The impact of ASR accuracy on the performance of an automated scoring engine for spoken responses

Derrick Higgins, Lei Chen, Klaus Zechner,
Keelan Evanini & Su-Youn Yoon
Educational Testing Service

Abstract:
Automated scoring to assess speaking proficiency depends to a great extent on the availability of appropriate tools for speech processing (such as speech recognizers). These tools not only must be robust to the sorts of speech errors and nonstandard pronunciations exhibited by language learners, but must provide metrics which can be used as a basis for assessment.

One major strand of current research in the area of automated scoring of spoken responses is the effort to develop deeper measures of the grammatical, discourse, and semantic structure of learners’ speech, in order to model a broader speaking proficiency construct than one which focuses primarily on phonetic and timing characteristics of speech. The quality of speech recognition systems is especially crucial to this research goal, as errors in speech recognition lead to downstream errors in the computation of higher-level linguistic structures that can be used to calculate construct-relevant features.

The goal of this paper is to provide a case study illustrating the effects of speech recognition accuracy on what can be achieved in automated scoring of speech. A comparison of speech features calculated on the basis of competing speech recognition systems demonstrates that scoring accuracy is strongly dependent on using the most accurate models available, both for open-ended tasks and for more restricted speaking tasks.

Introduction
Methods for automated scoring of speaking proficiency, or some aspect of it, have been known in the research literature for over a decade (cf. Bernstein, 1999; Bernstein, DeJong, Pisoni & Townshend, 2000; Franco et al., 2000; Eskenazi et al., 2007; Balogh et al., 2007). These methods have typically focused on more constrained speech (such as that elicited by asking examinees to read a sentence or passage aloud) rather than spontaneous speech that might be elicited through open-ended tasks. This is largely because the limitations of speech recognition technology degrade the signal upon which proficiency measures can be based to a lesser extent for constrained speech than for spontaneous speech. (For instance, pronunciation measures based on a fairly reliable hypothesized transcript of speech will be more reliable than those based on a transcript likely to contain numerous word errors (so that pronunciation characteristics of the speech signal may frequently fail to be associated with the correct target phone).
Concentrating on relatively constrained tasks, however, has resulted in a limitation of the types of variation that can be used to differentiate among examinees, and therefore in a narrowing of the speaking construct which can be assessed by automated means. While pronunciation and fluency can be directly assessed (at least in large part) based on speakers’ ability to reproduce a known text orally, such performance tasks do not provide direct evidence of other important aspects of proficiency in a spoken language, such as the ability to construct grammatically appropriate speech spontaneously, or the ability to organize an oral narrative in order to facilitate comprehension.

A goal of current research in the scoring of spoken performance tasks is to develop measures which address linguistically deeper aspects of speaking proficiency. These currently consist primarily of syntactic or grammatical features, but ultimately features related to discourse organization, semantic coherence, and social appropriateness of communication should be investigated as well, as technology allows.

An extensive body of previous research has established the usefulness of measures related to syntactic complexity and grammatically-informed fluency measures in identifying language proficiency based on written tasks (Homburg, 1984; Wolfe-Quintero, Inagaki & Kim, 1998; Ortega, 2003; Cumming, Kantor, Baba, Eounanzou, Erdosi & James, 2006; Lu; 2010; Lu, 2011). More recently, these approaches have been transferred to the speaking domain, and corpus linguistic studies of speech transcripts have demonstrated a relationship between speaking proficiency and the same measures (Halleck, 1995; Foster, Tonkyn & Wigglesworth, 2000; Iwashita, McNamara & Elder, 2001; Iwashita, 2006; Lu, forthcoming). While few existing studies using such measures have taken the automated scoring of spoken responses as their goal, some researchers have begun to apply these methods in the context of larger automated scoring systems (Chen, Tetreault & Xi, 2010; Bernstein, Cheng & Suzuki, 2010).

The key questions to be addressed in the current paper are to what extent current ASR technology can support the deeper linguistic features targeted by this direction of research, and whether there are important differences among currently available speech recognition systems which might be relevant to this question. On the basis of internal comparisons between speech recognition engines, we aim to provide a partial answer.

**Addressing the construct of speaking proficiency**

A discussion of the relevance of different sorts of linguistic features to assessing competence in spoken language presumes some clear conception of the construct of proficiency in speaking a foreign language, and in fact this can be characterized in different ways.

The Speaking section of the Test of English as a Foreign Language™ (TOEFL® iBT), responses from which have been used in much work to evaluate the SpeechRater™ automated scoring system (Higgins, Xi, Zechnner & Williamson, 2011; Zechnner, Higgins, Xi & Williamson, 2010), is based on a construct designed to reflect the judgments of ESL teachers and applied linguists about the important characteristics of competent English speaking in an academic environment. The rubrics for this test were designed based
on work described in Brown, Iwashita & McNamara (2006), and Figure 1 illustrates the overall structure of the TOEFL Speaking construct. As Figure 1 shows, The TOEFL Speaking test categorizes speaking skills into three broad domains: delivery, language use, and topic development. The skills which compose the latter two categories involve grammatical and semantic levels of competence that are commonly associated with spontaneous speech production, and in fact, many tests of speaking proficiency such as TOEFL and IELTS aim to assess these skills through fairly open-ended tasks which elicit free speech.

The inclusion of higher-level syntactic-semantic and discourse factors in the targeted speaking construct is consistent with other standards and testing programs, including the speaking section of the DoDEA English Language Proficiency Standards (Department of Defense Education Activity, 2009), the WiDA standards (WiDA Consortium, 2007) and the Cambridge ESOL speaking tests (Tyler, 2003).

A different approach is taken by Bernstein, Van Moere & Cheng (2010), in defining a construct of “facility in L2” to be assessed using restricted speaking tasks. As described by Bernstein et al., such a construct concentrates on “core spoken language skills” (p. 356), which in Hulstijn’s (2006) conception
focus on phonetic, phonological, lexical and morphosyntactic knowledge and skills rather than syntactico-semantic or discourse elements.

**The state of the art in speech recognition in support of language assessment applications**

Assuming a speaking test needs to assess such higher-level speaking skills, and that open-ended speaking tasks are necessary in order to assess them, the question arises what sort of features can be used to assess these skills, and how accurate a speech recognition system must be in order to support them. As noted above, one important direction that ongoing research is pursuing is the development of syntactic complexity features. Given this development, an important immediate question is how well-suited existing speech recognition technology is to supporting the kind of grammatical analysis which would be required for the reliable computation of syntactic complexity features.

This section aims to address this point on the basis of two case studies based on corpora and speech recognizers available at ETS. In each study, metrics were used both to assess the quality of individual speech recognition results with an eye toward supporting speech scoring applications, and to compare the results across speech recognizers.

**Case Study 1**

The first case study involves a set of 402 open-ended spoken responses collected as part of an on-line TOEFL practice test. These responses were collected from 72 different respondents, spanning a range of different native languages and proficiency levels (although generally within the proficiency range observed for TOEFL test takers). The time limit for spoken responses was either 45 or 60 seconds, depending on the task (as it is on the operational TOEFL test).

To assess the suitability of modern speech recognizers for calculating deeper syntactic-semantic features for assessment, two different speech recognizers were used to process the responses in this set. The first recognition system is a “commodity off the shelf” (COTS) speech recognition system which is commonly used in enterprise applications such as medical transcription. It is intended to reflect the level of results which can be expected when using a speech recognizer not actively developed as a research system reflecting the most recent and technologically advanced methods from the speech recognition literature, but which generally reflects the architecture and training methodology typical of a modern, commercially viable system. This system used a triphone acoustic model, and its off-the-shelf native speech model was adapted to this domain using approximately 1900 non-native student responses from the same on-line practice test (but distinct from the 402 responses used for evaluation). It used a trigram language model based on both external native speech corpora and transcriptions from the on-line practice test.

The second speech recognition model compared was a highly optimized commercial system (OCS) which is actively maintained and evaluated to ensure that it maintains the highest level of accuracy possible
given the current state of speech recognition research. Its training parameters were similar to the of the COTS model, in that it also used a triphone acoustic model, in which a native English baseline model was adapted to the set of 1900 non-native spoken responses, and its trigram language model was trained on both in-domain data and native speech resources.

As we do not yet have a broad set of candidate features for spoken response scoring which leverage deep syntactic-semantic information from the response, the evaluation of performance for this case study was purely at the level of word accuracy. Nevertheless, the evaluation is quite suggestive regarding the potential for speech recognition systems of this type to support such features. As Table 1 shows the two speech recognizers differ markedly in accuracy, with the highly optimized system demonstrating almost 40% higher word accuracy than the off-the-shelf system over the entire set of evaluation data. If we assume that methods of syntactic analysis can be applied to speech recognition output which are somewhat robust to errors, so that even responses with a word error rate of 20% or so could be handled appropriately, then 37% of responses could be processed to yield useful syntactic features using the OCS model—37% have a word accuracy of 80% or higher—while not a single response meets this criterion using the COTS model.

<table>
<thead>
<tr>
<th></th>
<th>WACC</th>
<th>WACC ≥ 90%</th>
<th>WACC ≥ 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>COTS</td>
<td>36.2%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>OCS</td>
<td>74.5%</td>
<td>5.2%</td>
<td>37.3%</td>
</tr>
</tbody>
</table>

Table 1: Speech recognition results from Case Study 1 (word accuracy, proportion of responses with 90% word accuracy, proportion of responses with 80% word accuracy)

One may question, of course, whether deep linguistic features will be useful in a scoring context if they can only be reliably calculated for 37% of responses. However, it is likely that these features will be of most use at the higher end of the scoring scale. Fluency and pronunciation features may differentiate fairly well at the low end of the score scale, while syntactic complexity, cohesion, and other such features already presuppose a certain mastery of speech delivery. Among responses which were awarded the highest possible score by human raters, the OCS speech recognition system was able to recognize 49.4% of responses with 80% word accuracy or higher. While this is still lower than we would like, it seems to be sufficient as a first approximation to develop syntactic-semantic features which can contribute to scoring empirically, and to the meaningfulness of those scores.

Case Study 2

The second case study uses a set of English spoken responses from 319 Indian examinees as a basis for evaluation. These examinees provided responses to three different types of speaking tasks: an open-ended speaking task, a task involving reading a passage aloud, and a task involving the oral repetition of a stimulus sentence presented aurally. The distribution of responses across these three tasks in the data set is presented in Table 2. Using a set of data including both open-ended and very constrained
speaking tasks allows for the effect of speech recognition accuracy on both types of tasks to be investigated.

<table>
<thead>
<tr>
<th></th>
<th>Open-ended</th>
<th>Read-aloud</th>
<th>Repeat sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers</td>
<td>319</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responses</td>
<td>1,274</td>
<td>10,134</td>
<td>1,275</td>
</tr>
</tbody>
</table>

Table 2: Number of speakers and responses of each type in spoken response data used for case study 2

The speech recognizer comparison performed in this case study was similar to, but not the same as, the comparison performed in case study 1. Where the OCS recognizer was compared to a COTS recognition system in the previous evaluation, in this study it is compared to an adapted version of the freely available HTK speech recognizer (Young, 1994). The OCS speech recognition model used for this experiment used a triphone acoustic model trained on approximately 45 hours of non-native English speech drawn from the same population of Indian speakers used for evaluation, and three different, item-type dependent trigram language models. These language models were trained on both in-domain data from the assessment itself, and other native English text resources. The HTK system used a triphone acoustic model with decision-tree based clustering of states, and was trained on the same non-native English speech data as the OCS system. It used an item-type specific bigram language model, which was trained exclusively on in-domain data.

As Table 3 demonstrates, speech recognition accuracy on this set of data is considerably lower than that observed in case study 1. This difference is largely attributable to the fact that the examinee population studied is less proficient overall than in the previous experiment. (This experiment drew its participants from the general body of English learners in India, where that experiment dealt with a population of examinees preparing for the TOEFL test, and therefore having some level of confidence that they were ready for that test).

The OCS speech recognizer exhibits very high accuracy on the constrained read-aloud task, somewhat lower accuracy on the sentence repetition task, and only about 50% word accuracy on the open-ended speaking task (compared to approximately 75% in case study 1). (While not indicated in the table, only about 3% of the open-ended responses had OCS word accuracies of 80% or greater.) The accuracy of the HTK recognizer was much worse on all item types, again displaying the substantial gap between off-the-shelf recognition capabilities and highly optimized systems for the challenging task of automated recognition of non-native speech.
As noted above, it was not possible to derive a set of proficiency features based on syntactic analysis of the speech recognition hypotheses produced for these data sets, so that the actual degradation in their predictive value could be observed as the word accuracy decreases. (In any case, given the significantly lower word accuracy rate observed in case study 2, it is not clear that such deeper syntactic features could be used for this data.) However, experimentation with other features used in proficiency estimation (cf. Higgins et al., 2011 and Chen, Zechner & Xi, 2009 for a list) revealed important differences in their behavior from one recognition system to another. In the final evaluation results of this paper, we consider two sets of features: one set of features based on the distribution of stress points within a response, and one set of measures related to pronunciation quality.

These two classes of features were generated for all of the responses in the evaluation data set described in Table 2, and the correlation between each feature and the score assigned to responses by trained raters was calculated for each item type. For visualization purposes, the correlations produced in this manner are displayed as heat maps in Figure 2 (stress-based features) and Figure 3 (pronunciation). A cell shaded in deep blue in the table indicates that the absolute value of the correlation between a given feature and human proficiency scores is very low, a cell shaded in white indicates a moderate correlation, and a cell shaded in purple indicates a correlation close to the highest value observed in the data set. (The color scales are set based on the range of the values in each figure, so they are not comparable between figures.) For reference, the highest correlation observed in Figure 2 is 0.47, and holds between human scores and the \texttt{relstresspct} feature produced by the OCS recognizer for the Repeating task. The highest correlation observed in Figure 3 is 0.51, and holds between human scores and the \texttt{L1} feature produced by the OCS recognizer for the Repeating task.

Comparing the first two columns of Figures 2 \& 3 (representing open-ended items), it is almost uniformly the case that the stress and pronunciation features examined here have higher correlations with human scores when calculated on the basis of the OCS recognizer output than when calculated using HTK. (The color gradient is from blue to purple, from dark blue to light blue, or from light purple to dark purple.) In fact, this pattern holds not only for the open-ended item type, but also for the read-aloud items (columns 3 and 4) and the sentence repeating items (with the puzzling exception of \texttt{L7} and \texttt{L6}). The generalization seems clear that the higher accuracy of the OCS speech recognizer improves the quality of the features derived from it for automated speech proficiency scoring.

<table>
<thead>
<tr>
<th></th>
<th>Open-ended</th>
<th>Read-aloud</th>
<th>Repeat sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTK</td>
<td>16.6%</td>
<td>57.9%</td>
<td>37.9%</td>
</tr>
<tr>
<td>OCS</td>
<td>49.9%</td>
<td>93.6%</td>
<td>74.3%</td>
</tr>
</tbody>
</table>

Table 3: Speech recognition performance (word accuracy) by speech recognizer and spoken response task type
This result is perhaps not remarkable, given the large gap in performance between the OCS and HTK systems described here. However, it bears note that modern speech recognizers are not interchangeable in speech assessment, even when the task addressed is a very constrained one, and even when the features in question do not involve particularly deep linguistic processing.

Figure 2: Heat map indicating relationship between stress features and human scores
Figure 3: Heat map indicating relationship between pronunciation features and human scores

Conclusions

We hope to have demonstrated three main points with this paper.

First, as researchers working on educational applications of speech recognition are keenly aware, even today’s most advanced speech recognition technology is still severely limited in its ability to reliably recognize language learners’ unconstrained speech, and to provide a basis for deeper natural language processing. Under favorable recording conditions, a state-of-the-art speech recognizer can provide fairly accurate recognition for high-proficiency non-native speakers, which may be sufficient to support the calculation and use of deeper features measuring aspects of the speaking construct such as grammatical accuracy. However, at lower proficiency levels and under less optimal recording conditions, recognition accuracy degrades significantly.

Second, the differences among modern speech recognizers, both commercial and freely-available systems, are quite substantial in this regard. The speech recognition underpinnings of a model developed to score learners’ speech or otherwise be used in educational settings must be selected carefully and evaluated in a task- and population-specific way as a prerequisite to system development.

Finally, the differences in speech recognizer performance are not limited only to the challenging case of scoring spontaneous spoken responses to open-ended tasks, but manifest themselves even for more restricted tasks involving read or repeated speech. These performance differences have substantial
effects even on relatively “shallow” features measuring aspects of speech delivery, such as pronunciation and fluency.

Future work will need to undertake the implementation of more sophisticated grammatical, semantic and rhetorical features for speech assessment on the basis of speech recognition output, and evaluate how valid and reliable they are as measures of speakers’ proficiency. Some work in this area (e.g., Chen et al., 2010) has already been undertaken, but much more is needed. And of course, the conclusions of this paper will need to be revisited and potentially modified as the field of speech recognition advances (especially as advances are made in the modeling of non-native speech).

References


