Speech Communication 99 (2018) 221–230

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

A comparison of grammatical proficiency measures in the automated assessment of spontaneous speech

Su-Youn Yoon[a,*], Suma Bhat[b]

[a] NLP and Speech Group, Educational Testing Service, Princeton, NJ 08540, United States
[b] Coordinated Science Laboratory, 1308 W. Main St. Urbana, IL 61801, United States

ARTICLE INFO

Keywords:
Automated scoring
Grammatical development
Natural language processing
Similarity measures
Syntactic complexity measures

ABSTRACT

We developed new measures that assess the level of grammatical proficiency for an automated speech proficiency scoring system. The new measures assess the range and sophistication in grammar usage based on natural language processing technology and a large corpus of learners' spoken responses. First, we automatically identified a set of grammatical expressions associated with each proficiency level from the corpus. Next, we predicted the level of grammatical proficiency based on the similarity in the grammatical expression distribution between a learner's response and the corpus. We evaluated the strength of the association between the new measures and proficiency levels using spontaneous responses from an international English language assessment. The Pearson correlation test results showed that compared to commonly used syntactic complexity measures the proposed measures had stronger relationships with proficiency. We also explored the impact of system errors from a multi-stage automated process and found that the new measures were robust against the errors. Finally, we developed an automated scoring model which predicted the holistic oral proficiency scores. The new measures led to statistically significant improvement in agreement between human and machine scores over the previous system.

1. Introduction

An automated scoring system can assess students' responses faster than human raters and at a lower cost. In addition, in contrast to human raters, it is robust against fatigue and emotional state, and the resulting scores are always consistent over time. These advantages have prompted a strong demand for high-performing automated oral proficiency scoring systems. In this study, we develop a new set of grammatical proficiency measures as part of an automated oral proficiency scoring system for non-native speakers' spontaneous speech. The automated scoring system produces a score which predicts the non-native speaker's holistic oral proficiency level.

Oral proficiency in a second language is widely regarded to be multi-componential. In particular, numerous studies (e.g., Hymes, 1972; Canale and Swain, 1980) have proposed multi-componential models of communicative language ability, which have served as important underlying models of the language proficiency for large scale language tests. These studies in second language ability (SLA) have succeeded in providing *general models* of constructs for global proficiency, and various traits such as grammar, vocabulary, accent, and pronunciation (e.g., Higgs and Clifford, 1982; Iwashita et al., 2008; Iwashita, 2010). Among them, grammar and vocabulary have been two core traits commonly selected by most studies. These traits were assessed along the dimensions of accuracy, fluency, and complexity (Foster and Skehan, 1996; Wolfe-Quintero et al., 1998; Lennon, 1990).

Despite knowing the importance of grammar as a core trait in proficiency, relatively fewer studies have explored the use of grammatical proficiency measures in the context of automated speech scoring. Among the studies that have explored automated scoring of restricted speech (e.g., Bernstein et al., 2000; Balogh et al., 2007) or tutoring of pronunciation and intonation, the focus has been on scoring and error detection of pronunciation (e.g., Witt and Young, 1998; Neumeyer et al., 2000) and scoring of fluency (e.g., Cucchiarini et al., 2002). They did not explore grammatical measures since these tasks did not require assessing non-native speakers' grammatical proficiency.

In contrast to these studies, Zechner et al. (2009) and Cheng et al. (2014) developed automated proficiency scoring systems to assess global oral proficiency from non-native speakers' spontaneous speech. However, despite the importance of the grammatical measures in assessing global oral proficiency, these systems included no grammatical proficiency measures or only a simple measure based on the language model score from the automated speech recognition (ASR) system.

* Corresponding author.
E-mail addresses: syoon@ets.org (S.-Y. Yoon), spbhat2@illinois.edu (S. Bhat).

https://doi.org/10.1016/j.specom.2018.04.003
Received 8 August 2017; Received in revised form 3 April 2018; Accepted 9 April 2018
Available online 12 April 2018

Recently, a few studies have begun to explore the use of grammatical measures primarily developed for essay scoring (e.g., Chen and Zechner, 2011; Chen and Yoon, 2011) with a focus on applying these existing grammatical measures to speech scoring.

In order for any measure to be used in an automated speech scoring system, it must be generated in a fully automated manner, which necessarily consists of multiple automated sub-processes, including speech recognition, automated clause or sentence boundary detection, and sentence structure analysis using an automated parser. With this design pipeline, it is to be expected that the errors from each automated subprocess can be accumulated and result in a drop in overall performance. Moreover, because of the shorter length of spoken responses compared to written responses, even one error from the automated clause boundary detection stage can seriously affect the accuracy of clause-based measures. Additionally, a major bottleneck in this multi-stage automated process is the ASR. The automated recognition of non-native speakers' spontaneous speech is a challenging task, as evidenced by the error rate of speech recognizers developed for this task. For instance, Chen and Zechner (2011) reported a 30% word error rate (WER) in speech recognition, and these frequent errors at the recognition stage negatively affect the subsequent stages of the speech scoring system in general. More specifically, these errors affect deep syntactic analysis such as sentence parsing, which operates on a long sequence of words as its context. Not surprisingly, Chen and Zechner (2011) found that the moderate associations between syntactic complexity measures and speech proficiency were substantially reduced when applied to the automatically recognized output. Looking ahead, although we can expect to reduce the extent to which such correlations are affected with improvements in WER of the ASR stage, measures that require complicated syntactic analysis (e.g., sentence structure analysis based on a parser) are not yet practical in a spontaneous speech scoring system.

In order to overcome this problem, we propose a set of new grammatical measures based on part-of-speech tagging derived from a large corpus of learners' spoken responses. Compared to measures from previous studies, our measures are unique in two important ways. First, unlike most measures that indirectly infer syntactic complexity based upon the length of the production unit, the new measures directly assess students' sophistication and range in grammar usage. Second, instead of using grammatical scales and metrics based on native speech production to score learners' speech, the proposed measures use a similarity-based metric obtained by comparing a response with a similar body of learners' speech. We show that the new grammatical measures are robust to the inevitable errors in a multi-staged automated scoring process, making them better indices of grammatical complexity from a system development perspective.

2. Syntactic complexity measures from applied linguistics

Grammatical ability is an important trait that strongly influences second language (L2) proficiency and has been further classified into syntactic complexity and grammatical accuracy. Syntactic complexity is "the range of forms that surface in language production and the degree of sophistication of such forms" (Ortega, 2003), and grammatical accuracy is the ability to generate sentences without grammatical errors.

Due to the strong influence of grammatical ability on L2 proficiency, many studies have focused on developing quantitative measures that can estimate grammatical proficiency levels. Wolfe-Quintero et al. (1998) and Ortega (2003)'s research syntheses includes the examination of over one hundred developmental measures explored in previous SLA studies. More recently, Lu (2010) selected 14 syntactic complexity measures that showed promising performance from among those examined in these research syntheses and classified them into 5 sub-types. *The length of production unit* type calculates the length of the production units, and is not tied to specific grammatical expressions. *The mean length of clauses* is a representative measure of this type. The second type (*Sentence complexity*) is comprised of one measure: *clauses per sentence*. The third (*Subordination*) and fourth (*Coordination*) types are designed to assess the amount of subordination and coordination, respectively. Finally, the fifth type (*Particular structures*) is related to the acquisition of specific morphosyntactic or grammatical expressions. *The number of complex nominals per clause* and *the number of verb phrases per T-unit* fall into this type.

The usefulness of these measures to assess non-native speaking proficiency has also been explored. Halleck (1995) found that some measures such as *the mean length of T-units, the mean length of error-free T-units,* and *the percentage of error-free T-units* consistently increased as the level of the oral proficiency increased. Iwashita (2010) analyzed a large body of monologic spontaneous speech from both learners of English as foreign language (EFL) and learners of Japanese as foreign language (JFL) with respect to these measures. While this study found significant differences in the grammatical accuracy measures between the high proficiency and the low proficiency groups in EFL, it found no such tendency in the JFL groups. Furthermore, a reverse trend was found with respect to the syntactic complexity measures: significant differences were found from the JFL group, but not from the EFL group. This suggests that the relationship between these measures and oral proficiency is inconclusive.

Studies have also reported that the discriminative ability of syntactic complexity measures with respect to oral proficiency levels is not strong. Iwashita et al. (2008) found that these measures could discriminate students' proficiency levels to some degree, but they could not make a fine-grained distinction between adjacent levels; there were large variations within a level, and the differences between the levels were not always statistically significant. These results suggest that we need measures of syntactic complexity that capture a wider set of grammatical structures than is captured using the existing measures. In addition to the weak association between existing syntactic complexity measures and proficiency, there are several practical difficulties one needs to overcome while applying them to spontaneous speech. First, dividing speech transcriptions into meaningful units is a challenging task in spontaneous speech. Most syntactic complexity measures were based on production units such as clauses and T-units, and segmenting spoken responses into these production units in a consistent and principled way is an essential task that has a significant impact on the results. In contrast to writing, spontaneous speech tends to include incomplete sentences and disfluencies such as repair and repetitions, which make the task even more challenging. In addition, non-native speakers' speech tends to include frequent grammatical errors. These result in incomplete and ungrammatical sentences that obscure the sentence structure and further increase the difficulty of this task.

Second, the short length of spoken responses (compared to written responses in essays) poses additional difficulties in obtaining measures in a reliable way. They tend to include only a few sentences or sentence-like units (a typical response in our data set had 10 clauses on average), and the resulting impact of an error in one unit can be large. For instance, one exceptionally long unit can inflate the measures based on the length of the production units substantially and may result in the overestimation of the learner's grammatical proficiency. This is supported by Chen and Yoon (2011)'s observation of a marked decrease in correlation between the measures and holistic proficiency scores as the amount of spoken material provided by the test taker decreased from six minutes to one minute.

Finally, in order for the measures to be used in an automated speech scoring system, they must be generated in a fully automated process which necessarily consists of multiple automated sub-processes, including speech recognition, clause or sentence boundary detection, part-of-speech (POS) tagging, and an optional sentence structure analysis using an automated parser. The errors in each stage of the automated process are cumulative and result in a drop in overall performance. With the spoken responses being particularly short, even one error from the automated clause boundary detection stage can seriously affect the accuracy of clause-based measures. In order to overcome this