Effective human computer interaction: an empirical study of the effects of a speech output modality of a goal-based dialog system on user satisfaction

Riccardo Bassani
6866840
r.bassani@students.uu.nl

Geanne Barkmeijer
6967280
a.g.barkmeijer@students.uu.nl

Samuel Meyer
5648122
s.j.meyer2@students.uu.nl

Wiebe de Vries
5713595
w.r.vries@students.uu.nl

Abstract

Is there a difference in user satisfaction when users interact with a goal-based dialog system containing a text-only output modality, compared to a goal-based dialog system containing a speech-based output modality? Current research shows that speech can influence users in various ways. In this paper we will examine if user satisfaction differs when the system's response is text-based or speech-based. We hypothesise that the user satisfaction of a text-only output modality differs from the user satisfaction of a speech-based output modality. To test our hypothesis, we conducted an experiment in which 20 participants were asked to use a recommendation system to find restaurants. After finishing each conversation with the dialog system, the participants filled in a survey. The results show no significant difference between user satisfaction of the text-only output modality and speech-based output modality. This suggests that a speech-based output modality does not affect user satisfaction. However, due to the limitations of the experimental design, it is suggested to further research the effects of a speech on user satisfaction, by simulating a realistic use case environment. This study aims to contribute to a growing body of research on the user satisfaction and output modalities of a goal-based dialog system.

Keywords

goal-based dialog system, text-to-speech, speech-based output modality, multi-tasking, divided attention, cognitive load, user satisfaction, cognitive-aware modality planning

1 Introduction

Nowadays, dialog systems are used extensively; hence more effort is expected to make human computer interaction as natural as possible [24]. A few examples of frequently used dialog systems are tourist information systems, navigation systems, weather or traffic information systems and recommendation systems. For a goal-based dialog systems to be successful, it is important to evaluate the satisfaction of its users. The modality through which a systems respond seem to be an important factor in the evaluation of the user satisfaction.

Along with the ability of a system to achieve specific goals, an important aspect of a dialog system resides in the communication modes it adopts. Spoken dialog systems, i.e. dialog systems are systems which integrate the transformation of a generated sentence into a spoken utterance which is understandable for human [18]. Systems which integrate natural language understanding are becoming more and more widespread.

The choice between a text-based system and a speech-based system does not have to be necessarily exclusive. On the contrary, multi-modal interfaces can offer more possibilities by enabling system output combining text and speech [19]. Adding speech as the output mode of a dialog system, could potentially have different effects on the user satisfaction. Research shows that users interact differently with systems which are mediated by speech, than with systems which are not mediated by speech[21]. This research showed that when a system is mediated by speech, user tend to attribute personality types and general knowledge to the system. Personality types and general knowledge are human qualities, which indicate that users are treating speech-based system as a person, even if they are aware that it is not a human being. This finding is also reflected by a research which shows that people instinctively act polite to systems which use voice as output modality [21]. Furthermore, the volume of the system's speech tend to influence the user's emotions and can have a positive effect on the user's memory [21]. These findings show that a speech-based output modality can influence the user's perception, emotions and memory, which could influence the user satisfaction.

A speech-based output modality could also have other effects. Research shows that when the system's response is preceded by a significant pause, the system is perceived to be

less trustworthy than when the system's response is quicker. Also, a bad quality of the system's voice correlate negatively to the user's perception of the quality of the information provided by the system [21]. These findings could influence the user satisfaction in a negatively way.

To examine if a speech-based output modality can be beneficial for a goal-based dialog system, not only the pros and cons discussed have to be taken into account, but also other important factors. For example, the task or goal of the dialog system and the social environment in which the system is used. These factors could influence the choice of communication modes [17]. It seems that a multi-modal dialog system can be a powerful solution for dealing with accessing information in dynamic environments, like cities or crowded locations [26]. For example, tourist need to access a large body of information about transportation timetables, programs of theatre shows or locations of restaurants according to their preferences, while they are moving through the city. This information is most valuable if it can be delivered in an efficient way. Multi-modal systems output could increase the efficiency of a dialog system. Therefore, the type of communication mode could influence the user satisfaction.

To summarise, current research shows that system's speech can have different effects human emotions, perception and memory. However, it is unclear in what way speech as output modality can influences the user satisfaction in case of a goal-based dialog system. Due to the growing demand in the spoken dialog systems, it is crucial to evaluate the user satisfaction on various modalities. To broaden the scientific knowledge on dialog system output modalities on user satisfaction, we will research the following question: 'Does a text-only and speech-based output modality of a goal-based dialog system lead to differences in user satisfaction?

In the next paragraph we will give a more detailed background on various conversational agents, the importance of user satisfaction as part of a evaluation framework and the potential benefit of integrating speech as an output modality for a goal-based dialog system. Then we will describe our study and the goal-based dialog system that we have developed in order to analyse how user satisfaction vary by comparing a text-based and a speech-based output modality. Finally, we will discuss the findings of this study, the implications for the theory, experiment design, limitations and suggestions for further research.

2 Theory

Background on various conversational agents

Conversation and communication are the core to the field of artificial intelligence. Conversational agents, like dialog systems, bring major components of artificial intelligence together: knowledge reasoning, natural language processing and learning. These conversation agents are often categorised into two classes: chatbots and goal-based dialog agents. In order to converse with a human both classes rely heavily on natural language understanding. Both classes are able to process the user input in various ways. The user input can be analysed while using a rule-based/frame-based approach, information retrieval-based approach or a neural network approach [11].

The fundamental difference between these two classes of conversational agents are the goals of the systems. Chatbots are created to maintain a conversation with a human for therapeutic purposes or entertainment and fun. Here, it is important for a chatbot to keep the user engaged. By doing so, the chatbot tries to mimic an informal human conversation. These chatbot-human conversations are characterised as long-winded and consists of more turns [11]. Dialog systems, on the other hand, are created to help the user to achieve a specific goal, like booking a flight or finding a restaurant. Here, it is important to keep the conversation as short as possible, since the purpose of the user is only to gain practical information [11]. In the past, chatbots and dialog systems were seen as two separate classes of conversation agents. Currently, more attention is paid to combining chatbots and goal-based dialog systems.

The importance of user satisfaction

The evaluation of the a conversational agent is crucial in order to know how to improve its performance. Estimating the performance is a central issue in developing an effective conversational agent. Therefore, coming up with the right evaluation method is an important research in itself. Not only are metrics such as task success, time efficiency and error rate important; also the user satisfaction is one very important metric for measuring the performance of a conversational agent in general [5]. The term 'User satisfaction' is defined as "the extent to which users believe the information system available to them meets their information requirements" [10], or "the perceived usefulness of information" [14].

User satisfaction as an evaluation metric is important, because a system not only need to act appropriate, adequate and correct; a system also needs to be accepted by the user. Currently, there are a lot of results available from research on conversational agents regarding user satisfaction. But since the large differences in the operationalisation of user satisfaction, these results are often not generalisable. However, we argue that the results of research on user satisfaction are still important. Although user satisfaction is often seen as only a by-product of the usability in the Human Computer Interaction literature [15], we argue that the user satisfaction can be seen as an indicator for the overall system performance. Our statement is build on the many findings in the literature that

show that systems with positive results on the objective evaluation metrics yield high subjective user satisfaction. This which implies a high overall system performance [9, 25].

When taking the user satisfaction as an important overall evaluation metric, we are able to analyse to which extent the objective evaluation metrics mirror the subjective user's needs. We believe that this approach will increase the progress in developing more sophisticated human computer interaction and in the understanding of the human mind.

Speech-based output modality for reducing cognitive load

For the development of a dialog system, it is important to understand why users are more satisfied of specific communication modalities in a given context. In an environment in which the user find it hard to pay attention to information offered via a specific modality, the user tends to go to another less error-prone input modality. This phenomenon is called 'contrastive functionality' [20]. Situations in which the user often experience difficulties with processing information and which are error-prone are typically dual tasks scenarios [22].

As mentioned earlier, goal-based dialog systems are typically used in dynamic urban environments, like cities or crowded locations. When technologies are used in dynamic urban environments, managing the user's attention will become one of the new challenges. When developing a technology guided by our knowledge and understanding of human cognition and attention, fatal situations can be avoided. For example, drivers could be distracted from driving when they are interacting visually and manually with their phones [6, 12]. In our case a tourist could navigate through the city while using the dialog system. One study shows that audio alerts can be a powerful method for getting the attention of a person who is performing secondary tasks, while driving in an autonomous car . It also indicates that processing audio stimulation doesn't interfere with performing other tasks, like navigating or driving.

These findings are in line with research on multi-tasking, or 'divided attention'. It appeared that people can do various tasks simultaneously, like walking and singing. But other kind of tasks, such as solve math exercises while maintaining a conversation, will be harder. This is because specific tasks compete for mental resources [22]. When tasks are different from each other, a person can perform multiple tasks simultaneously better [2].

Also different kind of tasks, such as talking and driving, can also compete for mental resources [13]. Intuitively that seems logic: when the traffic becomes complicated, a person will often pause or slow down a conversation [7, 8]. While specific tasks uses task-specific resources, general resources, such as the executive control, will be used for setting goals, priorities, create strategies and controls the order of cognitive processes [3]. The demand of the task-specific and general resources determines the cognitive load of performing various tasks simultaneously. It appeared that the combination of visual and audio modality lead to less cognitive load and better performance than other combinations of modalities [4]. Research show that cognitive load has a profound effect on the user satisfaction and performance when they are multi-tasking [23]. This suggest that when developing a dialog system, which is used especially in situations where the user's are multi-tasking, it is important to build a dialog system guided by cognitive-aware modality planning.

Built on previous research, we expect that speech as output modality can lead to a decrease in cognitive load. Since goal-based dialog systems are often used in dynamic urban environments, which requires divided attention, the decrease in cognitive load could contribute positively to user satisfaction. Therefore, we assume that a speech-based output modality will increase user satisfaction.

3 System description

The system we use for our experiment is able to recommend restaurants to users in the Cambridge area. The way it works is by interpreting user input, so called utterances, and responding to it. The dialog policy is implemented through a state transition function and the user intentions are interpreted associating a dialog act to each utterance. In order to do that, a machine learning classifier has been trained on a data set of dialogs from the second Dialog State Tracking Challenge (DSTC 2).

Through the dialog with the user the system firstly aims to acquire the uses preferences. The preferences the system looks for are food type, location and price range. The system starts the dialog welcoming the user and informing them about the preferences that can be expressed. If the user expresses a preference, the system registers it, otherwise it addresses the user with a question in which it randomly asks for the favourite type of food, area of the town or price range. Once the system has two of these preferences it will scan its known restaurants and recommend one of them that matches the user requests. If the user expresses indifference about food, area or price range, the system considers it in the count of the extracted preferences. When a recommendation has been made the user can respond with another request, for example for another restaurant with the same specifications. The user can also ask for more information about the recommended restaurant. Or the user can change its preference.

If we break this process down we get the following phases: Firstly the system greets the user and waits for them to tell the system them preferences. When the system has gathered enough preferences it will transition to the suggest

state. After the system has made a suggestion it will wait for the user to ask further information, to ask for a different suggestion or to change his preferences.

The dialog policy provides that the system answers modifying its response depending on the dialog act of the user utterance it is answering to. For instance, if during the dialog the user thanks the system, the response will start with "You are welcome". If the system fails in understanding the user intention, it will ask the user to repeat her utterance. When the user says goodbye to the system, the system replies and the dialog terminates. The system utterances are generated based on templates, either a simple string or a composition of strings. In order to mitigate the effect of user misspellings during the extraction of the preferences, the system relies on a JSON ontology containing correct terms for type of food, price ranges and town areas and on the calculation of the Levenshtein distance. The Levenshtein edit distance is defined as the number of edit operations necessary to convert a string into another string of the closest domain term. Using this tool, the system identifies words in the user utterances that could be typos or misspellings of known words. After having identified the possible mistakes, the system asks the user for confirmation by showing the edited string.

4 Method

Participants

Participants were 20 master students of Utrecht University, Utrecht, Netherlands, ranged in age from 20 to 29 years (M = 23.7, SD = 2.0 years). Subjects differed in gender (9 males, 11 females) and nationality.

The level of English knowledge was homogeneous among the participants (C1 Advanced) and all subjects were healthy and had normal hearing. Participants took part on voluntary basis.

Material

In order to determine how the user satisfaction varied between text-to-speech on and off, we asked participants to rate how much they agreed with three statements. The questions were based on the USE questionnaire conventions [16]. All questions were rated on a one to five-point Likert scale [1].

1. I am satisfied with this system.
2. I would recommend this system.
3. This system is fun to use.

Each participant received a paper containing the tasks to perform (see Appendix B).

Design

The experiment was run following a within-subject approach to prevent confounds due to the predisposition of different participants to be more or less generous in their rating.

There were two main conditions: In the first condition, participants had a conversation with the system using only text. In the second condition, participants had a conversation with the system using text, with the addition of text-to-speech. The randomness of the conversations and the possible related confounds were limited by giving each participant the same set of six tasks. Every task consisted of the participant finding a restaurant based on some given preference and obtaining related restaurant information. The tasks were designed to be feasible given the system.

task	description
1	Find a Spanish restaurant in the east of the town and get the phone number
2	Find a cheap restaurant in the south
3	Find an Italian restaurant in the south, then change your mind and ask for a Portuguese one
4	Find an expensive Chinese restaurant and ask for the location details
5	Find a moderate restaurant in the centre, refuse the suggestion and say bye
6	Find an Indian restaurant in the east of the town, ask for the address and the phone number

Table 1: Tasks description

Each task was performed in either one of the test conditions, with pre-determined sets of ordering. Four sets of ordering were assigned to four sets of participants, such that each subject interacted with the system 3 times with the text-to-speech and 3 times without. Moreover, to reduce the effect of the order in the results, counterbalancing has been applied. Condition orderings for 1 and 2 are inversions of each other, as are the orderings for 3 and 4. Task orderings have been established so that every task is performed exactly 10 times per condition.

Data collected during the survey was stored in a CSV file.

Procedure

Participants entered a quiet room and they and they sat in front of a computer screen displaying a starting prompt. They were informed about the nature of their task and how to

group	ordering
1	1T-2T-3S-5S-4T-6S
2	1S-2T-3S-4S-5T-6T
3	6S-5S-4T-3T-2S-1T
4	6T-5S-4T-2T-3S-1S

Table 2: Tasks ordering: the number represents the task; T: only test, S: text-to-speech

interact with the system. Namely they were said how to start a new dialog by pressing the Enter key and how to terminate a dialog by saying goodbye to the system. They were then given a sheet containing the tasks. The computer was located at a fixed distance from the seating position and the volume of the speaker was maintained on a predetermined level. The terminal was set to be always on top, so that the user could not accidentally leave the dialog. After each conversation the subjects were presented with a brief survey on their experience with the dialog agent. After the last question, a new prompt was offered to start the next dialog. The console was cleared after each dialog, after each question of the survey and after each instruction message. The utterances appeared as white text on a black background. The only difference between the two experimental condition was the presence of the sound reproducing the utterances. The whole process took around 10-15 minutes.

Measurements

The three questions in the survey measure a single concept, the general user satisfaction. For each condition, for each participant, an average of the answers to the three questions was calculated and the result constituted the dependent variable object of study.

A two-tailed paired t-test was performed to investigate significant differences in the results obtained in the "only text" condition and in the "text-to-speech" condition.

An alpha level of 0.05 for significance was used in the comparison.

5 Results

Figure 1 displays the rating distributions of the text condition and the text-to-speech condition. The left box plot shows the distribution of the ratings for the text condition. The median of the ratings is 4.25 on a scale from 1 till 5. The lower whisker and the lower half of the central box are longer than the upper ones. This indicates that the distribution of ratings of the text condition is skewed to the left.

The right box plot shows the distribution of the ratings for the text-to-speech condition. The median of the ratings is 3.75 on a scale from 1 till 5. The upper whisker and the upper halve of the central box are longer than the lower ones. This indicates that the distribution of the ratings of the text-to-speech condition is skewed to the right.

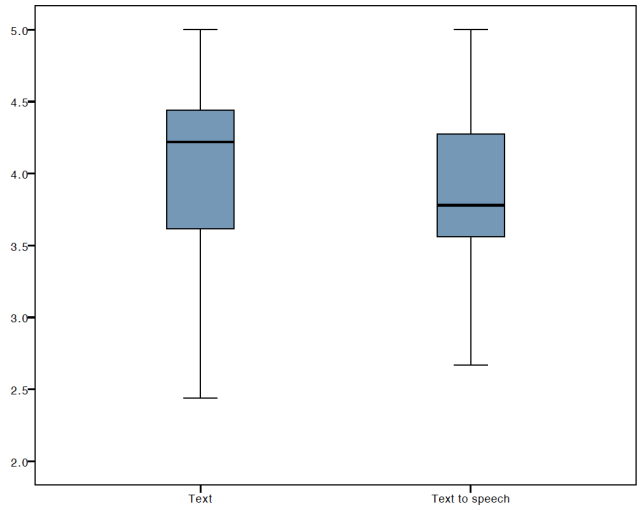

Figure 1: The rating distributions of the text condition and the text-to-speech condition

When testing if the two conditions differ significant from each other, we conducted a two-tailed paired t-test. The results of the two-tailed paired t-test show no significant differences between the two conditions ($t(19) = 0.979$, $p = .34$).

6 Discussion

Research question

In this research an experiment was performed in order to see if there is a difference in user engagement and satisfaction between two dialog models. One of these models uses only text to convey information, we call this model the text model. The other one uses both text and text-to-speech, we call this model the speech model. In order to test the user satisfaction in both models we conducted twenty experiments. The experiment consisted of six restaurant recommendation finding tasks. for three of these tasks the text with text-to-speech model was used. For the other three the model without text-to-speech was used. After each task the user evaluated the task by completing a short survey. The survey tested the user satisfaction by proposing three statements. For each statement the user rated how much they agreed with it on a one to five-point Likert scale. A two-tailed paired t-test on our results indicated that there is no significant difference between the two types of dialog systems.

Implications for theory

Contrary to our expectation, there appears to be no effect of the preferred output modality. This experiment does not provide evidence for a preference for speech output modality in a goal-based dialog system, since there is no significant difference in satisfaction between text and speech output modality. This means that there is no support for the theoretical argument that speech could positively affect the user satisfaction. Therefore, no support is found to state that speech would be beneficial and should be integrated as an output modality for a goal-bases dialog system.

This also means that in order to improve user satisfaction there is no reason for manufacturers of dialog systems to use the speech output. This is not to say that having both speech and text will be totally useless for manufacturers. When we think of these systems we often think of the typical user, who is able to read, therefore it is not likely that this user will benefit. Some dialog systems however will be aimed at people who cannot read. For these users it is very useful to be able to interact with the system through hearing. This results only suggest that within most users adding speech does not matter.

Limitations for experimental design

In this section we will take a look at possible limitations within our experiments design. We will highlight what we see as potential shortcomings and also address how big the impact of each limitation is on the overall validity of the study. The main limitation of this research is the text-to-speech component. The implementation of our text-to-speech does not feel and sound natural at all. This can have the effect that subjects would rather read the systems utterances instead of waiting for the text-to-speech to finish. If this effect does indeed occur the effect on the experiment can be viewed from two positions.

One position can be that this is the result of pairing text with text-to-speech. Because the experiment was made in order to test if there are positive effects on user satisfaction, the result can then still be seen as valid. In order to test this more research should be done on the influence on user satisfaction of text-to-speech. Specifically with text present and absent.

A factor to weight was the group of subjects that performed our experiment. The average age of the group was twenty three and a half year old. The people that performed the experiment were all internationals and mostly non-native English speakers. Most of them indicated to already have completed a higher education path. The result of having such a non diverse group of students is hard to reflect on. In our theory we have not found evidence that students perform better or worse on tasks such as the one in our experiment, but this does seem likely. Also having only young higher educated subjects will have a negative effect on the external validity. Future works should focus on getting a better representation of the overall population for their study.

A possible limitation could be the fact that our experimental setting is not reflecting the typical dynamic environment, in which the goal-based dialog system is typically used. Users often use goal-based dialog systems when they have less time, need to look up information about restaurants in the neighbourhood or train rail schedules fast, with possibilities of being distracted while consulting the dialog system. Without the simulation of this dynamic environment, the average cognitive load of the participants while using the applications is not taken into account. Therefore, our results can also not create a rebuttal against the theoretical argument of the potential benefits of a speech output modality on the cognitive load of the users.

Our experiment is also at risk of social desirability bias. This is because our experiments were mostly done in an informal setting. All participants got to see us face to face and we asked them if they wanted to participate personally. This can lead the participants to feel a connection with the researchers and thus have a positive or negative bias while answering the questions. In turn this can lead to a harmful effect on the validity of the data.

Due to the length of our experiment a potentially too large cognitive load can be laid on participants. This can result in participants growing bored, tired or simply less attentive. Which in turn can have a negative effect on our results. This can effect our data in two ways: Subjects can rate certain questions with a lower score due to being bored or frustrated, this will lead to a validity concern. They can however also simply make mistakes, which leads to a reliability concern in the data. In order to combat these biases we made sure to scramble the question orders, this will negate some of the above mentioned effects but it still does not tackle the problem of having to perform six tasks.

7 Conclusion

This study aims to contribute to the knowledge user satisfaction is a key performance metric when developing goal-based dialog systems. Current methods mostly use text to convey information to the user. Prior research indicates that speech as output modality can have a positive effect on user satisfaction due to a decrease in cognitive load when people are multi-tasking. Since a goal-based dialog systems, like recommendation systems, are often used in dynamic urban environments which requires divided attention, we examined if a speech-based output modality increase user satisfaction over a text-based output modality. In order to compare these two output modalities, we conducted an experiment in

which participants rated the two output modalities on a five-point Likert scale. Our results shows no statistical significant differences in user satisfaction between the text-based and speech-based output modality. Due to the mentioned limitations of the experimental design, further research is needed for further investigation on the effects of a speech-based output modality on user satisfaction.

References

[1] Allen, I. E. & Seaman, C. A. (2007). Likert scales and data analyses. *Quality progress*, *40*(7), 64–65.

[2] Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the division of attention: A disproof of the single channel hypothesis. *Quarterly journal of experimental psychology*, *24*(2), 225–235.

[3] Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive psychology*, *55*(1), 37–85.

[4] Cao, Y., Theune, M., & Nijholt, A. (2009). *Modality effects on cognitive load and performance in high-load information presentation*. ACM.

[5] Dafydd, G., Mertins, I., & Moore, R. K. (2000). *Handbook of multimodal and spoken dialogue systems*. Kluwer Academic Publishers.

[6] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, *113*(10), 2636–2641.

[7] Gaspar, J. G., Street, W. N., Windsor, M. B., Carbonari, R., Kaczmarski, H., Kramer, A. F., & Mathewson, K. E. (2014). Providing views of the driving scene to drivers' conversation partners mitigates cell-phone-related distraction. *Psychological science*, *25*(12), 2136–2146.

[8] Hyman Jr, I. E., Boss, S. M., Wise, B. M., McKenzie, K. E., & Caggiano, J. M. (2010). Did you see the unicycling clown? inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology*, *24*(5), 597–607.

[9] ITU (1994). Terms and definitions related to quality of service and network performance including dependability. *International Telecommunication Union*, *34*(2), 153–186.

[10] Ives, B., Olson, M., & Baroudi, J. J. (1983). *The measurement of user information satisfaction*. Information Systems Working Papers Series.

[11] Jurafsky, D. & Martin, J. H. (2018). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed. draft). To be published.

[12] Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine*, *370*(1), 54–59.

[13] Lamble, D., Kauranen, T., Laakso, M., & Summala, H. (1999). Cognitive load and detection thresholds in car following situations: safety implications for using mobile (cellular) telephones while driving. *Accident Analysis & Prevention*, *31*(6), 617–623.

[14] Larcker, D. F. & Lessig, V. P. (1980). Perceived usefulness of information: A psychometric examination. *Decision Sciences*, *11*(1), 121–134.

[15] Lindgaard, G. & Dudek, C. (2003). What is this evasive beast we call user satisfaction? *Interacting with computers*, *15*(3), 429–452.

[16] Lund, A. M. (2001). Measuring usability with the use questionnaire12. *Usability interface*, *8*(2), 3–6.

[17] Maguire, M. (2001). Methods to support human-centred design. *International journal of human-computer studies*, *55*(4), 587–634.

[18] Mctear, M. (2004). *Spoken dialogue technology: toward the conversational user interface*. Springer Science & Business Media.

[19] Mitkov, R. & Andre, E. (2012). *Natural Language in Multimodal and Multimedia Systems*. Oxford University Press.

[20] Oviatt, S. & Olsen, E. (1994). Integration themes in multimodal human-computer interaction. In *Third International Conference on Spoken Language Processing*.

[21] Reeves, B. & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge university press.

[22] Reisberg, D. (2015). *Cognition: exploring the science of the mind: sixth international student edition*. WW Norton & Company.

[23] Schmutz, P., Heinz, S., Métrailler, Y., & Opwis, K. (2009). Cognitive load in ecommerce applications: measurement and effects on user satisfaction. *Advances in Human-Computer Interaction*, *2009*, 3.

[24] Sreekanth, N., Pal, S. N., Thomas, M., Haassan, A., & Narayanan, N. (2009). Multimodal interface: Fusion of various modalities. *International Journal of Information Studies*, *1*(2).

[25] Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1998). Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech & Language*, *12*(4), 317–347.

[26] Zancanaro, M., Stock, O., & Strapparava, C. (1997). Multimodal interaction for information access: Exploiting cohesion. *Computational Intelligence*, *13*(4), 439–464.

Appendix A

Member contributions

Task	Performer(s)
Abstract	a.g.barkmeijer
Introduction	a.g.barkmeijer
Theory	a.g.barkmeijer
System Description	w.r.vries
Method	r.bassani and s.j.meyer
Results	r.bassani and a.g.barkmeijer
Conducting experimentation	all
Preparing program	r.bassani and s.j.meyer
Automatic data collection and survey implementation	s.j.meyer
User-friendly interface implementation	r.bassani
Discussion	w.r.vries
Conclusion	w.r.vries

The working load was evenly distributed between group members and all the parts of the report were proofread by all members.

Appendix B

Participant Task Instructions

The next pages contain the four instruction papers given to the participants, one for each group.

GROUP 1

Please fulfill the tasks listed below. After each task, answer a brief survey.

1) Find a Spanish restaurant in the east of the town and get the phone number.

2) Find a cheap restaurant in the south.

3) Find an Italian restaurant in the south, then change your mind and ask for a Portuguese one.

4) Find a moderate restaurant in the center, refuse the suggestion and say bye.

5) Find an expensive Chinese restaurant and ask for the location details.

6) Find an Indian restaurant in the east of the town, ask for the address and the phone number.

GROUP 2

Please fulfill the tasks listed below. After each task, answer a brief survey.

1) Find a Spanish restaurant in the east of the town and get the phone number.

2) Find a cheap restaurant in the south.

3) Find an Italian restaurant in the south, then change your mind and ask for a Portuguese one.

4) Find an expensive Chinese restaurant and ask for the location details.

5) Find a moderate restaurant in the center, refuse the suggestion and say bye.

6) Find an Indian restaurant in the east of the town, ask for the address and the phone number.

GROUP 3

Please fulfill the tasks listed below. After each task, answer a brief survey.

1) Find an Indian restaurant in the east of the town, ask for the address and the phone number.

2) Find a moderate restaurant in the center, refuse the suggestion and say bye.

3) Find an expensive Chinese restaurant and ask for the location details.

4) Find an Italian restaurant in the south, then change your mind and ask for a Portuguese one.

5) Find a cheap restaurant in the south.

6) Find a Spanish restaurant in the east of the town and get the phone number.

GROUP 4

Please fulfill the tasks listed below. After each task, answer a brief survey.

1) Find an Indian restaurant in the east and ask for the address and the phone number.

2) Find a moderate restaurant in the center, refuse the suggestion and say bye.

3) Find an expensive Chinese restaurant and ask for the location details.

4) Find a cheap restaurant in the south.

5) Find an Italian restaurant in the south of the town, then change your mind and ask for a Portuguese one.

6) Find a Spanish restaurant in the east of the town and get the phone number.