

Chapter 2

“...If I Had but the Time and You Had but the Brain...”: Computer-Centered Computing

Abstract Evolution of the smart home has, to a large extent, been driven by technological developments, frequently neglecting actual human needs and habits. To realize the vision of the Wise Home, a shift to a more human-centered approach is needed. In order to move to a more natural interaction, we must first accept that current interaction paradigms are unnatural. In nature, human communication is not based on a single modality, such as speech or gesture alone, but combines them as two separate but overlapping streams of communication.

Keywords Smart home • HCI • Wise home • Output modalities • Gesture-based interaction • Speech-based interaction • Sound-based interaction • Natural communication

In January, 2011, the US National Science Foundation gathered a group of 72 international researchers in Seattle to discuss the multidisciplinary problems involved in the future of networking smart tools. The discussion is summarized by Cook and Das [1]. The workshop and resultant paper focus on scalability as the key issue for the future. They broke down their concerns into eight subfields, only one of which directly focused on the human factors in HCI.

This focus on the machines rather than the people who should use them is a weakness in past and current trends in pervasive computing, despite the relatively long history of applying the perspectives of cultural anthropology and sociology to the adoption of cyber technology [2, 3]. It is a shame to think that this will continue into the future, but consider the opening sentence of the summary paper mentioned above:

The remarkable recent growth in computing power, sensors and embedded devices, smart phones, wireless communications and networking combined with the power of data mining techniques and emerging support for cloud computing and social networks has enabled researchers and practitioners to create a wide variety of pervasive computing systems that reason intelligently, act autonomously, and respond to the needs of the users in a context- and situation-aware manner [1].

The idea that intelligent agents should be making hidden decisions on behalf of humans is totally against the idea of Ubiquitous Computing (UC) mitigated with

Calm Technology (CT) as envisaged by Weiser. Despite that, and despite Weiser’s direct warning, as cited above, researchers have continued in their attempts to generate intelligent agents intended to make decisions so that humans don’t have to. Consider Diane J. Cook’s monograph in the March 2012 issue of *Science* [4], wherein she expresses an indiscriminate overlap between pervasive computing, ubiquitous computing and ambient intelligence, positing a home or work environment that is entirely under the control of intelligent agents. This is a far cry from the “gentle enhancement” of the natural environment with “self-effacing” interfaces that would “leave you feeling as though you did it yourself” posited by Weiser [5]. In fact, the idea of smart environments based around intelligent agents seems to be the inverse of Weiser’s idea of “Machines that fit the human environment instead of forcing humans to enter theirs” [6].

If it is accepted that using computers causes stress when the user feels that they are not in control, as per Riedl et al. [7], then it is a natural extension to assume that such stress would be an even greater threat in an immersive, computer-centered environment such as a smart home. Interviews and focus group sessions have shown that users prefer a centralized remote control to enable immediate interaction with a number of devices installed in a household [8]. While the concept of a control panel proved popular, as an interface [9], it is an artifact from what Weiser called the *Mainframe Era* of computing [5]. However, it may not be good to concentrate interaction with a smart environment into a single device for at least two reasons. First, if the central device does not function appropriately, the function of the whole system is affected. Second, in a conventional home, interfaces are distributed and used in combination depending on the task at hand. Given that technology were to develop appropriately, it might be a better alternative to have a system based on more or less equally powerful components distributed throughout the environment. This is what Norman called the appliances based approach, and it conforms to the current “apps” philosophy on smartphones and tablets, within which each component has a clearly defined and delimited function and responsibility.

Some researchers, such as Chan et al. [10], foresee the coming of either wearable or implantable systems to complement domotic control with the provision of biomedical monitoring sensors. While “apps” are state-of-the-art, it will be some time before these features can become ubiquitous. Chan et al. go on to stress that since smart homes promise to improve comfort, leisure and safety the interaction should be as natural as possible. If their proposed method of improvement is still developing technologically, our proposed method is built upon applying currently available technology in a novel manner.

2.1 Human-Computer Interaction

Early human-computer interaction was a multi-stage process, requiring that several specialists work on a single project. Those requiring computer assistance would consult these specialists, whose skill was the ability to communicate with the

machine. Since the machine, in those days, was essentially a series of on/off switches, all of the input mechanisms provided serial information; hundreds, thousands, even millions of noughts and ones.

The first area of specialization was the translation of the question into problems that could be presented to the computer. A question would have to be expressed as a series of logic problems, the sort that could be answered “yes” or “no”. The series of questions framed by the logicians had then to be translated into “machine language” by a group of translators, and then passed on to experts who created tapes or punch-cards. These were then passed on to the technicians who actually worked with the machine. According to one account:

This seems vastly more complex than the computer systems that we use now, but the only real difference is that most of the specialized steps in the process of human-computer interaction are now performed by the computer, rather than by the human [11].

To paraphrase Myers [12]; the change that allowed HCI to move from a field for experts to a field for common use was the realization that, through the addition of processing power, the machine could assume most of the expert roles. This was a great breakthrough in the proliferation of computers into day-to-day life. Unfortunately, as shown from the continued use of obsolete 400 codes for flagging errors (like the common “Error 404”), it gave the world a working model of computer-centered interaction that we have still not overcome. To evolve past computer-centered computing, one necessary step will be to stop our *Cross-Generational Habit* of designing interaction in accordance with obsolete technological standards like typewriters or TV screens [13]. In its place we should establish new human-centered standards based on a better understanding of the natural workings of our brains and on simple, observable facts. One example that is pertinent to HCI is the observable fact that human communication naturally involves complementing words with gestures.

2.2 A Gesture of Goodwill

In gestures we are able to see the imagistic form of the speaker’s sentences. This imagistic form is not usually meant for public view, and the speaker him- or herself may be unaware of it... [14]

As we have already said, all natural human interaction is multimodal; we constrain ourselves to a single modality only when required. When in a diving environment, scuba gear enables us to function without having to learn to breathe underwater, but formal communication is reduced to a single modality and becomes dependent on the use of strictly-defined and well-practiced gestures. When in a digital environment, the GUI interface enables us to function without having to learn machine language, but formal communication is reduced to a single modality and becomes dependent on the use of strictly defined and severely truncated words which have been removed from their usual ontological, cultural and environmental context.

Hurtienne et al. produced what they claim is “the first study looking into primary metaphors for gesture interaction in inclusive design” [15]. Their paper proposes the construction of physical gestures, based on the aggregate of twelve of what they called primary metaphors from other published studies. This list is quoted directly from theirs:

- (1) Important is central, unimportant is peripheral.
- (2) The future is in front, the past is behind.
- (3) Progress is forward movement, undoing progress is backward movement.
- (4) Similar is near, different is far.
- (5) Familiar is near, unfamiliar is far.
- (6) Considered is near, not considered is far.
- (7) Good is near, bad is far.
- (8) Good is up, bad is down.
- (9) More is up, less is down.
- (10) Happy is up, sad is down.
- (11) Virtue is up, depravity is down.
- (12) Power is up, powerless is down.

In each case, the authors quote spoken phrases that were used to support the metaphor in the original paper. At best these examples are facile, as in the pair “I feel close to him. He distances himself” used to illustrate number 5, which is easily countered with the common phrase “stranger in a strange land”. At worst, the chosen phrase does not mean what the authors seem to think. Consider the phrases used to support metaphor number 7: “Here is something interesting. There comes the difficulty.” “Here” and “there” are interchangeable in the first phrase, and the second phrase would only be correct English if “there” were replaced by “here”. This linguistic confusion is unfortunate, but it does not lessen the problem of trying to base a universal gesture on a non-representational subset of world languages. It is possible to generate lexicons of language- or culture-based gestures, as we have seen in the work discussed above. It is also possible that there will be some overlap between these gestures and any new lexicon of truly universal ones. Such an overlap, however, may be much smaller than one would initially anticipate.

If one were to pursue the idea of universal gestures, which is not the intent of this work, it might be better to turn away from simplified contrasting pairs like “up” and “down”. Consider that up and down can both mean, “at hand”, “within easy reach”, “within difficult reach”, “out of reach” and “far beyond reach”.

High or low, “out of reach” has a meaning that is universal. It is different from “within reach”, and both are different from “in-hand” and “unreachable”, but none of them are opposites. Neither are “hot and cold”. “Too hot” and “too cold” can both be mapped as cognitive vectors away from a concept of comfort. Maybe we can agree that they are both going through the realm of “discomfort”, towards a concept of “environmentally fatal”, but these would clearly all be best conceptualized as concentric spheres rather than 1-dimensional lines.

Leaving aside these logical flaws, Hurtienne et al. claim that their metaphors have not been influenced by technology. This claim is refuted in their examples and

through simple inductive reasoning. To wit: the experimenters and participants have all been influenced by underlying mental models in their most basic technological tools, such as the Cartesian increase in value when a switch is pushed either to the right or upward, and the decrease when the same switch moves in the opposite direction. Since these underlying concepts are applied globally, it would seem obvious to save time and trouble by simply using them as the basis of gestural interaction. Our attempt to do so, is described in Chap. 5.

The technology that supports gesture detection has greatly dropped in price and increased in popularity with the advent of gesture-based video game interfaces. This started with actively-broadcasting sensors in handheld gameplay controllers and active motion detectors in stationary consoles. Further improvements in the availability of gestural interaction have come from the development of smartphone applications that take advantage of the increasing presence and improved performance of accelerometers, magnetometers and gyroscopes. Despite the usability afforded by the increasing ubiquity of smartphones, empty-handed interaction is still a goal. Cohn et al. proposed a means for using the electromagnetic field generated by normal in-home wiring to detect the location, orientation, and hand and arm movements of participants in their own homes [16]. Refinements to their method, *Humantenna*, were presented the following year, with results suggesting multi-finger gesture recognition [17]. In both cases, though, the participants had to wear a backpack full of equipment. This volume is our response; a proposal for smartphone-based gestures and empty-handed interaction.

We do not want to create a gestural recognition system based on the false paradigm of single modality interaction. Instead, our gestures will amplify spoken word interaction, observing the same phonological synchronicity rules that have been observed when gestures consciously or unconsciously accompany speech in normal interaction [18]. Rather than succumb to the common behavior of designating a black box to encapsulate technological issues that have not yet been resolved but do seem to be imminently soluble, we have turned to an old triple-redundancy protocol that was used a generation ago on satellite control systems [19]. Our attempt, the S.N.A.R.K., is mentioned again below and discussed in detail in Chap. 4.

2.3 Speech and Sound

The control of networked and embedded systems through the use of automatic speech recognition has long been a feature of science fiction and fantasy interfaces, but the idea of implementing the technology in the real world predates the modern computer era, as reflected in the first volume of *Natural Communication with Computers*, from Woods et al. 1974 [20]. Even with the development of superior technology, attempts to translate the idea into real life have met only modest success, as is reflected in the title of a 2011 conference presentation by Oulasvirta et al., “Communication failures in the speech-based control of smart home systems” [21].

In 2008, Fleury et al. presented a study of speech and sound detection and classification ($n = 13$) that took place in the Health Smart Home in Grenoble, France [22]. The study is based on the use of 8 ceiling-mounted, omnidirectional microphones that are always turned on. Participants provided data in 3 ways. First they were asked to “make a little scenario” that involved closing a door, making a noise with a cup and spoon, dropping a box on the floor, and screaming the common French exclamation of pain “Aie”. This was repeated two more times. Next, each participant had to read 10 “normal” sentences and 20 “distress sentences”. Finally, the participant read a conversation into a telephone. Random selections from these separate noises and phrases were then chosen for testing the system. The authors report that the sound recognition results conform to results from laboratory conditions. Speech recognition results were “too low”, with 52.38% of the noises made with cup and spoon, and 21.74% of the screams of pain and 62.92% of the distress speech recognized as normal speech. The authors propose that one of the difficulties in speech recognition is that each participant pronounced each phrase differently each time they repeated it. Further weaknesses identified by the authors include background noise, the freedom of the participants to work at their own pace, and the uncontrolled orientation of the speaker vis-a-vis the mounted microphones. Fleury et al. sum up these weaknesses very well: “Thus our conditions are the worst possible, far from the laboratory conditions (no noise and the microphone just behind the subject)” [21]. These weaknesses were used as guidelines for developing our own testing.

In 2009, Hamill et al. also used ceiling-mounted microphones in their proposal of an automated speech recognition interface for use in emergencies [23]. They compared results with a single microphone to an array, in both noisy and quiet conditions ($n = 9$), and they tested a yes/no response dialog with four participants. Their array performed with 49.9% accuracy, and the single microphone with 29%. Recognition of the words “yes” and “no” was 93% accurate over their 3 scenarios, even though overall word recognition had an error rate of 21%. The researchers suggest that the “reason for this was because the system confirmed the user’s selection before taking an action.” The authors also report that background noises interfered with the performance of their microphone(s) and that speech recognition was greatly improved by limiting the user’s speech to two words. Again, we have been inspired in the design of our experimental protocol, to try and face the specific problems described herein. We were also influenced by another aspect of this study. In the discussion of future work, Hamill et al. mention that, in order to improve the robustness of their automated dialog system, they are developing a speech corpus recorded by an older adult speaking Canadian English. We went on to do the same.

In 2010, Chandak and Dharaskar reported an attempt to implement speech-based controls for a context-sensitive, content-specific Smart Home architecture based on natural language processing [24]. The key to their system was the ability of the user to customize the specific language or languages to be used for input. The paper itself seems incomplete, presenting none of the results promised in the introduction. In fact, no evidence is provided to indicate that the system was implemented at all.

That said, the premise of customizing input language on a dynamic, user-by-user basis informed our theoretical development and the implementation discussed in Chap. 5.

Two teams in France, GETALP in Grenoble and AFIRM in La Tronche, undertook an attempt to design a real-time smart home distress-detection system based on audio technology [25]. They started by testing speech-based detection of distress using a scenario based on a prepared corpus of phrases ($n = 10$) and reported an overall error rate of 15.6%. For their second experiment, four participants uttered prepared “distress sentences” while a radio news program played in the background. Distress went undetected 27% of the time. In 2010, the same research teams attempted to apply their more advanced sound and speech analysis system (AuditHis) to the recognition of Activities of Daily Living (ADL) [26]. They attempted to validate their stress-related keyword detection and their algorithm for suppressing background noise while using AuditHis (and installed sensors) to identify 7 ADLs. They again conducted their experiments in the Habitat Intelligent pour la Santé (HIS) Smart Home, in Grenoble, a site that they describe as “a hostile environment for information acquisition similar to the one that can be encountered in [a] real home”. Specifically, they note that uncontrolled noises outside and around the flat reduced their average signal to noise ratio to 12 dB, from the 27 dB measured in their laboratory setting. These normal, uncontrolled, noisy conditions inspired us to address the same issue in both a preparatory study and our final experiment.

As part of France’s new Sweet-Home project, Lecouteux, Vacher and Portet compared 7 sources of Automatic Speech Recognition (ASR) [27]. Twenty-one participants recorded pre-determined phrases. Each acoustic model was trained on “about 80 h of annotated speech”. In the end, they report that the array of seven microphones improved ASR accuracy and that Beamforming (as in their previous experiments) dropped the Word Error Rate (WER) from their baseline of 18.3% to 16.8%. They found that a Driven Decoding Algorithm (DDA) had only a 11.4% WER and provided slightly better results than the SNR-based ROVER system. Since the computational cost of the DDA is significantly less, and since the DDA would allow for the inclusion of a priori knowledge parameters which would significantly improve the results.

In 2011, Gordon, Passoneau and Epstein presented FORRSooth, a multi-threaded semi-synchronous architecture for spoken dialog systems as an improvement over CheckItOut, their previous pipeline-style architecture [28]. They reported on a pilot study suggesting that helping agents are helpful even when their speech recognition is not perfectly implemented. They addressed the important problem that most ASRs do not allow people to speak naturally during interaction. They suggest that a spoken dialog system (SDS) “should robustly accommodate noisy ASR, and should degrade gracefully as recognition errors increase.” This would allow “more nuanced grounding behavior from an SDS” and “help a system understand its user better.” These ideas supported our intent to create a dialog system that would help both the software and the user to understand each other better. This nuanced system is described in Chap. 5.

A different kind of sound-based output signal is reported by Bakker et al. [29]. They propose CawClock, an interactive system designed to allow a schoolteacher to set peripheral audio and visual cues by placing tokens depicting distinct animals and colors on the face of an analog clock. These placements cause sections of the clock's face to match the color of that particular token. So long as the minute hand is within the colored area, a background noise that corresponds to the animal is generated by the clock. The volume does not change but the number of animals making the sound increases as time passes, providing subtle cues that time is passing and that the end of the particular timeframe is approaching. Two prototypes were developed. An analog model was provided to the teacher in one classroom for 2 weeks. A mouse-enabled digital version was provided to another. As a reflection of the exploratory nature of the study, the teachers were taught how to use the device but asked to find their own uses for it. Two researchers then attended a 30–45 min classroom session during the second week, taking notes and recording the class on video. The teachers were interviewed singly and together at the end of the second week. Both teachers agreed that the sounds gave themselves and the children signs of passing time during periods of assigned work in the classroom. Both teachers also agreed that they had not noticed the increase in the number of animals over time. Interestingly, and in line with the fundamental understanding of peripheral perception, the ending of a marked period was reportedly more distinct when the background noise changed from one animal to another rather than when it simply ended.

Two unsolved questions in the realm of sound and speech detection have been whether or not to have constant sound detection and whether or not to have a live processor. This would mean a constant drain of both electrical power and processing power. Problems regarding processing power and the extension of battery life are easier to deal with quantitatively. The problems of accurately distinguishing sounds and recognizing speech are generally labeled “not insignificant” and replaced with a black box in flow charts and designs. As with our approach to resolving black box issues in gestural interaction, we have used an old triple-redundancy protocol as the basis of our S.N.A.R.K., a means of facilitating the accurate detection of user intent as communicated through natural means. This is discussed in detail in Chap. 5.

References

1. Cook DJ, Das SK (2012) Pervasive computing at scale: transforming the state of the art. *Pervasive Mob Comput* 8(1):22–35
2. Escobar A, Hess D, Licha I, Sibley W, Strathern M, Sutz J (1994) Welcome to cyberia: notes on the anthropology of cyberculture [and comments and reply]. *Curr Anthropol* 35(3): 211–231
3. Zayas-Cabán T (2002) Introducing information technology into the home: conducting a home assessment. In: *Proceedings of the AMIA symposium*. American Medical Informatics Association, p 924

4. Cook DJ (2012) How smart is your home? *Science* 335(6076):1579–1581
5. Weiser M (1993) Hot topics-ubiquitous computing. *Computer* 26(10):71–72
6. Weiser M (1991) The computer for the twenty-first century. *Sci Am* 265(3):94–104
7. Riedl R, Kindermann H, Auinger A, Javor A (2012) Technostress from a neurobiological perspective. *Bus Inf Syst Eng* 4(2):61–69
8. Lee S, Koubek RJ (2010) Understanding user preferences based on usability and aesthetics before and after actual use. *Interact Comput* 22(6):530–543
9. Rauterberg M (1996) Quantitative test metrics to measure the quality of user interfaces. In: *Proceedings of 4th European conferences on software testing analysis & review EuroSTAR96*, Amsterdam
10. Chan M, Estève D, Escriba C, Campo E (2008) A review of smart homes—present state and future challenges. *Computer Methods Programs Biomed* 9(1):55–81
11. Brown JNA (2004) A New input device: comparison to three commercially available mouses. Doctoral dissertation, University of New Brunswick
12. Myers BA (1998) A brief history of human-computer interaction technology. *Interactions* 5(2):44–54
13. Brown JNA “Expert talk for time machine session: designing calm technology “... as refreshing as taking a walk in the woods”,” 2012 IEEE international conference on Multimedia and Expo, vol 1, pp 423
14. McNeill D (1992) *Hand and mind: what gestures reveal about thought*. University of Chicago Press, Chicago
15. Hurtienne J, Stöbel C, Sturm C, Maus A, Rötting M, Langdon P, Clarkson J (2010) Physical gestures for abstract concepts: inclusive design with primary metaphors. *Interact Comput* 22(6):475–484
16. Cohn G, Morris D, Patel SN, Tan DS (2011) Your noise is my command: sensing gestures using the body as an antenna. In: *Proceedings of the 2011 annual conference on Human factors in computing systems*, vol 1. ACM, pp 791–800
17. Cohn G, Morris D, Patel SN, Tan DS (2012) Humantenna: using the body as an antenna for real-time whole-body interaction. In: *Proceedings of the 2011 annual conference on human factors in computing systems*
18. McEvoy SP, Stevenson MR, Woodward M (2007) The prevalence of, and factors associated with, serious crashes involving a distracting activity. *Accid Anal Prev* 39(3):475–482
19. Kaschmitter JL, Shaeffer DL, Colella NJ, McKnett CL, Coakley PG (1991) Operation of commercial R3000 processors in the Low Earth Orbit (LEO) space environment. *IEEE Transactions on Nucl Sci* 38(6):1415–1420
20. Woods WA, Bates MA, Bruce BC, Colarusso JJ, Cook CC (1974) Natural communication with computers. *Speech understanding research at BBN* (No. BBN-2976, vol I). Bolt Beranek and Newman Inc. Cambridge, Massachusetts
21. Oulasvirta A et al. (2007). Communication failures in the speech-based control of smart home systems. 3rd IET international conference on Intelligent Environments (IE 07), pp 135–143
22. Fleury A, Noury N, Vacher M, Glasson H, Seri JF (2008) Sound and speech detection and classification in a health smart home. *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th annual international conference of the IEEE*, pp 4644–4647
23. Hamill M, Young V, Boger J, Mihailidis A (2009) Development of an automated speech recognition interface for personal emergency response systems. *J Neuroengineering Rehabil* 6:26
24. Chandak MB, Dharaskar R (2010) Natural language processing based context sensitive, content specific architecture & its speech based implementation for smart home applications. *Int J Smart Home* 4(2):1–9
25. Vacher M, Istrate D, Portet F, Joubert T, Chevalier T, Smidtas S, Meillon B, Lecouteux B, Sehili M, Chahua P, Méniard S (2011) The sweet-home project: audio technology in smart homes to improve well-being and reliance. In: *33rd annual international IEEE EMBS conference*, Boston, Massachusetts, USA

26. Fleury A, Vacher M, Noury N (2010) SVM-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. *IEEE Trans Inf Technol Biomed* 14(2):274–283
27. Lecouteux B, Vacher M, Portet F (2011) Distant speech recognition in a smart home: comparison of several multisource ASRs in realistic conditions. In: *Interspeech 2011 Florence* pp 2273–2276
28. Gordon JB, Passonneau RJ, Epstein SL (2011) Helping agents help their users despite imperfect speech recognition. *AAAI symposium help me help you: bridging the gaps in human-agent collaboration*
29. Bakker S, van den Hoven E, Eggen B, Overbeeke K (2012) Exploring peripheral interaction design for primary school teachers. In: *Proceedings of the sixth international conference on tangible, embedded and embodied interaction*, pp 245–252

Building an Intuitive Multimodal Interface for a Smart
Home

Hunting the SNARK

Brown, J.N.A.; Fercher, A.J.; Leitner, G.

2017, XIII, 78 p. 22 illus., Softcover

ISBN: 978-3-319-56531-6