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[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[2] (Wang et al., 2013) or BULATS[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.

E: Do you work or are you a student
C: I'm a student in university er
E: And what subject are you studying

Figure 1: Testing dialogue excerpt between an IELTS human examiner (E) and a candidate (C) (Seedhouse et al., 2014).

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.[4] Second, there is a lack of optimal technical infrastructure. Existing SDS components such as speech recognizers will likely need modification to handle non-

[1] International English Language Testing System.

[2] Test of English as a Foreign Language.

[3] Business Language Testing Service.

[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[5] 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 trialogue-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 trialogue system. Our work differs in that we develop a dialogue rather than a trialogue 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[6], 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[7] (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-

[5] As 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.

[6] The 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).

[7] Thanks to Blaise Thompson for providing the system.

	Assessed												**Not**	**All**
	n	A1	A2	B1	B1B2	B2	B2C1	C1	C1C2	C2	Turns / Dial.	Wds. / Turn	n	n
L	21			1		4	6	7	2	1	11.48	3.9	4	25
R	14				2	8	3	1			6.36	5.5	6	20
B	20					1	2	10	6	1	13.65	2.6	2	22
C	55			1	2	13	11	18	8	2	10.96	3.6	12	67

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.

viously collected as part of the first Spoken Dialogue Challenge (SDC) (Black et al., 2011). However, our **Bus** corpus includes only a subset of the available SDC dialogues, namely non-native dialogues from the control condition. As in our AMT corpus collections, callers in the control condition received a scenario to solve over a web interface. 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 dialogues[8], we used AMT to collect similar numbers of dialogues with the other SDS. After removing problematic dialogues where the AMT task was completed but there was no caller speech or the caller turned out to be a native speaker, our final **Combined (C)** corpus contained 67 dialogues, distributed 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 benchmarking 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 recognition output was also made available. Since an expert in CEFR performed the assessment[10], dialogues were only scored by this single assessor. Sometimes the assessor assigned two adjacent levels 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 ordered 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[11] per user turn ("Wds./Turn") are also shown. With respect to RQ1, comparing the CEFR level distributions across rows suggests that different user recruitment methods do indeed yield different skill levels. Using AMT (the Laptop and Restaurant corpora) yielded more mid-level English speakers than the SDC method (the Bus corpus).[12] However, speakers in all three corpora are still biased towards the higher CEFR skill levels.

With respect to RQ2, not all dialogues could be assessed by the expert (as shown by the "Not" Assessed column of Table 1), often due to poor audio quality. Even for those dialogues that the expert was able to assess, human assessment was often felt to be difficult. When the SDS worked well, there was not very much user speech for making the assessment. When the SDS worked poorly, the dialogues became unnatural and speakers had to curtail potential demonstrations of fluency such as producing long sentences. Finally, only the Laptop and Bus systems recorded both sides of the conversation. Although the text for the Restaurant

[8] Only 22 of the 75 control callers were non-natives.

[9] The scores produced by the automatic grader described in Section 4.1 come with a mapping to CEFR.

[10] The Director of Academic Development and Training for International Students at Cambridge's Language Centre.

[11] The output of the speech recognizer for each SDS was used as only the SDC Bus download has transcriptions.

[12] A 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.

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

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.

system's prompts was made available, assessment was felt to be more difficult with only user speech.

4 Automated Assessment

After creating the SDS corpora with gold-standard speaker assessments (Section 3), we evaluated whether speech from such SDS interactions could be evaluated using an existing automated grader developed for prompted (non-dialogue) spontaneous speech (van Dalen et al., 2015).

4.1 The GP-BULATS Grader

The GP-BULATS automated grader (van Dalen et al., 2015) is based on a Gaussian process. The input is a set of audio features (fundamental frequency and energy statistics) extracted from speech, and fluency features (counts and properties of silences, disfluencies, words, and phones) extracted from a time-aligned speech recognition hypothesis. The output is a 0–30 score, plus a measure of prediction uncertainty. The grader was trained using data from Cambridge English's BULATS corpus of learner speech. Each of 994 learners was associated with an overall human-assigned grade between 0 and 30, and the audio from all sections of the learner's BULATS test was used to extract the predictive features. The speech recognizer for the fluency features was also trained on BULATS data. When evaluated on BULATS test data from 226 additional speakers, the Pearson's correlation between the overall grades produced by humans and by GP-BULATS was 0.83.

4.2 Applying GP-BULATS to SDS Speech

We transformed the expert CEFR ability labels (Table 1) to the grader's 0-30 scale, using a binning previously developed for GP-BULATS. The mean grades along with standard deviations are shown in the "Human" column of Table 2.[13] A one-way ANOVA with post-hoc Bonferroni tests shows that the Restaurant scores are significantly lower than in the other two corpora ($p \leq .01$).

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.

[13] GP-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

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41.

Suma Bhat, Huichao Xue, and Su-Youn Yoon. 2014. Shallow analysis based assessment of syntactic complexity for automated speech scoring. In *Proceedings 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1305–1315.

Alan W. Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, Jason D. Williams, Kai Yu, Steve Young, and Maxine Eskenazi. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of SIGDIAL*, pages 2–7.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *Ai Magazine*, 25(3):27.

CEFR. 2001. Common European framework of reference for languages: Learning, teaching, assessment. Cambridge University Press.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 722–731.

Lei Chen, Klaus Zechner, and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 442–449.

Keelan Evanini, Youngsoon So, Jidong Tao, D Zapata, Christine Luce, Laura Battistini, and Xinhao Wang. 2014. Performance of a trialogue-based prototype system for english language assessment for young learners. In *Proceedings Interspeech Workshop on Child Computer Interaction*.

Peter W Foltz and Mark Rosenstein. 2015. Analysis of a large-scale formative writing assessment system with automated feedback. In *Proceedings 2nd ACM Conference on Learning at Scale*, pages 339–342.

Kate Forbes-Riley and Diane Litman. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9):1115–1136.

Alexei V Ivanov, Vikram Ramanarayanan, David Suendermann-Oeft, Melissa Lopez, Keelan Evanini, and Jidong Tao. 2015. Automated speech recognition technology for dialogue interaction with non-native interlocutors. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 134.

Christopher M Mitchell, Keelan Evanini, and Klaus Zechner. 2014. A trialogue-based spoken dialogue system for assessment of english language learners. In *Proceedings International Workshop on Spoken Dialogue Systems*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing*, pages 794–799.

Rod D Roscoe, Danica Kugler, Scott A Crossley, Jennifer L Weston, and Danielle S McNamara. 2012. Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. In *FLAIRS Conference*.

Paul Seedhouse, Andrew Harris, Rola Naeb, Eda Üstünel, et al. 2014. Relationship between speaking features and band descriptors: A mixed methods study, the. *IELTS Research Reports Online Series*, page 30.

Vinay Shashidhar, Nishant Pandey, and Varun Aggarwal. 2015. Automatic spontaneous speech grading: A novel feature derivation technique using the crowd. In *Proceedings 53rd Annual Meeting of the Association for Computational Linguistics and 7th International Joint Conference on Natural Language Processing*, pages 1085–1094.

Pei-Hao Su, Chuan-Hsun Wu, and Lin-Shan Lee. 2015. A recursive dialogue game for personalized computer-aided pronunciation training. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(1):127–141.

Jesse Thomason and Diane Litman. 2013. Differences in user responses to a wizard-of-oz versus automated system. In *Proceedings of NAACL-HLT*, pages 796–801.

Rogier C. van Dalen, Kate M. Knill, and Mark J. F. Gales. 2015. Automatically grading learners' English using a Gaussian process. In *Proceedings Sixth Workshop on Speech and Language Technology in Education (SLaTE)*, pages 7–12.

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*.

Xinhao Wang, Keelan Evanini, and Klaus Zechner. 2013. Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 814–819.