Three Output Planning Strategies for Use in Context-aware Computing Scenarios

Gerrit Kahl[1] and **Rainer Wasinger**[2] and **Tim Schwartz**[1], and **Lübomira Spassova**[1]

Abstract. In everyday life, it is useful for mobile devices like cell phones and PDAs to have an understanding of their user's surrounding context. Presentation output planning is one area where such context can be used to optimally adapt information to a user's current situational context. This paper outlines the architecture of a context-aware output planning module, as well as the design and implementation of three output generation strategies: *user-defined*, *symmetric multimodal*, and *context-based* output planning. These strategies are responsible for selecting the best suited modalities (e.g. speech, gesture, text), for presenting information to a user situated in a public environment such as a shopping mall.
A central point of this paper is the identification of context factors relevant to presentation planning on mobile devices with finite resources to obtain a private and/or public output. We show via a working demonstrator the extent to which such factors can, with readily available technology, be incorporated into a system. The paper also outlines the set of reactions that a system might take when given context information on the user and the environment.

1 Introduction

Let us consider a typical shopping scenario: a customer goes to a supermarket and works through her/his shopping list, step by step. Some products have a range of attributes influencing a purchase, for example price and brand, and the customer may expend a lot of time in searching for just the right product. This process can easily become quite tedious.

One project that deals with shopping assistance is the *Mobile ShopAssist* (MSA) - a PDA based program that acts as a mobile shopping consultant (see [10]). With the MSA software, customers can inform themselves about the properties of different products and compare them with one another. The system provides different input and output modalities that can be used for human-computer interaction, including speech, handwriting and gesture, and customers can, during the interaction process, also reference both virtual data on the PDA's display along with real objects in the surrounding instrumented environment [10].

Customers may also be shopping for different products, ranging from medication (e.g. cold and flu tablets) to electronics (e.g. a digital camera), and the product type may well affect the selection of a modality used in communicating with the customer. For example, when dealing with products of type "medication", the system should avoid revealing any medical conditions the customer might have (*private output*).

The work in this paper can be seen to extend existing work on the Mobile ShopAssist, that focused largely on the recognition and interpretation of multimodal input, as described in section 3. An important aspect that has been extended is that of the generation and presentation of system utterances (described in section 4). In particular, system utterances are dynamically created, based on pre-defined sentence templates, to suit a customer's current context. In addition to this, the presentation of such utterances has been extended such that modalities like speech, gesture, and text, are mapped at runtime to individual semantic elements in the utterance.

2 Related Work

In order to demonstrate how to combine different output modalities, Elting [3] uses a virtual character. He explains that a multimodal presentation should take the current situation and context into account. His virtual character uses graphical and acoustic output modalities, and it is able to adapt to the preferences of the user as well as to the current context.

SmartKom [9, 8] is a system bearing resemblance to this work in that it focuses on context-aware computing and presentation output planning. This system is a multilingual (English and German), multimodal dialog system. Output is done with the help of a virtual character who can use the modalities of speech, graphics, gesture, and facial expressions. The output modalities are synchronized with each other and the system can furthermore choose which situations warrant speech-only output.

Our system is designed for use on mobile PDAs with limited resources. Although SmartKom takes the current context into account for selecting the appropriate output modalities, our system is capable of much finer mappings in which any combination of modalities can be mapped to individual semantic elements in an utterance. Furthermore, this work identifies a wide range of context factors that can have an influence on output generation and also defines the types of system reactions that might be used to adapt a system's output. Some system reactions include, for example, the ability to: modify the format and tempo of the speech output; change the display duration of the text output; and decide whether the output should be presented on- and/or off-device.

[1] DFKI GmbH, email: {Gerrit.Kahl, Luebomira.Spassova, Tim.Schwartz}@dfki.de
[2] Macquarie University, Australia, email: rainer.wasinger@mq.edu.au

3 Mobile ShopAssist Demonstrator

As described in the introduction, this work is based on the Mobile ShopAssist (MSA) demonstrator [10]. In addition to the user's PDA, the MSA uses some output devices situated in the environment to cater for *off-device* communication. Visual output, for example, can be displayed either on a large plasma screen next to the product shelf, or using a steerable projector system (*Fluid Beam* [2, 7]) that allows the creation of projected displays on arbitrary flat surfaces. Acoustic output is performed using a spatial audio system (*SAFIR*) as described in [6], which allows for the creation of virtual sound sources at any location in the environment.

The MSA was built following the concept of symmetric multimodality, which is defined in [8] to mean that "all input modes are also available for output, and vice versa". For the MSA application, the relevant modalities are: speech, handwriting/text and gesture. Gesture can refer to pointing actions on the touch screen of the PDA (see figure 1), and also to the act of taking a physical product out of the shelf. Each product in the shelf is fitted with an RFID tag. An antenna on the back of the shelf detects when a product is taken out of it. With the handwriting mode, the MSA recognizes the user's input by pattern matching it with dynamically loaded finite-state rule grammars.

4 Output Modalities

As user input, the MSA system is able to identify semantic elements like the name of an *object* (e.g. "PowerShot Pro1") and the name of a *feature* (e.g. "mega pixels"). The generated output additionally contains the *value* (e.g. 8) corresponding to the feature of the object. We make a distinction between output on the PDA (*on-device*) and output in the environment (*off-device*).

Speech output is generated in the form of natural language. The sentences are generated using the grammar stored in an XML file and speech is synthesized using ScanSoft RealSpeak Solo[3]. For off-device output the generated sentence is transmitted via wireless LAN to a server, which controls the public speakers, as mentioned in section 3.

As an alternative to the natural language output ("*The PowerShot Pro1 has 8 mega pixels.*"), there is also the possibility to output only the semantically-rich keywords: *"<object>, <feature>, <value>"*, e.g. *"PowerShot Pro1, mega pixels, 8"*. The advantage of this output is that it takes less time. Thus, only the "important" data is presented.

Text output on a PDA is often limited by display size constraints. In the MSA, an efficient text output algorithm was developed to counter this constraint. In our system, text output is shown on the bottom of the display in two rows (see figure 1). It always has the same structure, so that the user can easily find the part of interest. Off-device text output is displayed on a public screen and/or in the space that occurs when an object is taken out of the shelf. The so called *Product Associated Displays* (PADs, see [7]) provide visual feedback to the user in the form of projected images and text.

[3] http://www.scansoft.com/realspeak/mobility/

Figure 1. Gesture and text output of the MSA

Figure 2. Visualization of the used output modalities

Gesture output for the object consists of the drawing of a border around its image (see figure 1). For the feature, gesture output is achieved by highlighting the corresponding phrase in a scrollbar at the bottom of the screen. In this way, the user gets visual feedback that the system has recognized her/his input. Off-device gesture output is implemented only for *object* (and not for *feature* or *value* attributes). It is presented as a highlighted spot that is displayed on the object in the shelf (see [2]).

5 Output Planning Strategies

Three different output planning strategies are available in the MSA, namely: *user-defined*, *symmetric multimodal*, and *context-based*. The current modality settings are visualized in the right corner of the PDA display. In this way, the user has an overview of the currently used output modalities (see figure 2).

The letters S, T and G stand here for speech, text and gesture; F, O, V, onD and $offD$ stand for feature, object, value, on-device and off-device. When for example speech output for the object is selected, a coloured bar is displayed in the middle of the S-block, i.e. in the same row as the O. In order to make clear which of the three strategies mentioned above is currently being used, there are three different colours for the bars.

5.1 Symmetric Multimodal Output

A central point of this paper is the redefinition of the scope of the well-known term "symmetric multimodality" (see [8]), which in this work refers not just to the ability of using the same modalities for input as for output, but rather to the ability of using individual semantic-element to modality mappings for output as for input. In this way, the user can control which output modalities should be used without explicitly setting them. This means that the user controls the output by applying the corresponding input modality. As there is no input for the value attribute, the output modality for it is set to the input modality of the feature.

5.2 User-defined Output

With the user-defined output strategy, the user can explicitly select the output modalities. Similar to the symmetric output planning strategy described above, output modalities can be flexibly selected for all semantic elements or for any

combination of individual semantic elements to be presented to the user. In this strategy, the used output modalities are independent of the current situation and input modalities. This allows, for example, the user to manually select her/his favoured output modalities. The user can additionally exclude the use of certain modalities by not selecting them. A disadvantage of this strategy is that each modification of the user's preferences requires manual intervention by the user.

5.3 Context-based Output

The context-based output strategy makes use of different *factors* in the user's current situational context to generate an optimal output. These factors lead to different system *reactions* that influence the generation of the output. This context-based planning method represents one of the main contributions of this paper, and is novel with respect to the type and number of context influencing factors, the type and number of resulting system reactions, and the approaches used in determining what reaction to take for a given set of identified context factors, i.e. the output planning (see figure 3).

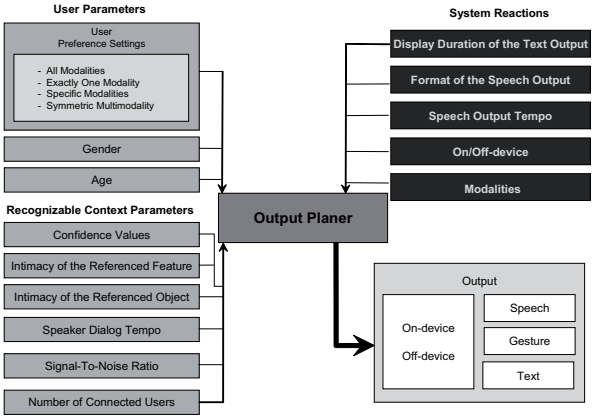

Figure 3. User parameters and system reactions for context-based output generation.

5.3.1 User and Context Parameters

This section describes different factors that can influence the final presentation of the output information. These factors can be classified into two groups: *user parameters*, which are specific to a given user, and *recognizable context parameters*, which are environment-specific.

Age: The age can be a criterion for the retentiveness of a user [5]. Elderly people might need more time for getting the presented information. Hence the text output should be visible for a longer period of time and the speech output tempo should be slower. The MSA system currently obtains the user's age from a user model managing system called UbisWorld [4].

Gender: The gender can for example influence the choice of the voice used for speech output. For instance, the system can choose a female voice for a male user and a male voice for a female user. Similar to the age parameter, gender is obtained through the UbisWorld service.

User Preference Modalities: In the context-based output planning strategy, the modalities selected by the user should be preferred to those not selected. The user can either select exactly one modality, or a combination of output modalities. The exclusive selection of only one modality is considered more important than a combined selection with only one modality. This means that the user prefers only this specific modality, so it should be used over other modalities.

Speaker Dialog Tempo: The speaker dialog tempo is calculated from the time that the user has taken to provide speech input for a pre-determined character length.

Signal-to-Noise Ratio: The signal-to-noise ratio (SNR) describes the signal strength compared to the background noise. The lower the SNR the more noise was recognized. In a noisy environment, the speech output might be hard to understand. To compensate for this, the system can raise the volume of the speech output or use an additional modality, e.g. graphical output. The SNR is detected by the speech recognizer.

Confidence Values: Confidence scoring refers to the process of attaching likelihood values to recognition results in an attempt to measure the certainty of finding a correct match to a user's input. For each of the modalities, a confidence value (Cf) is generated each time a user interacts with the system [12, 10].

Intimacy of the Object: Different objects have different intimacy levels. The intimacy level of the object is often highly user-dependent. Examples of objects with high intimacy levels might be medications, cosmetics, or contraceptives. For such products, unobtrusive output modalities (e.g. graphical output on-device) should be used.

Intimacy of the Feature: Similar to object intimacy, there are also different intimacy levels for features. An example of a feature with a high intimacy value might be the size of a particular item of clothing.

Number of Nearby People: If many people are nearby to the user, it might be undesirable to use off-device output: On the one hand, speech output of different users could overlap, and on the other hand, the user could feel uncomfortable by the speech output. The MSA system estimates the number of nearby people based on the number of people currently localized in the vicinity (see [1]).

5.3.2 Heuristically-derived System Reactions

After detecting the possible factors that might influence the system's reactions, the system uses these factors to generate an appropriate output.

Display Duration of the Text Output: As a baseline for the duration of the text output, the user can determine a preferred value. However, the ultimate display duration can be increased by the system according to the user's age and the SNR value. On the other hand, it can also be decreased depending on the level of intimacy of the object and/or the feature.

Format of the Speech Output: The decision to choose either natural language or short output depends on the one hand on the user's parameters, like preferences, age, and speech input tempo and on the other hand on environmental

parameters. like SNR value and the number of nearby people. A higher age and a higher SNR value for example favour the natural language output, whereas a large number of nearby people and a high speech input tempo would rather lead to the generation of short output.

Speech Output Tempo: Similar to the choice of the speech output format, the speed of the speech output depends on the user's age and speech input tempo, and the number of nearby people.

On-/Off-device: Whether the output is presented on- and/or off-device depends on the user settings, the intimacy of the object and feature attributes, the number of nearby people, and the confidence values. The more people that are in the vicinity of the user, the more off-device output should be avoided. Information about objects or features with a high intimacy level should not be presented in the environment. Moreover, unconfident system responses should not be presented off-device. The on-device modality is selected, if it is explicitly preferred in the user settings or if no off-device output is allowed.

Modality Selection: The gesture output modality is selected if it is either explicitly preferred by the user or if gesture input for the object was used. It is difficult to foresee any reason for not selecting gesture output when the user explicitly prefers it, and hence this preference is never ignored. When the user applies gesture for input, the system responds with gesture for output.

In the case of speech output, the environmental context plays an important role. In certain situations, it seems sensible not to use acoustic output, even if it is selected as preferred in the user settings. First of all, we consider the number of modalities selected by the user: if more modalities are selected, the speech modality can be resigned more easily. If *symmetric multimodal* is selected, the used input modality is also taken into account. As we found in an empirical study, speech output is more preferred by female users. Probably the most important factors for the choice of the speech output modality are the signal-to-noise ratio, the number of nearby people, and the intimacy of the currently selected object/feature. Additionally, the certainty with which the object input was recognized also plays a role, such that a more unobtrusive modality should be used if the object might not have been recognized correctly.

Similar to with gesture output, there is no reason to deselect the text output modality if it is explicitly selected by the user. As stated in [5], elderly people can recognize graphical output better than acoustic output. Therefore, for elderly users, text output is always displayed in our system.

6 Conclusions

The Mobile ShopAssist has undergone a number of usability studies in the past, primarily concerned with user preference for modality combinations (see [13] and [11]). Current work is now focused on an additional field study aimed at determining the accuracy and suitability of the presented output strategies for mobile users in a shopping domain. From a pilot study that has recently been conducted on the relevance of individual context factors, it has already been found, for example, that the intimacy of an object's features is considered less important than the intimacy of the object itself. Most of the interviewed participants in this study also highly rated the importance that the number of nearby people would have in a system selecting the current optimal set of output modalities.

In this paper, we have presented a context-aware output planning module and three accompanying strategies used for output generation in a shopping domain, namely: *user-defined*, *symmetric multimodal*, and *context-based*. In support of these strategies a range of context parameters relating to the user and the environment were identified (e.g. the user's age and gender; signal-to-noise ratio; the number of nearby users). Additionally, a range of possible system parameters used in determining appropriate reactions to take when presenting semantic information over a given set of modalities was identified (e.g. duration in which text is displayed; the format and speed of speech output). The outlined context parameters and system reactions can be seen to provide vital insight for all systems with a research focus on context-aware computing. Future work will now entail testing the degree of suitability of the proposed output planning strategies.

REFERENCES

[1] B. Brandherm and T. Schwartz, 'Geo referenced dynamic Bayesian networks for user positioning on mobile systems', in *Proceedings of the International Workshop on Location- and Context-Awareness (LoCA)*, pp. 223–234, (2005).

[2] A. Butz, M. Schneider, and M. Spassova, 'SearchLight: A Lightweight Search Function for Pervasive Environments', in *Proceedings of the 2nd International Conference on Pervasive Computing (Pervasive)*, pp. 351–356, (2004).

[3] C. Elting, 'What are multimodalities made of? - modeling output in a multimodal dialogue system', in *Workshop on Intelligent Situation-Aware Media and Presentations ISAMP 2002*, Edmonton, Alberta, Canada, (2002).

[4] D. Heckmann, *Ubiquitous User Modeling*, Ph.D. dissertation, Department of Computer Science, Saarland University, 2005.

[5] J. Jorge. Adaptive tools for the elderly: new devices to cope with age-induced cognitive disabilities, 2001.

[6] M. Schmitz and A. Butz, 'Safir: Low-cost spatial audio for instrumented environments', in *Proceedings of the 2nd International Conference on Intelligent Environments*, pp. 427–430, (2006).

[7] L. Spassova, R. Wasinger, J. Baus, and A. Krüger, 'Product Associated Displays in a Shopping Scenario', in *Proceedings of the 4th IEEE / ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 210–211, (2005).

[8] W. Wahlster, 'Smartkom: Symmetric multimodality in an adaptive and reusable dialogue shell', in *Proceedings of the Human Computer Interaction Status Conference*, pp. 47–62, (2003).

[9] W. Wahlster, N. Reithinger, and A. Blocher, 'Smartkom: Multimodal communication with a life-like character', in *Proceedings of Eurospeech*, pp. 1547–1550, (2001).

[10] R. Wasinger, *Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations*, Ph.D. dissertation, Saarland University, Department of Computer Science, 2006.

[11] R. Wasinger, A. Krüger, and O. Jacobs, 'Integrating Intra and Extra Gestures into a Mobile and Multimodal Shopping Assistant', in *Proceedings of the 3rd International Conference on Pervasive Computing (Pervasive)*, pp. 297–314, (2005).

[12] R. Wasinger, C. Stahl, and A. Krüger, 'Robust Speech Interaction in a Mobile Environment through the use of Multiple and Different Media Input Types', in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1049–1052, (2003).

[13] R. Wasinger and W. Wahlster, 'Multi-modal Human-Environment Interaction', in *True Visions: The Emergence of Ambient Intelligence*, 291–306, (2006).