, showing a significant relationship to the OPI levels $F(1, 168) = 269.23$, $p < .01$, $R^2_{Adjusted} = .613$. Five other variables were added to the model; however, since the aim of this analysis was to determine which predictor entered the model first, this section only focuses on the first predictor. The fact that SPSS stopped adding more variables to the model after a few steps indicated that not all 18 CAF variables are necessary to predict the OPI levels. The result suggested that the *effective speech rate* might be the primary predictor; however, it cannot be concluded yet without further investigation due to the multicollinearity issues. Given the strong correlations among the *Speed* variables, the primary predictor could be any of the five variables. By manually entering each of the *Speed* variables in the regression model, the multicollinearity issue was controlled. The five multiple regression models were then compared. The results showed that the *effective articulation rate* was the best among the *Speed* variables when used in combination with the *silent-pause ratio, repetition ratio, syntactical complexity,* and *error-free AS-unit*. Although this output model seems promising, there is one problem: the predictor includes the variable *silent-pause ratio*. As seen in Figure 1, this variable is strongly correlated with *Amount of speech* variables. Although no multicollinearity issues were found in this model, output models with different *Amount of speech* variables were compared to refine the model further. The results revealed that the model with the *silent pause ratio* demonstrated the greatest adjusted R^2 value with the least number of predictors.

After comparing several models and controlling for multicollinearity issues, it was found that a combination of the following five CAF variables can best predict the OPI levels ($F(5, 164) = 89.43$,

$p < .01$, $R^2 = .86$, $R^2_{Adjusted} = .723$). The regression equation is: *The OPI levels = 0.73 + 0.01 (effective articulation rate) - 0.04 (silent pause ratio) - 0.25 (repeat ratio) + 0.36 (syntactic complexity) + 0.01 (error-free AS-unit ratio).* In this model, approximately 72.3% of the variance in OPI levels can be explained by these five predictors. Table 2 summarizes the final model. The Variance Inflation Factor (VIF) values indicate that the correlations among the predictors are very low; therefore, there is no multicollinearity issue, and each variable makes a unique contribution to the model.

Table 2. Summary of Final Model

Variable	Cumulative R^2	R^2 Change	*b*	Beta	*t*	VIF
14. Effective articulation rate	0.447	0.451	0.01	0.39	8.43**	1.31
19. Silent pause ratio	0.565	0.119	-0.04	-0.24	-5.17**	1.32
40. Repeat ratio	0.654	0.090	-0.25	-0.23	-5.39**	1.15
45. Syntactic complexity	0.691	0.039	0.36	0.24	5.37**	1.25
48. Error-free AS-unit ratio	0.723	0.033	0.01	0.20	4.50**	1.23
(Constant)			0.73			

* $p < .05$, ** $p < .01$

5.3 Discussion

As a composite predictor, this final model captures oral proficiency in a multidimensional way, each representing an important aspect of speech production. The *effective articulation rate* represents speech speed and density, *syntactic complexity* represents the complexity of speech, and *error-free AS-unit ratio* represents the accuracy of speech in terms of grammar and vocabulary. Similarly, the *silent pause ratio* represents both the amount of speech and speech planning time, and the *repetition ratio* represents dysfluency. The interpretation of the final model is that as the OPI level advances, examinees can produce more meaningful, complex, and accurate speech at a faster rate, with less planning time and repetition. The fluency variables in the final model align with Skehan's (2009) proposed three categories of fluency: speed, breakdown, and repair. The *effective articulation ratio* represents speed (and density), the *silent pause ratio* for breakdown, and the *repetition ratio* for repair. Furthermore, the *effective articulation rate* was selected as the primary predictor, because it captures how fast a speaker can produce effective syllables if s/he is not interrupted by any pauses or dysfluency phenomena. The other four *Speed* variables include some portion of the pause or dysfluency in their calculation, overlapping with other variables. It is interesting to note that none of the *Pause-location* variables was included in the final model. According to the data used in this study, the amount of pause matters more than the location of the pause in predicting the OPI levels. For *Dysfluency*, the *repetition ratio* was the only measure among the four variables included in the model. For *Complexity* and *Accuracy*, although there was only one variable for each category, both variables were included as significant predictors in the final model. As in the literature, the findings

of this study indicate that CAF are indeed important components, and each makes a unique contribution to predicting L2 oral proficiency, to the extent that the OPI levels accurately represent it.

6. Conclusions and Recommendations

This study's findings support the notion that CAF must be considered together, not only in L2 pedagogy and research but also in L2 assessment. Although the composite predictor model found in this study still needs further refinement, it can explain 72.3% of the OPI levels with five CAF measures based on a speech sample of approximately one minute. One limitation of this study is the small number of measures used for representing *Complexity* and *Accuracy*; therefore, more variables representing *complexity* and *accuracy* should be incorporated in future studies. If the simple measures used in this study can still contribute significantly to the composite model, more fine-grained measures might yield even greater predictive power. Once a model with sufficient predictive power is found, it can greatly benefit classrooms, institutions, or high-stake tests by reducing the time and cost of conducting and grading/rating speaking tests.

Notes

1) A new major level of "distinguished" was added to the speaking guidelines in 2012; however, since the new level did not exist during the data collection period for the database, it is not mentioned in this study.
2) The Analysis of Speech Unit is "a single speaker's utterance consisting of an independent clause or subclausal unit, together with any subordinate clause(s) associated with it" (Foster et al. 2000, p. 365).
3) Intraclass Correlation Coefficients were computed for speech rate, articulation rate, and mean length of run; the average measure ICCs were .99 (95% C.I. = .97 - 1.00), .98 (95% C.I. = .93 - 1.00), and .93 (95% C.I. = .65 - .98), respectively.
4) Although the OPI level is an ordered-categorical variable, MLR was selected because it consists of nine categories and displays a normal-shape distribution (M = 5.07, SD = 2.00, Skewness = -0.02, and Kurtosis = -0.74).

References

桜木ともみ（2011）「「複雑さ・正確さ・流暢さ」指標の構成概念妥当性の検証―日本語学習者の発話分析の場合」『JALT journal』33（2）、157－173

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer program]. version 6.0.35, retrieved 16 October 2017 from http://www.praat.org/.

Chambers, F. (1997). What do we mean by fluency? *System*, 25(4), 535-544.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, 27(3), 317-334.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354-375.

Fukada, A., Hirotani, M., & Matsumoto, K. (2019). CAF calculator [Software], retrieved 27 February 2019 from http://tell.cla.purdue.edu/CAF-calculator/.

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.

Housen, A., Kuiken, F., & Vedder, I. (Eds.) (2012). Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA (Vol. 32). *John Benjamins Publishing*.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 579-589.

Leclercq, P., Edmonds, A., & Hilton, H. (Eds.) (2014). Measuring L2 proficiency: Perspectives from SLA (Vol. 78). *Multilingual Matters*.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387-417.

The National Institute for Japanese Language and Linguistics (2009). Japanese learners' conversation database [Online database], retrieved 27 October 2018 from https://mmsrv.ninjal.ac.jp/kaiwa/ (Vol. L2).

Skehan, P. (1996). Second language acquisition and task-based instruction. In J. Willis, & D. Willis (Eds.) Challenge and change in language teaching (pp. 17-30). Oxford: Heinemann.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510-532.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291-300.

Appendix

Category	Sub-category	Measures	Explanation / Calculations
Speech Quantity		9. Phonation time ratio	(Speech time) / (Total response time) * 100
Speed		10. Speech rate	(Total number of syllables) / (Total response time) * 60
		11. Articulation rate	(Total number of syllables) / (Speech time + Filled pause time) *60
	Speech density	12. Mean length run	(Total number of syllables) / (Number of runs) where a run is a sounding interval
		13. Effective speech rate	(Total number of effective syllable) / (Total response time) * 60
		14. Effective articulation rate	(Total number of effective syllable) / (Speech time – Dysfluency time) * 60
Pause		19. Silent pause ratio	Silent pause time as a percentage of Total response time
		20. Silent and filled pause ratio	(Silent pause time + Filled pause time) / (Total response time) * 100
	Pause location	29. Silent pause ratio within AS	(Silent pause time within AS-unit)/(Total response time)*100
		30. Silent and filled pause ratio within AS	(Silent pause time within AS-Unit + Filled pause time within AS-Unit)/(Total response time)*100
		31. Ratio of silent pause time between AS to total response time	(Silent pause time between AS-unit)/(Total response time)*100
		32. Ratio of silent and filled pause time between AS to total response time	(Silent pause time between AS-unit + Filled pause time between AS-unit)/(Total response time)*100
Repair/ Dysfluency		40. Repeat ratio	(Repeat time) / (Total Response Time) * 60
		41. Stutter ratio	(Stutter time) / (Total Response Time) * 60
		42. Self-correction ratio	(Self-correction time) / (Total Response Time) * 60
		43. Dysfluency ratio	(Dysfluency time) / (Total Response Time) * 60
Complexity		45. Syntactic complexity	Clause count / Number of AS-Units
Accuracy		48. Error-free AS-unit ratio	(Number of error-free AS-Units) / (Number of error-free AS-Units + Number of AS-Units with errors) * 100