A Trialogue-Based Spoken Dialogue System for Assessment of English Language Learners

Christopher M. Mitchell[1], **Keelan Evanini**[2], **and Klaus Zechner**[2]

[1]Department of Computer Science, North Carolina State University, Raleigh, NC, USA

[2]Educational Testing Service, Princeton, NJ, USA

cmmitch2@ncsu.edu, kevanini@ets.org, kzechner@ets.org

Abstract. Current automated assessment techniques for English language learners evaluate comprehension and synthesis skills via written text or one-turn spoken responses, measuring essential skills needed for academic and professional environments. However, as these current tests do not include dialogue-based components, they cannot provide insight into the conversational competence of the test-taker. This paper presents the first steps toward a dialogue-based, computer-driven assessment by introducing a spoken dialogue system designed to assess the English language skills of young children, and which provides scaffolding via its three-party trialogue-based structure. We show that it is possible to combine off-the-shelf components from the Olympus and Virtual Human Toolkits to quickly create a system with virtual characters and spoken language interaction for automated English language assessment.

1 Introduction

Automated assessment of English spoken language proficiency is an increasingly important goal in education and in business settings, with millions of English language learners completing assessments for credentialing and admissions purposes on an annual basis. Current automated assessments have been successful in gauging comprehension and synthesis skills of test-takers, but have not included meaningful assessment of skills necessary for conversational interaction. However, most spoken language takes place in an interactive context, so it is important for the validity of an assessment to incorporate more natural spoken communication scenarios. In recognition of this fact, some large-scale standardized language assessments do include interactive conversational tasks, but they are always mediated by a human interlocutor. Towards the goal of deploying such an automated spoken language assessment that contains naturalistic tasks, this paper presents a spoken dialogue system designed specifically for testing English language proficiency, built using freely available components from existing dialogue system toolkits. This dialogue system was developed as a proof-of-concept, demonstrat-

Proceedings of 5th International Workshop on Spoken Dialog Systems
Napa, January 17-20, 2014

ing the feasibility of using dialogue systems for English language assessment and presenting a three-party trialogue-based scenario in which conversational English skills can be tested.

2 Related Work

Current assessments of spoken language proficiency (both automated and human-scored) are primarily based on the stimulus-response model, in which the test taker is first presented with stimulus material (which could be an image, a video, a reading passage, a recorded conversation, etc.) and is then prompted to provide a spoken response; crucially, each prompt is not based on the preceding response provided by the test taker, and, thus, the assessment is not interactive. Automated systems for scoring these assessments have been shown to achieve promising performance levels, both for tasks eliciting restricted speech, such as reading a text out loud [1], and those eliciting spontaneous speech, such as summarizing a lecture [23]. However, none of these assessment systems have included interactive tasks that are able to evaluate a language learner's conversational skills.

Dialogue systems have shown great potential as an educational aid through the use of interactive tutoring systems in content domains such as physics [5,7], algebra [10], and computer literacy [6] as well as for the purpose of developing literacy skills [16]. In addition, there have been several applications that have employed spoken dialogue systems technology in the domain of foreign language learning to develop interactive tasks for improving various aspects of a language learner's proficiency. For the most part, however, the linguistic skills evaluated through these tasks have been limited to areas such as pronunciation (e.g., [20]) and vocabulary (e.g., [4]), and have not evaluated conversational skills that are necessary for interactive communication (although see [18] for an example of an interactive language learning task involving role-playing and problem solving with an automated agent and [12] for a system that assesses cultural skills that are necessary for successful second language communication). The goal of this project is to move beyond these relatively restricted types of tasks and design an interactive system that can be used to assess a language learner's communicative competence through their ability to participate successfully in an interactive conversation.

3 Trialogue Scenario Implementation

Interaction with the user in this system takes the form of a trialogue, i.e., a conversation between the user and two virtual agents. This form of interaction has been demonstrated to facilitate feedback and scaffolding in tutoring systems since it provides more opportunities for the user to assume different functional roles in the interaction than in two-party dialogues [8,13]. The scenario is designed for children in elementary school (grades 3-5), and thus takes place in settings which should be familiar to young students, such as classrooms and libraries, and involves communication with other students and teachers. The trialogue scenario begins with a teacher giving information to the user and to a virtual student, Lisa. The user must then relay the information to a second virtual student, Ron, who was not present for the teacher's announcement. As Ron asks questions to the user, Lisa is available to provide scaffolding support to the user, confirming correct answers by the user and responding appropriately to incorrect or off-topic responses. Figure 1 shows a potential sample interaction in the trialogue scenario (see [22] for further details about the trialogue materials used in this system).

	The teacher has explained that the students will be learning about weather in different parts of the world. Ron enters, having missed the teacher's explanation.
Ron:	What are we learning about today?
User:	**WEATHER**
Lisa:	Yes, but it's not about any weather. You need to tell Ron more.
Ron:	What are we learning about today?
User:	**WEATHER AROUND THE WORLD**
Lisa:	Yeah, that's right. We're learning about the weather around the world.

Figure 1. A sample interaction in the trialogue scenario as implemented in the system

In order to implement this trialogue scenario in a spoken dialogue system, several requirements had to be met, two of which are discussed here. The first major requirement was the capability to support multiple visually and aurally distinct agents within the system in order to represent the virtual agents in the scenarios: the teacher, Lisa, and Ron in this example. The second major requirement was the ability to preserve the dialogue history and to craft a custom dialogue manager using this history to provide suitable feedback and scaffolding during the scenario. After exploring the available toolkits for dialogue systems, we found that no single toolkit provided all of the needed functionality. The Virtual Human Toolkit (VHTK) [9] allows for quick creation of scenes with multiple computer-animated dialogue agents, but the default dialogue manager, the NPCEditor [14], is designed for building question-answering characters [19] and is not easily customizable to handle other dialogue structures. In particular, NPCEditor does not provide a straightforward way to find which parts of an expected response are missing, as in the trialogue illustrated in Figure 1. An alternative dialogue manager, FLoReS [15], is due to be included in a future release of VHTK, but was not available at the time of this writing. The Olympus framework [2] uses the more flexible RavenClaw dialogue manager [3], which is capable of handling a wide variety of dialogue structures, but the Olympus framework was designed with telephony in mind and does not support visual representations of dialogue agents or multiple system voices by default. So, components from both toolkits were combined in order to utilize the benefits of each.

Both toolkits consist of a collection of modules that communicate with each other via messages sent to a centralized hub (VHMessages in VHTK and Galaxy messages in Olympus). Thus, by creating an intermediary module to connect the two hubs, we were able to combine the two toolkits into the single dialogue system which we present here. This intermediary module was written in Java, utilizing the Java implementations of message handlers and senders for both message protocols. The next two sections present the components that were integrated from each toolkit into the dialogue system. Figure 2 shows the overall system architecture with the components from each toolkit. In essence, VHTK provides the front-end for the system, delivering visuals and audio to the user and accepting user input, while Olympus handles the back-end natural language understanding, dialogue management, and natural language generation.

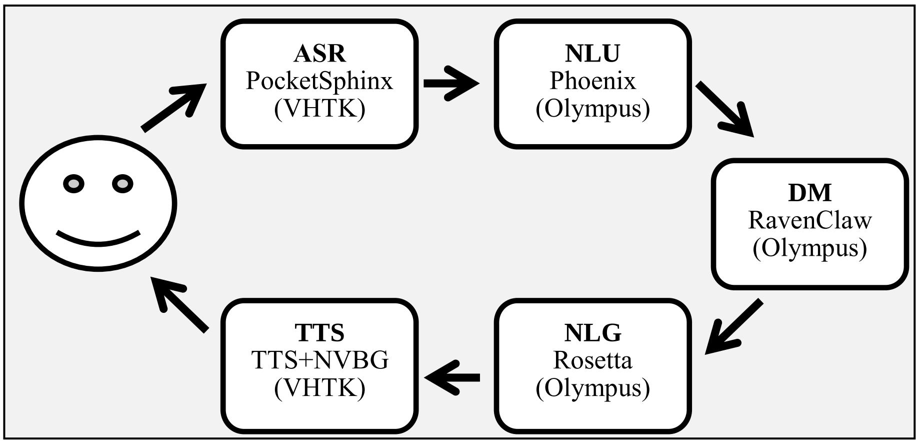

Figure 2. Dialogue system components

4 Virtual Human Toolkit

The Virtual Human Toolkit (VHTK) was developed at the USC Institute for Creative Technologies. Our system uses three modules from VHTK: the virtual human front-end, the automatic speech recognition module (PocketSphinx [11]), and the text-to-speech module.

The virtual human front-end is built using Unity[1], a multi-platform game engine and IDE, and consists of a number of virtual humans that can be placed into a scene, all of which are designed to respond to commands sent by VHTK. Three of these virtual humans were chosen (one each for the teacher, Lisa, and Ron), thus

[1] http://www.unity3d.com

preventing the costly time sink of modeling our own agents for the system. The virtual human front-end also integrates with the speech recognition module, allowing users to indicate when they are speaking, or to enter text and bypass the speech recognition module if necessary. The ASR output is sent to the Olympus components of the system, and an utterance is returned to VHTK along with its speaker. VHTK then generates the spoken output, either via TTS or via pre-recorded audio files. For our system, we use audio files which are pre-generated from TTS in order to reduce response time. VHTK also generates gestures and lip-syncing for the delivered utterances via its Non-Verbal Behavior Generator. Thus, the combined modules from VHTK allowed for rapid creation of high-quality visuals with vastly reduced effort compared to building a system from scratch.

5 Olympus

Olympus is a dialogue system framework created at Carnegie Mellon University. The trialogue system uses three components from Olympus: the Phoenix NLU module [21], the RavenClaw dialogue manager [3], and the Rosetta NLG module [17]. The Phoenix NLU grammar consists of a set of slots which represent answers to questions from the characters in the trialogue, with potential values for each slot listed in the grammar. The RavenClaw dialogue manager has a tree structure, with dialogue nodes executed in a depth-first, left-to-right order. The preconditions for entering and exiting each node determine the path through the tree based on user responses, i.e., the values of the slots from the Phoenix grammar. Each node then specifies to the Rosetta module the type of natural language output required, optionally calling C++ procedures to perform more complicated tasks as needed. In Rosetta, concrete realizations of each type of message are defined and, for the trialogue system, the speaker is prepended to the message so that the generated voice and animations will be sent to the correct character in the front-end.

6 Discussion and Future Work

The approach to designing a spoken dialogue system for interactive language assessment described in this paper enables the use of multiple interactive agents with distinct visual and aural characteristics combined with a robust custom dialogue manager that enables scaffolding and feedback to the test taker throughout the scenario. The presented system was designed using components from two publically available toolkits (VHTK and Olympus) along with a Java-based intermediary module to enable the communication hubs from the two systems to exchange messages.

Studies are currently being designed to evaluate the robustness of the presented system (in terms of ASR and NLU performance) as well as its usability (based on user feedback). After responses have been collected from participants, they will be provided with scores by expert human raters using scoring rubrics based on the language learner's successful completion of the communicative tasks and responsiveness to the feedback provided by the interactive agents. Subsequently, automated scoring features will be extracted both from the dialogue flow (for assessing task completion) and the spoken response (for assessing second language speaking proficiency), and a system for the automated prediction of these scores will be designed.

Acknowledgements

We would like to thank Youngsoon So and Diego Zapata for sharing the trialogue materials with us and Ben Leong for assistance using Olympus.

References

[1] Bernstein, J., Van Moere, A., Cheng, J. (2010). Validating automated speaking tests. *Language Testing* 27(3), 355-377.

[2] Bohus, D., Raux, A., Harris, T., Eskenazi, M., Rudnicky, A. (2007). Olympus: An open-source framework for conversational spoken language interface research. In *Proceedings of Bridging the Gap: Academic and Industrial Research in Dialog Technology Workshop at NAACL-HLT.*

[3] Bohus, D., Rudnicky, A. (2009). The RavenClaw Dialog Management Framework: Architecture and Systems. *Computer Speech and Language* 23(3), 332-361.

[4] Cai, C., Miller, R., Seneff, S. (2013). Enhancing speech recognition in fast-paced educational games using contextual cues. In *Proceedings of the 2013 Workshop on Speech and Language Technology in Education*.

[5] Chi, M., VanLehn, K., and Litman, D. J. (2010). Do micro-level tutorial decisions matter: applying reinforcement learning to induce pedagogical tutorial tactics. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, Pittsburgh, Pennsylvania, 224–234.

[6] D'Mello, S. K., and Graesser, A. (2012). AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. In *ACM Transactions on Interactive Intelligent Systems* 2(4).

[7] Forbes-Riley, K. and Litman, D. J. (2012). Adapting to multiple affective states in spoken dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on on Discourse and Dialogue (SIGDIAL)*, Seoul, South Korea, 217-226.

[8] Graesser, A.C., Britt, A., Millis, K., Wallace, P., Halpern, D., Cai, Z., Kopp, K. & Forsyth, C. (2010). Critiquing media reports with flawed scientific findings: Operation ARIES!, a game with animated agents and natural language trialogues. In J. Aleven, J. Kay, & J. Mostow (Eds.) *Lecture Notes in Computer Science*, 6095, 327-329. London: Springer.

[9] Hartholt, A., Traum, D., Marsella, S., Shapiro, A., Stratou, G., Leuski, A., Morency, L., Gratch, J. (2013). All together now: Introducing the Virtual Human Toolkit. In *Proceedings of the International Conference on Intelligent Virtual Humans*.

[10] Heffernan, N. T., and Koedinger, K. (2001). The design and formative analysis of a dialog-based tutor. In *Workshop on Tutorial Dialogue Systems*, 23–34.

[11] Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I. (2006). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held de-

vices. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[12] Johnson, W.L., Marsella, S., Mote, N., Vilhjalmsson, H., Narayanan, S., and Choi, S. (2004). Tactical language training system: Supporting the rapid acquisition of foreign language and cultural skills. In *Proceedings of the InSTIL/ICALL Symposium.*

[13] Lehman, B. A., D'Mello, S. K., Strain, A., Gross, M., Dobbins, A., Wallace, P., Millis, K., & Graesser, A. C. (2011). Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 171-178. Berlin: Springer-Verlag.

[14] Leuski, A., Traum, D. (2011). NPCEditor: A tool for building question-answering characters. In *Proceedings of the International Conference on Language Resources and Evaluation*.

[15] Morbini, F., DeVault, D., Sagae, K., Gerten, J., Nazarian, A., Traum, D. (2012). FLoReS: A forward looking, reward seeking, dialogue manager. In *Proceedings of the International Workshop on Spoken Dialog Systems*.

[16] Mostow, J., Nelson, J., & Beck, J. E. (2013). Computer-guided oral reading versus independent practice: Comparison of sustained silent reading to an automated reading tutor that listens. *Journal of Educational Computing Research*, 49(2), 249-276.

[17] Oh, A., Rudnicky, A. (2000). Stochastic language generation for spoken dialogue systems. In *Proceedings of the ANLP/NAACL workshop on conversational systems*.

[18] Seneff, S., Wang, C. Chao, C.-Y. (2007). Spoken dialogue systems for language learning. In *Proceedings of NAACL-HLT.*

[19] Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D., Lane, C., Morie, J., Aggarwal, P., Liewer, M., Chiang, J.-Y., Gerten, J., Chu S., and White, K. (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA).*

[20] Su, P.-H., Yu, T.H., Su, Y.-Y., Lee, L.-S. (2013). NTU Chinese 2.0: A personalized recursive dialogue game for computer assisted learning of Mandarin Chinese. In *Proceedings of the 2013 Workshop on Speech and Language Technology in Education*.

[21] Ward, W. (1994). Extracting information from spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing*.

[22] Zapata-Rivera, D., So, Y., Cho, Y., Vezzu, M. (2013). Using trialogues to measure English language skills. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco.

[23] Zechner, K., Higgins, D., Xi, X., Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51(10), 883-895.