Towards Using Conversations with Spoken Dialogue Systems in the Automated Assessment of Non-Native Speakers of English

Diane Litman
University of Pittsburgh
Pittsburgh, PA 15260 USA
dlitman@pitt.edu

Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen and David Vandyke
University of Cambridge
Cambridge, CB2 1PZ, UK
{sjy11,mjfg100,kmk1001,ko201,rcv25,djv27}@cam.ac.uk

Abstract

Existing speaking tests only require non-native speakers to engage in dialogue when the assessment is done by humans. This paper examines the viability of using off-the-shelf systems for spoken dialogue and for speech grading to automate the holistic scoring of the conversational speech of non-native speakers of English.

1 Introduction

Speaking tests for assessing non-native speakers of English (NNSE) often include tasks involving interactive dialogue between a human examiner and a candidate. An IELTS\textsuperscript{1} example is shown in Figure 1. In contrast, most automated spoken assessment systems target only the non-interactive portions of existing speaking tests, e.g., the task of responding to a stimulus in TOEFL\textsuperscript{2} (Wang et al., 2013) or BULATS\textsuperscript{3} (van Dalen et al., 2015).

This gap between the current state of manual and automated testing provides an opportunity for spoken dialogue systems (SDS) research. First, as illustrated by Figure 1, human-human testing dialogues share some features with existing computer-human dialogues, e.g., examiners use standardized topic-based scripts and utterance phrasing. Second, automatic assessment of spontaneous (but non-conversational) speech is an active research area (Chen et al., 2009; Chen and Zechner, 2011; Wang et al., 2013; Bhat et al., 2014; van Dalen et al., 2015; Shashidhar et al., 2015), which work in SDS-based assessment should be able to build on. Third, there is increasing interest in building automated systems not to replace human examiners during testing, but to help candidates prepare for human testing. Similarly to systems for writing (Burstein et al., 2004; Roscoe et al., 2012; Andersen et al., 2013; Foltz and Rosenstein, 2015), automation could provide unlimited self-assessment and practice opportunities. There is already some educationally-oriented SDS work in computer assisted language learning (Su et al., 2015) and physics tutoring (Forbes-Riley and Litman, 2011) to potentially build upon.

On the other hand, differences between speaking assessment and traditional SDS applications can also pose research challenges. First, currently available SDS corpora do not focus on including speech from non-native speakers, and when such speech exists it is not scored for English skill. Even if one could get an assessment company to release a scored corpus of human-human dialogues, there would likely be a mismatch with the computer-human dialogues that are our target for automatic assessment.\textsuperscript{4} Second, there is a lack of optimal technical infrastructure. Existing SDS components such as speech recognizers will likely need modification to handle non-

\textsuperscript{1}International English Language Testing System.
\textsuperscript{2}Test of English as a Foreign Language.
\textsuperscript{3}Business Language Testing Service.
\textsuperscript{4}Users speak differently to Wizard-of-Oz versus automated versions of the same SDS, despite believing that both versions are fully automated (Thomason and Litman, 2013).
native speech (Ivanov et al., 2015). Existing automated graders will likely need modification to process spontaneous speech produced during dialogue, rather than after a prompt such as a request to describe a visual (Evanini et al., 2014).

We make a first step at examining these issues, by using three off-the-shelf SDS to collect dialogues which are then assessed by a human expert and an existing spontaneous speech grader. Our focus is on the following research questions:

**RQ1:** Will different corpus creation methods influence the English skill level of the SDS users we are able to recruit for data collection purposes?

**RQ2:** Can an expert human grader assess speakers conversing with an SDS?

**RQ3:** Can an automated grader for spontaneous (but prompted) speech assess SDS speech?

Our preliminary results suggest that while SDS-based speech assessment shows promise, much work remains to be done.

## 2 Related Work

While SDS have been used to assess and tutor native English speakers in areas ranging from science subjects to foreign languages, SDS have generally not been used to interactively assess the speech of NNSE. Even when language-learning SDS have enabled a system’s behavior to vary based on the speaker’s prior responses(s), the skills being assessed (e.g., pronunciation (Su et al., 2015)) typically do not involve prior dialogue context.

In one notable exception, a triologue-based system was developed to conversationally assess young English language learners (Evanini et al., 2014; Mitchell et al., 2014). Similarly to our research, a major goal was to examine whether standard SDS components could yield reliable conversational assessments compared to humans. A small pilot evaluation suggested the viability of a proof-of-concept triologue system. Our work differs in that we develop a dialogue rather than a triologue system, focus on adults rather than children, and use an international scoring standard rather than task completion to assess English skill.

## 3 Computer Dialogues with NNSE

The first step of our research involved creating corpora of dialogues between non-native speakers of English and state-of-the-art spoken dialogue systems, which were then used by an expert to manually assess NNSE speaking skills. Our methods for collecting and annotating three corpora, each involving a different SDS and a different user recruitment method, are described below.

### 3.1 Corpora Creation

The **Laptop (L)** corpus contains conversations with users who were instructed to find laptops with certain characteristics. The SDS was produced by Cambridge University (Vandyke et al., 2015), while users were recruited via Amazon Mechanical Turk (AMT) and interacted with the SDS over the phone. To increase the likelihood of attracting non-native speakers, an AMT Location qualification restricted the types of workers who could converse with the system. We originally required workers to be from India, but due to call connection issues, we changed the restriction to require workers to **not** be from the United States, the United Kingdom, or Australia. In pilot studies without such qualification restrictions, primarily native speakers responded to the AMT task even though we specified that workers must be non-native speakers of English only.

The **Restaurant (R)** corpus contains conversations with users who were instructed to find Michigan restaurants with certain characteristics. The SDS used to collect this corpus was produced by VocalIQ (Mrkšić et al., 2015). Users were again recruited via AMT, but interacted with this SDS via microphone using the Chrome browser. Rather than using a location qualification, the title of the AMT task was given only in Hindi.

The **Bus (B)** corpus contains conversations with users who were instructed to find bus routes in Pittsburgh. Although the SDS was again produced by Cambridge University, the dialogues were pre-

---

5As explained in Section 3.1, this paper compares three corpora that were created in three different ways: via Amazon Mechanical Turk with worker qualification restrictions, via Amazon Mechanical Turk with non-English task titles, and via a Spoken Dialogue Challenge with SDS users from participant sites.

6The speech recognizer used in the off-the-shelf grader described in Section 4.1 was trained on speakers with Gujarti as their first language. The grader itself, however, was trained on data from Polish, Vietnamese, Arabic, Dutch, French, and Thai first-language speakers (van Dalen et al., 2015).

7Thanks to Blaise Thompson for providing the system.
viously collected as part of the first Spoken Dialogue Challenge (SDC) (Black et al., 2011). How-
however, our Bus corpus includes only a subset of the available SDC dialogues, namely non-native dia-
logues from the control condition. As in our AMT corpus collections, callers in the control condi-
tion received a scenario to solve over a web inter-
face. Furthermore, callers in the control condition
were spoken dialogue researchers from around the
world. Whether a caller was a non-native speaker
was in fact annotated in the SDC corpus download.

Since our Bus corpus contained 22 dialogues8, we
used AMT to collect similar numbers of dia-
logues with the other SDS. After removing prob-
lematic dialogues where the AMT task was com-
pleted but there was no caller speech or the caller
turned out to be a native speaker, our final Com-
bined (C) corpus contained 67 dialogues, dis-
tributed as shown in the “All” column of Table 1.

3.2 Manual Speaking Skill Assessment

Once the corpora were collected, the speaking
skill of the human in each dialogue was manually
assessed using the Common European Framework
of Reference for Languages (CEFR, 2001).9 The
CEFR is an international standard for benchmark-
ing language ability using an ordered scale of 6
levels: A1, A2, B1, B2, C1, C2. A1 represents
beginning skill while C2 represents mastery.

Assessment was done by a human expert while
listening to logged SDS audio files. Speech recog-
nition output was also made available. Since an
expert in CEFR performed the assessment10, dia-
logues were only scored by this single assessor.
Sometimes the assessor assigned two adjacent lev-
els to a speaker. To support a later comparison
with the unique numerical score produced by the
automatic grader discussed in Section 4.1, dual
assessments were mapped to a new intermediate
level placed between the original levels in the or-
dered scale. For example, if the expert rated a
speaker as both “B1” and “B2”, we replaced those
two levels with the single level “B1B2.”

The A1-C2 columns of the “Assessed” section
of Table 1 show the expert assessment results
for each corpus. The average number of user
turns per assessed dialogue (“Turns/Dial.”) and
the average number of recognized words per
user turn (“Wds./Turn”) are also shown. With re-

8Only 22 of the 75 control callers were non-natives.
9The scores produced by the automatic grader described in
Section 4.1 come with a mapping to CEFR.
10The Director of Academic Development and Training for
International Students at Cambridge’s Language Centre.

11The output of the speech recognizer for each SDS was
used as only the SDC Bus download has transcriptions.
12A statistical analysis demonstrating that the Restaurant
scores are significantly lower will be presented in Section 4.2,
after the CEFR labels are transformed to a numeric scale.

Table 1: Human CEFR dialogue assessments, average # of user turns per dialogue, and average number
of recognized words per turn, across corpora. L = Laptop, R=Restaurant, B=Bus, C=Combined.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td></td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>11.48</td>
<td>3.9</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>R</td>
<td>14</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>6.36</td>
<td>5.5</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>20</td>
<td></td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>13.65</td>
<td>2.6</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>C</td>
<td>55</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>11</td>
<td>8</td>
<td>2</td>
<td>10.96</td>
<td>3.6</td>
<td>12</td>
<td>67</td>
</tr>
</tbody>
</table>

272
Table 2: Mean (standard deviation) of human and automated grades, along with Pearson’s correlations between the human and automated individual dialogue grades, within each corpus.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Mean (SD) Grades</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n Human Auto R p</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>21 24.2 (3.1) 17.1 (1.9) .41 .07</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>14 21.5 (2.0) 11.6 (3.1) .69 .01</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>15 25.9 (1.9) 17.1 (1.7) -.11 .69</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>50 24.0 (3.0) 15.6 (3.3) .59 .01</td>
<td></td>
</tr>
</tbody>
</table>

For automatic dialogue scoring by GP-BULATS (trained prior to our SDS research as described above), the audio from every user utterance in a dialogue was used for feature extraction. The scoring results are shown in the “Auto” column of Table 2. Note that in all three corpora, GP-BULATS underscores the speakers.

The “R” and “p” columns of Table 2 show the Pearson’s correlation between the human and the GP-BULATS grades, and the associated p-values (two-tailed tests). With respect to RQ3, there is a positive correlation for the corpora collected via AMT (statistically significant for Restaurant, and a trend for Laptop), as well as for the Combined corpus. Although the SDS R values are lower than the 0.83 GP-BULATS value, the moderate positive correlations are encouraging given the much smaller SDS test sets, as well as the training/testing data mismatch resulting from using off-the-shelf systems. The SDS used to collect our dialogues were not designed for non-native speakers, and the GP-BULATS system used to grade our dialogues was not designed for interactive speech.

Further work is needed to shed light on why the Bus corpus yielded a non-significant correlation. As noted in Section 3.2, shorter turns made human annotation more difficult. The Bus corpus had the fewest words per turn (Table 1), which perhaps made automated grading more difficult. The Bus user recruitment did not target Indian first languages, which could have impacted GP-BULATS speech recognition. Transcription is needed to examine recognition versus grader performance.

5 Discussion and Future Work

This paper presented first steps towards an automated, SDS-based method for holistically assessing conversational speech. Our proof-of-concept research demonstrated the feasibility of 1) using existing SDS to collect dialogues with NNSE, 2) human-assessing CEFR levels in such SDS speech, and 3) using an automated grader designed for prompted but non-interactive speech to yield scores that can positively correlate with humans.

---

13GP-BULATS was unable to grade 5 Bus dialogues. For example, if no words were recognized, fluency features such as the average length of words could not be computed. There are thus differing “n” values in Tables 1 and 2.
Much work remains to be done. A larger and more diverse speaker pool (in terms of first-languages and proficiency levels) is needed to generalize our findings. To create a public SDS corpus with gold-standard English skill assessments, work is needed in how to recruit speakers with such diverse skills, and how to change existing SDS systems to facilitate human scoring. Further examination of our research questions via controlled experimentation is also needed (e.g., for RQ1, comparing different corpus creation methods while keeping the SDS constant). Finally, we would like to investigate the grading impact of using optimized rather than off-the-shelf systems.

Acknowledgments

This work was supported by a Derek Brewer Visiting Fellow Award from Emmanuel College at the University of Cambridge and by a UK Royal Academy of Engineering Distinguished Visiting Fellowship Award. We thank the Cambridge Dialogue Systems Group, the ALTA Institute and Cambridge English for their support during this research. We also thank Carrie Demmans Epp, Kate Forbes-Riley and the reviewers for helpful feedback on earlier drafts of this paper.

References


David Vandyke, Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain dialogue success classifiers for policy training. In ASRU.