

# User Engagement Study with Virtual Agents Under Different Cultural Contexts

Zhou Yu<sup>(✉)</sup>, Xinrui He, Alan W. Black, and Alexander I. Rudnicky

Carnegie Mellon University, Pittsburgh, USA  
zhouyu@cs.cmu.edu

**Abstract.** Human communication literature states that people with different culture backgrounds act differently in conversations. Currently most virtual agents are designed for a single targeted popular culture. We implemented two versions of a virtual agent targeting American and Chinese cultures. We found that users from different culture context express engagement differently.

## 1 Introduction and Related Work

Recent work on culture behavior difference mostly refer to the Hofstede's culture model which had six dimensions of visibilities [4]: small versus large power distance, individualism versus collectivism, masculinity versus femininity, weak versus strong uncertainty avoidance, long versus short term orientation and indulgence versus resilience. The major differences between American and Chinese cultures are that American culture is perceived to be individual, short-term oriented and low power distance, while Chinese culture is collective, long term oriented and high power distance.

Many studies found that people from different cultures behave differently during conversations. The CUBE-G project is one of the most extensive data-driven efforts to study German and Japanese cultures comparatively. Rehm et al. [7] collected a cross-cultural multimodal corpus of dyadic interactions and found that in most Japanese conversations, participants discussed the experimental setting while German subjects talked significantly more about social topics such as their studies or friends. Khaled et al. focused on cultural differences in persuasion strategies found that for short-term oriented cultures a stronger focus on the task itself can be expected, whereas for long-term oriented cultures a slower and more exhaustive way of problem solving can be expected, where every opinion is taken into account and harmony is at stake resulting in an increased frequency of contributions that are related to the communication management [5]. Matsumoto et al. found that people from Arab cultures gaze much longer and more directly than Americans [6]. In general, collective cultures, such as Arabian culture engage in more gazing and have more direct orientation when interacting with others.

Previous studies found that people in different cultures behave differently towards a task-oriented virtual agent as well. In a direction giving task, Arabic

**Table 1.** An example conversation with TickTock in mandarin

Partner	Utterance (Translation)
<i>User:</i>	那要不你给我唱首歌吧。(Then let me sing you a song)
<i>TickTock:</i>	让我想一下，就没有浪漫一点反应啊？(Let me see, don't you have something romantic to say?)
<i>User:</i>	什么鬼啊？我感觉受不了这个机器了。(What the heck, I feel I can't take the system anymore.)
<i>TickTock:</i>	你到底有没有爱过我？(Have you loved me at all?)
<i>User:</i>	什么情况？(What is the situation here?)
<i>TickTock:</i>	让我想一下，好吧。(Let me see, OK.)
<i>User:</i>	好吧。(OK.)
<i>TickTock:</i>	你有什么想说的吗？(Do you have anything to say?)
<i>User:</i>	爱过。(I loved you.)

and English native speakers interact with a virtual agent differently [3]. English natives had a higher frequency of using cardinals, pauses and intermediate information while Arab natives used units of distance, left/right turns and error corrections more frequently than English natives. Another comparison study suggested that Arabians trust more on the Arabian speaking robot who speaks language that is rhetorically well formulated, while the rhetorically factor is less important for Americans when they talk to an English speaking robot [1].

## 2 System and Data Descriptions

TickTock is a key-word based retrieval system with a set of conversational strategies [9]. The response of TickTock is retrieved from an open-domain dialog corpus, composed of various question-answer pairs using a keyword matching algorithm. The system selects strategies based on the retrieved response and the user model. The system reacts to the engagement level of the user. We built a computational model for user engagement using user verbal and nonverbal behaviors using simile procedure in [8]. Based on how engaged the user is, the system chooses between the four strategies: switch a topic, initiate an activity, tell a joke and refer back to a previous topic that user engaged.

The system is originally designed for American culture. The response generation model is trained on an open-domain dialog corpus formed by American popular social media, such as CNN interview corpus, TV show “Friends” and Reddit. Translated text usually appears unnatural in the target language, which may lead to less believability of the agent’s culture identity. Thus we used similar social media materials as the American version, but is originally created in Mandarin: Xinlang Aiwon (similar to Quora) and TV show “Love apartment”. We replaced the Google Automatic Speech Recognizer (ASR) and Flite Text-to-speech (TTS) [2] with Baidu ASR and Baidu TTS respectively to support automatic Mandarin recognition and synthesis. The agent has a cartoon face that signals its internal mind, such as smiling, and confused. The mouth of the agent moves while it speaks. However, it is incapable of complicated facial expressions,

let alone gestures. This design aims to avoid the uncanny valley dilemma, so that the users would not expect realistic human-like behaviors from the system. We did not change the appearance of the agent for different cultural contexts, as the agent’s appearance is designed to be culturally ambiguous. A demo can be found in <http://www.cs.cmu.edu/afs/cs/user/zhouyu/www/TickTock.html>.

We believe this is the first public data set that had audiovisual recordings of people from two different cultures interacting with a similar agent with their native languages. In order to isolate confounds, we balanced subjects for gender, education background and age. The average age of the American data set is 24.4 (STD = 2.53), and the Chinese data set is 21.4 (STD = 0.58). There are 11 people (7 males) in total in the American data set, and 21 people (12 males) in total in the Chinese data set. In the user study, American users interacted with the English speaking version of TickTock and the Chinese users interacted with the Mandarin speaking version. We recruited the participants and conducted the experiments in the country of the targeted culture. Participants are university students who were born and raised in the targeted culture. No participants interacted with a virtual agent before, however, they have varied familiarity with dialog systems, which may influence their behaviors during the interaction. We wish to control this factor in future studies. A Mandarin example conversation between TickTock and the user is shown in Table 1. We adopted the engagement annotation scheme in [9] and the inter-annotator agreement (kappa) is 0.93 and 0.74 in the American and Chinese data set respectfully.

### 3 Engagement Analysis

We used the same feature extraction technology in [8] to extract multimodal features from the two data sets. Then we perform a correlation analysis between each multimodal features with respect to its engagement score and report the tests with statical significance ( $p < 0.05$ ) in Table 2. We find that in both culture groups, word count correlate with user engagement positively. In other words, the more the user speaks, the more engaged the user is in both cultural contexts. We also find that in both cultures, the louder users speak, the more varied their loudness are, the more engaged they are.

We find that how frequently one smiles differs greatly between the two culture groups in terms of correlation with engagement. In American culture, more smiles indicates more engagement, while in Chinese culture, similes are less strongly correlated with engagement. The trend is similar for automatically predicted smiles. One possible explanation of the difference is that Chinese culture is a typical collective culture that seeks harmony between partners. Chinese may subconsciously treat the agent as one of their partner and try to harmonize the agent by using more positive affect. While American culture is a individualism culture according to Hofstede’s culture theory [4]. Americans would change less of their behaviors to harmonize their partners compared to Chinese. We conducted a qualitative analysis of the semantic contents of user responses and found 4 out of 9 Asian female and 2 out of 12 Asian male participants told the agent: “I like

**Table 2.** Engagement Behavior Correlation. The bold number indicates the correlation is statistically significant ( $p < 0.05$ ).

Features	American rho(p)	Chinese rho(p)
Word count (ASR)	<b>0.20(0.00)</b>	<b>0.15(0.02)</b>
Intensity mean	<b>0.15(0.00)</b>	0.12(0.08)
Intensity variance	<b>0.17(0.00)</b>	<b>0.16(0.02)</b>
Smile	<b>0.13(0.03)</b>	0.07(0.32)
System interruption	<b>0.19(0.00)</b>	-0.09(0.17)
User interruption	0.10(0.06)	0.01(0.85)
System response time	-0.01(0.82)	<b>-0.20(0.00)</b>

you, you are so cute.”, while none of the Americans expressed their likability towards the agent directly.

Another difference we find between the two cultures is that the frequency of system interruption correlates with higher engagement in American culture but not in Chinese culture. One explanation of such difference is that Americans are mostly individualists who tolerate interruptions much more than Chinese who seek for harmonized dynamics in conversations all the time according to the Hofstede culture model. On the other hand we find that user interruptions is not correlated with user’s engagement in both cultures. We find that in American culture, user response time is negatively correlated with user engagement, which indicates that the faster the user responds, the more engaged the user is, while in Chinese culture, the correlation is less significant. We find that in Chinese culture, system response time correlates with user engagement negatively, which indicates that the longer the system pauses, the less the user engages. While this phenomena is much less significant in American culture than Chinese culture. One possible explanation is that Chinese users care more about their interlocutors than American users. As Chinese is a more collective culture, a long pause from their partners makes Chinese participants less engaged.

## 4 Conclusion

This paper presents two versions of a virtual agent system that designed for American and Chinese cultures. We find that users from different cultures express engagement differently when they are interacting with virtual agents. This suggests that we should take culture into consideration when designing engagement sensitive virtual agents.

## References

1. Andrist, S., Ziadee, M., Boukaram, H., Mutlu, B., Sakr, M.: Effects of culture on the credibility of robot speech: a comparison between english and arabic. In: Proceedings of the HRI, pp. 157–164. ACM (2015)

2. Black, A.W., Lenzo, K.A.: Flite: a small fast run-time synthesis engine. In: 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (2001)
3. Gedawy, H., Ziadee, M., Sakr, M.: Variations in giving directions across arabic and english native speakers. In: Qatar Foundation Annual Research Forum (2012)
4. Hofstede, G.H., Hofstede, G.: *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Sage, Thousand Oaks (2001)
5. Khaled, R., Biddle, R., Noble, J., Barr, P., Fischer, R.: Persuasive interaction for collectivist cultures. In: Proceedings of the 7th Australasian User Interface Conference, vol. 50, pp. 73–80. Australian Computer Society Inc (2006)
6. Matsumoto, D.: Culture and nonverbal behavior. In: *Handbook of nonverbal communication*, pp. 219–235 (2006)
7. Rehm, M., André, E., Bee, N., Endrass, B., Wissner, M., Nakano, Y., Akhter Lipi, A., Nishida, T., Huang, H.-H.: Creating standardized video recordings of multimodal interactions across cultures. In: Kipp, M., Martin, J.-C., Paggio, P., Heylen, D. (eds.) *MMCorp 2008. LNCS (LNAI)*, vol. 5509, pp. 138–159. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-04793-0\\_9](https://doi.org/10.1007/978-3-642-04793-0_9)
8. Yu, Z., Gerritsen, D., Ogan, A., Black, A.W., Cassell, J.: Automatic prediction of friendship via multi-model dyadic features, pp. 51–60 (2013)
9. Yu, Z., Papangelis, A., Rudnicky, A.: TickTock: a non-goal-oriented multimodal dialog system with engagement awareness. In: *Proceedings of the AAAI Spring Symposium* (2015)