An Open-source Dialog System with Real-Time Engagement Tracking for Job Interview Training Applications

Zhou Yu[‡], Vikram Ramanarayanan[†], Patrick Lange[†] and David Suendermann-Oeft[†]

Abstract In complex conversation tasks, people react to their interlocutor's state, such as uncertainty and engagement to improve conversation effectiveness [2]. If a conversational system reacts to a user's state, would that lead to a better conversation experience? To test this hypothesis, we designed and implemented a dialog system that tracks and reacts to a user's state, such as engagement, in real time. We designed and implemented a conversational job interview task based on the proposed framework. The system acts as an interviewer and reacts to user's disengagement in real-time with positive feedback strategies designed to re-engage the user in the job interview process. Experiments suggest that users speak more while interacting with the engagement-coordinated version of the system as compared to a non-coordinated version. Users also reported the former system as being more engaging and providing a better user experience.

Key words: multimodal dialog systems, engagement, automated interviewing

1 Introduction and Related Work

Recently, multimodal sensing technologies such as face recognition, head tracking, etc. have improved. Those technologies are now robust enough to tolerate a fair amount of noise in the visual and acoustic background [3, 7]. So it is now possible to incorporate these technologies into spoken dialog systems to make the system aware of the user's behavior and state, which in turn will result in more natural and effective conversations [14].

Multimodal information has been proven to be useful in dialog system design in driving both low level mechanics such as turn taking as well as high level mechanics

[†] Educational Testing Service (ETS) R&D, San Francisco, CA, USA. e-mail: {vramanarayanan, plange, suendermann-oeft}@ets.org
[‡] Carnegie Mellon University, Pittsburgh, PA, USA. e-mail: {zhouyu}@cs.cmu.edu

such as conversation planning. Sciutti et al. [11] used gaze as an implicit signal for turn taking in a robotic teaching context. In [17], a direction-giving robot used conversational strategies such as pause and restarts to regulate the user's attention. Kousidis et al. [4] used situated incremental speech synthesis that accommodates users' cognitive load in a in-car navigation task, which improved user experience but the task performance stays the same. In Yu et al. [19], a chatbot reacts to the user's disengagement by generating utterances that actively invite the user to continue the conversation.

Thus we propose a task-oriented dialog system framework that senses and coordinate to a user's state, such as engagement, in real time. The framework is built on top of the HALEF[1] open-source cloud-based standards-compliant multimodal dialog system framework [8, 20]. It extracts multimodal features based on data that is streamed via the user's webcam and microphone in real time. Then the system uses these multimodal features, such as gaze and spoken word count to predict a user's state, such as engagement, using a pre-built machine learning model. Then the dialog manager takes the user's state into consideration in generating the system response. For example, the system could use some conversational strategies, such as positive feedback to react to the user's disengagement state.

With the advantage of being accessible via web-browser, HALEF enables users interact with the system whenever and wherever in their comfortable environment, thus making the data collection and system evaluation process much easier and economical.

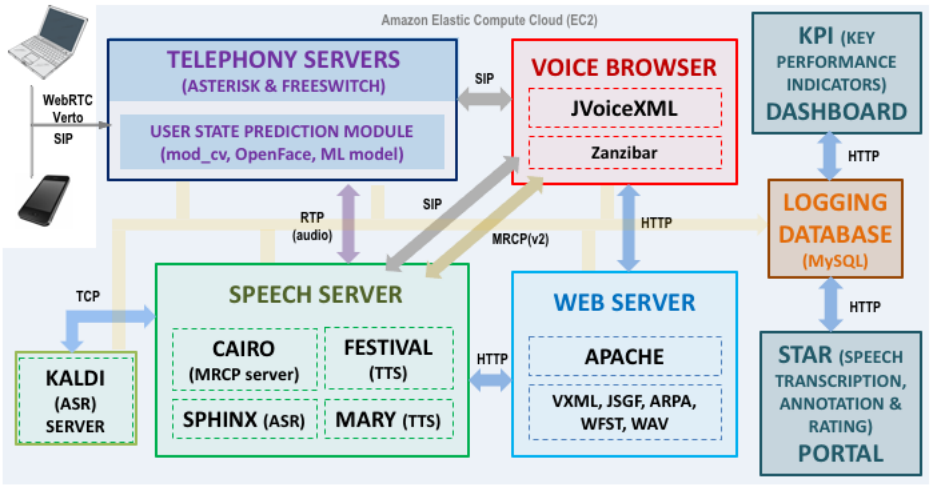

Fig. 1 System architecture of the User-Coordinated HALEF dialog system

[1] https://sourceforge.net/projects/halef/

2 The HALEF Framework with real-time engagement tracking

In this section, we describe the sub-components of the framework. Fig 1 schematically depicts the overall architecture of the HALEF framework.

2.1 The Multimodal HALEF Framework

FreeSWITCH, specifically versions above 1.6 [2], is a scalable, open-source and cross-platform telephony framework designed to route and interconnect popular communication protocols using audio, video, text or any other form of media. FreeSWITCH allows the experimenter to modify interaction settings, such as the number of people who can call in at any given time, whether to display the video of the user on the webpage, the resolution of the video, sampling rate of the audio, etc. The FreeSWITCH Verto protocol also allows users to choose between different I/O devices for recording. They can switch between different microphones and cameras connected to their computers by selecting appropriate options on the web-based graphical user interface. We use FreeSWITCH Verto to connect user to HALEF via web browsers.

HALEF leverages different open-source components to form a SDS framework that is modular and industry-standard-compliant: Asterisk, a SIP (Session Initiation Protocol) and PSTN (Public Switched Telephone Network) compatible telephony server [13]; JVoiceXML, an open-source voice browser that can process SIP traffic [9]; Cairo, an MRCP (Media Resource Control Protocol) speech server, which allows the voice browser to request speech recognition, speech synthesis, audio playback and audio recording from the respective components; the Sphinx automatic speech recognizer [5] and the Kaldi ASR system; Festival [12] and Mary [10] text to speech synthesis engines; and an Apache Tomcat-based web server that can host dynamic VoiceXML (VXML) pages and serve media files such as grammars and audio files to the voice browser. OpenVXML allows designers to author the dialog workflow as a flowchart, including details of specific grammar files to be used by the speech recognizer and text-to-speech prompts that need to be synthesized. In addition, dialog designers can insert "script" blocks of Javascript code into the workflow that can be used to perform simple processing steps, such as creating HTTP requests to make use of natural language understanding web services on speech recognition output. In order to react to the user's engagement, these "script" blocks retrieve and act upon the engagement score of the user in real time. The entire workflow can be exported to a Web Archive (or WAR) application, which can then be deployed on an Apache Tomcat web server that serves Voice XML documents.

We use the open-source database MySQL for our data warehousing purposes. All modules in the Multimodal HALEF connect to the database and write their log

[2] https://freeswitch.org/confluence/display/FREESWITCH/FreeSWITCH+1.6+Video

messages into it. We then post-process this information with stored procedures into easily accessible views. Metadata extracted from the logs include information about the interactions such as the duration of the interaction, the audio/video recording file names, the caller IP, etc. Additionally, we store participants' survey data and expert rating information. All the modules connected to the database have been implemented such that all information will be available in the database as soon as the interaction, the survey, or the rating task is completed.

2.2 User State Prediction Module

The user state prediction module is linked with FreeSWITCH via sockets to a standalone Linux server for automatic head tracking using OpenFace [1]. The head tracker receives raw images from FreeSWITCH and performs tracking of the user's head movement, gaze direction and facial action units. Visual information has been shown to be critical in assessing the mental states of the users in other systems as well [18]. So we include visual information in predicting the user state. The system uses the pre-trained machine learning model to predict the user's state. The user state module doesn't connect to HALEF directly, it passed the engagement information to the database first and then the application retrieves the user state information from the database. The application selects the conversation branch based on the user state information as well as the spoken language understanding results.

3 Example Application: Job Interview Training/Practice

In this conversation task, the system acts as the interviewer for a job in a pizza restaurant. The system first asks the user some basic personal information and then proposes two scenarios about conflicts that may happen in the workplace and asks the user how he/she would resolve them. We designed the task to assess non-native speakers' English conversational skills, pragmatic appropriateness of responses, and their ability to comprehend the stimulus materials and respond appropriately to questions posed during naturalistic conversational settings.

3.1 User Engagement Modeling

We first collected a set of data using a non-coordinated multimodal interviewer version. We then used this dataset to build supervised machine learning models to predict engagement in real time, as well as a baseline for the reactive version of the system. We collected 200 video recordings in all from crowdsourced participants recruited via the Amazon Mechanical Turk platform. To train the engagement

model we randomly selected 30 conversations which satisfied the following two quality criteria: (i) the face recognition system detects a face with high confidence (80%) throughout the interaction, and (ii) the automatic speech recognition output of the user utterances is not empty. There are in total 367 conversational exchanges in total over 30 conversations (note that we use the term conversational exchange here to denote a pair of one system turn and one user turn). We asked three experts to annotate user engagement for every conversational exchange based on the video and audio recordings. We adopted the engagement definition and annotation scheme introduced in [19]. For the purposes of our study, we defined engagement as the degree to which users are willing to participate in the task along three dimensions – behaviorally (staying on task and following directions), emotionally (for instance, not being bored by the task) and cognitively (maximizing their cognitive abilities, including focused attention, memory, and creative thinking) [16]. Ratings were assigned on a 1-5 Likert scale ranging from very disengaged to very engaged. We collapsed 1-2 ratings into a "disengaged" label, and 3-5 into an "engaged" label, because the system is designed to react to a binary signal in the conversational flow. The threshold was chosen because we would like to only regulate the extreme cases in our task, in order to keep the conversation to be effective. For other tasks, we recommend setting the threshold as an experimental parameter that decided through user preference. There are in total three annotators involved and they had an inter-annotator Cohen's κ agreement value of 0.82 on average on the binary engaged vs. disengaged rating task. For modeling purposes, we used the average label from all annotators as the ground truth or gold standard label. Among all the conversation exchanges, 75% were labeled as "engaged" and 25% were labeled as "disengaged". So while the current work only considers binary labels, future work will examine design policies that takes a finer grained 5-point engagement scale into consideration.

In order to train a vision-based engagement predictor, we extracted the following vision features: head pose, gaze and facial action units. After the user state module receives the real-time features, it simultaneously performs three computing steps for engagement detection [1]. It processes the mean and variance of the head pose change to determine the frequency of users changing their head pose. It also extracts the mean and variance of the action units that relate to smiles, in order to calculate the frequency of user smiles. It further computes the mean and variance of the gaze direction, to capture how frequency users shift their gaze. These features are computed per conversational exchange to form the feature set for engagement predictor training. Further, we also take verbal information into account for engagement prediction. In this interview training task, since there are a fixed number of conversation states, we calculated the mean word count of all the conversations that are labeled as disengaged in each state. We use this value as the threshold to decide if the user is disengaged or not for each state. The verbal engagement score is computed over the ASR output as soon as the user utterance is finished.

We used leave-one-conversation-out cross validation and a Support Vector Machine with linear kernel to train the model. Our experiments resulted in an F1 measure of 0.89 (the majority vote baseline was 0.72). We observed that the failures

occurred mainly due to the system turn-taking errors, such as system interrupting the user, which results a shorter user response and in turn leads to a low engagement score in the verbal channel. Note that the relative high baseline is due to the skewness of the data, as there are more engaged conversational exchanges than disengaged ones. During testing, we quantify vision features in a time window which is empirically determined by the engagement predictor's performance based on the task (we chose 2s, which happens to be the mean of the conversational exchange duration). Then we combine all the multimodal features mentioned above with weights obtained from the machine learning model to obtain a score that represents the visual engagement of the user. The dialog manager receives a score which is a weighted sum of all the modality-wise engagement scores with a set of weights determined empirically.

As a follow-up experiment to see how well this initial engagement detector performed, we then collected 54 conversations using the engagement-coordinated interview training application. Among them we found 23 recordings that satisfied the two quality criteria mentioned earlier as well as an additional criterion of analyzing data from unique speakers. Experts then rated the engagement at each conversational exchange where the system is required to make a dialog management decision based on the user's predicted engagement state.

Next, we used this data to retrain the linear regression weights assigned for the vision and verbal modalities in Equation (1); here x_{i1} stands for the visual-engagement value, x_{i2} stands for the verbal-engagement value and y_i stands for the ground truth label. The simple linear regression analysis performed a least-squares optimization of the following cost function:

$$min_{\alpha\beta} \sum_{i=1}^{n} (y_i - \alpha x_{i1} - \beta x_{i2})^2 \tag{1}$$

We found optimal regression coefficients of α= 0.63 and β = 0.37 and adjusted the weights in the model accordingly. We then collected another set of 50 conversations with adjusted weights. Among them 32 of them are valid videos for analysis according to the quality criteria. In the result analysis we used this batch data for analysis for the engagement-coordinated system.We found the F1 score for the trained engagement classifier was 0.86, a significant improvement over the majority vote baseline method of 0.74.

3.2 Conversational Strategy Design

The communication and education literature mention a number of conversational strategies which are useful to improve user engagement, such as active participation, mention of shared experience [15], and positive feedback and encouragement [6], among others. Particularly in job interview literature, researchers find that infusing positive sentiments into the conversation could lead to more self-disclosure from

interviewees. With this in mind, we designed a set of positive feedback strategies with respect to different dialog states. For example, in one dialog state, we asked about the interviewee's previous experience.

3.3 Coordination Policy

We implemented a local greedy policy to react to the user's disengagement in this interview training task. Once the dialog manager receives the signal from the end-pointing module reporting that the user has finished the turn, it queries the most recent engagement score from the database. If the user is sensed as disengaged, the positive strategy that is designed with respect to that conversational state will trigger, otherwise the conversation goes into next dialog state.

3.4 Results

We asked callers to fill out a survey after interacting with the system. We asked them how engaged they felt overall during the interaction as well as their overall conversational experience on a 1-5 Likert scale. We compared the user responses of the 30 conversations that were collected using the non-coordinated interview training method to the 32 conversations that are collected using the engagement-coordinated version, and found that the engagement-coordinated system received statistically higher overall ratings from the users in terms of both overall engagement and user experience (see Fig 2 for details).

Though the engagement-coordinated version had more system utterances than the non-coordinated one, the extra utterances are all statements (e.g., "I think the manager would do that.") instead of questions, so no extra user utterances were elicited. Nonetheless, when we calculated the number of unique tokens of all the uses' utterances based on the ASR output, we found that users who interacted with the engagement-coordinated version expressed significantly more information than users who interacted with the non-coordinated version (see Fig 2 for details).

We also found that there were three users who interacted with the engagement-coordinated interview item more than once. We found that their assessed average engagement improved (from 2.3 to 4.0) after interacting with the system several times. This gives us a positive indication that interacting with our system does have the potential to allow users to improve their conversational ability during job interviews.

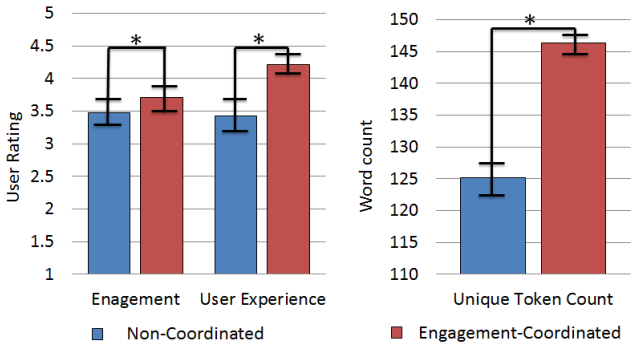

Fig. 2 Experiment results between non-coordinated and engagement-coordinated job interviewer

4 Conclusion and Future Work

We have proposed and implemented a real-time user-reactive system framework for task-oriented conversations. We implemented an example application based on the framework in the form of an engagement-coordinated interview training task. From the data collected using both the non-coordinated and engagement-coordinated versions of the interview training task, we found that the engagement-coordinated version was rated as more engaging and providing a better user experience as compared to a non-coordinated system.

In the future, we wish to design and implement improved reactive systems that are able to tackle more complex tasks, such as for conversational proficiency practice. We also wish to design better conversation policies that take dialog context information into consideration during conversation flow planning. We will also integrate more speech feature-based information into the engagement module in order to make the engagement prediction more accurate.

Acknowledgements

We would like to thank Robert Mundkowsky and Dmytro Galochkin for help with system engineering. We would also like to thank Eugene Tsuprun, Keelan Evanini, Nehal Sadek and Liz Bredlau for help with the task design and useful discussions.

References

1. T. Baltru, P. Robinson, L.-P. Morency, et al. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages

1–10. IEEE, 2016.
2. K. Forbes-Riley and D. J. Litman. Adapting to student uncertainty improves tutoring dialogues. In *AIED*, pages 33–40, 2009.
3. X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):328–340, 2005.
4. S. Kousidis, C. Kennington, T. Baumann, H. Buschmeier, S. Kopp, and D. Schlangen. A multimodal in-car dialogue system that tracks the driver's attention. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 26–33. ACM, 2014.
5. P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf. The cmu sphinx-4 speech recognition system. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, volume 1, pages 2–5. Citeseer, 2003.
6. B. Lehman, S. DMello, and A. Graesser. Interventions to regulate confusion during learning. In *International Conference on Intelligent Tutoring Systems*, pages 576–578. Springer, 2012.
7. L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the 7th international conference on Multimodal interfaces*, pages 18–24. ACM, 2005.
8. V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. Ivanou, Z. Yu, Y. Qian, and K. Evanini. Assembling the jigsaw: How multiple w3c 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.
9. D. Schnelle-Walka, S. Radomski, and M. Mühlhäuser. Jvoicexml as a modality component in the w3c multimodal architecture. *Journal on Multimodal User Interfaces*, 7(3):183–194, 2013.
10. M. Schröder and J. Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377, 2003.
11. A. Sciutti, L. Schillingmann, O. Palinko, Y. Nagai, and G. Sandini. A gaze-contingent dictating robot to study turn-taking. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 137–138. ACM, 2015.
12. P. Taylor, A. W. Black, and R. Caley. The architecture of the festival speech synthesis system. 1998.
13. J. Van Meggelen, L. Madsen, and J. Smith. *Asterisk: the future of telephony*. " O'Reilly Media, Inc.", 2007.
14. A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.
15. D. Wendler. Improve your social skills. *CreateSpace Independent Publishing Platform*, 2014.
16. J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.
17. Z. Yu, D. Bohus, and E. Horvitz. Incremental coordination: Attention-centric speech production in a physically situated conversational agent. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 402, 2015.
18. Z. Yu, D. Gerritsen, A. Ogan, A. W. Black, and J. Cassell. Automatic prediction of friendship via multi-model dyadic features. In *Proceedings of SIGDIAL*, pages 51–60, 2013.
19. Z. Yu, L. Nicolich-Henkin, A. Black, and A. Rudnicky. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016.
20. Z. Yu, V. Ramanarayanan, R. Mundkowsky, P. Lange, A. Ivanov, A. W. Black, and D. Suendermann-Oeft. Multimodal halef: An open-source modular web-based multimodal dialog framework. 2016.