15

Multi-modal Human–Environment Interaction

R. Wasinger and W. Wahlster

"The environment is everything that isn't me."

Albert Einstein

15.1 Introduction

AmI environments require robust and intuitive interfaces for accessing their embodied functionality. This chapter describes a new paradigm for tangible multi-modal interfaces, in which humans can manipulate, and converse with physical objects in their surrounding environment via coordinated speech, handwriting, and gesture. We describe the symmetric nature of human–environment communication, and extend the scenario by providing our objects with human-like characteristics. This is followed by the results of a usability field study on user acceptance for anthropomorphized objects, conducted within a shopping context.

The talking toothbrush holder is an example of a consumer product with an embedded voice chip. If activated by a motion sensor in the bathroom it says "Hey don't forget to brush your teeth!". Talking calculators, watches, alarm clocks, thermometers, bathroom scales, greeting cards, and the pen that says "You are fired" when one presses its button are all products that are often mass-marketed as gimmicks and gadgets (Talkingpresents, online; Jeremijenko 2001). However, such voice labels can also offer useful help for people who are visually impaired, since they can be used to identify different objects of similar shape or to supply critical information to help orientate the users to their surroundings (Talkingproducts, online). All these voice-enabled objects of daily life are based on very simple sensor–actuator loops, in which a recognized event triggers speech replay or simple speech synthesis.

This chapter presents a new interaction paradigm for Ambient Intelligence, in which humans can conduct multi-modal dialogs with objects in a networked shopping environment. In contrast to the first generation of voice-enabled artifacts described above, the communicating objects in our framework provide a combined conversational and tangible user interface that exploits situational context such as whether a product is in or out of a shelf, to compute its meaning.

15.2 Tangible Multi-modal Dialog Scenario

Our experimental scenario attempts to combine the benefits of both physical and digital worlds in a mixed-reality setting by targeting an in-store scene, but augmented by instrumented devices like a Personal Digital Assistant (PDA) and a shopping trolley with a mounted display. Whereas the PDA is used as a communication channel through which users can associate directly with the products rather than through a sales assistant, the shopping trolley is capable of offering shopping advice based on its current contents. An in-store setting encompasses the down-to-earth basics that only a traditional store in a real world and with real physical products can provide such as the sense of touch. When instrumented, it further provides the convenience inherent in digital worlds such as ubiquitous information access. The unification of these two worlds is achieved through a Tangible Multi-Modal interface (TMM) that is seamlessly integrated into existing shopping practices. TMMs are now being incorporated into a wide range of fields, up to and including safety-critical applications (Cohen and McGee 2004).

Tangible User Interfaces (TUIs) (Ullmer and Ishii 2001) couple physical representations (e.g., spatially manipulable physical objects) with digital representations, e.g., graphics and audio, yielding interactive systems that are computationally mediated. In our scenario, we use an intuitive "one-to-one" mapping between physical shopping items on the shelf and elements of digital information. The spatial relation of a physical token partially embodies the dialog state, which can be seen in our example in that a product can be either on a shelf, in the shopping trolley, or outside of these containers. The position of the product is mapped to a physical action of the user, where the physical movements of the artifacts serve as a means to controlling the dialog state.

The Mobile ShopAssist (MSA) is a demonstrator that aids users in product queries and comparisons. The goal is to provide rich symmetric multi-modal interaction and the ability for users to converse directly with the products. Using the MSA, a shopper interested in buying a digital camera would, for example, walk up to a shelf and synchronize its contents with their PDA. After synchronization, they may ask a product about its attributes, e.g., "*What is your optical zoom?*" or even compare multiple products together, e.g., "*<gesture> Compare yourself with this camera <gesture>*." Comparisons may be made among products from the physical world, digital world, or a mixture of both, i.e., mixed-reality.

When interacting with the digital cameras, the user may decide to communicate indirectly with the object "*What is the price of this camera <gesture>*" or directly "*What is your price?*". The input modalities available to the user include speech, handwriting, gesture, and combinations thereof. It is direct interaction and the concept of *anthropomorphization*, i.e., assigning inanimate objects human-like characteristics that we focus on, see also Sect. 15.5.

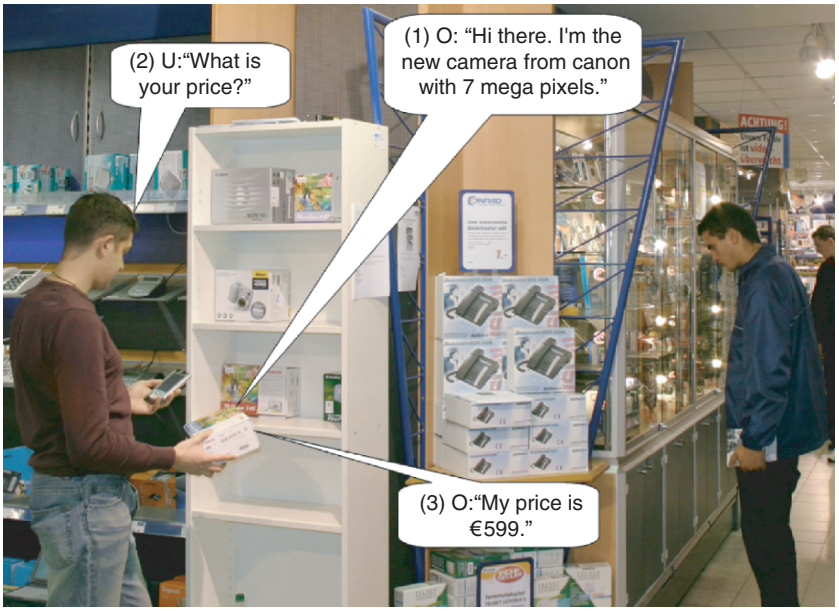

Fig. 15.1. Anthropomorphized object initiating a dialog

Assuming the user has chosen to interact directly with the objects, the objects will in return communicate directly with the user and may also initiate mini-dialogs when picked up or put down on a shelf or in a shopping trolley, similar to (1) in Fig. 15.1. Once the users have finished conversing with the products, they may decide to buy the product, or to simply take the information that they have downloaded back home with them to think about later on. The objects, not limited to digital cameras, can then be placed into the shopping trolley and taken to the cashier. On request, the user's interactions are logged and summarized in a personal shopping diary (Kröner et al. 2004).

The MSA is a mixed-initiative dialog system, which means that both a product and the user can start a dialog or take the initiative in a sub-dialog. For instance, when the product is picked up – and no accompanying user query is issued – the product will introduce itself. Another system-initiated dialog phase is that of cross-selling, which occurs when a product is placed into the shopping trolley. Such a dialog might give advice on accessories available for the product, for example: "*You may also find the NB-2LH batteries in the accessories shelf to be useful.*"

Instrumented environments containing RFID tagged products have till now primarily benefited the retailer through improved inventory management and tracking. Our scenario also highlights user benefits in the form of comparative shopping, cross-selling recommendations, and product information retrieval based on real physical indexes.

15.3 Instrumented Environment Infrastructure

The main infrastructure components that exist in our shopping environment include the mobile device, which is used as a communication channel, the containers, e.g., shelves, trolley, shopping products and the belonging to a shop; see Fig. 15.3. Each shelf is identified by an infrared beacon that is required when a user synchronizes the shelf's product data. The products are identified through the use of passive RFID tags, which allow a product to be classified as being either in or out of a container. Each container has an RFID antenna and a reader connected to it, and this allows the shelves and shopping trolley to recognize when products are put in or taken out of them.

The instrumented shelves may be scattered over several rooms, and communicate via a WLAN connection with the AmI server, as shown in Fig. 15.2 (similar instrumented shopping environments without a tangible multi-modal interface are the Metro Future Store and MyGrocer (Kourouthanasis et al., 2002)). It is the AMI server that maintains the product database, and the event heap (Fox et al. 2000), which is used for recording extra-gesture events. As described in Butz et al. (2004), a searchlight in the form of a steerable projector further allows the system to find and highlight products based on optical markers. This is important in establishing a link between the physical products and their digital counterparts (and vice-versa), which do not need to be sorted in the same way. Such a situation could for example arise when digital objects are re-sorted based on specific product features such as price or manufacturer, instead of their physical location.

After a client device such as a PDA has been synchronized with a shelf, it will maintain its own blackboard of events, on which it stores not only the extra-gesture interactions broadcast by the server, but also speech, handwriting, and intra-gesture interactions that the PDA is capable of recognizing and

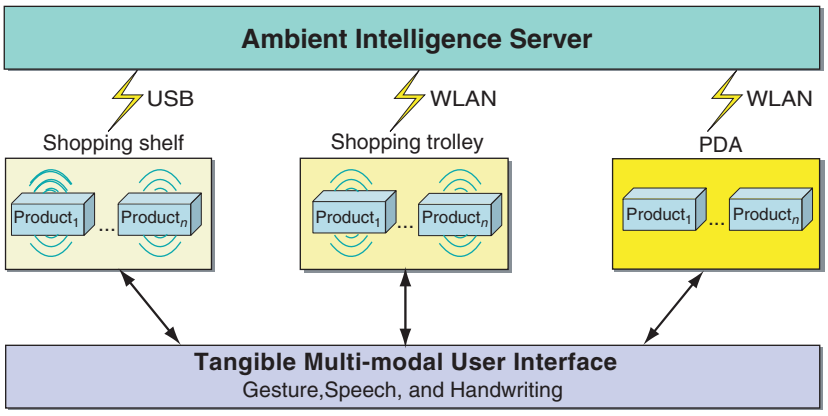

Fig. 15.2. Distributed architecture of the MSA

Fig. 15.3. Instrumented shopping environment

interpreting locally. When a shopping trolley is added to the scenario (Schneider 2004), the contained products are listed on a trolley-mounted display, and the trolley will offer advice on additional products that may be relevant to the user (see Fig. 15.3).

The data downloaded upon shelf synchronization is contained within the product database. This is located on the AmI server, and contains product feature–value lists for attributes like "price," "optical zoom," and "mega pixels." The database also contains images, links to URL manufacturer sites, RFID and optical marker values for the products, and a reference to the associated grammar file used for input recognition. This data is retrieved by SQL queries and transferred from the server to the mobile device in XML format.

The input grammar files contain a similar feature–value list, in which grammar entries for each feature are defined for the different modalities like speech and handwriting. The input grammars are assigned to a group of products based on their product type, which allows multiple products to share a single grammar file, as is the case for the product type "digital camera."

After a user has synchronized with a shelf and starts to browse through the products, internal data representations are created for both the objects, and feature keywords displayed on the PDA's PocketPC display. This representation is for example used by the modality of intra-gesture, first during input interaction as screen coordinates are mapped to underlying graphical objects and visual What-Can-I-Say (WCIS) keywords currently on the screen, and later during output presentation through the use of object and keyword lookup functions, which locate a particular reference and highlight it on the display.

15.4 Symmetric Multi-modal Interaction

As defined in Wahlster (2003), symmetric multi-modality refers to the ability of a system to use all input modes as output modes, and vice-versa. Empirical studies have shown that the robustness of multi-modal interfaces increase substantially as the number and heterogeneity of modalities expand (Oviatt 2002). Information provided by one or more sources can be used to resolve ambiguities or manage recognition and sensor uncertainties in another modality, thereby reducing errors both in the system's interpretation of the user's input, and the user's understanding of the system's output. Whereas modality fusion maps multi-modal input to a semantic representation language, the modality fission component provides the inverse functionality of the modality fusion component, since it maps a communicative intention of the system onto a coordinated multi-modal presentation.

Most of the previous multi-modal interfaces such as the ones presented by Oviatt and Wahlster (1997) and Maybury and Wahlster (1998) do not support symmetric multi-modality, since they focus either on multi-modal fusion, e.g., QuickSet (Cohen et al. 1997) and MATCH (Johnston et al. 2002), or multi-modal fission, e.g. WIP (Wahlster et al. 1993). Symmetric multi-modal dialog systems like SmartKom and the MSA create a natural experience for the user in the form of daily human-to-human communication, by allowing both the user and the system to combine the same spectrum of modalities for the input as for the output. The MSA represents a new generation of multi-modal dialog systems that deal not only with simple modality integration and synchronization, but cover the full spectrum of dialog phenomena that are associated with symmetric multi-modality. Symmetric multi-modality supports the mutual disambiguation of modalities, as well as multi-modal or cross-modal deixis and anaphora resolution.

15.4.1 Base Modalities

Multi-modal interaction in the MSA is based on the modalities: speech, handwriting, and gesture, whereby gesture can be further grouped into the types intra and extra. Intra-gestures refer to product and feature selections on the display of the PDA (intra_point), while extra-gestures refer to actions in the physical real world such as picking an object up from a shelf (extra_pick_up), or putting an object back onto a shelf (extra_put_down).

From this limited number of base modalities, a wide range of mixed and overlapped input combinations can be formed. Wasinger and Krüger (2004) outline a total of 23 input modality combinations that were tested within a laboratory setting for use with the system. The modalities included both unimodal (e.g., speech-only) and multi-modal (e.g., speech–gesture) combinations, as well as overlapped and non-overlapped modality combinations. Overlapped modality combinations are ones in which (possibly conflicting) information is provided multiple times in potentially different modalities, as seen in the following non-conflicting speech–gesture overlapped feature interaction: "*What is the price <intra_gesture = price> of the EOS10D?*" Such redundant information is useful for reference resolution.

All of these input modality combinations are however only one side of the interaction equation. The flipside encompasses the output modalities used by the anthropomorphized objects when replying to the user. Speech output for example is presented to the user via an embedded synthesizer. We currently use two synthesizers, one is a formant synthesizer which requires a small memory footprint (around 2 MB per language), while the other is a high quality concatenative synthesizer that has a much larger footprint, between (7 and 15 MB) per language for a single voice. Although the formant synthesizer sounds robotic, it provides far greater flexibility in manipulating voice characteristics such as age and gender, which is important in providing the anthropomorphized objects with their own personality (see Sect. 15.5.2). The output equivalent to handwriting is the use of system fonts that are displayed in a predefined location on the PDA's display. Intra-gesture output for object selection is achieved by drawing a border around the selected object, while intra-gesture output for feature selection is achieved by highlighting the active keyword within the visual WCIS text bar, which scrolls across the bottom of the PDA's display. Extra-gesture output is made possible through the use of a steerable projector, which provides for real-world product selection by placing the product under a spotlight. Figure 15.4 shows the use of the primary modalities within our system, for both input interaction and output presentation. This figure also shows how objects and features can be referred to within the modality types. The output for intra-gesture for example shows a selected feature and below it a selected object.

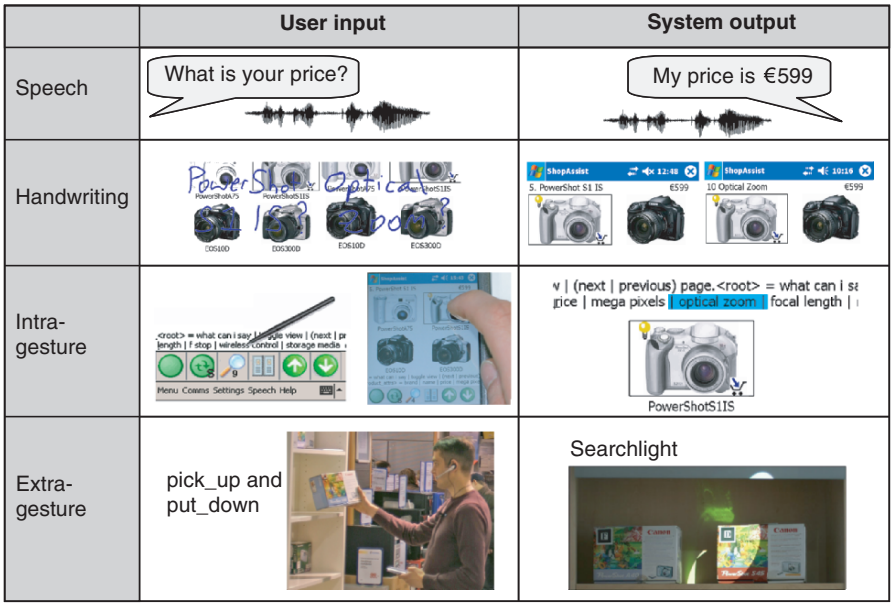

Fig. 15.4. The modalities used for both input and output

15.4.2 Symmetric Modality Combinations

Systems that support multi-modal interaction such as speech, handwriting, and gesture, require an efficient means of fusing the interactions together to form a single unambiguous dialog result, which can then be passed onto subsequent modules in the system such as a retrieval component. Multi-modal user input interaction within our system generally consists of a single feature and one or more object references, for example: "*What is your price* <*gesture* = *PowerShot* S70>?". Valid values for the feature tag include (in reference to digital cameras) "price," "optical zoom," and "mega pixels," while valid values for the object tag include "PowerShot S70" and "CoolPix 4300." Before such interactions can be parsed however, they must first be converted into a modality-free language. This language is formatted in XML and closely resembles the W3C EMMA standard (see www.w3.org/TR/emma/) in that each tag (i.e., FEATURE and OBJECT) contains a number of attributes like the modality type, timestamp, confidence value, and *N*-best list values.

On the flipside, multi-modal output from our anthropomorphized products must provide the resulting value information alongside reproducing the feature and object information, and be flexible enough to cater for both direct interaction: "*My price is* €500," and indirect interaction: "*PowerShotS50, price,* €*500.*"

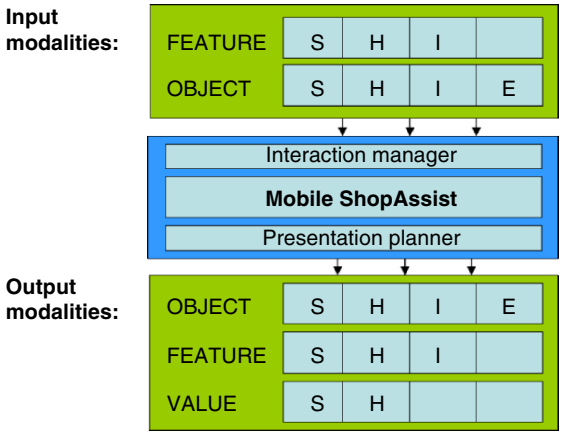

Fig. 15.5. Symmetric modality matching in the MSA

Figure 15.5 summarizes the potential range of modality combinations that exist for user input and anthropomorphized object output, when the modalities speech (S), handwriting (H), intra-gesture (I) and extra-gesture (E) are available. For input alone, possible multi-modal combinations can be seen to include: SS, SH, SI, SE, HS, HH, HI, HE, IS, IH, II, and IE. This figure does not consider multiple object referents, or overlapped input, which would create an even larger number of modality combinations to choose from. In this diagram, the interaction manager is responsible for recognizing and interpreting user interactions with the system, while the presentation planner is responsible for coordinating output for presentation back to the user. This output must be consistent not only in providing the correct information in response to user queries, but also in the choice of modality combinations that are used to present the information.

15.4.3 Output Modality Allocation Strategy

The output strategy within the MSA uses speech as a base modality to present object (O), feature (F), and value (V) information. Complementing speech is the modality of handwriting, which is used first to present the user with the transient information (O) and (F), and then a short time later with the non-transient information (V). Gesture is additionally used to show the selected (O) to the user as non-transient information, either solely via an intra-gesture on the PDA display or also as an extra-gesture via the searchlight, if it is available. Intra-gesture output for the feature (i.e., highlighting an active keyword from the scrolling WCIS text) only occurs if the scrolling text is currently visible. At the end of an interaction dialog, a user will have been presented with the same information in two complete modalities (speech and handwriting), and part of the modality gesture.

In comparison to recent usability tests in which our subjects stated that they preferred non-overlapped input modalities to overlapped ones ($\chi^2(2, N = 27) > 24.889, p < 0.000$), our users were keen to be provided with overlapped modalities for the output. Redundancy in the output modalities as used above compensated for the names of objects such as "PowerShot S1IS" and "EOS 10D" being pronounced incorrectly by the speech synthesizers, and for the transience required in presenting the written language as two separate events on the limited display space.

The current output strategy is just one of several possibilities. Other output allocation strategies include for example the exact replication of modalities used for input as for output (mimicking), user defined profiles, or profiles that limit the media to types that third person parties cannot observe (e.g., handwriting, intra-gesture, and speech output through a PDA-based headset), or that do not require a PDA (e.g., server-sided speech and extra-gestures).

15.5 Anthropomorphized Products

In this section, we outline the concept of anthropomorphization. We describe the difference between direct and indirect interaction, and also outline how we account for anthropomorphized objects in the MSA, with particular focus on the language grammars, the product personalities, and the state-based object models that define when our objects may initiate dialog interaction with the user.

15.5.1 The Role of Anthropomorphization

Anthropomorphism is the tendency for people to think of inanimate objects as having human-like characteristics. Many early cultures made no distinction between animate and inanimate objects (Todd 2002). Animism is looking at all Nature as if it were alive. It is one of the oldest ways of explaining how things work, when people have no good functional model. When users interact with AmI environments rather than with a desktop screen, there is a need for communication with a multitude of embedded computational devices in mass-marketed products. For human–environment interaction with thousands of networked smart objects, a limited animistic design metaphor seems to be appropriate (Nijholt et al. 2004); see also Chap. 14 of this book.

Although there are various product designs that use an anthropomorphic form (like the Gaultier perfume bottles that have the shape of a female torso), in the work presented here we stimulate anthropomorphization solely by the pretended conversational abilities of the products. Since the shopper's hands are often busy with picking up and comparing products, in many situations the most natural mode to ask for additional information about the product is the use of speech. When a product talks and answers the shopper's questions with its own voice, the product is being anthropomorphized.

There is a longstanding tradition among some HCI researchers against the use of anthropomorphism (Don 1992), because it may create wrong user expectations. This has lead to taboos like "Don't use the first person in error messages." People are however used to dealing with disembodied voices on the telephone, and our empirical user studies also provide evidence that most shoppers have little concern about speaking with shopping items such as digital cameras (see Sect. 15.6). In addition, through the world of TV commercials, shoppers are used to anthropomorphized products like "Mr. Proper," a liquid cleaning product that is morphed into an animated cleaning Superman, or the animated "M&M" round chocolates.

Of course, anthropomorphized interaction can be irritating or misleading, but our system is designed in such a way that it presents its limitations frankly. The WCIS mechanism in the MSA guides the users in their decision-oriented dialog and makes it clear that it has only restricted, but very useful communication capabilities. We contend that anthropomorphism can be a useful framework for interaction design in AmI environments, if its strengths and weaknesses are understood.

15.5.2 Adding Human-like Characteristics

Apart from the assortment of modality combinations available in the MSA, users may choose to interact either directly or indirectly with the shopping products. These products will in return also need to respond correspondingly. We derive the terms direct and indirect interaction from the mode of reference being made to the "person" segment of a dialog. In English for example, there exist the tenses: first person (the person speaking), second person (the person being spoken to), and third person (the person being spoken about). From an input perspective, direct interaction refers to the second person (e.g., "*What is your price?*"), while indirect interaction refers to the third person (e.g., "*What is the price of this/that camera?*"). From an output perspective, direct interaction (as used by the anthropomorphized objects) takes the first person (e.g., "*My price is* € 599"), while indirect interaction takes the third person (e.g., "*The price of this/that camera is* € 599").

Within the MSA, grammar files exist for each product type, such as "digital camera", and for both English and German. These grammar files define the recognizable input (e.g., product and feature information) for the modalities handwriting, gesture, and speech. Although the individual modalities may be used to communicate complete dialog acts (i.e., product and feature information), speech is the only modality in which complete sentences may be used. Three forms of speech input are accepted by the system, namely "keyword," i.e., speaking only the keyword, (e.g., "*price*"), "indirect", (e.g., "*What is the price of <product>*?"), and "direct," (e.g., "*What is your price?*"). The grammar files for each of the product types are downloaded

onto the PDA together with the product information, each time the user synchronizes with a particular shelf container. These files are then parsed by the PDA to create the individual grammars required for each of the recognizers.

Objects within the MSA are further personalized by one of five different formant synthesizer voice profiles (three male, two female, and all adult), which are based on parameters such as gender, head size, pitch, roughness, breathiness, speed, and volume. A limitation of our approach is that five different voices cannot provide each product in a shelf, let alone an entire store, with a unique voice. An alternative would be to use pre-recorded audio samples for each product, but this would require different magnitudes of storage space. A different approach might be to allow the PDA to assign the voices to products, which would allow at least the first five products interacted with to have a unique voice. Such an approach would also allow the use of personality matching strategies to better market products to specific user groups. Dynamic voice assignment would however also create the need for storing voice to product mappings for future use, so that returning users are not faced with anthropomorphized objects with multiple personalities.

15.5.3 State-Based Object Model

A further feature of our anthropomorphized objects is their ability to initiate interaction with the user when in a particular state; see Fig. 15.6. These states are based on variables such as a product's location, a recent extra-gesture action, and an elapsed period of time. The location of a product may be either "in a shelf," "out of a shelf," or "in a shopping trolley," and extra-gesture events include: "pick_up" and "put_down." Thus, the physical acts of the user like "Pick_Up (product007, shelf02)" and "Put_Down (product007, trolley01)" are mapped onto dialog acts like "Activate_Dialog_With (product007)" and "Finish_Dialog_With (product007)," respectively. In this case, the Put_Down

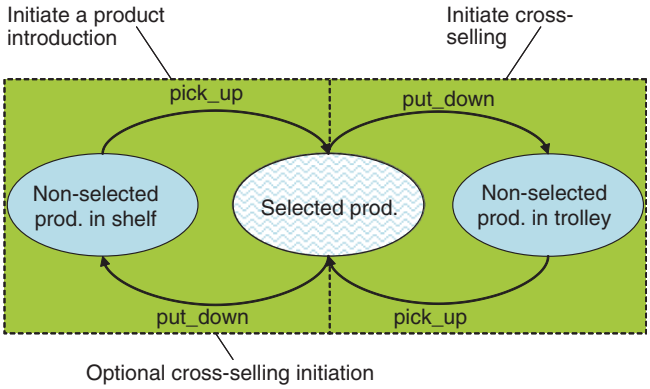

Fig. 15.6. Base product states used for object-initiated interaction

action reflects a positive buying decision as the product was placed inside the trolley, but the product could just as equally have been put down on the shelf instead, thus reflecting a negative buying decision. As an example, an object will initiate a dialog interaction if it is picked up from the shelf for the first time and no further user interaction is observed within a 5 s time frame. Silence as a powerful form of communication is well documented (Knapp 2000), and in our case such silence forces the product to introduce itself; see also (1) in Fig. 15.1. A product might also initiate an interaction when for example placed inside the shopping trolley, in order to alert the users of any further products (e.g., accessories) that they might be interested in purchasing, i.e., cross-selling.

15.6 Usability Study

Ben Shneiderman, as a prominent critic of anthropomorphized user interfaces, stated at a panel discussion documented by (Don 1992) "I call on those who believe in the anthropomorphic scenarios to build something useful and conduct usability studies and controlled experiments." That is exactly what we have done in the described research.

In this section, we describe an empirical field study on user interaction with anthropomorphized objects. The goal of the study - which was conducted at an electronics store of the "Conrad Electronic" - was to identify how accepting people would be to conversing with shopping products such as digital cameras. This study was part of a larger experiment designed to test modality preference and modality intuition. A total of 1,489 interactions were logged over the two-week test period, averaging 55 interactions per subject. Each test session generally took between 45 and 60 min to complete, during which time an average of 13.8 shoppers could be seen from the shelf's location.

15.6.1 Method

Our sample of test persons consisted of 27 people, 16 females, and 12 males, and ranging in age from 19 to 55 (mean: 28.3 years). We advertised the study by posting notices around the University of Saarland in Germany, and setting up a registration desk at the main cafeteria. Only two subjects were from the faculty of computer science. Our setup consisted of a shelf of digital cameras located in a prominent part of a local electronics store. Each participant was allocated a PDA and headset, and asked to stand in front of the shelf containing real-world camera boxes. The subjects were briefed on how to use the system and the individual modality combinations. They were then instructed to interact indirectly using the third person tense, e.g., "*What is the price of this camera?*" and then later on directly by using the second person tense, e.g., "*What is your price?*". In each case, the products responded in an aligned

manner, i.e., third and first person tenses, respectively. To ensure that our subjects spent enough time interacting each mode, they were given a series of smaller sub-tasks to complete, such as to find the cheapest camera on the shelf, or to find the camera with the largest number of mega pixels. During the test, system output was limited to a single female concatenative synthesizer voice. This configuration was chosen to minimize the effect that voice quality and limited number of voice types might have on the study. After having completed the practical component, the participants were given a small questionnaire.

15.6.2 Results

The first question that we asked our subjects was which of the two interaction modes they preferred best. The proportion of subjects that preferred direct interaction over indirect interaction (18 from 27, 66%) signifies a distinct trend for anthropomorphization, $\chi^2(1, N = 27) = 3.00, p = 0.083$. This result is seen clearer in men than in women, in which 10 from 12 men (83%) stated that they preferred direct interaction: $\chi^2(1, N = 12) = 5.22, p = 0.021$, which is significant. An advantage seen by several subjects with direct interaction was that the dialog interactions were shorter and simpler, e.g., "*What is your price?*" compared to "*What is the price of the PowerShot S50?*".

Following this question, we asked our subjects if they would reciprocate with direct interaction if the objects only spoke directly to them. Twenty-two from 27 subjects (81%) stated that they would allow themselves to be coerced into communicating directly: $\chi^2(1, N = 27) = 10.70, p = 0.001$, which is significant. Courtesy and conformity were cited reasons for this allowed coercion. Note that a "no" response to this question would result in incoherent language similar to the following:

U: "*What is the price of this <gesture> camera?*"
O: "*My price is* €599."
U: "*How many mega pixels does this camera have? <gesture>.*"

We then asked our subjects whether they would interact directly with a given range of products (soap, digital camera, personal computer, and a car), first as a buyer (B), and then as the owner (O) of the product. For brevity, we report only the resulting significance values obtained from our non-parametric χ^2 tests, where df = 1, and $N = 27$. Whereas only around 30% of people would interact directly with a bar of soap (as B: $p = 0.034$, as O: $p = 0.201$), around 70% of people said that they would interact directly with digital cameras (as B: $p = 0.034$, as O: $p = 0.033$), personal computers (as B: $p = 0.012$, as O: $p = 0.003$), and cars (as B: $p = 0.336$, as O: $p = 0.003$). Our subjects were more inclined to interact directly with the products as the owner rather than as a buyer, and this difference is best seen for the product type "car," in which a Wilcoxon signed rank test bordered on statistical significance

($z = -1.890, p = 0.059$). As the owner of the products "personal computer" and "car," men were more inclined than women to talk directly with the objects, with a Mann–Whitney U-test showing this trend in gender difference to be: $U(16, 12) = 40.5$, equating to $p = 0.072$ for both product types. Other objects that our subjects said they would consider talking directly with included plants, soft toys, computer games, and a variety of electronic devices like TVs and refrigerators.

Finally, we tested which modalities people would be comfortable using in a public environment, e.g., when surrounded by other shoppers, compared to a private environment, e.g., when no shoppers are around. Given the choice of "comfortable," "hesitant," and "embarrassed," the results showed that our subjects would feel comfortable using all modalities except speech when in a public environment ($\chi^2(2, N = 27) > 12.667, p < 0.002$). Moreover, they feel comfortable when using all modalities in a private environment ($\chi^2(2, N = 27) > 10.889, p < 0.004$).

15.6.3 Lessons Learnt

From this empirical study, our hypothesis that subjects would not simply reject the concept of anthropomorphized objects was confirmed, and indeed many of the subjects actually enjoyed the concept. The study has also shown that product type, e.g., toiletries, electronics, automobile, relationship to a product, e.g., buyer, or owner, and gender, (male, female) all have an effect on a person's preference for direct interaction with anthropomorphized objects. Future tests on the benefits of anthropomorphization could focus on a broader set of product types, the acceptance of cross-selling, and richer product personalities including distinct voices.

15.7 Conclusions and Future Work

This chapter has described a new interaction paradigm for instrumented environments based on tangible multi-modal dialogs with anthropomorphized objects. For this purpose, we introduced the concept of symmetric multimodality and applied it to speech, handwriting, and gesture. Finally, we showed via a usability field study that direct interaction with anthropomorphized objects is accepted and indeed preferred by the majority of users. Such findings have already been exploited in two other projects of our research group in which interactive installations for museums and theme parks are being developed.

Future work will now focus on scalability aspects of our approach, which will be particularly important if the system is to provide a shop full of differing products with rich forms of communication and personalities. The underlying grammars of this mobile system have currently been handcrafted for each

product type. This is acceptable when many products all have the same attributes, such as with digital cameras, but is less acceptable when many different product types exist, as would be the case when modeling the products of an entire store. We are currently developing a module to automatically generate the direct and indirect grammars based on keyword information available in the product database, and the type of question to be associated with the keyword, e.g., a wh-question (who, what, when, where, why, and how), or a yn-question (yes and no), and perhaps later also alternate and tag questions.