TOWARDS A SPEAKING STYLE-ADAPTIVE ASSISTANT FOR TASK-ORIENTED APPLICATIONS

Maria Schmidt & Patricia Braunger

Daimler AG
{maria.m.schmidt|patricia.braunger}@daimler.com

Abstract: This work presents the results of two user studies that help in modeling a speaking style-adaptive assistant for task-oriented applications, such as route guidance or setting the temperature inside a car. We have a look at the linguistic cues in the user's speaking style and at the desired adaptivity features in the system's speaking style. We investigate politeness, vocabulary, utterance length, and style of addressing. The results show that vocabulary and style of addressing are suitable to be modeled adaptively on the output side, and can also be deliberately used as triggers for the adaptive system behavior.

1 Introduction

To become more human-like as a spoken dialog system (SDS), one means is to adapt to its interlocutors. Apart from dialog strategy and information content the system could adapt its speaking style to the users, cf. Schmitt and Minker [1]. Consequently, when designing an adaptive voice assistant, one has to decide on two things: First, which of the linguistic features in the system's output should be modeled user-specifically. Second, which of the features in the user input the system should react to.

We call the latter *user input features* and the first *system output features*. More precisely, user input features include more or less static user properties such as age, gender, and level of experience with voice assistants, and the linguistics in the users' speaking style. System output features are linguistic features the system is able to vary, e.g., politeness or length of voice output. That is, the system produces rather short or long utterances or more or less polite ones depending on the users' preferences. In our work, we aim to identify potential triggers on the users' side for an adapted system speaking style. We address the following questions:

1. Which linguistic cues do we see in the users' speaking style?
2. Which linguistic features in the system's speaking style are relevant to be implemented adaptively?

Our analysis is based on data collected during two previously conducted user studies. The paper is structured as follows. First, we investigate which linguistic cues appear in the users' speaking style. Second, we inquire into which features in the system's speaking style are preferred by potential users. Third, we discuss on which triggers in the users' input the desired system's output depends and the challenges that arise by trying to extract those.

2 Related Work

In the literature there are several works on both adaptive SDSs ([1], [2], [3], [4]) as well as on speaking style ([5], [6]). Regarding adaptivity, Litman and Pan [2] discuss how to design and evaluate an adaptive SDS and Lemon and Pietquin [3] show methods for developing statistical

1. Listen to radio station SWR3
2. Play Michael Jackson Greatest Hits
3. Navigate to Stieglitzweg 23 in Berlin
4. Call Barack Obama on his mobile phone
5. Set temperature to 23 degrees
6. Send a text message to brother
7. Weather in Berlin today
8. Date of the European football championship final game
9. Population of Berlin
10. Score FC Bayern against VfB Stuttgart
11. Cinema program in Berlin today
12. Next Shell gas station

Figure 1 – Tasks

SDSs. Schmitt and Minker [1] define so-called adaptivity wheels in their work: the detection wheel models *on what to adapt?* with dynamic user properties (e.g., interest, emotional state), static user properties (e.g., age, gender, preferences), and interaction-related properties (e.g., user satisfaction). In parallel, the action wheel describes *how to adapt?*, that is characteristics of speech input (e.g., language model, acoustic model), dialogue strategy (prompt behavior, speaking style). Our work links to a variation of these wheels, we add *speaking style* to the detection wheel, for instance.

Talking about speaking style, most works which employ speaking style focus on spoken utterances and their prosodic and phonetic characteristics. Waibel et al. [5] report on speaker identification with the help of speaking style as a feature. In the latter, the distribution of the 50 most frequent words and parts of speech characterizes speaking style.

Our novel analyses of linguistic cues in the users' speaking style and of features in the system's speaking style are based on (transcribed) text and focus on politeness, vocabulary, utterance length, and style of addressing.

3 Linguistic Cues in the User's Speaking Style

Our analysis of the users' speaking style is based on data from a previous Wizard of Oz (WOZ) experiment in which users had to freely speak to an in-car spoken dialog system. The collected user utterances are examined in terms of the aforementioned linguistic features.

3.1 Study Design

This section explains the experimental setup. More details on the previously conducted experiment are described in [7].

In total, 45 German speaking subjects participated in the study. 54% of the participants are male and 46% are female. The average age is 39.5 years (standard deviation: 13.5). 55.6% of the participants are 20-39 years old, 26.6% are 40-59 and 17.8% are older than 60. More than two third has little to no experience with any kind of speech assistant and less than one third is experienced. Since we want to find out how users speak to a spoken dialog system while driving we put the participants in a simulated driving situation. In order to save time we decided to conduct a Wizard of Oz experiment.

Within the WOZ experiment, the participants were asked to solve predefined tasks via speaking to an in-car spoken dialog system. The system behavior was simulated by a human with the help of SUEDE [8]. The tasks the participants had to solve consist of six non-information seeking tasks (1-6) and six information seeking tasks (7-12), see Figure 1.

The tasks are described by pictures to not bias the participants. At the beginning of the study the pictures were pre-tested with the participants to find out if the desired interpretation was put in their mind. After a test drive lasting a few minutes the participants were randomly shown the pictures. In order to start the dialog the participants were told to activate the speech recognition engine via the phrase *Hallo Auto* (eng. *Hello car*). Afterwards they had to verbalize the intention given by the picture. The user input resulted in a visual and acoustic system feedback, e.g., the desired music started playing and the screen displayed the current radio station or title.

3.2 Analysis

In order to answer the question of which linguistic features in the users' speaking style are potential triggers for an adapted system speaking style we first analyze which linguistic phenomena appear when speaking freely to the system.

The following analysis is based on 540 utterances we collected from 45 participants. The utterances were manually transcribed and manually annotated.

3.2.1 Politeness

In the literature, there is no standard definition of what is polite. Pragmatic theories define politeness as a complex strategy that is not identifiable by single linguistic cues. Following Bublitz [9] it depends on the whole situation which utterance is perceived as being polite. For the purpose of this work we analyze the occurrences of the particle *please* and the sentence constructions in which the particle occurs. In order to complete the analysis of politeness we additionally rely on the empirical findings of Danescu-Niculescu-Mizil et al. [10], see also Braunger et al. [7]. With the help of a survey they characterize politeness marking in requests. Out of the 14 strategies which are perceived as being polite the following appear in our data:

- Counterfactual modal: *Could/Would* you
- Indicative modal: *Can/Will* you
- 1st person start: *I* search
- 1st person plural: Could *we* find

The distribution of these strategies and the distribution of utterances with the politeness indicator *please* is given in Figure 2. Most of the utterances (56.7%) do not contain any politeness indicator. 20.2% of the utterances contain a sentence-initial or sentence-medial *please*. 17.2% start with 1st person singular pronoun. The other strategies rarely occur.

Figure 3 indicates that most of the utterances that contain *please* are imperative sentences such as *Bitte rufe Barack Obama an* (eng. *Please call Barack Obama*). 36.6% of the utterances that contain *please* are infinite or verbless sentences, e.g. *Temperatur 23 Grad bitte* (eng. *Temperature 23 degrees please*. Only a few participants combined declarative and interrogative sentences with *please*.

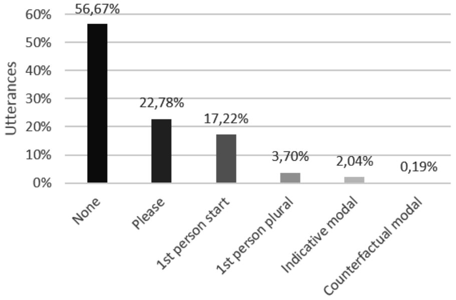

Figure 2 – Politeness indicators.

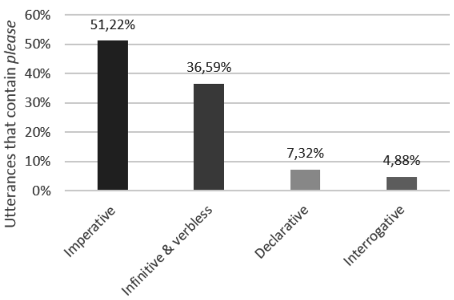

Figure 3 – *Please* within sentence types.

3.2.2 Vocabulary

As for the vocabulary the participants use, we found a great variety of synonymous expressions depending on the underlying semantic entity. Table 1 exemplarily shows the expressions we identified for five semantic entities. The semantic entity *mobile* was expressed by four different words in total. In contrast, the underlying entities *interior temperature* and *European football championship final* are each expressed by eleven different words. In addition, Table 1 shows that some user expressions are very simplified such as *Menschen* (eng. *humans*) instead of *Einwohner* (eng. *inhabitants*).

Table 1 – Vocabulary

Semantic entity	Synonymous expressions
[Mobile]	Handy, mobil, Mobiltelefon, Mobilfon
[Radio station]	Radiosendung, Sender, Radioprogramm, Radio, Radiosender
[Population]	Einwohnerzahl, Einwohner, Bevölkerung, Bewohner, Menschen
[Interior temperature]	Innenraumtemperatur, Temperatur im Innenraum,Temperatur, Innentemperatur, Klimaanlage, im Auto, Innenraum, Autotemperatur, im Wagen, Klimaautomatik, im Fahrzeug

3.2.3 Utterance length

Braunger et al. [7] already show that the length of utterances varies between 1 word and 19 words per utterance. We found that the length of user input strongly depends on the predefined task. Some tasks consist of only two keywords that necessarily have to be named to fulfill the task, e.g., task 1, and others consist of four keywords, e.g., task 4. Table 2 shows, the more keywords are required the longer the utterances. Additionally, we calculated the mean utterance length of each participant. Only a few participants uttered extremely long or extremely short sentences. 84.4% of the participants used between 6 and 8 words per utterance on average.

Table 2 – Utterance length

No. of required keywords	1	2	3	4
Mean utterance length	5.4	6.6	7.7	8.3

3.2.4 Style of Addressing

Most of the participants (74.4% of the utterances) do not address the system in any way, e.g. *Radio SWR3*. This is due to the fact that most of the information seeking tasks consist of direct questions such as *Wo ist die nächste Shell-Tankstelle?* (eng. *Where is the nearest Shell gas station?)*. 91.1% of the information seeking utterances do not contain any forms of addressing but only 58.9% of the non-information seeking utterances. 22% of all utterances implicitly address the system by the use of an imperative. Only 3.5% of the utterances contain explicit pronouns such as *du* (eng. *you*), referred to as *2nd person singular*. *3rd person singular* pronouns such as *Sie* (German polite *you*) and 1st person plural pronouns (*we*) occur three times in total.

In addition, we examine the variability of the aforementioned styles per user. Over all tasks, 14 participants stick to the same style of addressing the system. 27 participants used two styles, mostly implicit style and none, and 4 participants varied between three styles. The results are discussed in section 5.

4 Adaptive Features in the System's Speaking Style

In this section, we discuss features in the system's speaking style which should be adapted to users or user groups in order to make the user experience with an in-car SDS more natural and intuitive.

4.1 Study Design

We conducted a questionnaire-driven survey targeted at both German and US drivers. Our panel consists of regularly participating German and American subjects who received an invitation via e-mail. 1,100 German and 520 American participants completed the questionnaire aged from 20 to 94 years. The mean age was 59 years for the German set and 62 years for the US set. The relation of male to female participants was 79.1% to 20.9% across all participants, with a similar age distribution for both sexes. First, we presented the participants direct questions about the different adaptivity or personalization features. Following that, we gave them example dialogs regarding these different features to rate these implicitly.

4.2 Results

In this section, we analyze which adaptive features in the system's speaking style are favored by the subjects. Since the larger part of our analyses in this work focuses on characteristics in the German language, we limit the scope of our analyses to this part of the data set.

In order to guide quantitative evaluation, we created the following hypotheses and discuss them in the upcoming subsections. The system should be able to:

H_1 adaptively vary the politeness of its prompts H_3 adaptively vary the length of its prompts
H_2 adapt the terminology of the user H_4 adaptively vary the style of addressing different users

4.2.1 Politeness (H_1)

Politeness is linked to the length of an utterance to a certain extent, since the more politeness markers are contained in an utterance, the longer it tends to be. Therefore, we wanted to know whether the subjects prefer polite and therefore lengthier utterances or rather short and therefore less polite utterances. We posed this as a direct question. The subjects had to drag a toggle onto a 7-point slider bar. The results show that there's a tendency towards shorter and therefore less polite utterances with an arithmetic mean of 3.75 (N=1,100). We also investigated values for female and male subjects as well as six different age groups. Apart from minor deviations around the reported mean value, there are no significant differences between user groups.

4.2.2 Vocabulary (H_2)

In order to get to know which vocabulary variations the system should be capable to handle adaptively, we showed the subjects different exemplary system answers displayed in Figure 4 as suggestions to the user's non-information seeking request *Please set the inside temperature to 73 degrees*. The graphic shows that the most favored reply (46.68%) is S_1 which contains the same term *inside temperature* as opposed to other potential formulations such as *temperature in the vehicle* (23.02%) or *preferred temperature* (10.65%). 19.65% of the participants selected the option *"I don't care how it is formulated."*. Presumably S_1 is favored because the user feels fully understood by the system. In order to verify this effect for other car-specific use cases (beyond in-car settings) and other types of requests (e.g. information seeking ones), further user studies are needed. In addition to the examples, we posed two direct questions on speaking style adaptivity: The subjects rated both options *The system speaks with you the way* *you yourself speak* and *The system speaks with you the way you prefer*, similarly in the middle range of a 5-point Likert scale. But the latter option was preferred over "mirroring" as described in the first one.

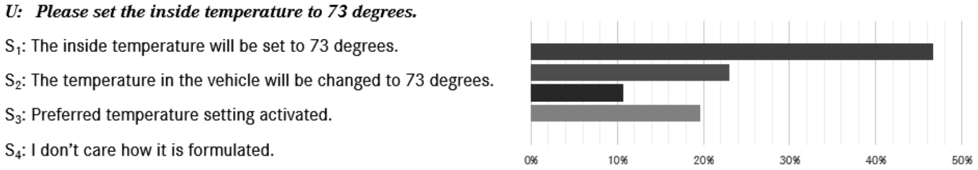

Figure 4 – Adapting to the user's vocabulary: preferred utterance variations

Table 3 – Which answer should the vehicle give?

Answer	**∅ DE**	**≤ 35**	**≤ 45**	**≤ 55**	**≤ 65**	**≤ 75**	**≤ 94**
Which track would you like to listen to?	**16.3**	**22.2**	**16.4**	**19.8**	**16.9**	**15.0**	7.3
Which track do you want to listen to?	7.8	11.1	7.5	9.2	6.8	8.5	4.8
What do you want to listen to?	2.6	2.8	2.7	1.5	2.6	3.7	2.4
Which track do you like to hear?	5.3	8.3	6.2	5.1	4.9	5.7	4.0
What do you want to hear?	3.7	5.6	3.4	2.9	3.9	3.3	5.6
What would you like to hear?	9.4	2.8	8.9	5.1	10.7	12.2	**12.1**
Please select a track.	**15.4**	0.0	**14.4**	**19.0**	**16.0**	**12.6**	**16.9**
Which track?	8.4	**13.9**	11.6	7.3	8.1	7.7	7.3
Which track should I play?	9.9	**16.7**	**12.3**	**13.6**	9.1	5.7	7.3
Which track should be played?	4.1	5.6	5.5	4.4	3.6	4.9	0.8
I don't care how it is formulated.	**17.3**	11.1	11.0	12.1	**17.3**	**20.7**	**31.5**

4.2.3 Utterance length (H_3)

Apart from the findings in 4.2.1 and 4.2.2, we investigated the subjects' preferences regarding utterance length by letting them choose from a wide variety of utterances. We asked the subjects to select their favorite system response(s) from Table 3 to their imaginary request *Play Michael Jackson's Greatest Hits*. Multiple selections of different lexical and syntactic variations were allowed. Based on the relative number of selections of these sentences – and the option *"I don't care how it is formulated"*, – we investigate the preferences of different user groups concerning utterance length of system utterances (cf. Table 3). The *"I don't care"* option is selected with a mean value of 17.3% among the top 3 options of all subjects. When looking at the different age groups, it is among the top 3 only for the 3 oldest groups, but among the top 5 for all groups. For 5 out of 6 groups *Which track would you like to listen to?* is the best answer. The second best one is *Please select a track.*, only the youngest group likes *Which track?* better. The third best voting shows an age shift: While the 3 youngest groups up to 55 years select the self-referring *Which track should I play?*, the 3 oldest groups choose "I don't care" as third choice. One can see that both a quite long utterance and a short one are chosen with similar rates across most user groups.

Despite the clearly preferred utterances on average across most user groups, there is no priming effect due to order since we displayed the sentences in a random order to each subject.

4.2.4 Style of addressing (H_4)

In the beginning of the survey we asked the German participants in a direct question, if the system should address them by using the pronoun *du* (eng. *you*) or the more polite *Sie* (eng. *you*, grammatically 2nd person plural). After this question was answered, all upcoming examples of system utterances or dialog sequences were shown to the subject containing their selected style of addressing. The results show that the younger the subjects, the more likely they vote for being called *du*. On average over all German participants, 57.2% favor being called *du*. If we only take into account the subjects between 20 and 65, 63.3% prefer *du*. A closer investigation shows that subjects between 56 and 65 only prefer *du* with 60.8% while subjects between 66 and 75 already drop to 46.7%.

5 Discussion

First, we saw that there is no straightforward definition of **politeness**. Consequently, there are different strategies to express politeness. But even though *please* is a politeness marker, not all sentences containing it will be perceived as polite. In our WOZ data we see that more than 50% of the utterances do not even contain any politeness marker, and only few sentences are truly polite with a declarative construction or a counterfactual modal. The presented politeness indicators can potentially serve as triggers for a more or less polite system response. At the same time, the subjects of our online survey indicate that they only slightly prefer shorter sentences over highly polite ones. There are no significant differences between age groups or sexes (and no strong tendencies either). It might be the case that subjects would rate politeness differently, if the question item only included politeness and not utterance length in addition.

Concerning **vocabulary** we showed that subjects use multiple different terms for the same entity or intent. While it could be beneficially to mirror the users' vocabulary in order to create rapport, not all users would want to have the system mirroring them. Furthermore, it is questionable whether mirroring is fruitful for the system's appeal. If the user employs simplified terms like *Menschen (eng. humans)* to express *Einwohner (eng. inhabitants)* this could have a bad effect of the user's perception of the system. For the presented use case H_2 may be true. But the system's response has to be modeled in a very cautious way to not flaw the overall user experience (cf. simplified vocabulary).

Utterance length cannot be analyzed independent of the use case since an utterance becomes longer the more obligatory entities have to be named. Most of the subjects utter 6 to 8 tokens and only 15.6% of them use fewer or more items. Looking at the top two preferred utterances of the online survey, one can see that both a quite long utterance and a short one are chosen with similar rates across most user groups. In combination with the aforementioned, H_3 turns out to be false at this point. We might investigate this again at a later point, if there is an underlying data set containing both user input and the selections of system utterances of the same subjects. Furthermore, we assume that the likability depends more on which sentence type is favored than on length as such: e.g., wh-question (*Which track would you like to listen to?*), direct request (*Please select a track.*), wh-question with self-referring *I* (*which track should I play?*).

Regarding **style of addressing** we found that 74.4% of the WOZ utterances do not include any addressing, and only 3.5% of the utterances include explicit pronouns. In total only 14 subjects used the same style of addressing, the others switched styles. If an SDS system adapted to this kind of style switching between different pronouns, this would hardly be advantageous for the user experience. The results of the online survey show a tendency that the younger the subjects, the more of them want to be called *du*. But at the same time in every age group there are subjects that want to be called *Sie*. Therefore, H_4 holds true and style of addressing should be modeled adaptively. But since there is not much explicit information in the users' utterances, one should rather derive the needed information from static user properties. Furthermore, this approach needs further verification through another user study whether potential users accept adaptively configured addressing style.

6 Conclusion

To conclude, we investigated linguistic cues in the users' speaking style on the one hand, and adaptivity features in the system's speaking style on the other hand. We analyzed two different user studies and discussed how their findings can be combined to design a speaking-style adaptive voice assistant for in-car usage: both **vocabulary** and **style of addressing** should rather be modeled adaptively by the SDS than politeness and utterance length – on the basis of the two studies at hand. Furthermore, for vocabulary as well as style of addressing, the reported features in the users' input can be carefully used as a trigger to elicit the respective system behavior adaptively.

References

[1] SCHMITT, A. and W. MINKER: *Towards adaptive spoken dialog systems*. Springer Science & Business Media, 2012.

[2] LITMAN, D. J. and S. PAN: *Designing and evaluating an adaptive spoken dialogue system. User Modeling and User-Adapted Interaction*, 12(2), pp. 111–137, 2002.

[3] LEMON, O. and O. PIETQUIN: *Data-Driven Methods for Adaptive Spoken Dialogue Systems: Computational Learning for Conversational Interfaces*. Springer Publishing Company, Incorporated, 2012.

[4] RIESER, V., O. LEMON, and S. KEIZER: *Natural language generation as incremental planning under uncertainty: adaptive information presentation for statistical dialogue systems. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(5), pp. 979–994, 2014.

[5] WAIBEL, A., M. BETT, F. METZE, K. RIES, T. SCHAAF, T. SCHULTZ, H. SOLTAU, H. YU, and K. ZECHNER: *Advances in automatic meeting record creation and access*. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 1, pp. 597–600 vol.1. 2001. doi:10.1109/ICASSP.2001.940902.

[6] HORI, T., D. WILLETT, and Y. MINAMI: *Language model adaptation using wfst-based speaking-style translation*. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 1, pp. I–228–I–231 vol.1. 2003. doi:10.1109/ICASSP.2003.1198759.

[7] BRAUNGER, P., W. MAIER, J. WESSLING, and S. WERNER: *Natural language input for in-car spoken dialog systems: How natural is natural?* In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 137–146. 2017.

[8] KLEMMER, S. R., A. K. SINHA, J. CHEN, J. A. LANDAY, N. ABOOBAKER, and A. WANG: *Suede: A wizard of oz prototyping tool of speech user interfaces*. In *Proceedings of the 13th Annual ACM Symposium on User interface Software and Technology*. 2000.

[9] BUBLITZ, W.: *Englische Pragmatik. Eine Einführung*. Schmidt, 2002.

[10] DANESCU-NICULESCU-MIZIL, C., M. SUDHOF, D. JURAFSKY, J. LESKOVEC, and C. POTTS: *A computational approach to politeness with application to social factors*. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*. 2013.