Understanding User Perceptions of Robot’s Delay, Voice Quality-Speed Trade-off and GUI during Conversation

Figure 1: The robot “Xiaowei” used in our study.

Zhenhui Peng
The Hong Kong University of Science and Technology
Hong Kong, China
zpengab@connect.ust.hk

Kaixiang Mo
AI Group, WeBank
Shenzhen, China
nashmo@webank.com

Xiaogang Zhu
AI Group, WeBank
Shenzhen, China
xiaogangzhu@webank.com

Junlin Chen
AI Group, WeBank
Shenzhen, China
journeychen@webank.com

Zhijun Chen
AI Group, WeBank
Shenzhen, China
zhijunchen@webank.com

Qian Xu
AI Group, WeBank
Shenzhen, China
qianxu@webank.com

Xiaojuan Ma
The Hong Kong University of Science and Technology
Hong Kong, China
mxj@cse.ust.hk

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
DOI: https://doi.org/10.1145/3334480.3382792

Abstract
Conversational robots face the practical challenge of providing timely responses to ensure smooth interactions with users. Thus, those who design and implement robots will need to understand how different levels of delay in response may affect users’ satisfaction with the conversation, how to balance the trade-off between a robot’s quality of voice and response time, and how to design strategies to mitigate possible negative effects of a long delay. Via an online video-prototype study on a service robot with 94 Chinese participants, we find that users could tolerate up to 4s delay but their satisfaction drops at the 8s delay during both information-retrieval conversations and chitchats. We gain an in-depth understanding of users’ preference for the trade-off between the voice quality and the response speed, as well as their opinions on possible robot graphic user interface (GUI) design to alleviate negative user experience with response latency.

Author Keywords
Human-Robot Interaction; delay; voice; speech synthesis; graphic user interface

Introduction
Service robots that provide voice-based services and play the role of shop assistants [8], snack deliverers [7], and companions [3], to name a few, are increasingly popular in

Figure 2: The robot Xiaowei displays a text chat window during conversation. The left bubbles are robot's words, while the right one is user's utterance.

Figure 3: The computational models of robot Xiaowei's dialogue system and corresponding graphic user interface (GUI) design.

our daily life. To maintain a natural conversational experience, robots should respond to a user's request within an appropriate time range that meets the user's expectation [11, 13]. For example, Shiwa et al. found that when ordering a robot to carry a trash can, users had the highest level of satisfaction when the robot responded with a delay of one second compared to zero, two, and three seconds [13]. However, little work tries to understand users' perception of the robot's response delay during information-retrieval (IR) conversations or chitchats – two common tasks in service industries, e.g., robot receptionists, shop assistants [10], etc. Since user expectation of the delay could vary by types of conversational contexts [11], the conclusions in [13] may not be readily applicable to IR and chitchat tasks. Such an understanding is thus practically important as it can help robot teams to better allocate computational resources of the robot and design proper mitigating strategies in case of inevitable long conversational delay.

When a robot has a long response delay, enhancing the speed of computational models of its dialogue system is a direct way to improve user experience. However, the robot teams may encounter some trade-offs in implementing the dialogue system under a limited budget. For example, deploying an offline speech synthesis (SS) model to avoid network transmission latency could make the robot respond faster than evoking an online cloud-based SS service. Yet, the quality of offline synthetic voice may not be as good as the online one due to the limited computational resources offline. It would be interesting to know user preferences for the voice quality-speed trade-off, which can guide the engineers to decide which SS model to use.

In situations where long delays in a robot's service are unavoidable, previous research on Human-Robot Interaction has looked into the design of voice strategy (e.g., "uh..." [13]) and human-like gaze aversion [1] to moderate negative impressions toward the robot. Nevertheless, with more and more robots embedded with a screen [5], the design space of robot graphic user interface (GUI) in maintaining user experience during response latency needs further exploration. Shi et al. showed that a proper GUI design (e.g., text body movement vs. voice waveforms) can evoke stronger user engagement with the voice agent in smartphones [12]. It is thus useful to collect user opinions of how a robot's GUI design may reduce the impairment of a long conversational delay.

In this preliminary case study, we explore the following research questions: **RQ1)** How would a robot's delay at different levels affect user's satisfaction with the IR conversations and chitchats? **RQ2)** Do users prefer a robot that responds more slowly but with a higher-quality voice or a robot in the reversed condition? **RQ3)** What are users' opinions on the robot's GUI design in case of a long delay? We carry out an online video-prototype study with 94 Chinese participants on a robot "Xiaowei" which embeds a screen and provides services in a bank (Figure 1). Our results show that users' satisfaction with both the IR conversation and chitchat maintains up to 4s delay but drops at 8s delay. In general, they prefer to chat with a robot that has a high-quality voice, even though its response has a longer latency than that with a relatively low-quality voice. The common GUI design recommended by our participants is to add indicators (e.g., a "thinking" emoji) of delay on the screen.

Our contributions to HCI communities are three-fold. First, we identify the range of acceptable robot's delay in two common conversational contexts. Second, we provide insights into choosing a robot's speech synthesis model under the voice quality-speed trade-off. Third, we offer design considerations for robot GUI to handle a long delay.

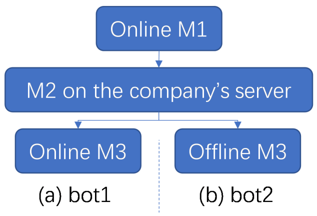

Figure 4: The two methods of implementing a speech synthesis model (M3) in the robot: (a) bot1 invokes M3 online, which has a higher-quality voice but is slower; (b) bot2 invokes an offline M3, which is faster but has a lower-quality voice.

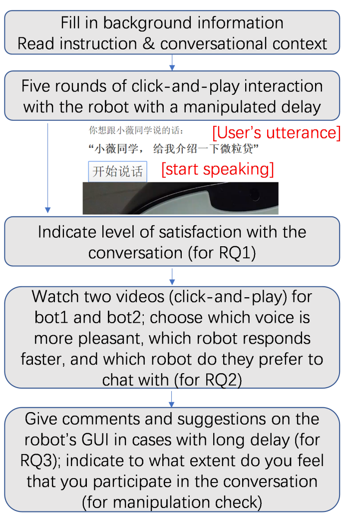

Figure 5: The overview and the procedure of the online video-prototype study.

Methods

The Robot "Xiaowei", its Dialogue System and GUI

The robot "Xiaowei" is designed to offer multiple functions in a company including banking-related information retrieval and social chitchats. It is 124 cm tall and has a 13.3 inch, 1920 $\times$ 1080 screen that displays a text chat window during conversations (Figure 2). As shown in Figure 3, the robot's dialogue system consists of three parts and its GUI is designed to reveal their status [12]. After being woken up by the command "Xiaowei Tongxue", the robot will say "I am here" and the icon on the screen will indicate that "I am listening" while the user is speaking. With the input speech, the speech recognition model (M1) gets back the text utterance, which is displayed on the screen once the process is finished. After being processed by the natural language understanding model and dialogue manager (M2), a textual response shows up on the screen. Next, the speech synthesis model (M3) generates the phonic response and the robot speaks it out with the "mouth" of the icon. The current version of Xiaowei invokes online services for M1 and M3 and implements M2 on the company's server (Figure 4a, denoted as bot1). However, the robot team of Xiaowei found that it would suffer from a long delay (e.g., the worst case is 7 **s**econds in our study) if the textual response is too long and the network is unstable. This makes it necessary for us to examine the RQ1 about user satisfaction. A faster technical solution is replacing the online M3 by an offline one (Figure 4b, denoted as bot2), which can reduce the delay to 4s in the same worst case. Although bot2 can technically respond faster, its voice is different from the voice of bot1 and has lower quality due to limited computational resources offline [4]. It motivates us to explore the RQ2 about user preference for the voice quality-speed trade-off. To gain insights into the GUI design during response latency, we further investigate RQ3 about user's opinions and suggestions on the GUI.

Overview of the Online Video-Prototype Study

We conducted an online video-prototype study to address the RQs raised in the Introduction (Figure 5). As a start of the experiment, participants filled in a consent form and their background information, read instructions and conversational contexts in the study website. They then went through three parts of the study. The **first part** is a 4 (level of delay: 1, 2, 4, 8s) $\times$ 2 (context: IR, chitchat) between-subject design (for RQ1). The participants randomly interacted with the robot (note: bot1) for five rounds in one of the eight conditions. To simulate an immersed conversational experience with the robot, we present the HRI scenarios in the videos rather than the figures and scripts as in [6]. Specifically, the participant clicks a button to "speak out" each listed utterance. The recorded video plays immediately, in which the "user" is chatting with the robot. After indicating their levels of agreement on "I am satisfied with the conversation with Xiaowei" (1 - strongly disagree, 5 - strongly agree), the participants proceeded to the **second part**, a within-subject (bot1 vs. bot2) design (for RQ2). They watched two videos with the same content for bot1 and bot2 following the context in their first part. They were asked to choose whose voice is more pleasant, which robot responds faster and which robot do they prefer to chat with (options: bot1, bot2, hard to tell). We counterbalanced the order of videos for bot1 and bot2 to check the order effect in this part. In the **third part**, we asked them to comment on the current GUI and suggest the GUI design in cases with a long delay (for RQ3). They also indicated how much they agree with "I feel that I am participating in the conversation with the robot" (for manipulation check).

Conversational Contexts and Video Prototypes

We test the information-retrieval (IR) and chitchat contexts [1, 11]. We prepare five utterances for the user in each scenario as does in [1]. In the IR scenario, the user encounters

Context	IR	Chitchat
bot1	5.8s (1.2)	3.8s (0.2)
bot2	3.4s (0.2)	3.2s (0.1)

Table 1: The average (SD) delay of responses in the three-round conversations with bot1 and bot2 in the contexts of information retrieval (IR) conversation and chitchat. Unit: **s**econd.

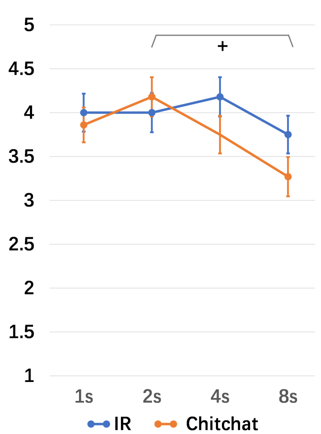

Figure 6: The ratings of "I am satisfied with the conversation with Xiaowei" (1 - strongly disagree, 5 - strongly agree) at different levels of delay during information-retrieval (IR) conversation and chitchat. $+ : 0.05 < p < 0.1$.

a robot in a bank and wants to learn the bank's products. The sample utterances are "what are the products in the bank" and "introduce [product A] to me". In the chitchat context, the user is waiting in line in a bank and wants to have a small talk with the robot. Sample utterances include "I am unhappy today" and "Your words are funny". At the beginning of each video-prototype scenario, an actor approaches the robot (note: bot1) that displays an instruction to wake it up for each round of conversation, i.e., call me "Xiaowei Student".The actor then starts chatting with the robot following the order of utterance in the list. We prepare both the male and female actor's voices to match the gender of the participant. We use the DJI Osmo Pocket (an anti-shake camera) to record the videos via a projective viewpoint of the actor, as it can avoid the bias of actor's appearance and has been shown to have considerable external validity [9]. We cut the videos into five rounds so that the participant can chat with the robot in a round-by-round manner online.

Manipulation of Robot's Delay
Through multiple conversational tests on the robot (note: bot1), we found that the longest delay is around 7 **s**econds. Considering that improving the models of the dialogue system can reduce the delay, we choose to test four levels of delay (1, 2, 4, 8s) in our between-subject study for RQ1. The length of a delay is defined as the period right after the user's speech and right before the robot's phonic response. We use Adobe Premiere Pro CC 2019 to adjust the length of the delay. We equally allocate the manipulated delay to M1, M2, and M3, i.e., 1/3, 2/3, 4/3, 8/3s for each model.

Manipulation of the Speed-Voice Tradeoff
We record another three rounds of conversations with bot1 and bot2 in both IR and chitchat contexts. In each context, the videos are the same for bot1 and bot2 except the voice and the delay of response. Compared to bot1, bot2 responds technically faster but its voice quality is lower. We do not cut the videos round by round nor edit the delay of responses. Table 1 shows the average delay in our within-subject study for RQ2. The robot's responses in the IR context are longer than those in chitchat and take more time through the online speech synthesis service.

Participant
We recruited 94 Chinese participants (52 males, 42 females) through posting on social media (e.g., wechat) and word-of-mouth. Their age ranges from 17 to 40 ($Mean = 24.1, SD = 3.7$). There are at least 11 and at most 14 participants in each of the eight conditions. 68 of them are university students, and others work as designers, IT engineers, teachers, etc. Most of the participants sometimes have voice interaction with physical robots ($M = 3.3, SD = 0.9$; 1 = "never", 5 = "very frequently"), but they use the voice agent in their smartphones frequently ($M = 4.1, SD = 0.9$).

Analysis and Results

Manipulation check. Overall, participants are positive that they were participating in the conversation with the robot ($M = 3.5, SD = 1.0$). Via data analysis, we address the three research questions regarding the conversational delay of service robots in the rest of this section.

RQ1: How would a robot's delay at different levels affect user's satisfaction with the IR conversations and chitchats?
After confirming the assumption of equal variance by the Levene's Test, we ran a two-way (levels of delay and contexts) ANOVA on the ratings of user's satisfaction with the conversation (Figure 6). There is a marginally significant difference in user satisfaction with the conversation at different levels of delay; $F(3, 86) = 2.6, p = 0.06, \eta^2 = 0.08$. The Bonferroni post-hoc test reveals that only the user sat-

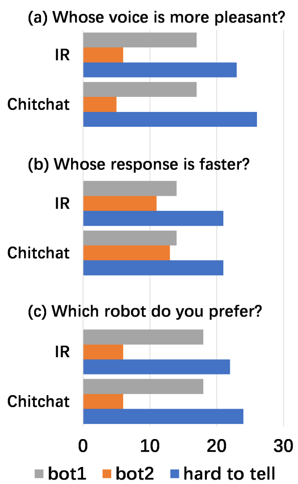

Figure 7: User preference for bot1 and bot2 in two contexts.

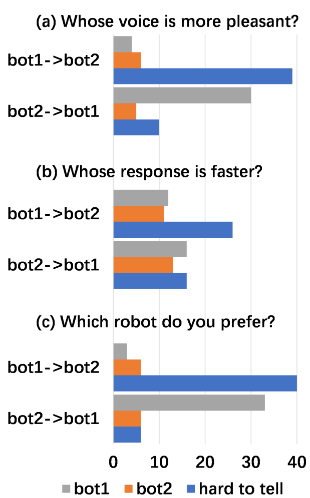

Figure 8: User preference for bot1 and bot2 when they interact with the robots in different orders.

isfaction at 2s delay ($M = 4.1, SD = 0.6$) is marginally significantly higher than that at 8s delay ($M = 3.5, SD = 1.0, p = 0.07$). Interestingly, user satisfaction with the IR conversation is the highest at 4s delay while that with chitchat is the highest at 2s delay. This may reflect user's different expectations for these two contexts [1, 11]. No significance is found for both the main effect of the conversational context ($p = 0.16$) and its interaction effect with the levels of delay ($p = 0.43$). However, there is a trend that user satisfaction drops more quickly at 4s ($M = 3.8, SD = 0.8$) and 8s ($M = 3.3, SD = 0.9$) delays in chitchat than those (4s: $M = 4.2, SD = 0.4$; 8s: $M = 3.8, SD = 1.1$) in IR conversation. Overall, these results suggest that user satisfaction with the conversations could maintain up to 4s delay but drop at 8s delay. Moreover, users in the chitchat conditions were slightly less satisfied with 4s and 8s delays than those in the IR conditions.

RQ2: Do users prefer bot1 that responds more slowly but with a higher-quality voice or bot2 in the reversed condition?
We counted how many participants chose "bot1", "bot2" or "hard to tell" in each question of the within-subject study. Figure 7 shows the distributions of user preferences in two contexts. Participants in both contexts generally agree that the voice of bot1 is more pleasant ($N = 36 = 18$(IR) $+18$ (Chitchat)) or at least not worse (i.e., hard to tell; $N = 46 = 22 + 24$) than bot2's voice. Interestingly, in both contexts, more participants perceived that bot1 responds faster ($N = 28 = 14 + 14$) or at least not slower ($N = 42 = 21 + 21$) than bot2, while bot2 is actually faster (Table 1). In general, more participants prefer to chat with bot1 ($N = 34 = 17 + 17$) rather than bot2 ($N = 11 = 6 + 5$). These results suggest that the quality of a robot's voice could be more important to user experience than the speed of its response in our case of 3-6s delay. The voice of higher quality may even let the users feel that the robot responds faster.

We further checked if the order of interacting with bot1 and bot2 affects user's preferences (Figure 8). The result of the perceived speed is similar in both groups (bot1 $\rightarrow$ bot2, bot2 $\rightarrow$ bot1). However, the order of interaction may impact users' preferences for voice and robot. When the users chat with bot1 first, most of them feel that it is hard to tell whose voice is more pleasant ($N = 40/49$) and which robot they prefer ($N = 39/49$). However, when they interact with bot2 first, most of them choose that bot1's voice is more pleasant ($N = 33/45$) and they prefer to chat with bot1 ($N = 30/45$). These results indicate that improving the quality of voice could enhance the experience of old users. If users have experienced the high-quality voice, reducing the delay with sacrifice in voice quality could not improve user experience.

RQ3: What are users' opinions on the robot's GUI design in case of a long delay?
To gain insights into GUI design, we first summarized the users' comments on the current GUI design of robot Xiaowei (Table 2). 14 participants indicated that displaying textual utterances can help them check the input and understand the robot's response. Also, they felt that Xiaowei's GUI is easy to follow ($N = 20$). *"It is simple, just like the text chat window on our smartphones".* However, seven participants commented that the GUI lacks proactivity. *"It is not human-like enough. It would be nice to proactively display 'what else can I help you'".* The other drawbacks are the platitude background ($N = 35$) and small font size ($N = 6$). *"The background is not appealing and the text is hard to read".*

We then categorized user suggestions on possible GUI design in case of a long delay (Table 3). Most of the participants mentioned that they would expect some indicators for the delay, including text ($N = 16$) such as "just a moment",

Pros
Help to verify the input and understand the response (14); Simple and easy to follow (20)
Cons
Lack of proactivity (7); Platitude/Cold background (35); Small font size (6)

Table 2: Pros and cons of the current Xiaowei's GUI design. (N): N participants have similar opinions.

Suggestions on GUI design in cases with long delay
Indicators: text (e.g., "just a moment") (16), progress bar (7), "thinking" emoji (15); **Funny animations**, e.g., special video effects, to distract user's attention (16); **Touchable options** on the screen, e.g., "skip" button (8)

Table 3: Suggestions on Xiaowei's GUI design to handle cases with a long delay. (N): N participants have similar suggestions.

progress bar (7), and emoji (15) showing that "I am thinking". 16 participants said that the screen can display some funny animations to distract user attention. Another eight participants suggested that the robot can offer touchable interaction during the response delay. *"I would like some touchable buttons such as 'skip' and 'ask again'".*

Discussion

Set a Context-Based Threshold for Robot's Delay
Our work extends the work of Shiwa et. al [13] by exploring how different levels of robot's delay affect user's satisfaction with information-retrieval (IR) conversations and chitchats. We found that it would be safe to have a delay up to 4s in both contexts (Figure 6). Specifically, users have the highest level of satisfaction at 4s delay during IR conversation but at 2s delay during chitchats. These could be accounted for users' different expectations towards these two contexts [11]. It indicates that robot designers could set a threshold of delay based on the conversational context, e.g., 4s for IR and 2s for chitchats. When the delay exceeds the threshold, the designers should consider improving the speed of computational models or exploiting mitigating strategies.

Be Cautious about Reducing Delay with Change of Voice
The quality of a robot's voice has been shown to affect user behaviors and perceptions. For example, Walters et. al found that the robot with a synthesized voice induces significantly further approach distances of users than the robot with a high-quality male or female voice [14]. Our study supplements that even with a drawback of low response speed, users seem to still prefer the robot with a high-quality voice (Figure 7). It is suggested that robot designers should be cautious about the change of voice quality when they consider improving the speed of the speech synthesis model. If the voice becomes less pleasant, it would not improve user experience even the response is faster (Figure 8).

Add Expressive Indicators of Delay on Robot's Graphic UI
To handle the cases of a long delay, the filler "uh..." and human-like gaze aversion proposed in previous works [1, 13] can be viewed as indicators of the robot's delay. As recommended by our participants, the indicators of delay can also be displayed on the robot's screen (Table 3). Considering that the current design of Xiaowei's GUI is easy to follow but somehow platitude (Table 2), we suggest that the added indicators should be expressive. For example, the robot can display a facial expression of thinking [5] or a robot icon with a "reflection" animation, which could maintain user expectation better than plain text.

Limitation and Future Work
Our work has two main limitations. First, we carry out the experiment on the robot with a screen – one type of common robot design [5], and we adopt a GUI design similar to the chatbots in smart devices. Robot designers need to take the appearance and GUI of their robots into considerations when they refer to our results. Also, the conversation is conducted in Chinese, which may affect people's sense of time and delay [2]. It would be interesting to compare our results to that from other countries.

Conclusion

This paper provides preliminary results on user perceptions of the robot's delay during conversations. Via an online study with 94 Chinese participants, the results show that user satisfaction with the conversation maintains up to a 4s delay in both information-retrieval conversation and chitchat. We further present user preference for the tradeoff between robot's voice quality and response speed, and we summarize the expected GUI design for cases of a long delay. Our work has practical insights for robot designers.

REFERENCES

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI '14)*. ACM, New York, NY, USA, 25–32. DOI: http://dx.doi.org/10.1145/2559636.2559666

[2] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. Association for Computing Machinery, New York, NY, USA, 157–164. DOI:http://dx.doi.org/10.1145/2696454.2696464

[3] K Dautenhahn, M Walters, S Woods, K L Koay, C L Nehaniv, hertsacuk E A Sisbot, and R Alami. 2006. How May I Serve You? A Robot Companion Approaching a Seated Person in a Helping Context. (2006). http://delivery.acm.org/10.1145/1130000/1121272/p172-dautenhahn.pdf?ip=175.159.124.16

[4] iFLYTEK. 2019. Offline Speech Synthesis - iFLYTEK Open Platform. (2019). https://www.xfyun.cn/services/offline_tts.

[5] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. 2018. Characterizing the Design Space of Rendered Robot Faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM, New York, NY, USA, 96–104. DOI: http://dx.doi.org/10.1145/3171221.3171286

[6] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully Mitigating Breakdowns in Robotic Services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction (HRI '10)*. IEEE Press, Piscataway, NJ, USA, 203–210. http://dl.acm.org/citation.cfm?id=1734454.1734544

[7] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. 2012. Ripple Effects of an Embedded Social Agent: A Field Study of a Social Robot in the Workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 695–704. DOI: http://dx.doi.org/10.1145/2207676.2207776

[8] FIONA PARKER FOR THE DAILY MAIL. 2018. Shop hires robot assistant... then fires it after just a week: Fabio the ShopBot irritates and confuses customers with vague replies and bad directions. (2018). http://www.dailymail.co.uk/news/article-5295837/Shop-hires-robot-assistant-fires-just-week.html.

[9] Richard Nisbett, Craig Caputo, Patricia Legant, and Jeanne Marecek. 1973. Behavior as Seen by the Actor and as Seen by the Observer. *Journal of Personality and Social Psychology* 27 (08 1973), 154–164. DOI: http://dx.doi.org/10.1037/h0034779

[10] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 98, 13 pages. DOI: http://dx.doi.org/10.1145/3290605.3300328

[11] Steven C. Seow. 2008. *Designing and Engineering Time: The Psychology of Time Perception in Software* (1 ed.). Addison-Wesley Professional.

[12] Yang Shi, Xin Yan, Xiaojuan Ma, Yongqi Lou, and Nan Cao. 2018. Designing Emotional Expressions of Conversational States for Voice Assistants: Modality and Engagement. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article LBW557, 6 pages. DOI: http://dx.doi.org/10.1145/3170427.3188560

[13] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2008. How Quickly Should Communication Robots Respond?. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI '08)*. ACM, New York, NY, USA, 153–160. DOI: http://dx.doi.org/10.1145/1349822.1349843

[14] Michael Walters, Dag Sverre Syrdal, Kheng Koay, Kerstin Dautenhahn, and Rene Boekhorst. 2008. Human approach distances to a mechanical-looking robot with different robot voice styles. *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN* (09 2008), 707 – 712. DOI: http://dx.doi.org/10.1109/ROMAN.2008.4600750