In: Crocker, M. W.; Siekmann, J. (Eds.): Resource-Adaptive Cognitive Processes, Heidelberg, Berlin, Springer, 2010, pp.205-237

The Shopping Experience of Tomorrow: Human-Centered and Resource-Adaptive

Wolfgang Wahlster, Michael Feld, Patrick Gebhard, Dominikus Heckmann, Ralf Jung, Michael Kruppa, Michael Schmitz, Lübomira Spassova, and Rainer Wasinger

1 Introduction

What would the shopping experience of tomorrow look like? In this chapter we propose several human-centered and resource-adaptive ideas to this question. Throughout the whole chapter we explain our ideas with the recurrent theme of a shop that consists of instrumented shelves, public displays, audio systems, and mobile devices for each user. The shelves are fitted with RFID antennas and allow for sensing implicit user interactions with RFID-labeled objects, such as picking up a product or putting it back into the shelf. We will present the novel interaction paradigm of "Talking Objects", which involves multimodal interaction with instrumented objects, spanning the modalities of speech, gestures, sound and haptics. Imagine talking objects in shopping malls with which individuals or groups are able to interact. This means accomplishing shopping tasks by offering an intuitive interface to a complex environment. Furthermore, these talking objects will be associated with personalities by the means of controlling speech attributes and behavior. In addition to this anthropomorphism, we will provide these objects with the abilities to sense their state, e.g., whether they are in or outside the shelf, or whether a user is turning, squeezing, or shaking them. The novel concept of "Product Associated Displays" is a way of providing visual feedback to users interacting with physical objects in an instrumented shop. These projected public displays are created at locations that can be intuitively associated with the objects they show information about. Furthermore, a life-like character lives as a "Virtual Room Inhabitant" in our smart shop. The novel concept of "Personalized Ambient Audio Notification" describes a notification service that allows users to monitor information with less distraction of attendees in their surrounding. The ambient notification service works with personalized non-speech audio cues that can be embedded in aesthetic background music depending on the event and the current position of the user. Areas of applications are shops where employees can receive information (e.g., a cashier

W. Wahlster (✉)
DFKI GmbH and Department of Computer Science, Saarland University, 66123
Saarbrücken, Germany
e-mail: wahlster@dfki.de

M.W. Crocker, J. Siekmann (eds.), *Resource-Adaptive Cognitive Processes,*
Cognitive Technologies, DOI 10.1007/978-3-540-89408-7_10,

is needed in the point of sale area) without arousing the customer's attention. At the same time the background soundscape has a comfortable effect on customers. Another important point of a shopping activity is the preparation of a shopping list, which helps the user to remember the things that need to be bought. We will present an implemented web-based agenda, where users or user groups can enter tasks, such as buying certain articles for a party. Typical existing organizers, such as PDAs or smartphones, only provide an alarm function to remind one user at a certain time. Our novel "Ubiquitous Agenda Service" allows the user to specify a place, or select a semantic category, such as a supermarket or brand. As soon as the user is nearby one of the specified places, a location-aware mobile device can present a reminder. Inside the shop, public displays will recognize the mobile device wirelessly via bluetooth and adapt their advertisement to the user's shopping list and general interests. In this chapter, we will put a focus on the role of cognitive and affective states for the adaptation of information presentation in instrumented environments. We will present how to recognize the resources of users with specialized "Dynamic Bayesian Networks" that probabilistically estimate the cognitive load and the time-pressure of users on the basis of their symptomatic behavior and physiological data that is derived from bio-sensors that measure for example the heart rate, the muscle tension, the electrodermal activity, or the eye movements. Finally, we will present an ontological approach to model and share the limited cognitive resources of users between different resource-adaptive applications. The "Ubiquitous User Model Service" provides contextual information on the users' actions, characteristics, and locations, while the users are enabled to access and control their profiles via a sophisticated web interface which integrates the necessary privacy issues.

1.1 Overview Described Within a Motivating Scenario

In order to get an impression of what shopping might look like in the future, imagine a fictitious shopping scenario in an instrumented environment in which Mrs. Smith and her husband are consumers. In preparation for her shopping, Mrs. Smith creates an electronic shopping list using a web interface. At the entrance of the supermarket, Mrs. Smith connects to her current shopping list with the tablet PC mounted at the handle of her shopping cart. She is navigated through the supermarket by an indoor navigation system (see chapter *Seamless Resource-Adaptive Navigation*). Meanwhile, Mr. Smith remembers that he has invited a friend for dinner and recognizes that he has no more wine at home. Thus he adds the entry "some French wine" to his wife's electronic shopping list which immediately appears on the screen of her shopping cart. When she sees the new entry, Mrs. Smith heads for the wine department to which she is guided by a projected virtual character that moves along the shelves and walls of the supermarket (see Sect. 6). When she enters the wine department, Mrs. Smith notices an elderly lady standing in front of the interactive wine information kiosk asking for some wine that suits her taste. On the basis of her speech input, the kiosk system recognizes that the questioner is an elderly woman and recommends preferably sweet wines in a slow and comfortable voice (see Sect. 7). As Mrs. Smith does not have much knowledge of wine (this information can be retrieved from an ubiquitous user model, see Sect. 8) and the kiosk is already occupied, she uses the Mobile ShopAssist on her PDA to get some information about the different wines she considers buying (see Sect. 3). The Mobile ShopAssist monitors the user's choice and matches her actions to an affective state model (see Sect. 9). Beside other interaction modalities for the system output, a wine bottle that Mrs. Smith takes out of the shelf can "answer" her questions explaining her its features (see Sect. 2). In this way, Mrs. Smith can learn more about the specific features of the wines and compare them with each other. The required information can also be presented visually on projected displays that automatically appear at the back surface in the shelf when a bottle is taken out of it (see Sect. 4). If Mrs. Smith is still undecided which wine to buy after some time, an ambient sound notification system seamlessly informs an employee of the shop who is a wine expert that there is a customer in the wine department who might need some help (see Sect. 5).

2 Dialogue Shell of Talking Products

One important design goal of our interactive shopping assistance is to support arbitrary users, particularly computer novices, who are not able or willing to learn the use of such a system. We therefore have to find a solution that provides a natural interaction, requiring minimal effort of a user to understand and utilize the assistance system. Nijholt et al. [43] suggest that a limited animistic design metaphor seems to be appropriate for human–environment interaction with thousands of networked smart objects. People often tend to treat objects similar to humans, according to findings of Reeves and Nass [50], which allows users to explain the behavior of a system if they lack a good functional conceptual model. In consequence, we decided to employ a natural language system, which enables the user to talk to each product.

Our group conducted a usability study of a multi-modal shopping assistant [62]. The implemented system allows users for instance to request product information in a combination of speech and selecting gestures (i.e., taking a product out of the shelf). Findings of this study showed among others that users generally preferred direct over indirect interaction, i.e., by asking "What is your price?" instead of "What is the price of this camera?" which encouraged us to pursue this approach.

Previous studies have shown that interacting with embodied conversational agents that have consistent personalities is not only more fun but also lets users perceive such agents as more useful than agents without (consistent) personalities [42, 20]. It is further shown that the speech of a consistent personality enables the listener to memorize spoken contents easier and moreover reduces the overall cognitive load [23, 42]. Thus we emphasized the anthropomorphic aspect of this interaction pattern by assigning personalites to products, which are reflected by the spoken responses of a product.

Product manufacturers benefit as well, since the personalization of the product provides a new channel to communicate a brand image or distinct attributes of a certain product. A study within the context of marketing research showed that if in radio advertisements a voice fits the product, it helps the listener to remember the brand, the product and claims for that product [44].

2.1 Modelling Personality in Voices

In a first step we created voices that reflect certain personalities according to Aaker's brand personality model [1] only by adjusting prosodic parameters. We chose this model over the (rather similar) five factor model [37] commonly used in psychology, since we are applying the concept of talking objects in the shopping domain. However, both models are rather similar and to a certain extent exchangable.

We changed the four prosodic parameters pitch range (in semitones), pitch level (in Hz), tempo (as a durational factor in ms), and intensity (soft, modal, or loud as in [57]) according to our literature review [55]. For example, a competent voice has a higher pitch range (8 semitones), a lower pitch level (–30%), a 30% higher tempo, and a loud intesity compared to the baseline voice. In [55] we also evaluated whether it is possible to model different personalities with the same voice by adjusting these prosodic parameters, such that listeners will recognize the intended personality dimension. The study has shown that there are clear preferences for our prosody-modeled speech synthesis for certain brand personality dimensions. But not all personality dimensions were perfectly perceived as intended, such that we have to amplify the effect.

Personality is certainly not only expressed in qualitative attributes of a voice, other properties of a speech dialogue are also essential, like the used vocabulary or the general discussion behavior. For this reason we created a dialogue shell that incorporates these aspects.

2.2 Expressing Personality in Dialogues

The widely adopted personality model by Costa and McRae [37] constitutes five dimensions of human personality: Extraversion, Agreeableness, Conscientousness, Neuroticism, and Openness on a scale from 0 to 100. Obviously, differentiating 100 levels in a dimension is far too much for our goals, therefore we simplified this model by discriminating three levels in each dimension:

- low: value between 1 and 44 (31% of population)
- average: values between 45 and 55 (38% of population)
- high: values between 56 and 100 (31% of population)

Related work, e.g., by Andre et al. [2] limited their personality modeling to only two of the five dimensions, namely extraversion and agreeableness, since these are the most important factors in interpersonal communication. Nevertheless, we discovered considerable influences of openness and conscientousness to speech, therefore we incorporated these two dimensions as well. The effect of the dimension neuroticism is mainly to describe the level of susceptibility to strong emotions, both positive and negative ones [17]. It is further shown that the level of neuroticism is very hard to determine in an observed person [26]; thus we decided that four dimensions will suffice for our work.

We conducted an exhaustive literature review on how speech reveals different personality characteristics. Among numerous other resources, two recent research papers provided essential contributions to our work: Pennebaker and King's analysis in *Journals of Personality and Social Psychology* [48] and Nowson's *The Language of Weblogs: A Study of Genre and Individual Differences* [45]. In both studies a large number of text blocks were examined with an application called Linguistic Inquiry and Word Count[1] (LIWC), which analyzes text passages word by word, comparing them with an internal dictionary. This dictionary is divided into 70 hierarchical dimensions, including grammatical categories (e.g., noun, verb) or affective and emotional processes. Pennebaker determined in a study the 15 most reliable dimensions and searched for them in diary entries of test persons with LIWC. With these results together with the given personality profiles of the probands (according to the five factor model), he identified correlations between the two. Nowson performed a similar study and searched through weblogs for the same LIWC factors.

Based on these results, we provided a set of recommendations on how responses of a talking object with a given personality should be phrased. For instance, for a high level of extraversion these recommendations are given:

- Prefered bigrams: *a bit, a couple, other than, able to, want to, looking forward*, and similar ones.
- Frequent use of terms from a social context or describing positive emotions
- Avoidance of *maybe, perhaps*, and extensive usage of numbers
- Usage of colloquial phrases, based on verbs, adverbs, and pronouns
- Comparably more elaborate replies

Following these principles we implemented basic product responses (greetings, inquiries for product attributes, farewell) for several personalities. All possible replies of our dialogue shell are stored in one XML-file, which we named the *Anthropomorphic Fundamental Base Grammar*. All entries include an associated personality profile, for example:

```
<reply
     query="hello"
     reply="Hello, nice to meet you!"
     ag="1" co="2" ex="1" op="1">
<\reply>
```

[1] http://www.liwc.net/

which means that this is the greeting of a product with average agreeableness, extraversion, and openness and a high value in conscientousness. Another example:

```
<reply
    query=”hello”
    reply=”Hi! I’m sure I can help you! Just tell me
        what you need and I bet we can figure
        something out!”
    ag=”2” co=”2” ex=”2” op=”2”>
<\reply>
```

All entries that do not regard any particular personality should have average personality values in all dimensions.

A central product database with all products and their attributes is extended by the assigned personality profile, i.e., the values in each of the four dimensions. When the application starts up, the dialogue shell retrieves the product data of each product instance and extracts the appropriate entries from the base grammar to build the custom product grammar. If there are no entries that exactly match the given profile, the one that has the most identical values will be chosen. This dialogue shell generates a consistent speech interface to a product by knowing its attributes and a given personality profile, for instance preset by the manufacturer.

3 Mobile ShopAssist

The Mobile ShopAssist (MSA) is a platform originally designed to demonstrate a wide range of different multimodal interaction possibilities in everyday contexts, and particularly those contexts in which a user is mobile [61]. Created for use in mobile and ubiquitous environments, the ShopAssist application allows shoppers, accompanied by a PDA, to enquire about product features and to compare different products with one another. This is achieved through the use of input modalities like speech, handwriting, and selection gestures. Figure 1 shows the ShopAssist application in use during field studies conducted at Conrad Electronic in Saarbrücken.

Since its conception, the MSA has become a test bed for a number of research focuses including mobile multimodal interaction, on- and off-device input recognition, on- and off-device presentation output planning, anthropomorphisation, and public associated displays. The architecture of the platform, as implemented for demonstration of mobile multimodal interaction, can be seen in Fig. 2.

Multimodal interaction refers to “the means for a user to interact with an application using more than one mode of interaction” [60]. Such interaction might occur sequentially or simultaneously in time, and may also contain semantically overlapped information in which certain semantic constituents (such as a shopping product’s price) is provided multiple times by similar or different modalities (such as speech and handwriting).

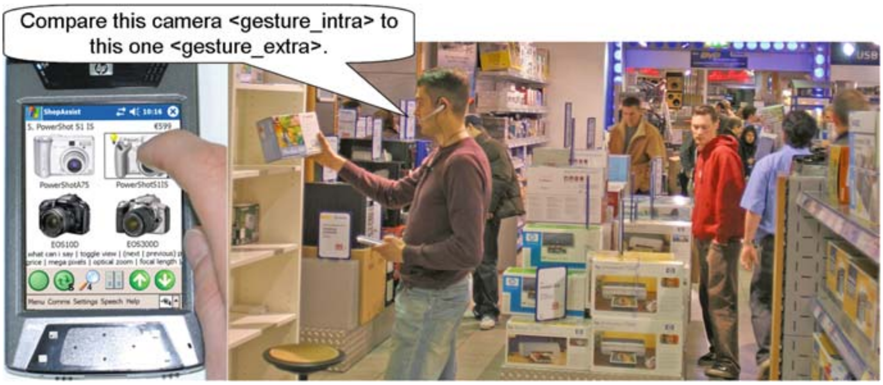

Compare this camera <gesture_intra> to this one <gesture_extra>.
ShopAssist
S. PowerShot S1 IS
€599
PowerShotA75
PowerShotS1IS
EOS10D
EOS300D
what can i say | toggle view | (next | previous)
price | mega pixels | optical zoom | focal length
Menu Comms Settings Speech Help

Fig. 1 Mobile ShopAssist interaction, as used during field studies at Contrad Electronic in Saarbrücken

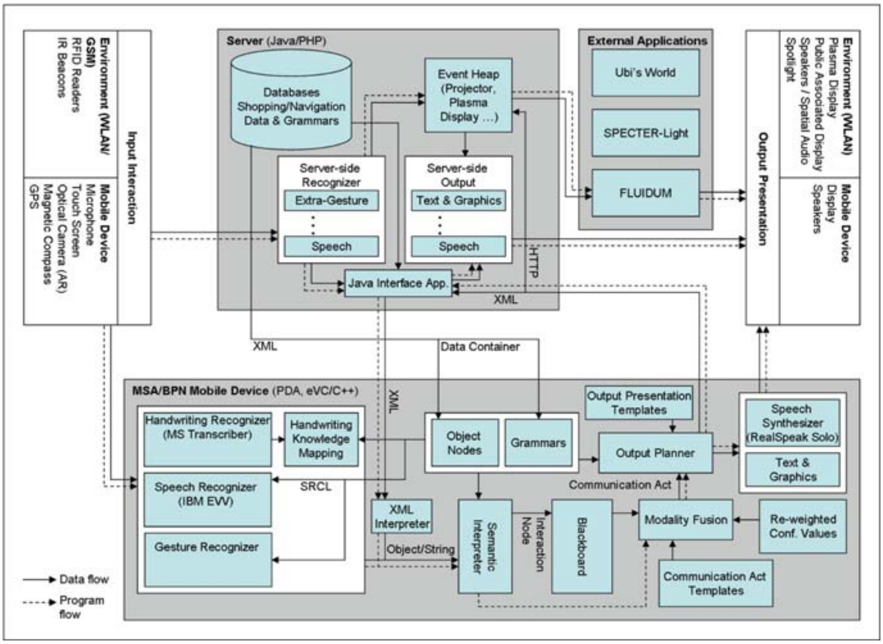


Fig. 2 The MSA architecture showing the data flow between different components

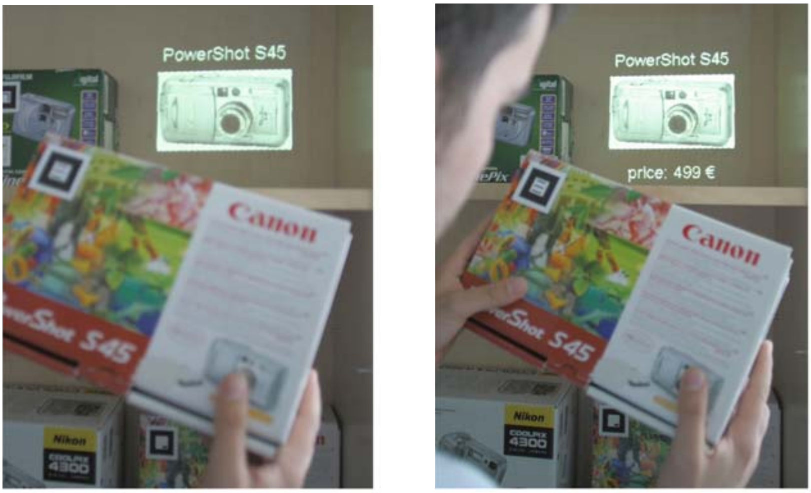
The MSA supports an Always Best Connected (ABC) methodology such that interaction components located in a publicly instrumented environment, for example, distributed speech recognizers and gesture recognizers, can be made accessible to a user for the purpose of enhanced application functionality like improved recognition accuracy and support for larger vocabularies. This adaptation to available resources (i.e., recognition results from multiple recognizers) has been made possible through results from field studies that were conducted to determine correlations between recognizer confidence and recognition accuracy. The benefit of such correlations is that they allow for recognizers that inherently process signals differently (e.g., speech and handwriting) to have their results compared with one another. This also applies to same-type recognizers (e.g., server-side and embedded speech engines) in which confidence values are generally based on entirely different factors like acoustic models, grammars, and associated computing power.

Similar to the recognition of user input, output presentation planning in the MSA is also resource-adaptive. In particular, user parameters from UbisWorld (e.g., age, gender, and modality preference) as well as context parameters (e.g., spoken dialogue tempo, SNR, ambient light level, and number of surrounding people) are used to determine particular system reactions. Typical reactions of the system are for example the display duration of textual output, the format and tempo of speech output, and whether output is to be presented on-device or off-device (i.e., on the mobile PDA, or on devices located in the surrounding instrumented environment).

4 Product Associated Displays

PADs (Product Associated Displays) are projected virtual displays which are created at locations that can intuitively be associated with the products the user is currently interacting with. Instead of displaying product information on a stationary screen, which can be installed, e.g., beside the product shelf, and which the user might not be aware of because of the spatial distance to the product, a projected PAD presents the relevant product information in the gap left in the shelf when the product is taken out of it. As in the process of taking a product, its former location in the shelf remains in the user's peripheral view, a PAD that occurs there immediately after the action is very likely to catch the user's attention. In this way, a spatial mapping between the physical location and the displayed information is established and a relationship between the product and the corresponding information on the PAD arises automatically.

In our shopping scenario, PADs are projected using the Fluid Beam system [59]. Its hardware part consists of an LCD projector and a digital camera placed in a moving yoke in which they can be rotated horizontally (pan) and vertically (tilt). In this way, the projector beam can be directed at almost any surface in the room. In order to avoid image distortion due to oblique projection, the Fluid Beam software implements a method described in [49]. It is based on the fact that projection is a geometrical inversion of the process of taking a picture given that the camera and the projector have the same optical parameters and the same position and orientation. The implementation of this approach requires an exact 3D model of the environment, in which the projector is replaced by a virtual camera. By synchronizing the movements of the steerable projector in the physical environment and the virtual camera in the 3D model, the image delivered by the virtual camera appears undistorted when it is projected in the physical environment. Thus virtual displays showing images, videos, or video streams can be placed in the 3D model and they are projected at the corresponding locations in the physical world when the virtual camera (and respectively the steerable projector) is directed at them. In this way, a sort of virtual layer is created that covers the surfaces of the physical environment, on which projected virtual displays can be placed and moved.



(a) Initial display showing a picture of the product and its name

(b) Price information displayed on a PAD as an answer to the user's request

Fig. 3 Product associated display

The event of taking a product out of a shelf or putting it back is recognized by means of passive RFID tags attached to the products and an RFID antenna placed behind the shelf. If the user takes out a product, this action is recognized by the system and a corresponding event is generated and sent to an Event Heap [29]. The Mobile ShopAssist [63] receives the event from the heap and sends a command to the steerable projector to display a PAD at the appropriate location showing a picture of the removed product and its name (see Fig. 3a). After that, the user is given the opportunity to ask for additional information about the product (e.g., price) using speech, handwriting, or intra-gestures on his or her PDA (see Sect. 3). The answer to the user's request can then be displayed on the PAD if this is allowed by the user's preference settings (see Fig. 3b).

5 Personalized Ambient Soundscape Notification

In most instrumented environments the visual sense is the primarily used of all human senses. Usually, audio signals are limited to simple warning cues and system feedbacks that are in most cases intrusive because of their dissimilarity compared to the environmental noise. That has the effect that persons present in the room will be distracted from their current tasks. To prevent the disturbing effect of traditional notification signals we developed the novel concept of non-speech audio notification embedded in ambient soundscapes to provide a method for multi-user notification in a more discreet and non-disturbing way.

5.1 Introduction to Ambient Audio Notification

In 1978, English musician Brian Eno coined the term ambience in combination with music in the notes to his longplayer *Ambient 1: Music For Airports*. This type of music has a calming effect and can be listened to either actively, that means the focus of attention lies on the music, or it can be listened to peripherally without paying attention to the music. This effect is also known as the auditive figure-ground phenomenon which decribes human's ability to pay attention to an auditory stream (figure) while at the same time any other sound is listened to peripherally (ground) [10]. Already in the year 1953, Colin Cherry described this effect in his famous "cocktail party" experiment when he found out that the auditive perception is associated with the attention of a person [16]. The allocation of the limited resource attention depends on a variety of factors like the stimuli that act on the person and his current mental and social conditions [21].

Perception of auditive signals can be divided into the physiological phenomenon of hearing and the semantic sound processing which leads to the personal interpretation of the signal, influenced by the experiences of each individual listener. The intensity and complexity of environmental noises influence whether we perceive a single sound or whether it is masked which depends on multiple factors like loudness and the frequency of the noises. Traditional audio notification signals are mostly stand-alone cues that attract the attention of everybody in a room because they are not integrated into the natural sound environment [46]. That works fine for high-priority notifications (e.g., fire alarm), but often a more personal and discreet notification is desirable.

We had two main goals for the design of our notification signals. On the one hand we want to seamlessly integrate the notification signal into background music without arousing the attention of other people, but on the other hand the target person must become aware of the signal.

Auditory experiences can be permanently extended and trained [3]. We use this fact to make the listener more sensible to his specific auditory signals that we use for attracting his attention. These audio cues are used to provide the listener with information that he links with the specific auditory signal. The user can choose which sound he wants to link with which information, so we get an individual and personalized notification that respects the user's preferences.

Since only the user knows which sound he selected for which information, this type of notification also slightly fulfills the privacy aspect.

5.2 Ambient Soundscapes and Audio Notification Cues

The main problem with traditional stand-alone notification signals is the distraction of other present persons, especially in multi-user environments. Indeed, popular non-speech audio cues like earcons [6] and auditory icons [11, 14] can provide a perceptible type of notification but they are also separated from environmental noise.

To introduce more privacy and confidentiality, we decided to integrate notification instruments with respect to the musical compositions seamlessly into background music, the *ambient soundscape*, which serves as the musical envelope [13]. Instead of attracting the listener's attention, the soundscape should have a calming and mood influencing effect (see also [4, 35]).

To reach this goal we composed and recorded three ambient soundscapes and suitable notification instruments by ourselves. We took some perceptual constraints such as the auditive Gestalt laws and several studies dealing with musical perception into consideration as described in [15, 51, 52, 19]. Table 1 gives a brief overview of the emotional impact of compositional parametes on the listener's mood. In our shopping scenario, the ambient soundscapes create a more friendly atmosphere for consumers.

Table 1 Categorization of musical parameters, including range and emotional impact [12]

Category	Parameter	Range	Emotional impact
Time	Speed	fast – slow	pleasant – calm
	Phrasing	staccato – legato	lively – gently
	Rhythm	firm – smooth	serious – dreamy
	Dynamic	cresc. – decresc.	animated – relax
	Meter	even – odd	dignified – restless
Pitch	Mode	major – minor	bright – plaintive
	Frequency	high – low	exciting – sad
	Melody	ascending – descending	dignified – serene
	Note Range	≥octave – ≤octave	brilliant – mournful
	Harmony	consonant – dissonant	serene – ominous
Texture	Volume	forte – piano	animated – delicate
	Orchestration	instrumentation	majestic – grotesque

In the second phase we add *notification instruments* to the ambient basic soundscape and play them with slightly increased volume at the current position of the task person by using an indoor positioning system [58] and a spatial audio framework [54]. The *Always Best Positioned* (ABP) mobile localization system called *LORIOT* uses RFID technology in combination with infrared beacons to find out what the user's current position is. The calculation is done on the PDA by using Dynamic Bayesian Networks (DBN's) [8]. More information about the positioning system can be found in the chapter "Seamless Resource-Adaptive Navigation" of this book. *SAFIR* (Spatial Audio Framework for Instrumented Rooms) is used to play the audio cues at the loudspeaker that is the nearest to the target person's postition.

Since the notification instruments will be seamlessly integrated in the ambient soundscape this has the effect that an occurring notification could be perceived after a while. To prevent the effect of ignoring a notification, we also provide a hierarchy of notification signals that are grouped by "level of intrusiveness" [33], depending on the importance of the occurring event.

1. High-Priority
   Signals: Arousing Noises (e.g., beep, siren, and bell).
   Immediate and intrusive notification which is independent of the current compositional context.
2. Medium-Priority
   Signals: Ambient Noises (e.g., birds, rain, water- and wind noises).
   Immediate and context independent, but still an ambient notification with natural sounds.
3. Low-Priority
   Signals: Notification Instruments (e.g., drums, cymbals, guitar, piano, and violin).
   Seamless integration of melodic patterns, played by natural instruments, into the ambient soundscape (compositional-context-awareness).

The user can choose a soundscape that matches his personal preferred music style and an instrument or ambient noise that he can easily recognize. His personal preferences can be stored in his user profile of UbisWorld, a user model ontology [28] that is also described in this chapter. The profile information can be accessed by the notification system via http request.

The effectiveness of the peripheral perception was successfully tested in a user study with 25 persons [30] where we especially checked whether the users percept the notification instruments and the elapsed time to recognize the notification (delay time). The study was subdivided into a computer-based test and a questionnaire to get a subjective and personal feedback of the participants' opinion about the soundscapes and this new concept of notification.

5.3 Applications and Shopping Scenario

The *Personal Ambient Audio Notification* service (PAAN) handles an audio server, the data exchange to UbisWorld, the indoor positioning system (LORIOT), and the spatial audio system (SAFIR). Figure 4 shows the hardware that we use for PAAN. The audio hardware includes Hi—Fi amplifiers and loudspeakers that are connected to a multi-channel soundcard. For a scenario where the user changes his position, we use the Always Best Positioned system LORIOT that uses a PDA equipped with wireless LAN and an RFID reader card, active RFID tags that are mounted on the ceiling of the room and optional infrared beacons mounted at shelves or walls.

In our shopping scenario, the ambient soundscape can be selected by an employee of the wine store by browsing available soundscapes in the web interface on his computer. The soundscapes are stored on an audio server and managed by an audio database. Search queries for the audio database can include the name of the soundscape or *GEKOS*[2]-keywords that describe each soundscape by its compositional elements. Employees of the store can change their personal audio notification

[2] GEKOS: **G**enre, **E**xpression, **K**ey, **O**rchestration, **S**ignature

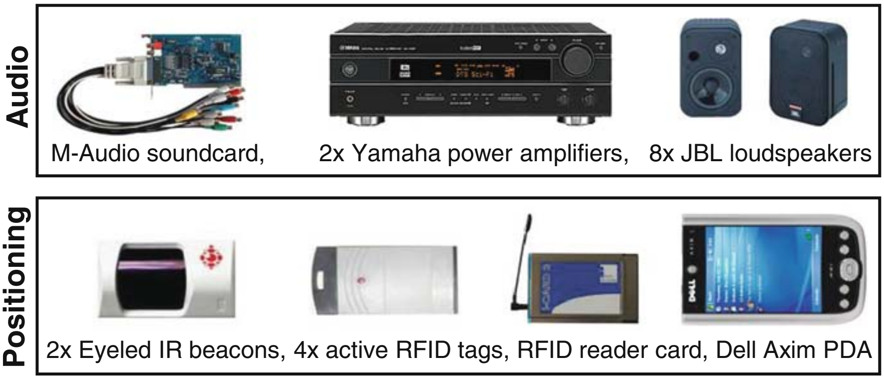
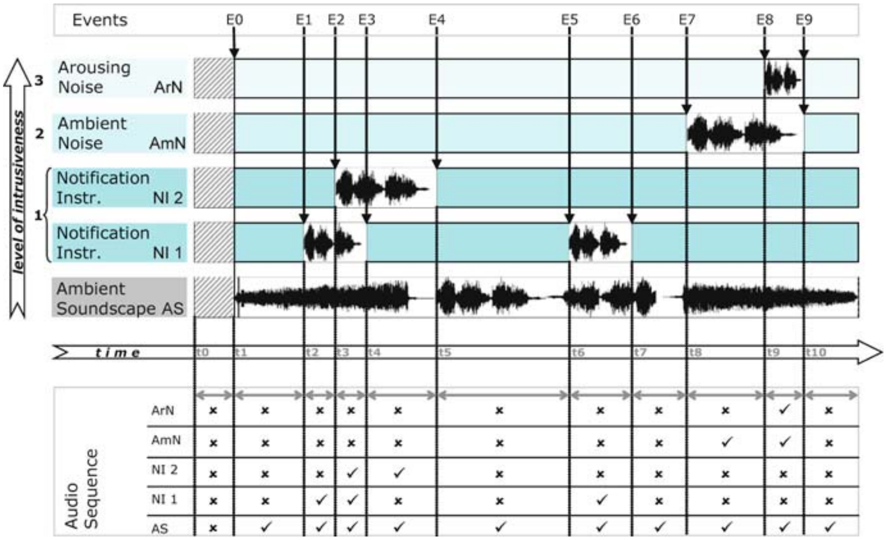


Fig. 4 Audio- and positioning hardware requirements for PAAN

instrument by updating their UbisWorld account. The IP address of the user's PDA can also be specified and exported in an XML file which can be accessed by PAAN via http request to route an event notification to the appropriate task person [31].

Figure 5 shows an audio sequence of a possible shopping scenario with two selected Notification Instruments (NI 1, NI 2) assigned to two employees, an Ambient Noise (AmN) for group notification and an Arousing Noise (ArN) for high-priority notification. The notification service reacts to relevant events (E_i) by mixing the adequate notification in the playing soundscape at the right time [32].

The Ambient Soundscape (AS) starts automatically when a registered user, namely an employee, enters the instrumented area of the shop with his PDA ($E_0(t_1)$).



ArN	×	×	×	×	×	×	×	×	×	✓	×
AmN	×	×	×	×	×	×	×	×	✓	✓	×
NI 2	×	×	×	✓	✓	×	×	×	×	×	×
NI 1	×	×	✓	✓	×	×	✓	×	×	×	×
AS	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Fig. 5 Audio sequence example

The preselected notification signals that are assigned to authorized users are in stand-by mode (muted).

Assume a consumer who enters the wine department of a supermarket with the intention of purchasing a French red wine while at the same time the employees are out of sight and do not notice the appearance of the potential consumer. After his arrival, the customer ($E_1(t_2)$) can be detected for example by the instrumented shopping cart [56] or a location-aware PDA [58]. The notification system determines the salesman's current position by checking the positioning coordinates of his PDA, matches them to the nearest loudspeaker and informs him about the presence of a person by starting to play his personal notification instrument (NI 1) with slightly increased volume at his position. The salesman notices that his instrument (e.g., piano) starts playing in the soundscape and that this is the appointed signal for a potential consumer in the wine department. Back in the wine area, it turns out that he cannot satisfy the consumer's wish because he is looking for a specific red wine and needs the advice of a wine specialist focused on French red wines. Thereupon, the salesman calls the specialist by starting NI 2 (e.g., drums), which is the signal for the French wine specialist to come to the wine department ($E_2(t_3)$). After whose arrival, the salesman and the specialist stop their notification instruments by pressing a GUI button on their PDAs ($E_3(t_4)$, $E_4(t_5)$). The instruments leave the music seamlessly and only the basic soundscape is still playing. Event $E_5(t_6)$ describes a similar scenario where the salesman receives a call to come to the department where he stops the notification after his work is done.

Event $E_7(t_8)$ is an example for group notification and could occur if the head of the wine department wants to call his colleagues for a meeting by sending an email to the department's mail account. The Ambient E-Mail Notification system (*AEMN*) periodically checks the mail server for incoming messages and filters them for predefined keywords. The important announcement email triggers an event with the corresponding group notification signal in the form of an ambient water noise (AmN) which immediately starts playing in the whole department. Unfortunately, not every staff member noticed the ambient notification after a while, so AEMN sets the level of intrusiveness to the highest level and starts playing an additional Arousing Noise (ArN), e.g., a beep sound ($E_8(t_9)$). After the arrival of the remaining employees, the two notification sounds were stopped on the PDA or on the department's desktop PC at time t_{10}.

The introduced personalized ambient notification is an effective and non-intrusive concept to provide users with information location-aware and under low-privacy aspects.

Now, music is no longer a pure emotion mediator, but rather contain emotion and content, whereby the sum of these two factors results in the information content of the music.

6 Virtual Room Inhabitant

The Virtual Room Inhabitant (VRI) is a virtual character capable of guiding and following a user throughout physical spaces. The main purpose of the VRI in our shopping scenario is to welcome the user when entering the shopping area and to guide the user to a particular shelf within the room. Unlike traditional virtual characters, the VRI is not limited to the narrow boundaries of a display or a fixed projection. Instead, the VRI is free to move along arbitrary surfaces in its surroundings and to appear at any location that will allow for a projection. From a technical point of view, the VRI utilizes a steerable projector (as described in Sect. 4) and a spatial audio system (see Sect. 2). This combination allows to visually locate the VRI within physical spaces and to locate the origin of the characters voice at the same location, hence conveying the impression of the character actually standing at that location.

In order to allow the VRI to follow the user throughout the room, it is necessary to sense the users location and to react to the users movements accordingly. The position is acquired using a positioning technology based on PDA localization. The position is determined by using a combined system of active RFID tags and infrared beacons. Since the infrared beacon technology demands a direct line of sight between sender and receiver, we also get a rough estimate of the users orientation (the details of the indoor positioning system are described in [9]). The positioning information is stored on the so-called Event Heap which implements a blackboard architecture and allows clients to post and retrieve data by signing in to particular information channels. The character engine which constitutes the central part of the VRI registers itself to the positioning information channel on the Event Heap. If the user enters a particular region in the shop, the situation is recognized by the character engine. In order to start the VRI, the character needs to get access to the necessary hardware. Therefore, it posts a request to the presentation manager which, in combination with a device manager, grants access to all registered devices (see Fig. 6 for details). Each device in our hardware set up has to register itself and services it offers at the device manager. The presentation manager handles all device requests from concurrently running applications in the environment. The presentation manager decides, whether a particular user may access a device or

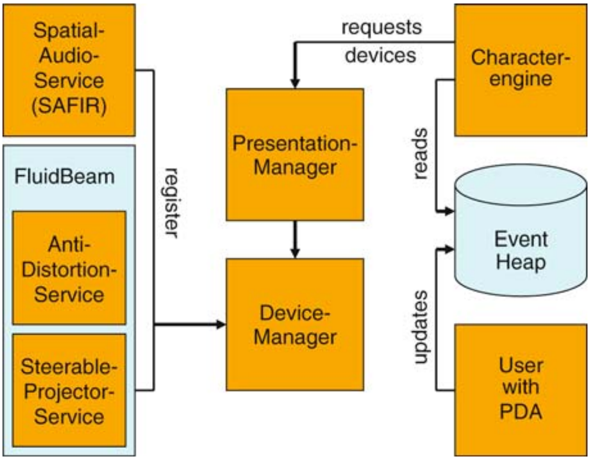


Fig. 6 The system components of the Virtual Room Inhabitant

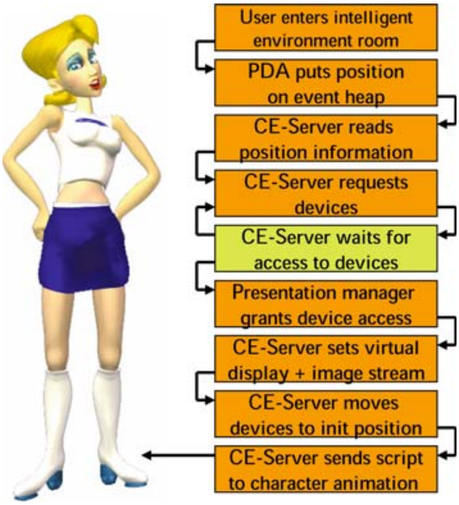

User enters intelligent environment room
PDA puts position on event heap
CE-Server reads position information
CE-Server requests devices
CE-Server waits for access to devices
Presentation manager grants device access
CE-Server sets virtual display + image stream
CE-Server moves devices to init position
CE-Server sends script to character animation

Fig. 7 The initialization sequence for the Virtual Room Inhabitant

whether the request is denied or delayed. The remote access mechanism is realized with Java Remote Method Invocation[3] objects, which allow arbitrary applications to control remote devices as if they were locally connected. The whole device access mechanism is depicted in Fig. 7. The character engine consists of two parts, namely the character engine server (CE-server) written in Java and the character animation, which was realized with Macromedia Flash MX.[4] These two components are connected via an XML-Socket-Connection. The CE-server controls the Flash animation by sending XML commands/scripts. The Flash animation also uses the XML-Socket-Connection to send updates on the current state of the animation to the CE-server (i.e., whenever a part of an animation is started/finished). The character animation itself consists of ∼9,000 still images rendered with Discreet 3D Studio Max[5] which were transformed into Flash animations. To cope with the immense use of system memory, while running such a huge Flash animation, we divided the animation into 17 subparts. While the first consists of default and idle animations, the remaining sixteen are combinations of character gestures, like for example, shake, nod, and look behind. Each animation includes a lip movement loop, so that we are able to let the character talk in almost any position or while performing an arbitrary gesture. We have a toplevel movie to control these movie parts. Initially, we load the default movie (i.e., when we start the character engine). Whenever we have a demand for a certain gesture (or a sequence of gestures), the CE-server sends the corresponding XML script to the toplevel Flash movie which than sequentially

[3] http://java.sun.com/products/jdk/rmi/

[4] http://www.macromedia.com/software/flash/

[5] http://www4.discreet.com/3dsmax/

loads the corresponding gesture movies. The following is a short example of an XML script for the character engine:

```
<VRI-script>
    <script>
        <part>gesture=LookFrontal sound=welcome1.mp3</part>
        <part>gesture=Hips sound=welcome2.mp3</part>
        <part>gesture=swirl sound=swirlsound</part>
    </script>
    <script>
        <part>gesture=LookFrontal sound=cart.mp3</part>
        <part>gesture=PointDownLeft sound=cart2.mp3</part>
        <part>gesture=swirl sound=swirlsound</part>
    </script>
    <script>
        <part>gesture=PointLeft sound=panel.mp3</part>
        <part>gesture=swirl sound=swirlsound</part>
    </script>
</VRI-script>
```

Each script part is enclosed by a script tag. After a script part was successfully performed by the VRI animation, the CE-server initiates the next step (i.e., move the character to another physical location by moving the steerable projector and repositioning the voice of the character on the spatial audio system, or instruct the VRI animation to perform the next presentation part). In order to guarantee a smooth character animation, we defined certain frames in the default animation as possible exit points. On these frames, the character is in exactly the same position as on each initial frame of the gesture animations. Each gesture animation also has an exit frame. As soon as this frame is reached, we unload the gesture animation, to free the memory, and instead continue with the default movie or we load another gesture movie, depending on the active XML script.

In addition to its animation control function, the CE-server also controls the spatial audio device, the steerable projector, and the antidistortion software. The two devices, together with the antidistortion software are synchronized by commands generated by the CE-server, in order to allow the character to appear at any position along the walls of the room, and to allow the origin of the character's voice to be exactly where the character's visual representation is.

Presentations are triggered by the user's movements. As soon as a user enters the instrumented room, the CE-server recognizes the relevant information on the Event Heap. On the next step, the CE-server requests access to the devices needed for the VRI. Given access to these devices is granted by the presentation manager (otherwise the server repeats the request after a while), the CE-server generates a virtual display on the antidistortion software and starts a screen capture stream, capturing the character animation, which is then mapped on the virtual display. It also moves the steerable projector and the spatial audio source to an initial position.

As a final step, the CE-server sends an XML script to the character animation, which will result in a combination of several gestures, performed by the character

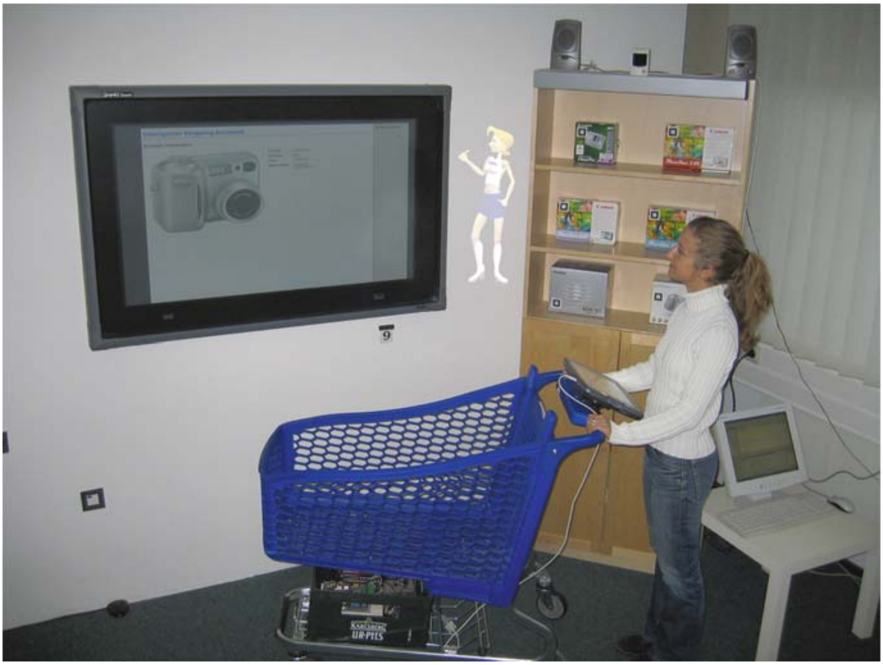


Fig. 8 The Virtual Room Inhabitant in action

while playing synchronized mp3 files (synthesized speech) over the spatial audio device. The whole initialization process is indicated in Fig. 7.

The main purpose of the VRI is to guide the user while exploring a shop; however, it may also perform references to physical objects, for example, to help the user to locate a particular product. In the example in Fig. 8, the VRI is standing between a wall-mounted display and a product shelf. At this location it may point both at objects on the screen as well as on products in the shelf.

7 Live Acquisition of User Profile Data from Speech

As of today, the most common case in modeling shop visitors is that no predefined information is available about the individual subject, possibly because it is their first visit to that specific shop, or it could be that they did not spend the time creating a profile yet. But even if the visitor does have a user profile provided to the shop by a compatible service such as *UbisWorld* (see Sect. 8), he may not be immediately authenticated to the system when he enters the shop, be it for privacy reasons or just because he just does not see the necessity yet. It resembles a behavior that is well known from the web, where users often do not log into websites unless a privileged action requires them to do so.

There are a number of advantages with having more information about the visitor. First and foremost, more profile data allow for a more personalized experience. However, requiring the user to enter some minimal information manually, which is certainly an option, carries the risk of lowering the overall customer satisfaction, as having to fill out a form (e.g., on a mobile device) is at least time-consuming. For some types of visitors such as elderly people, it may be even more inconvenient and more likely a reason to avoid such a shop or completely refrain from using its digital services.

Instead, in this case where no previous knowledge about the visitor is present, it is desirable that all available sources of user characteristics that do not require direct interaction of the person be utilized to draw as many conclusions as possible in order to bootstrap a user profile. Speech is one such source. For this purpose, a set of speech-based classifiers have been developed in the AGENDER project, which can classify the age, gender, and language of a speaker with relatively high accuracy.

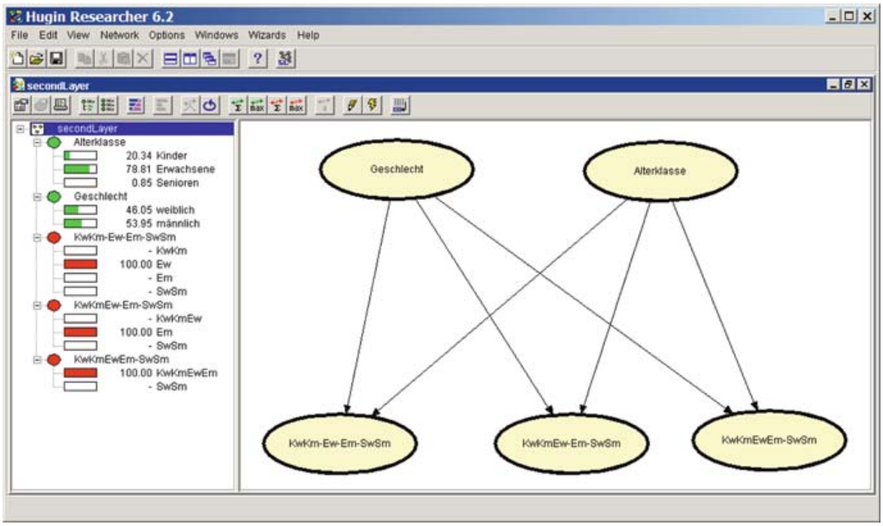
A person's recorded voice contains a lot of information about the speaker. Literature studies conducted by Müller [40, p. 43] suggested that voices do not only differ between the genders and between different ages, but that also a gender-specific vocal aging can be witnessed. The information is conveyed in prosodic features such as *pitch*, *jitter*, *intensity*, *shimmer*, and *harmonics-to-noise ratio*. Using methods from signal processing implemented in the tool *Praat*,[6] common statistics based on these features (e.g., mean and standard deviation) were extracted on large corpora of labeled speakers such as *Timit* [24] and *BAS*[7] [53]. A Gaussian probability distribution analysis performed on the extracted data revealed the subgroups of features that were most promising in being a discriminator for certain classes. In subsequent tests, it also became apparent that some classes, including those spanning different speaker properties, could be grouped to form a single combined class that resulted in a better overall performance for a specific feature set. For example, one classifier discriminates between the three classes "children," "female adults," and "male adults or seniors."

Based on these features, classifiers for each of the combined classes were trained using different machine learning algorithms. Most of them were from the *WEKA*[8] machine learning toolset. The current implementation uses a fast Gaussian Mixture Model for classification. The results from multiple classifiers extracted on a "first layer" can be combined on a "second layer" using a Dynamic Bayesian Network such as the one depicted in Fig. 9 [7]. This method can also be used to exploit the aforementioned fact of gender-dependant vocal aging by modeling the probability of a gender-specific age classifier as dependant on the probability output of the gender classifier. This improves performance because currently, gender can be classified with a much higher accuracy than age on unfiltered data. Additionally, the aspect of time is incorporated into the network when multiple utterances of the same speaker

[6] `http://www.praat.org`

[7] Bayerisches Archiv für Sprachsignale

[8] Waikato Environment for Knowledge Analysis



```
secondLayer
  Alterklasse
     20.34 Kinder
     78.81 Erwachsene
      0.85 Senioren
  Geschlecht
     46.05 weiblich
     53.95 männlich
  KwKm-Ew-Em-SwSm
         - KwKm
    100.00 Ew
         - Em
         - SwSm
  KwKmEw-Em-SwSm
         - KwKmEw
    100.00 Em
         - SwSm
  KwKmEwEm-SwSm
    100.00 KwKmEwEm
         - SwSm
```

Fig. 9 A Bayesian Network which is part of the second layer in AGENDER is being evaluated with the application *Hugin Researcher*

are considered, so that the final probability should converge and reduce classification errors.

While there are no technical limitations relating to the number of age classes, only classification modules with two, three, and four age classes have been evaluated until now, with the four-class-classifier discriminating children (0–13 years), teenagers (14–19 years), young adults (20–64), and seniors (65+). One reason for this choice lies in the vocal significance of the chosen age borders. Another reason is the fact that with an increasing number of classes, it is considerably harder to come by an adequate amount of training material for each class. Table 2 shows the covariance matrix for an eight-class-scenario (combining age and gender, e.g., *Cf* = *Children female*). The average accuracy in that case is 63.5%.

For language, a different approach had been taken because prosodic features do not convey sufficient information for accurate language classification. The idea of a

Table 2 Confusion matrix for the 8-class-problem with an ANN. The total accuracy is 63.5% with a chance level of 12.5%. The diagonal axis is given bold

	8-class-problem					total accuracy 63.5%		
	Cf	Cm	Yf	Ym	Af	Am	Sf	Sm
Cf	**76.09**	4.07	13.6	5.06	0.54	0.05	0.44	0.15
Cm	54.25	**12.37**	12.52	15.51	1.13	0.25	3.78	0.2
Yf	54.15	2.41	**27.44**	13.16	1.28	0.1	1.37	0.1
Ym	20.08	3.98	6.33	**59.25**	1.03	1.13	4.96	3.24
Af	0.25	0	0.2	0.54	**84.73**	3.44	6.92	3.93
Am	0	0	0	0.74	3.53	**87.87**	1.57	6.28
Sf	0.59	1.13	0.15	2.5	3.78	0.93	**77.07**	13.84
Sm	0	0.05	0	1.67	1.18	1.47	12.47	**83.16**

phonotactics model in combination with a pseudo-syllable model first sketched in [41] has since been developed and extended. A large corpus containing speech samples in all languages to be considered was used to form the background model for the language classification problem. Then, a variable number of MFCC-based vector quantization front-ends were trained with the *HTK*[9] toolset on the background model or parts of it. Each of these front-ends is trained with different parameters and/or different speaker classes and can output a sequence of subphonemes (thus works as a segmenter, similar to a phoneme recognizer). From the output of the front-ends, n-grams were computed with $n = 1$, 2, and 3. In order to reduce the size of the models, only the statistically most significant n-grams were used, e.g. only 20% of the trigrams, but 100% of the unigrams. The actual feature used in the classifier is the relative count of the individual n-grams. Through experiments, the best front-ends were selected. Using all data of the background model, a normalization model was computed, with rank normalization providing the best results. For each language, the Support Vector Machine *SVM-Light*[10] was trained with a fixed training set using the corresponding (normalized) feature vector. The distribution of classes can be chosen freely. In a last step, using a test set, each of the classifiers was evaluated separately with different decision thresholds, which modify the output score, to minimize the mean error. The best-performing decision threshold was applied to the result of the final classifier. Decision thresholds can also be manually adjusted to improve the overall performance in scenarios where classes are not equally distributed.

The classifiers are implemented in a high-performance C++ library, which can be trained and configured at design-time to include any number of classifiers for the required classes. Using a tool named *SBC Development Platform*, these classifiers are compiled into so-called “embedded classification modules,” which consist of mostly binary code that represents the classifier. There is also the option for including post-processing layers such as a Bayesian Network in these modules. This approach has already been successfully applied for gender-dependant recognition of age and is described in [40, p. 181].

With these classification utilities at hand, the next step to an application scenario is to identify a source from which speech will be taken and a suitable setup for the classification engine. In the shopping scenario, this could be a voice-controlled mobile application on a PDA like the *ShopAssist* (see Sect. 3) that guides the user through the shop and provides interaction with virtual item listings and actual products in a shelf. The input features that are used to classify the user are the utterances which make up the voice commands given to the *ShopAssist*. Using the same speech samples for classification is straightforward as it requires no additional effort on the user’s side. An alternative to the PDA would be a device built into the shopping basket or a stationary microphone installed at a shop shelf facilitating communication with “Talking Products” (see Sect. 2). Also, a conversation with the Virtual

[9] Hidden Markov Model Toolkit, `http://htk.eng.cam.ac.uk`

[10] `http://svmlight.joachims.org`

Room Inhabitant (see Sect. 6) character could be used. In another typical scenario for AGENDER, which is automatic call forwarding in a call-center [22, p. 133], the voice recorded in an initial speech prompt serves as the classification input. It should be stressed that in a concrete situation, the input can be used at any time and in any order, or it may not be present at all. It is also possible to use multiple input sources if available. While there is the possibility to run most parts of Agender on a mobile device, a client/server-based approach where there is a single server handling all classification requests for all users is favored in scenarios that support it, because it will usually result in lower classification times. In the shopping scenario, the shop could set up a server running AGENDER and stream voice data used to interact with the *ShopAssist* from the PDA over wireless LAN or Bluetooth to that server.

While it does not lie within the domain of AGENDER how the obtained information is used, one common application that is also referred to in the shopping scenario from the beginning is user adaptation. By creating or updating a user profile, smart services consulted by the same user may be adjusted to the user's characteristics. For example, the virtual character may be chosen from the same age group as the speaker and answer in the correct language. For elderly people, it may be a good idea to reduce the rate of speech for easier understanding, which applies to the "Talking Products" as well. Another common scenario is to adapt the choice of products suggested to the user in other applications or advertisements to match the user's demographics, or even exert influence on the path created by indoor navigation systems that are part of the shopping experience. Adapting applications should not rely on the classification to be as reliable as information entered by the user himself, and – depending on the way the input is acquired – should provide fallback solutions if some information is not present, e.g., because the user did not provide any speech yet.

8 Ubiquitous User Modeling with UbisWorld

In order to realize user modeling for intelligent environment and ubiquitous computing as indicated by this future shopping scenario, the concept of *ubiquitous user modeling* has been proposed in [27]. This concept contains a RDF-based general user model ontology GUMO and a context markup language UserML that lay the foundation for inter-operability using Semantic Web technology. GUMO and UserML enable decentralized systems to communicate over user models as well as situational and contextual factors. The idea is to spread the information among all adaptive systems, either with a mobile device or via ubiquitous networks. UserML statements can be arranged and stored in distributed repositories in XML, RDF, or SQL. Each mobile and stationary device has an own repository of situational statements, either local or global, dependent on the network accessability. A mobile device can perfectly be integrated via wireless lan or bluetooth into the intelligent environment, while a stationary device could be isolated without network access. The different applications or agents produce or use UserML statements to represent the user model information. UserML forms the syntactic description in the knowledge exchange process. Each concept like the user model auxiliary “has Property” and the user model dimension timePressure points to a semantical definition of this concept which is either defined in the general user model ontology GUMO, the UbisWorld ontology, which is specialized for ubiquitous computing, or the general SUMO/MILO ontology. Figure 10 shows Basic User Dimensions in the GUMO ontology.

GUMO collects the user’s dimensions that are modeled within user-adaptive systems like the user’s age, the user’s current position, the user’s birthplace, or the user’s gender. In the GUMO ontology, long-term user model dimensions are categorized as demographics. Ontologies provide a shared and common understanding of a domain that can be communicated between people and heterogeneous and widely spread application systems. Since ontologies have been developed and investigated in artificial intelligence to facilitate knowledge sharing and reuse, they should form the central point of interest for the task of exchanging situation models. Figure 11 shows an example of an ubiquitous user model in the UbisWorld.

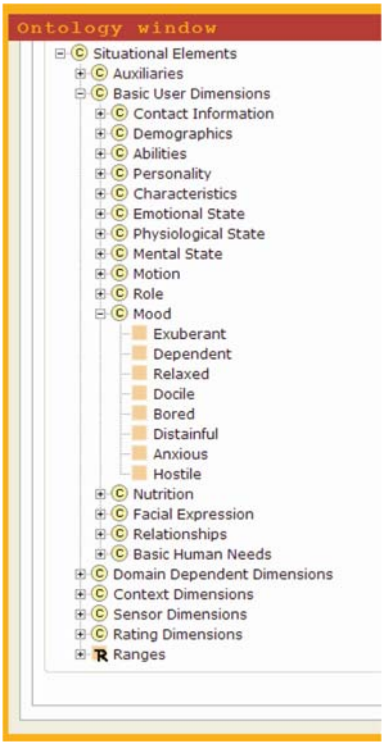
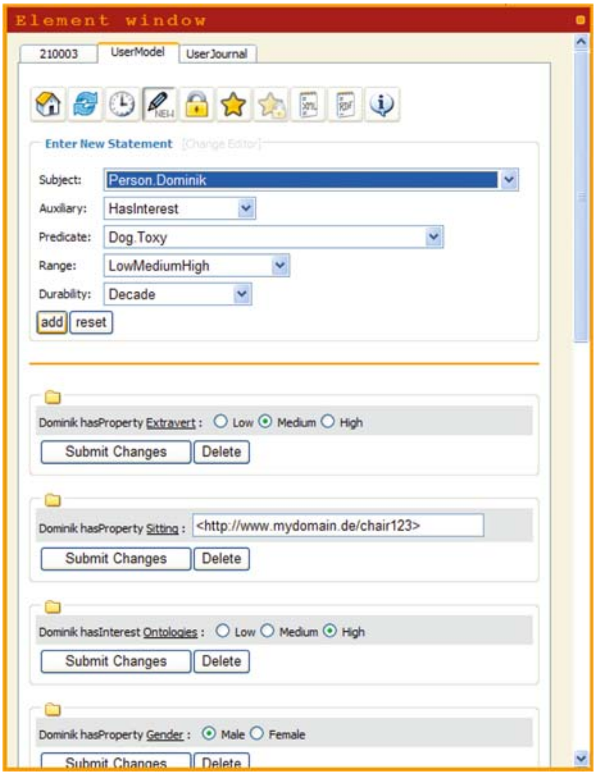


Fig. 10 Selected basic user dimensions in the GUMO ontology



Enter New Statement [Change Editor]

Subject: Person.Dominik
Auxiliary: HasInterest
Predicate: Dog.Toxy
Range: LowMediumHigh
Durability: Decade

Dominik hasProperty Extravert : ○ Low ● Medium ○ High

Dominik hasProperty Sitting : <http://www.mydomain.de/chair123>

Dominik hasInterest Ontologies : ○ Low ○ Medium ● High

Dominik hasProperty Gender : ● Male ○ Female

Fig. 11 User model inspection and editing in UbisWorld

The web ontology language OWL has more facilities for expressing semantics. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. Thus, OWL is our choice for the representation of user model and context dimension terms and their interrelationships. This ontology should be available for all user-adaptive and context-aware systems at the same time, which is perfectly possible via internet and wireless technology. The major advantage would be the simplification for exchanging information between different systems. The current problem of syntactical and structural differences between existing adaptive systems could be overcome with such a commonly accepted ontology. UbisWorld (Fig. 8) enables users to annotate their user models with the GUMO ontology. UbisWorld represents persons, objects, locations as well as times, events and their properties and features. UbisWorld could be understood as a virtual colored blocks world where each color represents a different category in the ontology. The main focus of this approach lays on research issues of ubiquitous computing and user modeling. Apart from the representational functionality, UbisWorld can be used for simulation, inspection, and control of the real world.

9 Modeling Affect

The understanding of an users affective state surely enables new ways of how to adapt to a users specific situation. This is especially important in sales scenarios, where affect plays a major role. According to the current affective state of a person, she/he decides whether to accept a specific offer. By extending the underlying user model, respectively the ubis world ontology, by a fine grained model of affect more adapted sales dialog will be possible.

The representation and real-time simulation of affect appraises a users actions in the described scenario. As a result of this process possible short-term emotions and long-term moods will be computed.

9.1 Affect Taxonomy

The affect classes of the ubis world ontology is designed to represent and simulate affect types as they occur in human beings. As suggested by Krause [34] affect can be distinguished by the eliciting cause, the influence on behavior, and its temporal characteristics. Based on the temporal feature, we use the following taxonomy of affect:

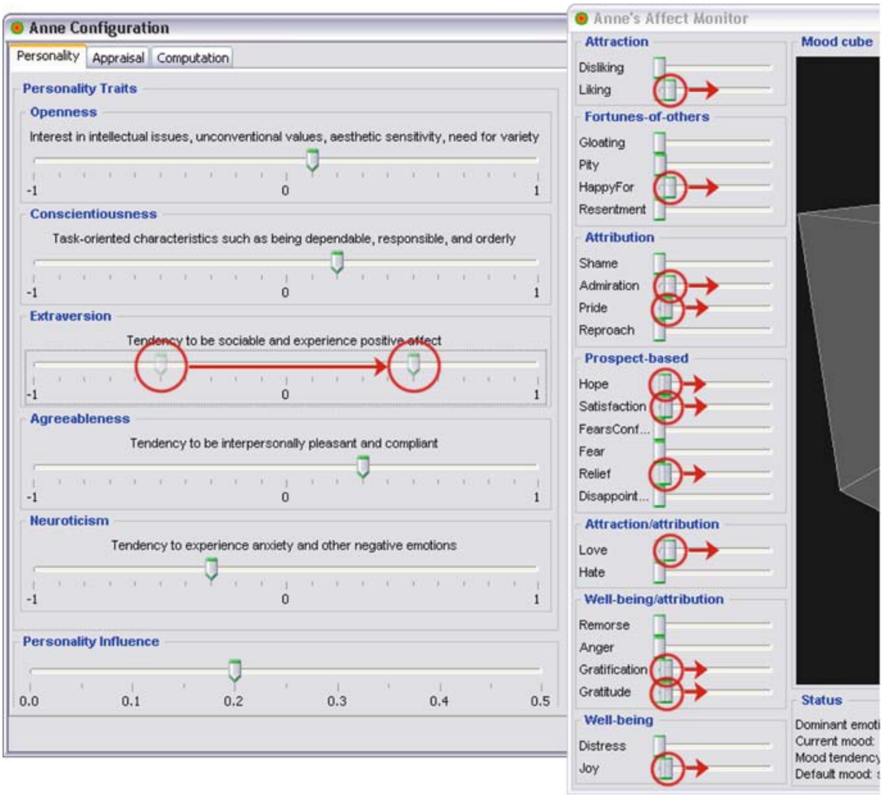
(1) Emotions reflect short-term affect that decays after a short period of time. Emotions influence facial expressions, facial complexions (e.g., blush), and conversational gestures. (2) Moods reflect medium-term affect, which is generally not related with a concrete event, action, or object. Moods are longer lasting affective states, which have a great influence on humans' cognitive functions [39, 18]. (3) Personality reflects long-term affect and individual differences in mental characteristics [36].

As known by psychological research, those different types of affect naturally interact with each other. Personality usually has a strong impact on the emotions' intensities [5, 64]. The same applies to moods [18]. With our computational model we want to simulate the interaction of the different affect types in order to achieve a more consistent overall simulation of affect.

9.2 Affect Computation

Our work is based on the computational model of emotions (ALMA) described in [25]. It implements the model of emotions developed by the psychologists Ortony, Clores, and Collins (OCC model of emotions) [47] combined with the five factor model of personality [18] and a simulation of mood, to bias the emotions' intensities. All five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism) influence the intensities of the different emotion types. We therefore adopted essential psychology research results on how personality influences emotions to achieve a more human-equivalent emotion simulation. Watson and Clark [64] and Becker [5] have empirically shown that personality, described through the big-five traits, impacts the intensity of emotions. They discovered, e.g., that extravert people experience positive emotions more intensely than negative emotions. In our computational model this is realized by the change of an emotion's basic intensity, the so-called emotion intensity bias. Note that, the intensity of elicited emotions cannot be lower than the emotion intensity bias. When the personality is defined by a graphical user interface one can directly observe the impact on the emotions intensity bias, see Fig. 12.

Figure 12 consists of two screen shots showing the direct impact of the change of the extravert personality trait on emotions' intensity bias. In the example the extravert trait value is increased by moving the slider to the right side. As a consequence the basic emotion intensities of positive emotions increase. Note that not all emotions are biased in the same way. This depends on the fact that personality traits potentially bias emotion intensities at different strengths. Also the intensity biases are influenced by a person's current mood, see next section. The OCC cognitive model of emotions is based on the concepts of appraisal and intensity. The individual is said to make a cognitive appraisal of the current state of the world. Emotions are defined as valenced reactions to events of concern to an individual, actions of those



Anne Configuration

Personality | Appraisal | Computation

Personality Traits

Openness

Interest in intellectual issues, unconventional values, aesthetic sensitivity, need for variety

Conscientiousness

Task-oriented characteristics such as being dependable, responsible, and orderly

Extraversion

Tendency to be sociable and experience positive affect

Agreeableness

Tendency to be interpersonally pleasant and compliant

Neuroticism

Tendency to experience anxiety and other negative emotions

Personality Influence

Anne's Affect Monitor

Attraction: Disliking, Liking

Fortunes-of-others: Gloating, Pity, HappyFor, Resentment

Attribution: Shame, Admiration, Pride, Reproach

Prospect-based: Hope, Satisfaction, FearsConf..., Fear, Relief, Disappoint...

Attraction/attribution: Love, Hate

Well-being/attribution: Remorse, Anger, Gratification, Gratitude

Well-being: Distress, Joy

Mood cube

Status

Dominant emoti

Current mood:

Mood tendency

Default mood: :

Fig. 12 Impact of personality traits on emotion intensities

she/he considers responsible for such actions and objects/persons. ALMA is able to compute all 24 emotions that are defined by the OCC theory.

The employed computational model of moods is based on the psychological model of mood (or temperament) proposed by Mehrabian [38]. Mehrabian describes mood with the three traits pleasure (P), arousal (A), and dominance (D). Each trait represents a specific mood component. Pleasure describes how much an individual enjoys the actual situation. Arousal stands for the excitement of an individual in the actual situation. Dominance describes to what extent an individual controls the actual situation. The three traits are nearly independent, and form a three-dimensional mood space. A PAD mood can be located in one of eight mood octants. A mood octant stands for a discrete description for a mood: $+P+A+D$ is exuberant, $-P-A-D$ is bored, $+P+A-D$ is dependent, $-P-A+D$ is disdainful, $+P-A+D$ is relaxed, $-P+A-D$ is anxious, $+P-A-D$ is docile, and $-P+A+D$ is hostile. Generally, a mood is represented by a point in the PDA space.

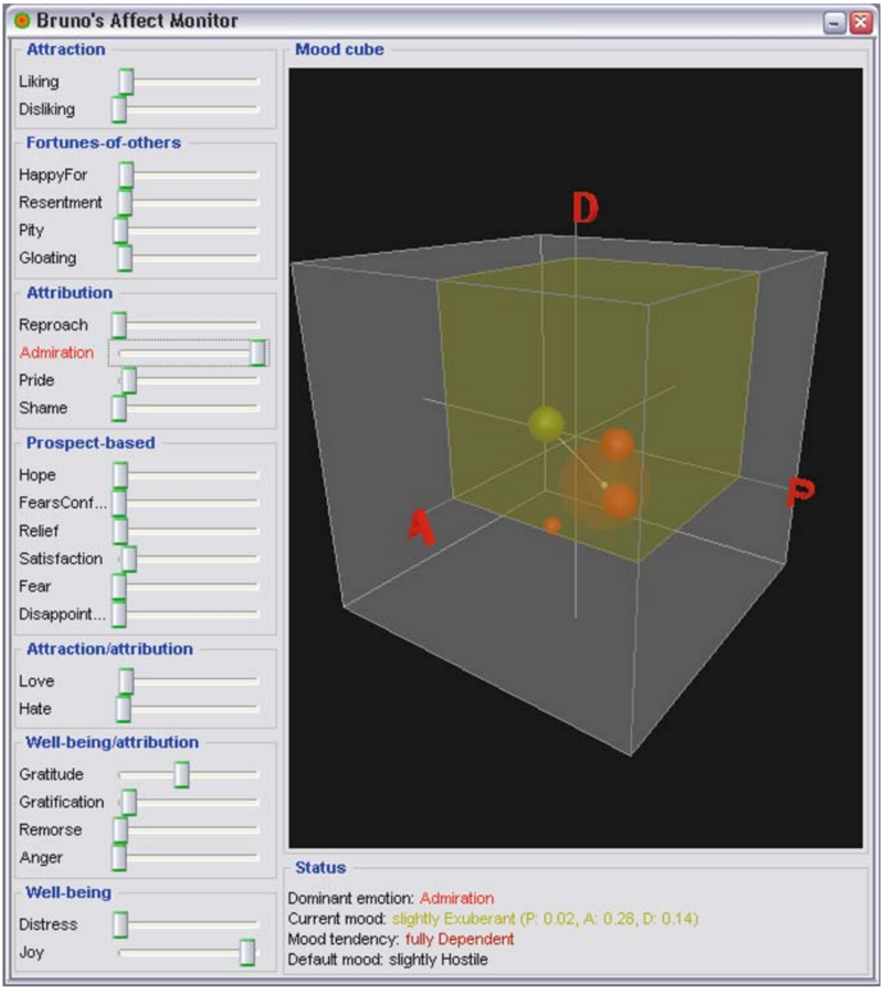
For a mood computation, it is essential to define a person's default mood. A mapping, empirically derived by Mehrabian [20], defines a relationship between the big five personality traits and the PAD space. The big-five traits can be obtained by a UbisWorld user profile. If they are not defined, a neutral mood will be assumed.

We define the mood strength by its distance to the PAD zero point. The maximum distance is $\sqrt{3}$. This is divided into three equidistant sections that describe three discrete mood intensities: slightly, moderate, and fully. Using the above mentioned mapping and the mood strength definition, a person whose personality is defined by the following big five personality traits: openness $= 0.4$, conscientiousness $= 0.8$, extraversion $= 0.6$, agreeableness $= 0.3$, and neuroticism $= 0.4$ has the default mood slightly relaxed (pleasure $= 0.38$, arousal $=$–0.08, dominance $= 0.50$).

An AffectMonitor, shown in Fig. 13, is used to visualize a person's current mood and mood changing emotions. The left side of the AffectMonitor shows all emotions and their intensities. Newly elicited emotions are marked dark gray (red). The right side shows a three-dimensional PDA mood cube displaying the current mood (the highlighted octant stands for the discrete mood description, whereas the light gray (yellow) ball reflects the actual mood) and all active emotions (dark gray (red) balls). Below, the affective state, including the current dominant emotion, and the default as well as the current mood, is displayed. The current mood also influences the intensity of active emotions. The theory is that the current mood is related to personality values that interfere with a person's personality values. Based on the current mood, the most intense related personality trait is identified. The actual value of this trait blends over the person's original personality trait value and is used to regulate the intensity of emotions. This increases, for example, the intensity bias of joy and decreases the intensity bias of distress, when a person is in an exuberant mood.

9.2.1 Mood Changes

According to Morris [39, p. 24] conditions for mood changes can be divided into (a) the onset of a mildly positive or negative event, (b) the offset of an emotion-inducing event, (c) the recollection or imagining of an emotional experience, and (d) the



Dominant emotion: Admiration
Current mood: slightly Exuberant (P: 0.02, A: 0.28, D: 0.14)
Mood tendency: fully Dependent
Default mood: slightly Hostile

Fig. 13 AffectMonitor visualizes a person's mood and emotions

inhibition of an emotional response in the presence of an emotion-inducing event. To keep the modeling of mood changes as lean as possible, we take elicited emotions as the mood changing factor. In order to realize this, emotions must be somehow related to a specific mood. While using the PAD space for modeling mood, it is obvious to relate emotions to the PDA space too.

We rely on Mehrabian's mapping of emotions into the PAD space [38]. However, not all 24 emotion types of the OCC-Modell are covered by this mapping. For those that lack a mapping, we provide the missing pleasure, arousal, and dominance values by exploiting similarities to comparable emotion types [25].

Our approach to the human-like simulation of mood changes relies on a functional approach. We concentrate on how the intensity of emotions influences the change of the current mood. Moreover, we consider the aspect that the more experiences a person makes that support a specific mood, the more intense this person's mood gets. For example, if a person's mood can be described as slightly anxious and several events let the person experience the emotion fear, the person's mood might change to moderate or fully anxious.

9.2.2 Appraisal Based Affect Computation

In our cognitively inspired affect computation, the first step is to evaluate relevant input by imitating a person's own subjective appraisal of a current situation. The situation is appraised according to two different concepts: (1) situational appraisal: events, actions, and objects and (2) interaction appraisal. The appraisal is realized by specific tags that relates to the appraisal concepts: (1) basic appraisal tags and (2) act appraisal tags. All appraisal tags are defined in a person's ubis world ontology.

Basic appraisal tags express how a person appraises the event, action, or object in the current focus. There are 12 basic tags for appraising events, e.g., the tag GoodEvent, which marks an event to be good according to the subjective view of the one which does the appraisal. The other event tags are BadEvent, GoodEventForBadOther, GoodEventForGoodOther, BadEventForGoodOther, BadEventForBadOther, GoodLikelyFutureEvent, GoodUnlikelyFutureEvent, BadLikelyFutureEvent, BadUnlikelyFutureEvent, EventConfirmed, and EventDisconfirmed. For appraising actions, there are four basic appraisal tags: GoodActSelf, BadActSelf, GoodActOther, and BadActOther. And finally there are two basic tags for appraising objects: NiceThing and NastyThing. All basic appraisal tags together are the basic set of a high-level appraisal language which can be used for a subjective appraisal of situations. These tags can be used to appraise dialog acts and other affective signals. For each of these types the appraisal language provides specific tags: act appraisal tags and affect display appraisal tags. Act appraisal tags represent the underlying communicative intent of an utterance, e.g., tease or congratulate.

Generally, the output of the appraisal process is a set of emotion eliciting conditions. Based on them active emotions are generated that in turn influence a person's mood. On the technical side, each person has their own ALMA process, which processes affect input. The input consists of appraisal tags, dialog act input, emotion and mood input, information about who is speaker, addressee, and listener. The computed affect (emotions and mood) is then passed to the other modules of the application.

The evaluation of this computational model of affect shows that nearly all affect types are plausibly represented in dialog scenarios.

10 Conclusions

The project BAIR has been concerned with research about user adaptation in instrumented rooms, with user-centered approaches in respect of the limitation of cognitive and technical resources. One focus was set on the role of cognitive and affective states of the user for generating affective responses from the instrumented environment and for adapting the presentation of information. We developed an affective layer between the user services and the physical environment. In BAIR, we investigated the introduced concepts and novel interaction methodologies for proactive and user-centered support in multi-user instrumented environments. Research findings were also used in collaboration with other projects.

Acknowledgments The research project BAIR was supported by the German Research Foundation (DFG) in its Collaborative Research Center on Resource-Adaptive Cognitive Processes, SFB 378, Project EM 4. We also would like to thank the Transfer Unit on Cognitive Technologies for Real-Life Applications, TFB53, Project TB2, RENA for their cooperation in the research on indoor navigation solutions.

References

1. Aaker, J. Dimensions of brand personality. Journal of Marketing Research, 34(3):342–352 (1997).
2. André, E., Klesen, M., Gebhard, P., Steve Allen, T.R. Integrating models of personality and emotions into lifelike characters. In A. Paiva, C. Martinho (Eds.), Proceedings of the Workshop on Affect in Interactions – Towards a New Generation of Interfaces in Conjunction with the 3rd i3 Annual Conference (pp. 136–149). Italy: Siena (1999).
3. Arons, B. A review of the cocktail party effect. Journal of the American Voice I/O Society, 12:35–50 (1992).
4. Barrington, L., Lyons, M., Diegmann, D., Abe, S. Ambient display using musical effects. In IUI '06: Proceedings of the 11th International Conference on Intelligent User Interfaces (pp. 372–374). ACM Press, New York, USA (2006).
5. Becker, P. Structural and relational analyses of emotion and personality traits. Zeitschrift für Differentielle und Diagnostische Psychologie, 22(3):155–172 (2001).
6. Blattner, M., Sumikawa, D., Greenberg, R. Earcons and icons: Their structure and common design principles. Human Computer Interaction, 4:11–44 (1989).
7. Brandherm, B. Eingebettete dynamische Bayessche Netze n-ter Ordnung. PhD thesis, Computer Science Institute, Saarland University, Germany (2006).
8. Brandherm, B., Schwartz, T. Geo referenced dynamic bayesian networks for user positioning on mobile systems. In Proceedings of the International Workshop on Location- and Context-Awareness (LoCA), LNCS 3479, volume 3479/2005 of Lecture Notes in Computer Science (pp. 223–234). Springer-Verlag Berlin Heidelberg, Munich, Germany (2005).
9. Brandherm, B., Schwartz, T. Geo referenced dynamic Bayesian networks for user positioning on mobile systems. In T. Strang, C. Linnhoff-Popien (Eds.), Proceedings of the International Workshop on Location- and Context-Awareness (LoCA), LNCS 3479, volume 3479/2005 of Lecture Notes in Computer Science (pp. 223–234). Springer-Verlag Berlin Heidelberg, Munich, Germany (2005).
10. Bregman, A. Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press (1990).
11. Brewster, S. Using non-speech sounds to provide navigation cues. ACM Transactions on Computer–Human Interaction, 5:224–259 (1998).
12. Bruner, G.C. Music, mood, and marketing. Journal of Marketing, 54:94–104 (1990).
13. Butz, A., Jung, R. Seamless user notification in ambient soundscapes. In IUI '05: Proceedings of the 10th International Conference on Intelligent User Interfaces (pp. 320–322). ACM Press, New York, NY, USA (2005).

14. Buxton, W., Gaver, W., Bly, S. Tutorial chapter 6: The use of non-speech audio at the interface. In Proceedings of Computer Human Interaction (CHI). ACM Press; Addison-Wesley, New Orleans (1991).
15. Camurri, A., Leman, M. Gestalt-based composition and performance in multimodal environments. In Joint International Conference on Cognitive and Systematic Musicology (pp. 495–508) (1996).
16. Cherry, E.C. Some experiments on the recognition of speech, with one and two ears. Journal of the Acoustic Society of America, 25:975–979 (1953).
17. Costa, P., McCrae, R. The Neo Personality Inventory Manual. Odersa, FL: Psychological Assessment Resources (1985).
18. Davidson, R. On Emotion, mood, and related affective constructs. The nature of Emotion: Fundamental Questions (pp. 51–55). New York: Oxford University Press (1994).
19. Davies, J. The Psychology of Music. London: Hutchinson (1978).
20. Duggan, B., Deegan, M. Considerations in the usage of text to speech (tts) in the creation of natural sounding voice enabled web systems. In ISICT '03: Proceedings of the 1st International Symposium on Information and Communication Technologies (pp. 433–438). Trinity College Dublin, Ireland (2003).
21. Eysenck, M., Keane, M. Cognitive Psychology: A Student‘s Handbook. Hove: Psychology Press (2005).
22. Feld, M. Erzeugung von Sprecherklassifikationsmodulen für multiple Plattformen, Diploma Thesis, Soarland University (2006).
23. Fiske, S., Taylor, S. Social Cognition. New York: McGraw-Hill (1991).
24. Garofolo, J. e. A. DARPA TIMIT CD-ROM: An Acoustic Phonetic Continous Speech Database. Gaithersburg, MD, USA: National Institute of Standards and Technology (1998).
25. Gebhard, P. Alma – a layered model of affect. In F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. P. Singh, M. Wooldridge (Eds.), Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (pp. 29–36) (June 2005).
26. Gill, A.J., Oberlander, J., Austin, E. The perception of e-mail personality at zero-acquaintance. Personality and Individual Differences, 40:497–507 (2006).
27. Heckmann, D. Ubiquitous user modeling. PhD thesis, Department of Computer Science, Saarland University, Germany, DISKI 297, ISBN 3898382974, Aka Verlag (2005).
28. Heckmann, D., Schwartz, T., Brandherm, B., Kröner, A. Decentralized User Modeling with UserML and GUMO (pp. 61–66). Scotland: Edinburgh (2005).
29. Johanson, B., Fox, A. The event heap: A coordination infrastructure for interactive workspaces. In Proceedings of the Workshop on Mobile Computing Systems and Applications (2002).
30. Jung, R., Butz, A. Effectiveness of user notification in ambient soundscapes. In Proceedings of the workshop on Auditory Displays for Mobile Context-Aware Systems at Pervasive 2005 (pp. 47–56). Munich, Germany (2005).
31. Jung, R., Heckmann, D. Ambient audio notification with personalized music. In Workshop on Ubiquitous User Modeling at ECAI 2006 (pp. 16–18). Riva del Garda, Italy (2006).
32. Jung, R., Schwartz, T. A location-adaptive human-centered audio email notification service for multi-user environments. In J.A. Jacko (Ed.), Human–Computer Interaction, vol. 4552 of LNCS (pp. 340–348). New York: Springer (2007).
33. Jung, R., Schwartz, T. Peripheral notification with customized embedded audio cues. In Proceedings of the 13th International Conference on Auditory Displays (pp. 221–228). Schulich School of Music, McGill University, Montreal, Canada (2007).
34. Krause, R. Affekt, Emotion, Gefühl (2nd edn., pp. 30–36). Stuttgart: Kohlhammer (2002).
35. Legaspi, R., Hashimoto, Y., Moriyama, K., Kurihara, S., Numao, M. Music compositional intelligence with an affective flavor. In Proceedings of Conference on Intelligent User Interfaces (IUI) (2007).
36. McCrae, R., John, O. An introduction to the five-factor model and its applications. Journal of Personality, 60:175–215 (1992).

37. McRae, R., John, O. An introduction to the five-factor model and its applications. Journal of Personality, 60:175–215 (1992).
38. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology: Developmental, Learning, Personality, Social, 14:261–292 (1996).
39. Morris, W.N. Mood The Frame of Mind. New York: Springer (1989).
40. Müller, C. Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender]. PhD thesis, Computer Science Institute, University of the Saarland, Germany (2005).
41. Müller, C., Feld, M. Towards a multilingual approach on speaker classification. In Proceedings of the 11th International Conference "Speech and Computer" SPECOM 2006 (pp. 120–124). Anatolya Publishers, St. Petersburg, Russia (2006).
42. Nass, C., Isbister, K., Lee, E.-J. Truth is beauty: Researching embodied conversational agents. Embodied conversational agents (pp. 374–402). Cambridge, MA: MIT Press (2000).
43. Nijholt, A., Rist, T., Tuijnenbreijer, K. Lost in Ambient Intelligence? Panel Session. In Proceedings of CHI'04 (pp. 1725–1726). ACM, New York (2004).
44. North, A., MacKenzie, L., Hargreaves, D. The effects of musical and voice "fit" on responses to advertisements. Journal of Applied Social Psychology, 34(8):1675–1708 (2004).
45. Nowson, S. The langauge of weblogs: A study of genre and individual differences. PhD thesis, University of Edinburgh. College of Science and Engineering. School of Informatics. (2006).
46. O'Conaill, B., Frohlich, D. Timespace in the workplace: dealing with interruptions. In CHI '95: Conference Companion on Human Factors in Computing Systems (pp. 262–263). ACM Press, New York, NY, USA (1995).
47. Ortony, A., Clore, G.L., Collins, A. The Cognitive Structure of Emotions. Cambridge, MA: Cambridge University Press (1988).
48. Pennebaker, J., King, L. Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77:1296–1312 (1999).
49. Pinhanez, C. The everywhere displays projector: A device to create ubiquitous graphical interfaces. Lecture Notes in Computer Science (2001).
50. Reeves, B., Nass, C. The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places. Cambridge, MA: CSLI Publications and Cambridge university press (1996).
51. Reybrouck, M. Gestalt concepts and music: Limitations and possibilities. In Joint International Conference on Cognitive and Systematic Musicology (pp. 57–69) (1996).
52. Scherer, K., Zentner, M. Music and Emotion: Theory and Research (Chapter 16, pp. 361–392). Oxford, England: Oxford University Press (2001).
53. Schiel, F. Speech and speech-related resources at BAS. In Proceedings of the First International Conference on Language Resources and Evaluation (pp. 343–349). Granada, Spain (1998).
54. Schmitz, M., Butz, A. Safir: Low-cost spatial audio for instrumented environments. In Proceedings of the 2nd International Conference on Intelligent Environments. Athens, Greece, (2006).
55. Schmitz, M., Krüger, A., Schmidt, S. Modelling personality in voices of talking products through prosodic parameters. In Proceedings of the 10th International Conference on Intelligent User Interfaces (pp. 313–316) (2007).
56. Schneider, M. A smart shopping assistant utilising adaptive plan recognition. In Proceedings of the Workshop 'Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen' (ABIS) at LLWA (2003).
57. Schröder, M., Grice, M. Expressing vocal effort in concatenative synthesis. In Proceedings of the 15th International Conference of Phonetic Sciences (pp. 2589–2592) (2003).
58. Schwartz, T., Brandherm, B., Heckmann, D. Calculation of the user-direction in an always best positioned mobile localization system. In Proceedings of the International Workshop on Artificial Intelligence in Mobile Systems (AIMS). Salzburg, Austria (September 2005).

59. Spassova, L. Fluid Beam – A Steerable Projector and Camera Unit. Student and Newbie Colloquium at ISWC/ISMAR (2004).
60. W3C-EMMA. EMMA: Extensible MultiModal Annotation markup language. W3C Working Draft, 9 April 2007, http://www.w3.org/TR/emma/, last accessed: 31.10.2007 (2007).
61. Wasinger, R. (Ed.) Multimodal Interaction with Mobile Devices: Fusing a Broad Spectrum of Modality Combinations. Akademische Verlagsgesellschaft, Berlin, Germany (2006).
62. Wasinger, R., Krüger, A., Jacobs, O. Integrating intra and extra gestures into a mobile and multimodal shopping assistant. In Proceedings of the 3rd International Conference on Pervasive Computing (pp. 297–314) (2005).
63. Wasinger, R., Krüger, A., Jacobs, O. Integrating intra and extra gestures into a mobile and multimodal shopping assistant. International Conference on Pervasive Computing (2005).
64. Watson, D., Clark, L.A. On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. Journal of Personality, 2(60): 441–476 (1992).