Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 67 (2015) 42–57

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic assessment of syntactic complexity for spontaneous speech scoring

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

[a] Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign, IL, USA
[b] Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Received 12 November 2013; received in revised form 7 July 2014; accepted 12 September 2014
Available online 26 September 2014

Abstract

Expanding paradigms of language learning and testing prompt the need for developing objective methods of assessing language proficiency from spontaneous speech. In this paper new measures of syntactic complexity for use in the framework of automatic scoring systems for second language spontaneous speech, are studied. In contrast to most existing measures that estimate competence levels indirectly based on the length of production units or frequency of specific grammatical structures, we capture the differences in the distribution of morpho-syntactic features across learners' proficiency levels. We build score-specific models of part of speech (POS) tag distribution from a large corpus of spontaneous second language English utterances and use them to measure syntactic complexity.

Given a speaker's response, we consider its similarity with a set of utterances scored for proficiency by humans. The comparison is made by considering the distribution of POS tags in the response and a score-level. The underlying distribution of POS tags (indicative of syntactic complexity) is represented via two models: a vector-space model and a language model.

Empirical results suggest that the proposed measures of syntactic complexity show a reasonable association with human-rated proficiency scores compared to conventional measures of syntactic complexity. They are also significantly robust against errors resulting from automatic speech recognition, making them more suitable for use in operational automated scoring applications. When used in combination with other measures of oral proficiency in a state-of-the-art scoring model, the predicted scores show improved agreement with human-assigned scores over a baseline scoring model without our proposed features.

Keywords: Language testing; Automated scoring; Speaking proficiency; Computer-aided language learning; Syntactic complexity; Objective measures

1. Introduction

The expansion of natural language and speech processing capabilities have created new areas of application for use with the expanding paradigms of human–computer interaction. Today, language learning is gradually moving away from tutor-based or language-lab based scenarios, to become computer-aided. The obvious advantages of this emerging paradigm are both its potential to make language learning materials accessible to a wider range of learners at reduced costs (as compared to using human tutors) and being more ubiquitous for use on a more flexible schedule. With more opportunities for computer-aided language learning (CALL) interfaces being created today, there is an increased need to endow CALL systems with the ability to assess language ability automatically. While the resulting technology could be used for automated scoring in a testing scenario or for providing diagnostic feedback to the learner, efforts are being made to develop objective methods of assessing language ability from spontaneous speech.

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

http://dx.doi.org/10.1016/j.specom.2014.09.005
0167-6393/

Overall spoken proficiency in a target language can be assessed by testing the abilities in various areas including fluency, pronunciation and intonation, grammar and vocabulary, and discourse structure. Currently, speech-enabled dialog systems allow learners to practice their speaking and listening with a virtual interlocutor (e.g., SpeakESL), to receive feedback on their pronunciation [e.g., Carnegie Speech, or *Native Accent* (Eskenazi et al., 2007), *EduSpeak* from SRI (Franco et al., 2000)].

These and other spoken response scoring systems work on restricted speaking tasks such as reading a passage or answering questions with a limited range of responses (Bernstein et al., 2000; Balogh et al., 2007). In contrast to these systems that score restricted speech, scoring unstructured, unrestricted, and spontaneous responses poses a much harder problem. In addition, if the systems target learners with diverse levels of second language proficiency and varied first language backgrounds, the difficulty increases substantially.

The state-of-the-art system for scoring spontaneous speech in a testing scenario is SpeechRater^{SM} (Zechner et al., 2009). Although the current capability is sufficiently advanced to allow it to be used for the scoring of TOEFL® Practice Online (TPO), a low-stakes practice test product, there is room for improving its feature set by expanding the coverage of important aspects of speaking proficiency and modifying others. For instance, aspects of grammar and vocabulary sophistication are only being measured indirectly (more details on this later in this paper) and a more direct approach to measuring these aspects is necessary.

Taking the challenges posed in processing spontaneous speech automatically into consideration, we propose a set of measures of grammatical competence. This paper describes the measures and their potential of being used in a state-of-the-art spontaneous scoring system. In Section 2 the problem being studied is placed into the context of previous work done in the related areas of written and spoken language assessment. A description of the measures studied in this paper if found in Section 3. In Section 4, we delve into the details of the implementation of our proposed measures. A description of the data is provided in Section 5. The experimental details comprise the material in Section 6 and the results are presented in Section 7. In Section 8 we discuss the results of data analyses and highlight some extensions to the study. Finally, a brief summary of the major findings of the paper is presented in Section 9.

2. Motivation

2.1. Assessment of syntactic competence in second language learning

Numerous studies in related second language acquisition literature reveal that syntactic complexity and grammar accuracy are regarded as some of the key skills that strongly influence second language proficiency. Thus, the study of measures that reflect language learners' command of these influential skills has been the central theme of various studies in the area of second language acquisition.

In related literature, Ortega (2003) indicates that "the *range* of forms that surface in language production and the degree of *sophistication* of such forms" are two important areas in grammar usage collectively termed, "syntactic complexity". A vast majority of measures of syntactic complexity have been used as indicators of levels of acquisition of syntactic competence, and in turn, are suggestive of proficiency levels in ESL writing (e.g. Wolf-Quintero et al., 1998; Ortega, 2003; Lu, 2010). These measures have been broadly classified into two groups (Bardovi-Harlig and Bofman, 1989). The first group is related to the acquisition of specific grammatical expressions corresponding to various stages of language acquisition. Frequencies of negation or relative clauses – in terms of whether these expressions occurred in the test responses without errors, fall into this group (hereafter, the expression-based group). The second group, not tied to particular structures, is related to length of clauses or the relationship between clauses (hereafter, the length-based group). Representative measures in the second group include the *mean length of clause unit*, the *ratio of dependent clauses to the total number of clauses*, and the *number of verb phrases per clause*.

In contrast with syntactic complexity, grammatical accuracy is the ability to generate sentences without grammatical errors. The measures in this group can be classified into two groups. Global accuracy measures include those that count all errors in sentence production and are calculated as normalized values, e.g., the percentage of error-free clauses among all clauses (Foster and Skehan, 1996). A second group of measures is more focussed on specific types of constructions such as verb tense, third-person singular forms, prepositions, and articles, and calculate the percentage of error-free clauses with respect to these constructions (Robinson, 2006; Iwashita et al., 2008).

In the area of spoken language assessment, researchers have sought the application of measures of syntactic competence and grammatical accuracy. In particular, Halleck (1995)'s study found that in the context of English as a foreign language (EFL) assessment, holistic oral proficiency scores were highly correlated with three quantitative measures (mean length of T-units,[1] mean error-free T-unit length, and percentage of error-free T-units). Again, the results from a similar study that included both English and Japanese foreign language assessment, confirmed the utility of these and other quantitative measures that assess grammatical accuracy and syntactic complexity, in addition to vocabulary, pronunciation, and fluency (Iwashita et al., 2008; Iwashita, 2010). However, the results were inconclusive about the strength of the relationship between the measures and the proficiency scores. Strong data

[1] Hunt (1970) proposed the idea of a T-unit which is a main clause with a subordinate clause and non-clausal units. It is different from a clause since it does not consider a subordinate clause as an independent unit.