Adaptation in Speech Dialogues – Possibilities to Make Human-Computer-Interaction More Natural

Markus Berg, Antje Düsterhöft
University of Wismar
{markus.berg, antje.duesterhoeft}@hs-wismar.de

Bernhard Thalheim
University of Kiel
bernhard.thalheim@is.informatik.uni-kiel.de

Abstract: Adaptation is an important issue for the creation of pleasant user interfaces. In this paper we identify characteristics that can be adapted in speech dialogues. In order to realise adaptive speech dialogues we first have to develop a model that enables us to easily define dialogues and that supports adaptivity. Hence we propose the first step of a backend-oriented dialogue model.

1 Introduction

Human Computer Interaction is a successful, expanding and also hyped and controversially discussed area of research. Today nearly every device and even clothes already are or can be equipped with small microprocessors, resulting in ad hoc networks, sensor arrays and remotely controllable devices. But apart from the technical aspect, we also need interfaces to be able to interact with these devices and to benefit from the information they offer. We have permanent internet access and we use computers several times a day. As technology finds its way into our living rooms and also involves new user groups, conventional user interfaces become inconvenient. Usability suffers from inappropriate user interfaces and alternatives have to be found. Most prominently Touch- and Visualisation Technologies as well as Speech Processing Systems. In this paper we focus on the latter.

Text-based language interfaces already entered the market in the 1960s. With the turn of the millennium, speech recognition technology reached an acceptable quality and became self-evident in the field of telephony, resulting in so called Interactive Voice Response Systems. Although disliked by many customers, companies gladly used this new technology in order to save money. But not only in callcenters Speech Technology can help us to improve effectivity, convenience and usability. Think of a speech enabled living room or the possibility to search for holiday trips just by saying *"I'd like to go to Paris with my wife for one week, could you find me a cheap hotel near the city centre?"*. Besides, also in the field of eInclusion Speech Technology becomes an important factor.

The lack of today's systems is not the underlying technology but a poor human-like be-

haviour and a deficient dialogue strategy. Users are often disappointed about the limited understanding today's systems offer and the missing feature of adaptivity. Like humans adapt to their dialogue partner, also the system should adapt to the user. In this paper we identify different possibilities to adapt a speech dialogue and point out how we can reach this aim.

2 Related Work

During the last years we have worked on different dialogue systems. In the *eOhr* (electronic Ear) and later *MAIKE*[1] [13] projects we developed a system to control rooms by voice. Unlike other systems we not only focussed on a command-and-control system but on a natural dialogue. In the *Travel Consult* project [12] we developed a text-based booking system. In contrast to current chatbots like *IKEA Anna* we realised a mixed-initiative dialogue that enables the user to freely choose what to say. As we received different opinions on the style of the user interface, we conducted a user study [14] to answer the question if people from different age groups and with different abilities use a different style of speaking. We could not infer a general rule set or a one-to-one relation between any of the regarded criteria. Instead we believe that the style of speaking is a matter of personal preference. Some people prefer systems which make use of full sentences and social elements (like greeting, thanking,...) and others like a telegraphic style claiming that one does not need to be polite when speaking to a machine.

Also Bell [11] finds that "speaking styles and dialogue strategies vary from one user to another". Edlund [8] explains this with the help of the point of view with regard to the system. From the point of the *interface metaphor* the dialogue system is perceived as a machine, from the *human metaphor* it is perceived as a human-like creature. As we have to support different user types with different preferences and skills, models for adaptive dialogue systems help to simplify the development of those systems and improve user satisfaction.

We don't want to hide the fact that some researchers don't support the idea of natural dialogues. They claim that reliability is more important than naturalness. Usability is defined by success rate, time and simplicity – not by the level of anthropomorphism. Unfortunately anthropomorphism can make it even worse as "users attribute the computer more intelligence than is warranted to it". This leads to "unrealistic expectations of the capabilities of the system" [7]. Brennan and Oheari [5], too, concluded that the anthropomorphic style is undesirable for dialogue systems because it encourages more complex user input which is harder to recognise and interpret [10]. This results in recognition errors and misunderstandings. In the worst case the user is not able to reach his aim, or at best needs much time and many attempts. Stent points out that "users choose not to interact with dialogue systems as they would with other humans" resp. that "humans adapt to the interaction style of their conversational partners" [1]. A reason why people adapt to systems is given

[1] Mobile Assistive Systems for Intelligent and Cooperating Rooms and Ensembles, in cooperation with the University of Rostock

by Brennan [4] who suggested that "users adopt system's terms to avoid errors, expecting the system to be inflexible" [10].

As you can see, the negative attitude is connected with inflexible systems, negative experience and systems which are not able to process natural dialogue. This strengthens our belief that we need to focus on the system behaviour. As there are different types of users with different demands, we need to support adaptive dialogues. The mere fact that man is able to adapt to a system does not guarantee that this form of communication is the most effective.

3 Adaptation in Speech Dialogues

Adaptation is the adjustment of software behaviour during runtime. We differentiate between *adaptability* which refers to a manual adaptation by the user and *adaptivity* which is an automatic, self-adjusting adaptation of the system. Besides users often have the possibility to manipulate settings before runtime. In this case we speak of *configurability*. Both terms are closely related as it is possible for most attributes to change them before and during runtime.

The function of a dialogue does not refer to its form. The following example leads to the same result, yet having a completely different realisation.

S: How can I help you?
U: I'd like to go to Paris with my wife.
S: Oh, Paris, what a nice city. And do you already have a date in mind?
U: Easter would be great.
S: Ok, um, let me see.
S: I have fou... yeah... I have found three wonderful hotels.

S: Where do you want to go to?
U: Paris
S: Did you say Paris?
U: Yes
S: Departure date?
U: April, 20th
S: Return date?
U: April, 24th
S: How many persons?
U: Two
S: I will now list three hotels.

Our user study [14] has shown that different users have different preferences. As most people are conditioned to graphical user interfaces, they also try to use speech in the same way. We call this a *menu-oriented mindset*. Like Brennan [4] we also believe that many people avoid the use of human-human-like language in order to avoid errors. Also we do believe that they are just not aware of the system's capabilities. Consequently most participants claimed that they would have used a more human-like language style if they had known of this possibility. Interestingly unease led to the use of informal language in full sentences, which indicates that people concentrate on not using this form of language when speaking to the system. Only in certain situations they fall back to their natural style of interaction. All these observations confirm the need of adaptive dialogue systems as they embody an efficient and easy usable interface. Apart from the manual configuration of the dialogue by the dialogue designer also a self-adjusting dialogue system would be

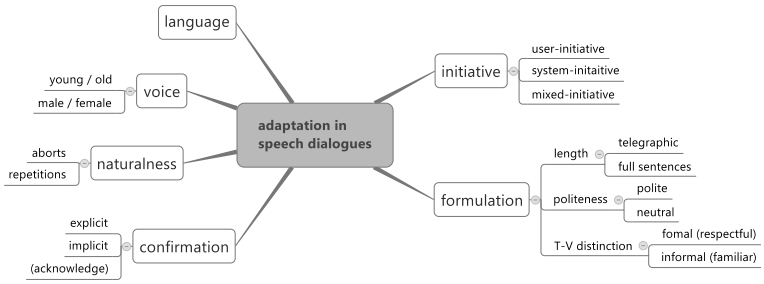

Figure 1: Adaptation in speech dialogues

beneficial. Every dialogue could start in a formal and mixed-initiative way and adapt to the user's style of speaking over runtime.

When speaking of speech dialogues, there are several characteristics to adapt: initiative, formulation, style, politeness, confirmation, naturalness, voice and language, as you also can see in figure 1. We can classify all attributes as *form-related* or *behaviour-related*.

Initiative

The initiative describes if the user is only passive and responds to questions or if he can actively influence the dialogue flow. An adaptation makes sense when errors occur. A dialogue could start with an open-ended question like *"How can I help you?"*. An inexperienced user may not know what to say or uses the wrong words. In most systems a recognition error leads to the repetition of the same question which of course is not of any help. Instead the system should create a direct question: *"Where do you want to go to?"*. Another example of an initiative-change can be seen in the following example:

S: When do you want to travel?
U: For two weeks starting next Monday.
S: I did not understand you. Please say the start date!
U: 12th of May.
S: Now specify the return date!
U: 26th of May.

The system expects answers in the form "from X to Y" and is thus not able to interpret the user answer. Consequently the system generates a more specific question. A change of the type of initiative also makes sense in the context of different user types, as "certain users are likely to voluntarily give a spoken dialogue system feedback throughout the dialogue, while others have to be explicitly asked to provide the same information" [11]. So the system has to adapt its style of asking questions to the user's way of responding to questions.

Confirmation

Confirmation is a method to ensure that the system correctly understood a user's answer. This is crucial for the usability of a system. Different from graphical user interfaces you don't automatically have feedback. At the same time feedback is much more important because natural language is fuzzy and harder to recognise than keyboard input and thus more error-prone. Despite its importance, confirmation in speech interfaces can easily be annoying. That's why we use different confirmation strategies based on the confidence of the recognition result. In the following example S1 represents explicit confirmation and S2 represents implicit confirmation. S3 and S4 use no confirmation at all whereas S3 produces the illusion that the system understood the user instead of only asking the next question. S3 can be seen as acknowledging the user's answer.

S: Where do you want to go?
U: Rome
S1: You want to go to Rome, correct?
S2: When do you want to come back from Rome?
S3: Ok, and when do you want to come back?
S4: When do you want to come back?

Formulation

The formulation of system prompts extremely contributes to the appearance, or *hear & feel*, of a voice interface. Some people prefer telegraphic sentences or even single words like *"Destination?"* and others expect full sentences with obeying politeness like *"Which city do you want to fly to?"* or *"Would you please inform me about your destination city?"*. In many languages you also have to obey T-V-distinction[2], which refers to respectful[3] or familiar[4] formulation. While common systems use a predefined formulation which is only synthesised by the system (TTS), we aim at a concept-to-speech-component (CTS) which automatically generates language.

Moreover the formulation style of the system has influence on the style of the user answers. Hence we can "influence users to behave in a certain way, for instance by implicitly encouraging a speaking style that improves speech recognition performance" [11]. Jokinen suggests an inclusion of context information in the case of low confidence levels. This refers to different confirmation styles depending on the confidence. In her example [9] the answer to the question *"When will the next bus leave for Miami?"* could be *"2.20pm"*, *"It will leave at 2.20pm"* or *"The next bus to Miami leaves at 2.20pm"*.

Further Attributes

Apart from these characteristics, also the language or voice can be adapted to the user's wishes. A language identification algorithm could automatically switch the language when

[2] lat. tu/vos
[3] german: "siezen"
[4] german: "duzen"

it recognises that the language of the user differs from the current interface language. Moreover we can adapt the gender or age of the synthesised voice. Another interesting approach is to increase naturalness or human-likeness by introducing typical speech phenomena like repetition or abortion. Moreover the use of coughing, hawking or filler words like “um” or “err” could improve acceptance within some user groups.

4 Dialogue Modelling

The task of the dialogue manager is the evaluation of the user goal and response generation [9], i.e. to find out *what* the user wants and *how* this goal can be realised with regard to different context information. In order to create an adaptive dialogue system, we first have to realise the interpretation of the user goal. Thus we need a dialogue model which allows an easy dialogue definition that also supports adaptivity. As we have mentioned, the dialogue logic is independent from the form of the dialogue. That’s why we plan to separate form and function with the help of language generation algorithms. Another important way of supporting this goal is dialogue acts. Dialogue acts base on speech acts which have been invented by Austin [2] and advanced by Searle [15]. They define what a user *does* by *saying* words. In other words they generalise utterances to their function or the user’s intention. By saying *“hello”* we greet, by saying *“Mind the gap”* we warn or by saying *“Do you know where James is?”* we ask. Speech Acts have been revised by different researchers (e.g. Harry Bunt [6]) resulting in Dialogue Acts. They have been used in many research projects like *Verbmobil* or *Trains* and embody an important tool in dialogue modelling.

In our research we focus on *task-oriented dialogues*. This includes *information-seeking* dialogues (e.g. travel booking systems), *question-answering dialogues* and *command dialogues* (e.g. smart room control). In these dialogue types we focus on the understanding of the question or command (we comprise this as *concern*) and the generation of an appropriate answer or the intended action (what we summarise as *reply*) [3]. We delimit from conversational or small talk systems and also from human-human-communication, i.e. we are only interested in information that helps us to support the user. We want to realise this as natural and human-like as possible but we explicitly do not focus on the exchange of opinions or the telling of stories.

When analysing existing dialogue act tagsets, we observe that they are very detailed and also include parts we excluded from our considerations, namely commitments, promises, oaths, nominations or threats. Also the recognition of turn-taking and stalling is not important to understand the intention of the user. In task-oriented dialogue systems the user has a specific goal. This goal is defined within the range of services that the system offers. Thus we limit the dialogue acts to those which directly refer to the backend. In [3] we identified three basic backend functions: *getInfo*, *setInfo* and *do* as you can see in figure 2. A *getInfo*-call puts a request to the backend, *setInfo* changes or sets the value of a variable in a frame and *do* executes a function (switch the light on, start a presentation software). The first two methods are information-related while the last one refers to all other actions. These backend-functions influence the selection of dialogue acts, so we propose the

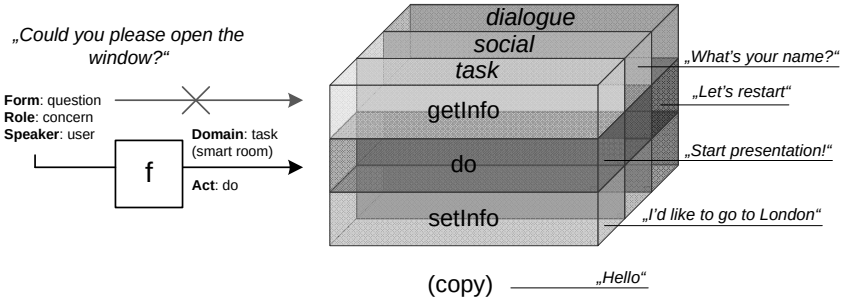

Figure 2: Backend-based dialogue acts in task-oriented dialogues

following top-level dialogue-acts: *information seeking*, *information providing* and *action requesting*. We must not mix up these categories. An instruction like *"Search for hotels in London that offer breakfast"* is not an action request but an information seeking act.

We now have the *role*-attribute (concern and reply) and the *act* or goal of the user (information seeking, information providing, action requesting). Apart from that, we can also define *domain*, *speaker* and *form*. Even a single-task system consists of different domains, i.e. task domain, social domain and dialogue domain. Whereas all task-related requests obviously belong to the task domain, requests like *"Could you repeat that?"* or *"Let's restart"* are dialogue-related (i.e. they influence the behaviour of the dialogue). Here we also distinguish between getInfo, setInfo and do. A dialogue-oriented do-request is especially interesting in the context of adaptability because it allows the user to influence dialogue attributes, e.g. *"Please use indirect confirmation"* or *"Set the voice to female"*. The social domain represents social obligations that need to store or retrieve information like *"Hello, my name is Markus"* (setInfo) resp. *"Do you remember my name?"* (getInfo). Of course most social obligations don't need any access to the backend. Most acts are symmetric adjacency pairs like openings and closings (*greet*, *say goodbye*). Hence we introduce a special dialogue act – *copy* – that doesn't need backend access. As already mentioned, another attribute is the *speaker*. Here we distinguish *user* and *system*. The *form* of an utterance is the most basic information. It can be seen as the foundation for any further analysis. But since form and function don't build a one-to-one relationship, the inference of the function based on form and context is an eminent task. Sometimes we regard form as *secondary illocution* and function as *primary illocution*.

The resulting dialogue model consists of role, speaker, act/function, form and domain, e.g. *"Could you please open the window?"* can be formalised as a quintuple: *(concern, user, action request, question, smart room)*. Of course this information only allows us to call the correct type of backend function, i.e. we identify the general user goal. The actual proposition has to be modelled in the next step.

5 Conclusion and Future Work

The adaptation of speech dialogues is important for the usability of the system. We have shown several possibilities how to adapt dialogues, i.e. when adaptation should take place and what characteristics can be adapted. Our examples include the adaptation of the initiative in case of errors or bad recognition confidence, the adaptation of the style or formulation, change of language and voice and the adaptation of the confirmation strategy.

In order to be able to create an adaptable dialogue system, we first need a dialogue model which allows us to define dialogue systems in an abstract way. As you can see in figure 3, in this paper we have realised a model for the *general user goal*. This is the first step for creating an adaptive dialogue system. In our future work we will address further modules, i.e. the inference of the proposition, the language generator and of course the dialogue manager that determines *how* communication takes place. All of these parts have to interact with the adaptation context for the system to be able to adapt properly.

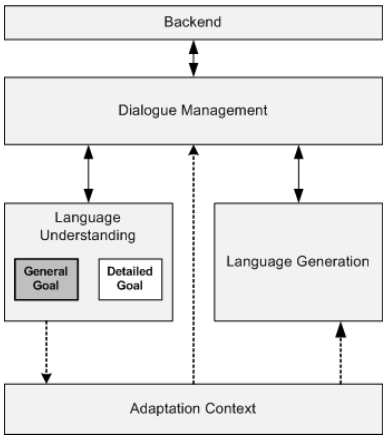

Figure 3: Interaction of the modules with the adaptation context

References

[1] Amanda J. Stent. *Dialogue Systems as Conversational Partners: Applying Conversation Acts Theory to Natural Language Generation for Task-Oriented Mixed-Initiative Spoken Dialogue*. PhD thesis, University of Rochester, 2001.

[2] J. L. Austin. How to do things with words. 1962.

[3] Markus Berg, Bernhard Thalheim, and Antje Düsterhöft. Dialog Acts from the Processing Perspective in Task Oriented Dialog Systems. In *Workshop on the Semantics and Pragmatics of Dialogue*, Los Angeles (USA), 2011.

[4] Susan E. Brennan. Lexical entrainment in spontaneous dialogue. In *Proceedings of the International Sympo- sium on Spoken Dialogue*, pages 41–44, 1996.

[5] Susan E. Brennan and Justina O. Ohaeri. Effects of message style on user's attribution toward agents. In *Proceedings of CHI'94 Conference Companion Human Factors in Computing Systems*, pages 281–282, 1994.

[6] Harry Bunt et al. Towards an ISO standard for Dialogue Act Annotation. In Nicoletta Calzolari et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[7] Byron Long. Natural Language as an Interface Style. `http://www.dgp.toronto.edu/people/byron/papers/nli.html`, 1994.

[8] Jens Edlund, Mattias Heldner, and Joakim Gustafson. Two faces of spoken dialogue systems. In *Interspeech 2006 - ICSLP Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems.*, 2006.

[9] Kristiina Jokinen and Graham Wilcock. Adaptivity and response generation in a spoken dialogue system, 2007.

[10] Ivana Kruijff Korbayova and Olga Kukina. The effect of dialogue system output style variation on users' evaluation judgments and input style. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, SIGdial '08, pages 190–197, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[11] Linda Bell. *Linguistic Adaptations in Spoken Human-Computer Dialogues*. PhD thesis, KTH Stockholm, 2003.

[12] Markus Berg and Antje Düsterhöft. Website Interaction with Text-based Natural Language Dialog Systems. In *7. Wismarer Wirtschaftsinformatiktage*, Wismar (Germany), 6 2010.

[13] Markus Berg, Nils Weber, Gernot Ruscher, and Sebastian Bader. Maike: Mobile Assistenzsysteme für intelligente kooperierende Räume und Ensembles. In *5. Kongress Multimediatechnik*, Wismar (Germany), 10 2010.

[14] Markus Berg, Petra Gröber, and Martina Weicht. User Study: Talking to Computers. In *3rd Workshop on inclusive eLearning*, London (UK), 9 2010.

[15] John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.