Adapting Spoken and Visual Output for a Pedestrian Navigation System, based on given Situational Statements

Rainer Wasinger, Dominika Oliver, Dominik Heckmann,
Bettina Braun, Boris Brandherm, Christoph Stahl
Saarland University
D-66123, Saarbrücken, Germany
{wasinger, heckmann, stahl}@dfki.de, brandherm@cs.uni-sb.de, {dominika, bebr}@coli.uni-sb.de

Abstract

As mobile devices become more and more complex, there is an increasing desire for these devices to adapt to their users. This paper identifies parameters for different input sources (user, device and environment), and the parameters of media output (speech, graphics, sound and text), that may be modified to tailor user presentation in a pedestrian navigation system. We also provide an initial insight into some of the causal relationships between our input and output parameters, with a specific focus on the effects that speech can contribute to the presentation of media output.

1 Introduction

Human-computer-interaction in mobile environments has long been a difficult research area. This has been due to the limited resources available to portable computing devices, the dynamic nature of the environments in which they are used (in contrast to stationary indoor desktop settings), and the difficulty in adapting the device to the changing needs of the user.

The focus of this paper is to outline parameter information for a user, a mobile device and an environment, and to describe how these parameters can be harnessed into providing *user-adapted visual and audio output*. In section 2, we describe the pedestrian navigation system, on which our research is being implemented. This is followed in section 3 and 4 by a look at the different types of input and output parameters that may be used in determining how best to adapt spoken and visual output to the user. In section 5, we provide two examples that illustrate the type of effect that input parameters may have on speech output, and take this as a stepping-stone in defining a set of causal rules for adapting presentation output to the user. We conclude in section 6 with a description of our proposed future work.

2 Pedestrian Navigation Scenario

2.1 Functionality

The pedestrian navigation system allows the user to plan one or more routes over the Internet, and to download these routes (either indoor or outdoor) to their PDA [Wasinger *et al.*, 2003]. Upon downloading the routes, the user can select the one that they are currently interested in, and be directed along the route through the use of speech and graphics output. Aside from the navigation mode, the user can at any time switch into an exploration mode, in which they have the ability to explore their surrounding environment and query what it is that they see. Speech output can be of type *instruction* (e.g. "Walk 210 meters, and then turn right into Max-Diamand-Street"), or *description* (e.g. "The building 36.1 is the location of Professor Wahlster's professorship"). Graphics output is in the form of street and building maps. Sounds such as system beeps, and text such as 'next street name' are also used to support speech and graphics output. The interface as seen by the user is shown in Figure 1 below.

Figure 1: Pedestrian navigation interface.

2.2 Supporting Technology

The pedestrian navigation system comprises a navigation server and a Pocket PC. The Pocket PC component developed in C/C++, incorporates the IBM Embedded ViaVoice[1] speech synthesizer and recognizer. The 2D/3D graphics are generated via the embedded Cortona VRML[2] browser. GPS provides the user's geographical coordinates when outside, while infrared beacons are used to locate a user when inside. Outdoor navigation is based on commercially available material from NavTech, but data on indoor floor plans and detailed landmark information (e.g. opening hours, cost and description) have to be modeled by hand.

[1] IBM Embedded ViaVoice, http://www.ibm.com/software/pervasive/products/voice/vv_enterprise.shtml
[2] Virtual Reality Markup Language, http://www.w3.org/MarkUp/VRML/

2.3 Architecture for Incorporating the Input and Output Parameters

An improvement to the system currently being considered is that of using situational input statements to modify the presentation of audio and visual media output, to better suit a user. The situational input parameters relate to the user, the device and the environment, while the targeted media output is that of speech, graphics, sound and text. The set of input parameters is modeled in a program called Ubi's World[3], and then made accessible to the navigation system as an XML file. The collection and creation of these input parameters is described in [Heckmann, 2003b]. Based on the situational input parameters and the media output parameters, the presentation of route descriptions and landmark queries can be adapted to the user. This architecture is shown below in Figure 2.

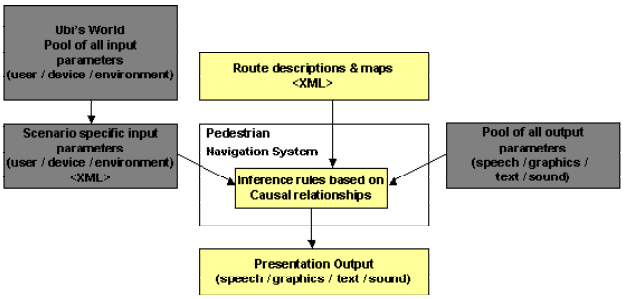

Figure 2. Data flow in adapting the media output presentation.

3 Adaptation Input

The adaptation input such as user model entries, environmental contexts and the devices' resource-limitations are uniformly represented as situational statements [Heckmann, 2003a]. For design simplicity, these parameters are currently statically defined. It is however foreseen that these input parameters and their associated values be dynamically updateable in the future.

3.1 User Model

User model parameters refer to characteristics of the current user of the Pocket PC. The list of user model parameters that we envisage in our scenario are as follows:

- Role of user (tourist, business person).
- Age (young, middle-aged, elderly).
- Gender (male, female).
- Walking speed (slow, normal, fast).
- Eye sight and special needs.
- Emotions (anger, distress, happiness).
- Cognitive load and time pressure.
- Social environment, carrying something.
- User interests and preferences.

The usability issues learnt from designing for elderly / average-aged people [Müller and Wasinger, 2002. Müller et al., 2003] will also play a role.

[3] http://www.u2m.org/ubisworld.htm

3.2 Device Resources

Device resource parameters refer to parameters that justly model the current state of the device. Such parameters may include:

- Remaining time of use.
- GPS coverage.
- Speaker loudness range.
- Map size and zoom factor.
- Screen size and contrast.
- Working memory and storage.

3.3 Environmental Context

Environmental context is especially important for mobile indoor / outdoor navigation systems. A list of contributing parameters follows:

- Noise-level and light-level.
- Quality of street surface.
- Street crowdedness.
- Weather conditions (outdoor).

While some of the above mentioned input parameters are static over time (e.g. gender), others are highly dynamic (e.g. emotions). Some are easy to detect, while yet others are hard to detect. For the purpose of this paper, we treat all parameters equally, and as already available to the system.

4 Adapting Presentation Output

Although we focus on spoken output in this paper, there are a total of four defined sets of output parameters that contribute to the presentation of media in our system. The two primary sets are that of speech and graphic output, while the two support sets are that of sound and text (see Figure 4). All are essential for multi-modal output presentation, and largely influence how natural and understandable dialog with a user can be.

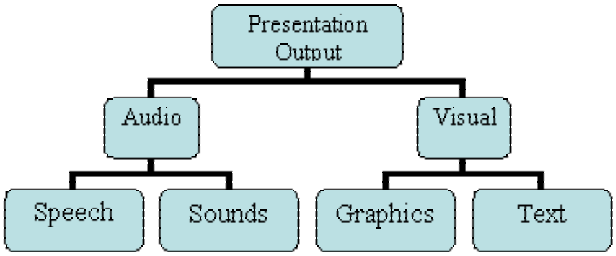

Figure 4: Categorization of the media presentation outputs.

4.1 Speech Parameters

The formant-synthesis used in our system allows for a variety of acoustic parameters to be modeled. These can be explicitly set by the user (e.g. through the program's menu or via speech), or implicitly defined by our causal relationships. These parameters can be modified to improve the quality and intelligibility of speech for a given context and for different locations along a user's trip (see Figure 5). The parameters are listed below:

- Different languages and dialects (e.g. UK English).

- Physical voice characteristics (e.g. pitch baseline, speed, volume).
- Speaking style (e.g. whisper, monotone).
- Prosody (e.g. emphasis, pauses, tones).

Adapting speech output is achieved by specifying a *language* and its *dialectal form*, along with the gender and age. The available *voice characteristics* include a range of physical factors like: pitch baseline and fluctuations, level of breathiness or roughness, as well as speech rate and volume. Additionally, depending on the context and environment, *speaking style* can be adapted to suit a format such as whisper or monotone.

To enhance clarity, or to avoid ambiguity of instructions, additional *prosodic cues* can be introduced. Such cues include emphasis, additional pauses, tones and phrase accents, which can all be assigned to the input. Prosodic signals such as cues for prominence and phrasing are crucial. We split the utterance into reasonable chunks and assign variables (e.g. pitch contour), appropriate for the type of output. In the examples provided in Section 5, we manipulate physical voice characteristics and prosody, based on sentence type (i.e. instruction or description).

4.2 Sound Parameters

Sound represents audio other than speech such as system beeps that are used for example, in the confirmation of accurately recognized user input. Sounds can be used to both alert the user of landmarks in the vicinity, and also (if in navigation mode) as to how far away they are from their destination. Similar to text, it is a complementary form of output because it cannot itself provide all the information that is required by a user when navigating to a destination, or exploring an area.

4.3 Graphical Parameters

We similarly categorize the graphical parameters into the following groups:

- Level of detail (e.g. object filters).
- Reference symbols (e.g. landmarks, POIs).
- Accessibility (e.g. menus, toolbars).
- Color and contrast.

Mobile devices need careful consideration as to the *level of detail* that may be supported on a single display. For our scenario, this not only includes the density of roads per m^2, but also reference symbols. *Reference symbols* allow a user to more easily locate their position on the map, and refer for example to prominent landmarks (e.g. buildings, parks, monuments), and Point Of Interests (POIs), such as bus stops and money dispensers. Other reference symbols may also refer to the type of route (e.g. 'rough road', 'cycle-way') and upcoming hazards (e.g. 'stairs ahead'). *Accessibility* refers to the placement and size of graphical elements like menus and toolbars, and the functionality that they provide, such as different zoom levels, 2D/3D views, map-orientation (e.g. true north vs. walking direction), and the status of connected devices (e.g. GPS and mobile phones). A final important aspect is the *color* and differing levels of *contrast* of the objects on the display. This can be used to denote focus and to highlight different object categories, both of which make object differentiation easier for a user. Figure 5 shows the use of some of the above-mentioned graphical parameters.

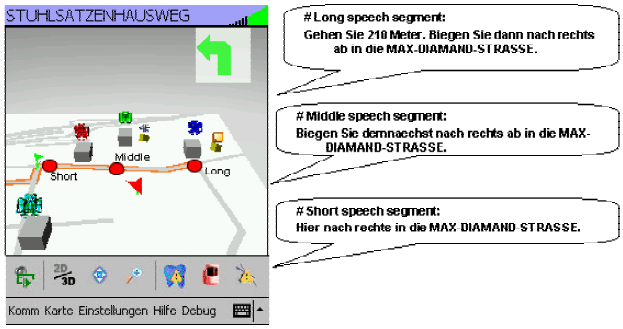

Figure 5: Speech and graphical presentation output.

4.4 Text Parameters

In mobile environments where display space is limited, textual output mainly takes the role of supporting the other forms of output like speech and graphics. This is usually in the form of overlapped or additional information such as previous, current and next street names, or the names of prominent nearby landmarks. We group relevant factors in the presentation of textual output as follows:

- Textual format (e.g. font size, type, style).
- Presentation format.

Textual format includes physical characteristics of the text such as font size, type, style, spacing between characters, color, and contrast to the background. The textual format also refers to whether parts of the text may be abbreviated to save space, or whether specific words may be removed or added to better adapt user output. The *presentation format* defines the amount of display space allocated to text, and the type of display space (e.g. text area, text field, scrollbars, semitransparent text). It also defines the time period in which text should remain visible before it is replaced by other text, and whether it should capture the user's attention while on the screen (e.g. suddenly appearing text, scrolling text or blinking text).

5 Identifying Causal Relationships

In this section, we describe our initial attempts in defining causal relationships between the input parameters and the speech output. As an example, if the user input parameter 'cognitive load' is set to 'high', then the speech output parameter 'speed' may be set to 'slow'. We do not look at how the input parameters may affect the other forms of media output (graphics, text, audio), nor at the complexities involved in combining multiple media forms. This is left for future work.

We discuss our motivations for changing speech parameters, based on existing studies of a general nature, studies on special needs of users, and psychological studies. We also provide two examples demonstrating how the input parameters and speech output can affect a presentation.

Two *general* studies that contributed to the adaptation of speech output in our pedestrian navigation system, were that of synthesized speech perception [Ní Chasaide and Gobl, 2001. Nass and Lee, 2001], and German intonation [Grice and Baumann, 2002]. Both studies highlighted areas where it may be useful to modify the speech parameters defined in Section 4.1. Combined with the authors' own linguistic knowledge, the following examples illustrate the results. The utterances (1) and (2) are of type 'description' and 'instruction ' respectively.

(1) `vs60 Das Gebäude `vs50 36.1, ist der Sitz des Lehrstuhls `vs60 Professor `vs50 `2 Wahlster `/.
The building 36.1 is the location of Professor Wahlster's professorship.

(2) `vb60 Gehen `0 Sie 210 `ar Meter `%, `vs60 biegen Sie dann nach `vs40 rechts ab `%%, `vs50 in die `0 Max `2 Diamand-Strasse `/.
Walk 210 meters, and then turn right into Max-Diamand-Street.

In (1) we increase the tempo of known tokens ('vs60 Das Gebäude), put additional emphasis on the accented word ('2 Wahlster) and assign a large pitch fall at the end of the sentence for a more perceived finality ('/).

In (2) we have a longer sentence, which has been split up into two chunks. Similarly, important pieces of information are pronounced slower ('vs40 rechts) than the rest. The pitch baseline has been lowered to achieve a more factual and instructional tone ('vb60). In the compound (Max-Diamand-Strasse), the middle token is additionally emphasised ('2 Diamand) and the first token is de-accented ('0 Max). At the end of the first phrase, we assign a rising tone to the last content word ('ar Meter) and also add a pitch rise (%). We also add a phrase-final continuation rise (%%), which functions as a cue to the listener that more information follows.

Apart from achieving general clarity and ease of comprehension, speech parameters can also be modified to suit people with *special needs* such as the elderly. For example, studies show that the understanding of synthetic speech decreases with age, reaching around 60% loss for the older part of the population [Eskenazi and Black, 2001]. This loss can however be minimized through changes in speech parameters such as speech rate and volume. In the above study, it was also reported that in a semantically restricted domain, the ability to predict keywords within speech segments, remained constant with age. This is advantageous, because even if an elderly user does not understand all that was said (due to shortcomings in the generated synthetic speech), we can still rely on the users own cognitive ability to recognize the primary keywords in our navigation domain.

Psychological studies suggest that people use voice characteristics to assess personality [Nass and Lee, 2001]. When exposed to synthetic (clearly non-human) voices, people assign personalities to the voices. Furthermore, the study showed that people seem to be attracted to voice characteristics exhibiting 'personality' markers similar to their own. The voice characteristics found to be responsible were intensity, mean fundamental frequency, frequency range, and speech rate. By linking these parameters to the user model, we could in practice evoke trust or liking in the user.

6 Future Work and Conclusions

This paper described different types of input and output parameters that may be used to better adapt presentation output. In the future, we will need to evaluate the current implementation through our own user studies. Similar to speech, we will need to identify causal inference relationships between the input parameters and the graphics, text and audio modalities. A final and very important area for future work will be the modeling of our causal relationships via Bayesian networks.

Acknowledgements

This work is supported by the European Post Graduate College "Language Technology and Cognitive Systems", and the BMBF funded project COLLATE.

References

[Eskenazi and Black, 2001] Maxine Eskenazi and Alan Black, A study on speech over the telephone and aging, *Proc. of Eurospeech 2001*.

[Grice and Baumann, 2002] Martine Grice and Stefan Baumann, Deutsche Intonation und GtoBI, *Linguistische Berichte*, pp. 267-298, 2002.

[Heckmann, 2003a] Dominik Heckmann. Introducing "Situational Statements" as an integrating Data Structure for User Modeling, Context-Awareness and Resource-Adaptive Computing, *ABIS*, 2003.

[Heckmann, 2003b] Dominik Heckmann. UbisWorld, a Blocksworld for Ubiquitous Computing, *Submitted to Ubicomp*, 2003.

[Müller and Wasinger, 2002] Christian Müller and Rainer Wasinger. Adapting Multimodal Dialog for the Elderly, *ABIS-Workshop on Personalization for the Mobile World*, 2002.

[Müller et al., 2003] Christian Müller, Frank Wittig and Jörg Baus. Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs. *Proc. of Eurospeech 2003 Special Session on Spoken Language Processing for e-Inclusion*, 2003.

[Nass and Lee, 2001] Clifford Nass and Kwan Min Lee. Does Computer-Synthesized Speech Manifest Personality? Experimental tests of Recognition, Similarity-Attraction, and Consistency-Attraction, *Journal of Experimental Psychology Applied*, 2001.

[Ní Chasaide and Gobl, 2001] Ailbhe Ní Chasaide and Christer Gobl. Voice quality and the synthesis of affect. In E. Keller, G. Bailly, A. Monaghan, J. Terken and M. Huckvale (eds.), *Improvements in Speech Synthesis*, pp. 252-263, 2001.

[Wasinger *et al.*, 2003] Rainer Wasinger, Christoph Stahl, Antonio Krüger. M3I in a Pedestrian Navigation & Exploration System, *Proc. of the Fourth International Symposium on Human Computer Interaction with Mobile Devices*, 2003.