Development of an Audiovisual Database of Human-Machine Conversations for Educational Learning and Assessment Applications

Vikram Ramanarayanan[†], David Suendermann-Oeft[†], Patrick Lange[†],
Robert Mundkowsky[‡], Alexei V. Ivanov[†], Zhou Yu[⋆], Yao Qian[†] and Keelan Evanini[‡]
Educational Testing Service R&D
[†]90 New Montgomery St, #1500, San Francisco, CA
[‡]660 Rosedale Road, Princeton, NJ
[⋆]Carnegie Mellon University, Pittsburgh, PA
vramanarayanan@ets.org

Abstract

Due to the increasing use of English as lingua franca in the workforce and academia, there is a need for valid assessments of English conversational skills for non-native speakers as well as instructional materials that enable English learners to improve these skills. Current large-scale assessments of non-native English speaking proficiency (such as TOEFL iBT[1], TOEIC[2], Pearson Test of English Academic[3]) typically contain isolated test questions that elicit monologues from the test taker. In this *prompt-response* model, the test questions are fixed, and the presentation of questions does not depend on the answers provided by test takers to previous questions. Crucially, since there is no interactive dialogue, these types of test questions are not able to elicit the full range of English speaking skills (such as turn taking abilities, politeness strategies, pragmatic competence) that are required for successful communication.

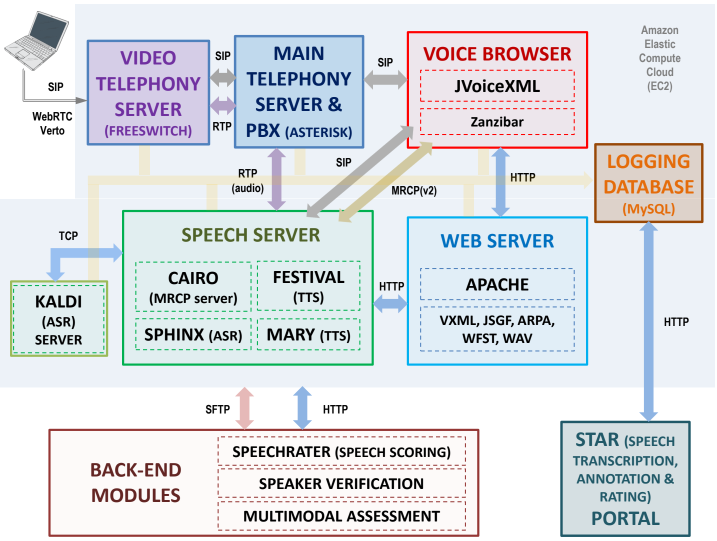

Figure 1. The HALEF multimodal dialog system architecture that supports educational learning and assessment applications.

To this end, we present the current state of the art of an educational-domain multimodal dialog system and associated database of audiovisual interactions between the system and human interlocutors. We leverage HALEF (Help Assistant–Language-Enabled and Free)[4], an open-source cloud-based standards-compliant multimodal dialog system to collect data in a crowdsourced manner using Amazon Mechanical Turk. The HALEF architecture and components (see Figure 1) have been described in detail in prior publications [4, 3, 6]. Crowdsourcing, and particularly Amazon's Mechanical Turk, has been used in the past for the assessment of spoken dialog systems (SDSs) as well as for collection of interactions with SDSs [2, 5, 1].

We deployed six goal-oriented conversational tasks (see Table 1) from common workplace communicative scenarios for the purposes of this data collection; these scenarios included responding to an offer of food, scheduling a meeting, interviewing for a job, taking a customer's order, and requesting a meeting with boss or friend. Figures 2 shows an example dialog workflow for one example item. See [3] for details regarding other example workflows. In addition to reading instructions and calling into the system, users were

[1] http://www.ets.org/toefl
[2] https://www.ets.org/toeic
[3] http://pearsonpte.com/
[4] http://halef.org

Table 1. *The six conversational tasks deployed. Along with the number of interactions collected for each task, we also list the total duration of audio and video collected as well as the average handling time, i.e. average duration of the human-machine interaction.*

item	n	total audio/h	total video/h	AHT/s
Food Offer	1047	16.9	9.2	58.0
Schedule	323	9.9	0.0	110.5
Job Interview	891	56.1	40.0	226.8
Customer Order	1058	35.5	20.6	120.9
Meeting Request Boss	713	14.5	14.0	73.2
Meeting Request Friend	661	12.2	12.1	66.4
total	4693	145.2	96.1	111.4

requested to fill out a 2-3 minute survey regarding the interaction. Participants were mostly native speakers of American English from all over the continental United States, the ratio of non-native speakers was about 17%. In all, we collected 4693 conversations over about nine months, with 145 hours of audio out of which 96 hours also contained video dialog data.

Collecting multimodal streams of information, including video, of learner- and test-taker interactions, is useful for several reasons. In addition to making the interactions more natural as compared to traditional *prompt-response* model-based questions, the combination of audio and video cues can be used to determine how engaged users are, analyze their facial expressions and potentially their body language, all of which can be used to inform dialog management routines affecting how the automated system interacts with the users in real-time. Another important application is the post-analysis of the audio-video data to measure, manually or automatically, speaking and behavioral skills of the user for formative or summative assessment or learning purposes.

In the future, we plan to continue collecting data from human participants through crowdsourcing to both improve the current items as well as develop new ones. We also plan to place a stronger effort in developing statistical language understanding, dialog management, and computer vision modules in order to enhance system performance, user experience, and feedback capabilities.

The authors would like to thank Eugene Tsuprun, Lydia Rieck, Elizabeth Bredlau, Katie Vlasov, Juliet Marlier, Phallis Vaughter, Nehal Sadek, Veronika Laughlin and Tanner Jackson for their help in designing the conversational tasks, and Hillary Molloy and Zydrune Mladineo for their help in data collection.

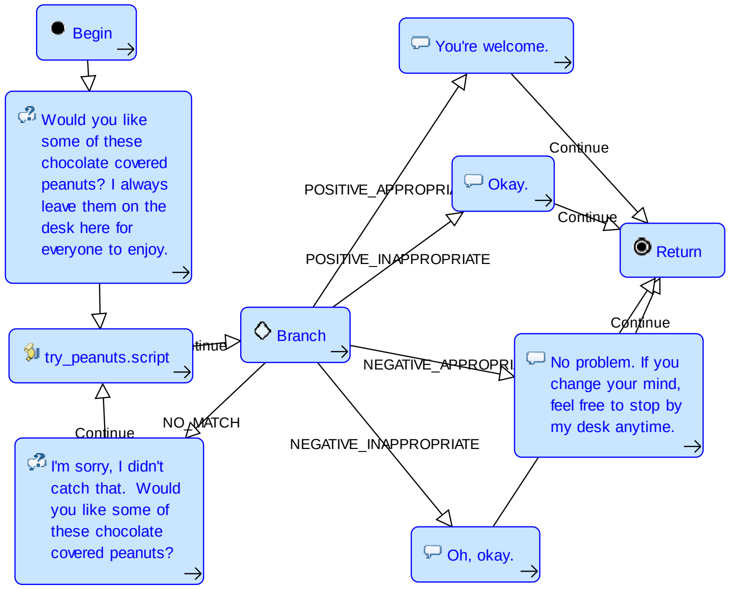
Begin
Would you like some of these chocolate covered peanuts? I always leave them on the desk here for everyone to enjoy.
try_peanuts.script
tinue
Branch
Continue
NO_MATCH
I'm sorry, I didn't catch that. Would you like some of these chocolate covered peanuts?
POSITIVE_APPROPRI
POSITIVE_INAPPROPRIATE
NEGATIVE_APPROPRI
NEGATIVE_INAPPROPRIATE
You're welcome.
Okay.
Continue
Continue
Return
Continue
No problem. If you change your mind, feel free to stop by my desk anytime.
Oh, okay.

Figure 2. Example design of a workplace pragmatics-oriented application targeted at non-native speakers of English where the caller has to accept or decline an offer of food (peanuts, in this case) in a pragmatically appropriate manner.

References

[1] F. Jurcıcek, S. Keizer, M. Gašic, F. Mairesse, B. Thomson, K. Yu, and S. Young. Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk. In *Proceedings of INTERSPEECH*, volume 11, 2011.

[2] I. McGraw, C.-Y. Lee, I. L. Hetherington, S. Seneff, and J. Glass. Collecting voices from the cloud. In *LREC*, 2010.

[3] V. Ramanarayanan, D. Suendermann-Oeft, A. Ivanov, and K. Evanini. A distributed cloud-based dialog system for conversational application development. In *16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015), Prague, Czech Republic*. 2015.

[4] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. Ivanou, Z. Yu, Y. Qian, and K. Evanini. Assembling the jigsaw: How multiple open standards are synergistically combined in the HALEF multimodal dialog system. In *Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything*, page to appear. Springer, 2016.

[5] M. Rayner, I. Frank, C. Chua, N. Tsourakis, and P. Bouillon. For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language call application. In *Speech and Language Technology in Education*, 2011.

[6] D. Suendermann-Oeft, V. Ramanarayanan, M. Teckenbrock, F. Neutatz, and D. Schmidt. HALEF: An Open-Source Standard-Compliant Telephony-Based Modular Spoken Dialog System: A Review and An Outlook. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 53–61. Springer, 2015.