Testing Strategies For Bridging Time-To-Content In Spoken Dialogue Systems

Soledad López Gambino, Sina Zarrieß and David Schlangen

Abstract What should dialogue systems do while looking for information or planning their next utterance? We conducted a study in which participants listened to (constructed) conversations between a user and an information system. In one condition, the system remained silent while preparing a reply, whereas in the other, it "bought time" conversationally, using strategies from previously recorded human interactions. Participants perceived the second system as better at responding within an appropriate amount of time. Additionally, we varied between mid- and high-quality voices, and found that the high-quality voice time-buying system was also seen as more willing to help, better at understanding and more human-like than the silent system. We speculate that participants may have perceived this voice as a better match for the more human-like behavior of the second system.

1 Introduction

A common pattern in spoken human-machine interaction consists of a request for information by the human followed by presentation of this information by the system. Retrieval of this information may take time (e.g., for queries to remote databases). What should a system do while it prepares its reply?

A simple approach would be to remain silent until it can present information. However, this is not what humans do in such a situation. [9] show that people have a variety of resources available for "buying time", such as producing fillers (*uhm* or *uh*) [6], repeating parts of the interlocutor's request [5], explaining the reasons for the delay, etc.

How would users perceive an automatic system which produces such an array of resources instead of adopting a more traditional "please hold the line" type of approach? Would this system be viewed as a more human-like conversational part-

Soledad López Gambino
CITEC, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany e-mail: m.lopez_gambino@uni-bielefeld.de

ner? Or, to the contrary, would this behavior strike listeners as too unusual for an automatic system? To answer these questions, we conducted an overhearer study in which participants compared two (simulated) systems: The WAIT system asked users to wait, and then remained silent until it was able to present information, whereas the TIME-BUYING system produced behaviors similar to those observed in humans (see Fig. 1). Results showed that participants perceived the TIME-BUYING system as capable of finding a result within a more appropriate time period than the WAIT system, even though the actual time elapsed was the same for both conditions. Furthermore, as long as the system's voice was high quality, the TIME-BUYING system was also perceived as more willing to help, better at understanding and more human-like than the WAIT system. This, however, was not the case when the system used a mid-quality voice (see Section 3).

2 Method

DESIGN The main factor was WAIT vs. TIME-BUYING (see above). We conducted two runs of the study, with two different speech synthesizers, the first more easily identifiable as a machine and the second sounding more natural (see *MATERIALS* below). Participants listened to four recordings, two for each condition, in random order.

PARTICIPANTS Recruitment was carried out on the crowdsourcing platforms Amazon Mechanical Turk and Crowdflower and limited to workers in Germany. Forty-two subjects participated in the first run (16 female and 26 male, aged 20 to 69) and 39 in the second run (15 female and 24 male, aged 21 to 63). The study was published in the form of a questionnaire on the online platform SoSciSurvey.[1]

Fig. 1 Example dialogue for each of the two experiment conditions (original utterances in German in bold; English translation provided below in italics)

[1] URLs: `https://www.mturk.com/`, `https://www.crowdflower.com`, `https://www.soscisurvey.de/`

MATERIALS For the first run, the system's utterances were synthesized using MaryTTS, whereas Cereproc was used for the second run.[2] For MaryTTS we chose an HSMM voice, which resulted in (subjectively) less natural sound than the second one, a commercial professional voice. We used a male voice and the same utterances in both runs. The utterances were also the same for all participants. In order to produce them, we implemented a simple "time-buying generator" which produced a sequence of five time-buying utterances and then announced having found a flight. The system used the time-buying categories described in [9]. Some examples are **filler** (*uh, uhm*), **echoing** (A: *I need a flight to Bristol.* B: *Okay, a flight to Bristol...*) and **justification** (*The system is very slow today.*). At each step, the system chose one of these categories and produced one out of a set of canned utterances belonging to that category. The choice of category depended on: a) the previous system utterance and b) the time elapsed since the beginning of the time-buying stretch. Given these two parameters, the system selected a category by sampling from a probability distribution over all possible categories. The probabilities were trained on the DSG-Travel Corpus, a corpus of human interactions simulating a travel agency scenario [9]. The full recordings, as presented to the participants, consisted of a customer's request for a flight, followed by the system's time-buying utterances and final announcement of having found a result (as illustrated in Fig. 1).

PROCEDURE The participants first provided some demographic data, did a brief German language check, and read the task instructions. Participants then listened to recordings of enacted phone conversations between a human customer and an automatic system at a travel agency.[3] The human customer asked for a flight meeting certain criteria and the system pretended to look for an option which satisfied the customer's needs (see Fig. 1). After a while, the system announced having found an appropriate flight. The time between the end of the customer's request and the system's announcement was approximately 12 seconds. [4]. The behavior of the system during this period varied according to the experimental condition:

- **WAIT**: The system asks the customer to wait by producing an utterance such as *Bitte einen kleinen Moment Geduld* (Please be patient for a moment), and then remains silent until it announces having found the flight.
- **TIME-BUYING** The system produces a variety of utterances separated by short pauses, thus "buying time" until it has found a flight.

After each recording, participants rated the corresponding system on a 1-5 scale (5 meaning "strongly agree") with respect to five statements (here in translation):

1. The system understood the caller well.
2. The system took an appropriate amount of time to find a flight.

[2] `http://mary.dfki.de/`, `https://www.cereproc.com/`

[3] The customers' utterances were taken from the DSG-Travel corpus [9].

[4] We considered 12 seconds to be a realistic waiting period a relatively lengthy lookup might take, yet not so long that the WAIT strategy would obviously be disadvantaged

3. The system sounds as if willing to help.
4. The system acts the way I would expect a person to act.
5. If I had to buy a flight on the phone, I would use this system.

3 Results

We compared the ratings between the WAIT and the TIME-BUYING strategy. We test significance of differences through a paired-samples t-test and Wilcoxon signed-rank test, using Bonferroni adjusted alpha levels (.05/5 = **.01**, .01/5 = **.002**, .001/5 = **.0002**). In the first run (Mary-TTS voice is used), mean ratings for TIME-BUYING are higher than for WAIT, for all five statements. However, the difference only proved significant in the case of statement 2, "The system took an appropriate amount of time to find a flight" ($t(83) = 3.22$, $p<.002$; $W = 244.5$, $p<.002$).

Table 1 Mean ratings, standard deviations and medians for both conditions in statement 2, in the first run of the study

Condition	Mean	Std. Dev.	Median
WAIT	3.7	0.99	4
TIME-BUYING	4.07	0.94	4

In the second run (Cereproc Text-to-Speech is used), the TIME-BUYING strategy was rated better for each of the five statements, and differences were highly significant in all cases (see Table 2).

Table 2 Statistics for the statements (see Section 2); high-quality voice run

State-ment	*M* WAIT	*M* TB	*Mdn* WAIT	*Mdn* TB	t-test	Wilcoxon
1	3.91 (*SD*=0.85)	4.47 (*SD*=0.71)	4	5	$t(77)=6.11$, $p<.0002$	$W=111$, $p<.0002$
2	3.21 (*SD*=1.17)	4.38 (*SD*=0.77)	3	5	$t(77)=9.38$, $p<.0002$	$W=52$, $p<.0002$
3	3.33 (*SD*=1.02)	3.98 (*SD*=0.91)	4	4	$t(77)=5.67$, $p<.0002$	$W=163.5$, $p<.0002$
4	3.03 (*SD*=1.03)	3.7 (*SD*=1.09)	3	4	$t(77)=5.03$, $p<.0002$	$W=248$, $p<.0002$
5	2.85 (*SD*=1.04)	3.42 (*SD*=1.17)	3	4	$t(77)=5.5$, $p<.0002$	$W=132$, $p<.0002$

4 Discussion

The results presented above show that an information-providing dialogue system which can use speech to avoid long gaps after a user's request—similarly to what humans usually do—can make a better impression on overhearers than a system which asks the user to wait and then remains silent until it can provide an answer. In the first run of our study, participants found waiting times to be more appropriate in the TIME-BUYING system than in the WAIT one, even though the actual times remained constant across conditions. Additionally, the second run revealed that overhearers also perceived the TIME-BUYING system as more willing to help,

better understanding of the user's request, and more human-like than the WAIT system. Finally, participants preferred the former over the latter for their own use. These results suggest that dialogue systems could benefit from the incorporation of time-buying capabilities.

Additionally, the differences between the results of both study runs open up questions regarding the interplay of voice quality and time-buying strategy. One possible interpretation is that participants may have found the more human-like voice in the second run a better match for the more human-like behavior of the TIME-BUYING system. This could be connected to the idea of the *metaphors* involved in humans' perception of dialogue systems. Edlund et al. [7] draw a distinction between the *interface metaphor*, in which the system is perceived as a machine, and the *human metaphor*, in which the system is viewed as an interlocutor with whom speech is the natural interaction channel, and highlight the need for internal coherence between the metaphor selected and the behavior of the system.

From this perspective, one could argue that a system seeking to buy time like humans should use a voice as similar as possible to that of a human. However, deciding what kind of voice is best for a dialogue system is not always so straightforward, and other considerations also need to be taken into account. One of them is flexibility. Many commercial TTS systems sound relatively human-like but do not offer many options for acoustic modification (other than general emotion tags, etc). Systems like MaryTTS, on the other hand, offer both unit selection and HSMM voices, and the latter grant the possibility, for example, to adjust the frequency and duration of each phone to specific values [12]. It is therefore necessary to take this trade-off between human-likeness and flexibility into account, and prioritize depending on the aims and specificities of the dialogue system under construction.

5 Related work

Our results for both study runs are compatible with the idea that "filled time" is perceived as shorter than "unfilled time". This is, however, a somewhat contested assumption: Although there is research suggesting its validity [14, 8], it has also been postulated that what creates a perception of shorter waiting time is not the fact that the time is filled, but rather the nature of the information which is used to fill it. An example could be information about the waitee's place in the queue, which may convey a feeling of progressing towards the goal [10] or information about the estimated total duration of the wait [1].[5] This seems to be connected to a need for transparency regarding the state of the interaction. Such considerations are highly relevant when it comes to incorporating more conversational time-buying utterances in a system, since these utterances may also enable the system to provide justification for the wait and convey a sense of progress towards the desired goal. Finally, we highlight the importance of time-buying mechanisms within the area of

[5] In this study, information about duration of the wait did not make perceived waiting time shorter than actual waiting time, but it did reduce overestimation of its length in comparison to other experimental conditions.

incremental speech processing, since a number of studies have shown the benefits of systems with the ability to start producing some speech even before they have a full plan of the information to present [13, 11, 4, 2, 3].

6 Conclusion and future work

We have presented an overhearer study in which participants rated two information systems: one which asked the interlocutor to wait and remained silent while looking for the information to present, and another one which produced utterances during the wait. We found that participants perceive the time elapsed between the interlocutor's request and the system's response as longer in the first condition. Additionally, if the synthesized voice is relatively human-like, the system producing utterances is also perceived as more willing to help, better understanding of the user's request, and more human-like. In the future, we plan to incorporate time-buying capabilities into an actual dialogue system and explore the effects of different time-buying strategies in an interactive scenario, with regard to users' preferences as well as to more objective measures of task performance [15, 16, 3].

7 Acknowledgments

This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

1. Antonides, G., Verhoef, P., van Aalst, M.: Consumer perception and evaluation of waiting time: A field experiment. In: Journal of Consumer Psychology, vol. 12 (3), pp. 193–202 (2002)
2. Baumann, T., Schlangen, D.: Open-ended, extensible system utterances are preferred, even if they require filled pauses. In: Proceedings of Short Papers at SIGdial 2013 (2013)
3. Betz, S., Carlmeyer, B., Wagner, P., Wrede, B.: Interactive hesitation synthesis and its evaluation (2017). Preprint at `https://www.preprints.org/manuscript/201712.0058/v1`
4. Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., Schlangen, D.: Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 295–303 (2012)
5. Byron, D., Heeman, P.: Discourse marker use in task-oriented spoken dialog. In: Proceedings of Eurospeech 97 (1997)
6. Clark, H., Fox Tree, J.: Using uh and um in spontaneous speaking. In: Cognition, vol. 84 (1), pp. 73–111 (2002)
7. Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. In: Speech Communication, vol. 50, pp. 630–645 (2008)

8. Hirsch, I., Bilger, R., Heatherage, B.: The effect of auditory and visual background on apparent duration. In: American Journal of Psychology, vol. 69 (1950)
9. Lopez Gambino, S., Zarrieß, S., Schlangen, D.: Beyond on-hold messages: Conversational time-buying in task-oriented dialogue. In: Proceedings of SIGdial 2017 (2017)
10. Munichor, N., Rafaeli, A.: Numbers or apologies? Customer reactions to telephone waiting time fillers. In: Journal of Applied Psychology, vol. 92 (2), pp. 511–518 (2007)
11. Schlangen, D., Skantze, G.: A general, abstract model of incremental dialogue processing. In: Dialogue and Discourse, vol. 2 (1), pp. 83–111 (2011)
12. Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. In: International Journal of Speech Technology, vol. 6, pp. 365–377 (2003)
13. Skantze, G., Hjalmarsson, A.: Towards incremental speech generation in dialogue systems. In: Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '10, pp. 1–8. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
14. Tom, G., Burns, M., Zeng, Y.: Your life on hold: The effect of telephone waiting time on customer perception. In: Journal of Direct Marketing, vol. 11 (3), pp. 25–31 (1997)
15. Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with PARADISE. In: Natural Language Engineering, vol. 6 (3-4) (2000)
16. Whittaker, S., Walker, M.: Evaluating dialogue strategies in multimodal dialogue systems. In: D.L. Minker W. Bühler D. (ed.) Spoken Multimodal Human-Computer Dialogue in Mobile Environments. Text, Speech and Language Technology, vol. 28 (2005)