Choosing a dialogue system's modality in order to minimize user's workload

Adam Chýlek, Luboš Šmídl, and Jakub Nedvěd [*]

NTIS - New Technologies for Information Society
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{chylek,smidl,nedvedj}@ntis.zcu.cz

Abstract. The communication during human-machine interaction often happens only as a secondary task that distract the user's main focus on a primary task. In our study, the primary task was driving a vehicle and the secondary task was an interaction with a dialogue system on a tablet device using touch and speech. In this paper we present the design and the analysis of a study that can be used to create an optimal strategy for a dialogue manager that takes into consideration several metrics. These include the type of the information we require from the user, the expected cognitive load on the user, the expected duration of a user's response and the expected error rate.

Keywords: dialogue system, choice of modality, lane change test

1 Introduction

Multimodal dialogue systems start to play a role in cooperative robotics in industry and in interactive systems in our day-to-day lives. They also present a distraction from some tasks, such as checking your surrounding when walking, operating industrial machines or driving a car.

We will focus our attention on secondary tasks that require touch or speech as their input modality. Our motivational use case is filling an electronic journey log while driving (e.g. logging the arrival at a destination or the offloading of a cargo). The electronic logging happens via a device with a touchscreen or using an automatic speech recognition system.

The driver's main focus here should, of course, be on the driving, but we also want to make sure that the log is also filled in a timely manner. This leaves us with a problem of correctly choosing the types of an input that we want from the user and the correct modality that won't distract the user too much and that also won't cause too many problems with the actual dialogue (like error corrections, recognition timeouts, etc.).

[*] The final authenticated version is available online at https://doi.org/10.1007/978-3-319-99579-3_12. This work was supported by the European Regional Development Fund under the project Robotics for Industry 4.0 (reg.no.CZ.02.1.01/0.0/0.0/15_003/0000470).

Since driving is the primary task in our use case, we have used the ISO 26022 standard [4] for the assessment of the impact of secondary tasks on a driver of a motor vehicle. This standard provides a lane change task in a simulated environment, so we can safely test several workload-heavy tasks and later analyse the recorded data.

The results of the analysis will allow us to create a situation-aware dialogue system that uses the right modality for the given situation.

2 Related work

Lane change test (also often called lane change task) is commonly used to assess the effect of visual-manual interaction on the primary task of driving [1, 6, 8]. Similarly to us, the authors of [6] have evaluated several different styles of visual presentations on handheld devices, but a speech interface was not tested.

In [10] a spoken interaction was compared to a visual interaction using questionnaires. The spoken interaction was preferred by the subjects and their perceived cognitive load was lower in that case. We can further improve these findings by analysing in which situations would the visual interaction be beneficial and back these findings by changes in a performance on simulated tasks. Other researchers have focused on incremental dialogue processing [5] that allows the dialogue system to continually monitor the state of the environment and adjust the interactions with the user accordingly.

The point of this study is not to decide whether the primary task is influenced by either of the modalities, as it's been already shown in e.g. [3, 2] that both of the modalities do in fact have an impact on the driving. Our goal is to have the basis for a strategy that could minimize the impact on a cognitive load individually for different types of input information and different requirements from the dialogue manager (a duration of the response, an expected error rate).

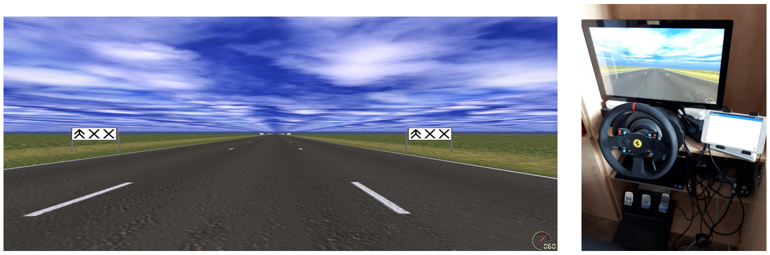
Related to our concept of a dialogue is also a multimodal system that requires fusion of both modalities (as opposed to us using only a single modality at a time). The analysis of modality choices with increasing load was done in [7]. The authors concluded that with increasing difficulty of the tasks users started to prefer more the multimodal interactions over the unimodal.

3 Experiment design

Our experiment was designed to resemble the motivational example - a simulation of a car and a simulation of a dialogue system on a touchscreen device.

3.1 Hardware and software setup

The hardware part of the setup consisted of a PC, 26" LCD display with speakers, a gaming steering wheel with pedals and an Android tablet for the dialogue



(a) Software (b) Hardware

Fig. 1: Setup of the experiment.

system (fig. 1b). On the tablet, there was an offline automatic speech recognition (ASR, based on [9]) system that processed the spoken input on the device itself and an offline text-to-speech (TTS, [11]) system. The tablet presented to the user a graphical user interface (GUI, figure 2) for the touch interactions.

The PC was running a simulation program called LCTSim[1] (fig. 1a) that had been set up according to the ISO standard (the lane change test). The position of the vehicle on the track as well as steering wheel angle and speed were recorded from the simulator at approximately 200 records per second.

The events from a subsystem that handled the secondary task were recorded separately and they were later merged with the simulation's log. The following types of events were used: a task was displayed, a user's answered, the answer was correct or incorrect, the task timed out.

3.2 Primary task

We have chosen a lane change test that conforms to the standard ISO 26022. This test consists of a 3 km three-lane straight road with equally spaced road signs. These road signs appear every 150 meters and indicate to which lane the participant should change. At most 18 lane changes were possible and the subject was expected to finish the scenario and the primary task before the track's end. The simulator limited the speed to 60 km/h and the participants were instructed not to slow down.

3.3 Secondary task

The secondary task was designed to represent an interaction with a dialogue system. It consisted of inputting several pieces of information one at a time using the available modalities. We have prepared following templates for the GUI to test common types of input elements (fig. 2): a short list that fits on a screen, a long list with a search field, a text input, a date input as a spinner, a time input as a spinner, a grid of images and a dialogue window with buttons.

[1] Downloadable from https://isotc.iso.org/livelink/livelink?func=ll&objId=11560806

Several tasks were created based on these templates. These tasks allowed filling the information using either this GUI or an ASR and will be listed in section 3.4.

For each task, the user would see the objective on the screen as well as hear the same text synthesized using TTS. In order to mimic real-world conditions, the users did not have any microphone nor headphones on them. The ASR used the built-in microphone from the tablet and the TTS used the tablet's speaker. Another speaker was connected to the PC and the simulation program emulated the sound of an engine. The entirety of the test (instructions, tasks, the TTS and the ASR) was in Czech.

3.4 Scenarios

The experiment was divided into 5 sessions. The participants first had to perform a training session. They drove on a track without any secondary tasks to get comfortable with the controls of the vehicle, with the appearance of the signs and with the primary task of changing lanes when instructed by the sign. They were instructed that changing the lanes as quickly and accurately as possible has the highest priority during the rest of the sessions.

The subjects could start the next session at their own discretion. The order of the lane changes during the second session was different from the previous one. This session was also without any secondary task. This way we could obtain a reference drive (we have recorded the participant's reactions to the signs without any workload from a secondary task).

The rest of the sessions (3 to 5) continued on the same track (with the same order of lane changes) but now with secondary tasks that had following restrictions:

During the third session, the participants were forced to use only the GUI to fulfil the objective. After the last task was completed the same track was loaded again from the start and a set of tasks for the fourth session started.

This time the participant had to complete the tasks using only speech. The ASR had a constrained language model in order to recognize only the options that were presented (e.g. colours for 1st task, numbers for 2nd, etc.). After completing this set of tasks the same track was loaded for the last time.

The choice of the modality for the last set of tasks was up to the users. To complete the task they could use the GUI or the ASR without any restrictions. This also meant that when the ASR failed to recognize their commands they could use the GUI to complete the task and vice versa.

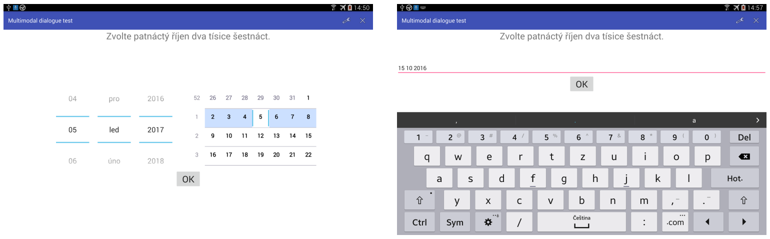
The participant had 20 seconds to perform the given task. Otherwise, the next task was shown. If an incorrect input was made the subject was notified and could try again until the task succeeded or timed out.

The tasks were shown always in the same order but with different values to be filled each time during the test (to mitigate habituation). Throughout the paper, we will refer to them using their order of appearance. The tasks' objectives were as follows:

1. choose a colour from a grid

2. input a number into a text field
3. choose from two buttons
4. input a date using a text field
5. choose a picture from thumbnails in a grid
6. input a time using a text field
7. choose from a short list of items
8. choose from three buttons
9. choose from a long list of items with an active search field
10. input a date using system's date input method (a spinner)
11. choose from a short list of items
12. input a time using system's native time input method

These tasks were designed not only to test all the basic input types on a smart device but also to test whether the amount of the information that is shown or that needs to be typed has any effect on the results. This is why a text field, an image grid, a list of items and buttons are included multiple times. Concretely, the 1st task was designed as an easier image selection version of task 5. The text input in the 2nd task is an easier version of the tasks 4 and 6. The task 3 is simpler than the 8th task. The list of items in the 7th and the 11th task contained fewer items than the task 9. Also, the native date and time input methods (10, 11) were supposed to be easier than typing into a text field (4, 6).



Zvolte patnáctý říjen dva tísice šestnáct.

Zvolte patnáctý říjen dva tísice šestnáct.

15 10 2016

(a) Native date input type (b) Textual input of a date

Fig. 2: Example of different input types used for secondary task.

3.5 Participants

There were 20 participants between 21 and 62 years of age (mean age 32.7, standard deviation of 9.7 years). All participants were native Czech speakers familiar with driving a car and using a touchscreen device.

4 Results

For the purpose of our analysis, we chose as a reference a drive through the track without secondary tasks. It is also possible to create a theoretical "ideal"

Table 1: The overall statistics for each type of scenario. Mean difference from a reference pass of each participant and mean duration of a scenario (from the start until all the tasks of the scenario have been finished).

modality	touch	voice	user's choice
mean difference [m]	1.05	0.76	0.73
mean duration [s]	132.3	127.98	123.3

drive based on the position of the signs and a fixed distance needed for a lane change. The results using these references differed only slightly and after manual comparison of the results of several sessions, we have concluded that the ideal reference corresponded with the reality less than the chosen reference drive.

In the following paragraphs, we will have to distinguish two types of positions on a track. We define the position between the lanes as an offset from the centre of the middle lane (shortly "offset") and position on a track length-wise as a "distance".

Several metrics will be evaluated to measure the impact of the secondary tasks on the performance during the primary task. These metrics can later be used by a dialogue manager to create a strategy based on the expected impact. A mean of differences between the offset of a reference drive and the drive with a secondary task (referred to only as a "mean difference") will be one of the metrics we assess. The duration of the task (until successfully finished or until it timed out) was chosen as another metric and finally, the error rate of the answers is the last metric.

We have included the overall results regarding mean duration and mean difference for each modality in table 1. We can see that if a simple strategy is needed we can leave the choice of the modality to the user, as it offers the best overall performance. But this would require the dialogue that uses this strategy to have similar composition as our scenarios. Because that would not often be the case, we will take a closer look at each individual type of a task in the following sections.

4.1 Comparing mean offset differences

Although the overall results can be interesting on their own, we wanted to analyse each kind of an input separately. We compared a mean difference for each given task across all the subjects. These results can be seen in the table 2. The task numbers refer to the order in which the tasks were shown to the user, as defined in section 3.4. A smaller difference is better.

From these results, we can see that using only the touch for the interaction resulted in a bad performance for tasks 3 to 12 (tasks 5 to 10 are significantly the worst with $p < 0.05$). This metric clearly does not favour using touch, with one exception - the 1st task. On one hand, using touch for the first task of selecting a colour was significantly better ($p = 0.1$) than using speech. On the other hand, choosing a more complex image from a grid (task 5) proved to be

Table 2: Mean offset from the reference drive (in meters) for each task based on modality. A standard deviation is in the brackets, best performing modality is in bold and $*$ marks significant difference from the next best performing modality ($p < 0.05$).

task	1	2	3	4	5	6	7	8	9	10	11	12
voice	0.51 (0.36)	0.92 (0.83)	0.76 (0.68)	**0.73** (0.50)	0.98 (0.97)	0.94 (0.95)	**0.82** (0.50)	**0.61*** (0.52)	**0.51** (0.34)	0.75 (0.51)	**0.89** (1.06)	**0.71** (0.63)
touch	**0.37** (0.30)	0.82 (0.57)	0.85 (0.65)	0.92 (0.45)	1.30 (1.17)	1.29 (1.00)	1.53 (1.15)	1.27 (1.54)	1.10 (0.82)	1.07 (0.90)	1.11 (1.33)	0.94 (0.42)
user's choice	0.40 (0.25)	**0.71** (0.39)	**0.73** (0.67)	0.90 (0.65)	**0.77*** (0.65)	**0.67*** (0.69)	0.89 (0.99)	0.78 (0.31)	0.61 (0.32)	**0.59** (0.29)	0.95 (0.95)	0.82 (0.49)

more demanding. For the possible human-machine interaction we could argue that the use case would be more often similar to the more complex fifth task than to the 1st one. From this, we can say that forcing the user to use a touch interface does not look like a viable strategy for any of the input types.

Leaving the choice of the modality up to the user proved to be marginally beneficial in 3 tasks and significantly better in 2 tasks. It also never was the worst performing setup. The types of input had a common theme - short or simple methods of input.

One might think that the user would willingly choose a modality that causes fewer problems during the primary task. But we can argue that some of the users must have chosen a modality that was not optimal - otherwise, the results for the user's choice of a modality would be similar to one of the forced modalities. This was clearly not the case since the spoken input was marginally better in 4 cases and even significantly better than the rest in 1 case (most of these were input types that would require a lot of typing or visual searching). We can conclude that the users can choose a modality that does not always result in the least amount of cognitive load.

Comparing performance based on the amount of information presented (as discussed in section 3.4) for the inputs of the same type we can see that presenting fewer information results in better performance. The same goes for typing, as inputs that required more typing increased the mean difference.

4.2 Comparing task duration

We will now focus on another important aspect of an input in the secondary task - its duration. The results for each run are in table 3 (shorter duration is better). Here we can see an interesting difference from the previous metric: using the touch interface is significantly faster in 5 cases, marginally in 1. These types of input where touch was faster had in common that they did not require many touch events (like typing or tapping a spinner).

If the choice of a modality is left up to the user, it is with the exception of task 11 better than the worst performing modality. Using speech is significantly

Table 3: Mean duration (in seconds) of each task based on modality. A standard deviation is in the brackets, best performing modality is in bold and $*$ marks significant difference from the next best performing modality ($p < 0.05$).

task	1	2	3	4	5	6	7	8	9	10	11	12
voice	7.2 (3.4)	9.8 (3.3)	8.3 (2.7)	12.2 (3.2)	8.9 (3.0)	**10.0*** (2.9)	7.6 (0.8)	8.45 (0.4)	**8.6** (0.8)	13.3 (7.4)	7.8 (0.3)	**13.0** (5.4)
touch	**3.0** (0.7)	8.4 (4.1)	**4.0*** (1.1)	16.0 (3.9)	**7.2*** (4.8)	13.3 (8.6)	**6.1*** (1.9)	**4.7*** (1.8)	10.6 (3.6)	13.8 (7.4)	**5.1*** (1.6)	18.5 (5.9)
user's choice	3.3 (1.1)	**6.4** (1.3)	7.3 (1.0)	**11.5** (2.0)	8.3 (1.2)	11.0 (2.6)	7.2 (0.9)	7.8 (1.0)	9.2 (2.5)	**12.7** (4.6)	8.6 (2.8)	14.9 (6.4)

faster only in 1 task (filling a date into a text field), marginally in 2 tasks. The worse performance can be partly due to the lag of an ASR system that has to process the input and partly because the participants occasionally had to repeat the input several times because of the errors the ASR makes, as we will show later.

We can again compare the tasks with elements of the same type that contain less information versus the ones with more information (e.g. the short list in task 7 versus long list in task 9). The tasks with less information are completed faster if using touch. Using speech these differences are less pronounced.

4.3 Comparing modality choices and error rates

During the last session, the user had a free will at choosing a modality. In this section, we will analyse which modality the subject preferred for which task. The detailed results are in table 4.

We can clearly see that using speech as the input method was preferred in most of the tasks, with the first task being the only exception. Interestingly this theoretically very simple task of choosing a colour resulted in the highest error rate in both modalities. In contrast to this, the similar 5th task (choosing an image) had the lowest error rate. Reasons for this phenomenon could not be found. From the data, it is clear that touch input, although less prone to errors, is not preferred by the users and they are willing to try and repeat the input

Table 4: Which modalities did the subject choose and what error rate the modality caused.

task	1	2	3	4	5	6	7	8	9	10	11	12
voice input [%]	25	95	70	95	100	100	65	55	95	85	75	90
touch input[%]	75	5	30	0	0	0	35	45	0	15	20	5
input timed out [%]	0	0	0	5	0	0	0	0	5	0	5	5
voice error rate [%]	64.3	13.6	12.5	17.4	4.8	20.0	7.1	21.4	26.9	43.3	28.6	55.0
touch error rate [%]	16.7	0	0	0	0	0	0	0	0	0	0	0

several times using speech. This knowledge is important for a dialogue strategy where we expect the recovery from recognition errors to be difficult. Forcing the use of a touch interface instead of speech in these situations will result in lower error rates.

Table 5: Error rates when the user was forced to use one of the modalities.

task	1	2	3	4	5	6	7	8	9	10	11	12
voice error rate [%]	25.93	40.00	20.83	25.00	16.67	9.09	4.76	0	4.76	40.00	4.76	35.48
touch error rate [%]	0	9.52	4.76	14.29	0	16.67	9.09	0	0	17.39	0	10.53
voice timed out [%]	0	10	5	10	0	0	0	0	0	5	0	0
touch timed out [%]	0	5	0	70	0	25	0	0	10	5	0	15

Our last analysed metric was an error rate of the forced modalities. The detailed results are included in table 5. The results of the voice input were expected to contain errors in the conditions of the test. The 4th task (typing a date) involved a lot of interaction with a virtual keyboard and most of the users were unable to finish the task in time. From the perspective of a dialogue strategy, this data can provide a valuable insight into an expected error rate of a touch interface. Whenever the user is forced to use a keyboard we should expect increased error rates or longer response times. Choosing from a grid of images or buttons should be preferred.

5 Conclusion

The acquired data and the presented analysis allow us to create a strategy for a dialogue manager that either forces the user to use a certain modality or gives the user a free choice of the modality. Such strategy can be based on several factors that can be used to infer the expected impact on the primary task. For our purposes, this impact was measured as a mean offset from a reference drive without any secondary task, an error rate on the secondary task and as a time needed to accomplish a task. The factors that the manager may take into account are the types of the input (e.g. a choice from a list, a date), the amount of presented data (e.g. choice from two versus twenty images), the requirements on an expected error rate or a limit on the expected duration of the input. The strategy does not have to be based solely on the results of this study. For example, it can be further improved on the fly based on the interaction with the user. If a simple strategy is required, the best overall performance was achieved when the user had a choice of a modality.

In the upcoming future, a dialogue manager that uses the data from the experiment as the basis of its strategy will be created and evaluated. This will also allow us to analyse whether the knowledge acquired using driving as a primary task is transferable to other primary tasks (e.g. operating a robotic hand).

References

1. Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., Montanari, R.: Driver workload and eye blink duration. Transportation Research Part F: Traffic Psychology and Behaviour **14**(3), 199–208 (2011). https://doi.org/10.1016/j.trf.2010.12.001
2. Curin, J., Labský, M., Macek, T., Kleindienst, J., Young, H., Thyme-Gobbel, A., Quast, H., König, L.: Dictating and editing short texts while driving. Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '11 p. 13 (2011). https://doi.org/10.1145/2381416.2381418, http://dl.acm.org/citation.cfm?doid=2381416.2381418
3. He, J., Chaparro, A., Nguyen, B., Burge, R.J., Crandall, J., Chaparro, B., Ni, R., Cao, S.: Texting while driving: Is speech-based text entry less risky than handheld text entry? Accident Analysis and Prevention **72**, 287–295 (2014). https://doi.org/10.1016/j.aap.2014.07.014
4. Road vehicles – Ergonomic aspects of transport information and control systems – Simulated lane change test to assess in-vehicle secondary task demand. Standard, International Organization for Standardization, Geneva, CH (Sep 2010)
5. Kousidis, S., Kennington, C., Baumann, T., Buschmeier, H., Kopp, S., Schlangen, D.: A Multimodal In-Car Dialogue System That Tracks The Driver's Attention. Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14 pp. 26–33 (2014). https://doi.org/10.1145/2663204.2663244, http://dl.acm.org/citation.cfm?doid=2663204.2663244
6. Louveton, N., McCall, R., Koenig, V., Avanesov, T., Engel, T.: Driving while using a smartphone-based mobility application: Evaluating the impact of three multi-choice user interfaces on visual-manual distraction. Applied Ergonomics **54**, 196–204 (2016). https://doi.org/10.1016/j.apergo.2015.11.012, http://dx.doi.org/10.1016/j.apergo.2015.11.012
7. Oviatt, S., Coulston, R., Lunsford, R.: When do we interact multimodally? Cognitive load and multimodal communication patterns. International conference on multimodal interfaces pp. 129–136 (2004). https://doi.org/http://doi.acm.org/10.1145/1027933.1027957, http://dl.acm.org/citation.cfm?id=1027957
8. Pitts, M.J., Skrypchuk, L., Wellings, T., Attridge, A., Williams, M.A.: Evaluating user response to in-car haptic feedback touchscreens using the lane change test. Advances in Human-Computer Interaction **2012** (2012). https://doi.org/10.1155/2012/598739
9. Pražák, A., Psutka, J.V., Hoidekr, J., Kanis, J., Müller, L., Psutka, J.: Automatic online subtitling of the Czech parliament meetings. In: Text, Speech and Dialogue. pp. 501–508. Springer (2006)
10. Silvervarg, A., Lindvall, S., Andersson, J., Esberg, I., Jernberg, C., Frumerie, F., Jonsson, A.: Perceived usability and cognitive demand of secondary tasks in spoken versus visual-manual automotive interaction. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH pp. 1171–1175 (2016). https://doi.org/10.21437/Interspeech.2016-99
11. Tihelka, D., Stanislav, P.: ARTIC for Assistive Technologies: Transformation to Resource-Limited Hardware. WCECS 2011 **I**, 581–584 (2011)