RC22542 (W0208-038) August 13, 2002
Computer Science

IBM Research Report

Autonomic User Interface

Arun Hampapur, Andrew W. Senior, Sharathchandra Pankanti,
Ying-Li Tian, Gopal S. Pingali, Rudolf M. Bolle
IBM Research Division
Thomas J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Autonomic User Interfaces

Arun Hampapur, Andrew Senior, Sharat Pankanti, Ying-Li Tian, Gopal Pingali,Ruud Bolle.
IBM Thomas J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
arunh@us.ibm.com

August 14, 2002

Abstract

The object of autonomic computing is essentially to minimize human supervision, i.e. the computer system must at some level manage all its processes and seek human attention only to resolve high level issues. In today's computer systems, interfaces are by far the most demanding aspect of computers– we spend more time using the "backspace" key than changing broken disks. Thus for a computer system to be truly autonomic it must possess an "autonomic user interface (AUI)". An autonomic user interface must provide users with a much higher level of service than today's interfaces while at the same time being self-aware, aware of its environment, adapting to changes and being self-healing. While tasks like wordprocessing and programming can be carried out efficiently with today's inflexible keyboard and mouse interface, many tasks that involve querying, controlling and instructing a computer can be completed more easily with an autonomic, multi-modal interface, including speech and visual inputs, and more complex output modes than the traditional monitor. The paper discusses a natural interface to a computer through the use of multiple cameras and microphones as sensors while focussing on ways of achieving autonomic characteristics in such interfaces through the use of multiple sensors, cross modality in input and output, learning algorithms and models of system architecture, the user and the environment.

1 Introduction

Autonomic computing essentially revolves around the concept of "don't bother me", i.e, the computer system must at some level manage all its processes and seek human attention only to resolve high level issues. A "botherless" interface should achieve efficient communication: with the computer understanding what the user wants to communicate, the computer effectively presenting information to the user and this two-way communication being accomplished efficiently. Communication between a user and computer can be achieved through engineered devices like today's keyboards, or through emerging speech based interfaces or through natural interfaces. Natural interfaces are those which allow users to communicate with computers as they would with other humans. The key to enabling natural interfaces is a synergistic combination of natural communication channels (like speech and vision) with context awareness. For instance, consider the following scenario. My coworker and I walk into my office discussing the design draft, as I walk in I realize I should check on the important note I am expecting, and say "Do I have any mail from John?", and the computer responds with "Yes two notes from John Doe". We continue discussing our design draft, and my coworker remarks, "We should change the web interfaces." I turn towards my computer and say, "please

make a note of that." The meeting ends, and as soon as my coworker leaves, I ask for John's email which gets projected onto my white board so I can dictate a response to John.

What may empowering computers with context involve? Since the human-centric contexts typically revolve around a 4-dimensional space of answers to the questions who, what, where and when, it may entail enabling the computer to sense, understand and relate these attributes of the users and of the world, possibly at different scales of time and space. Specifically the computer system must have the capabilities to answer the following questions:

- What is the physical space in which the interface operates? Knowledge of the type of environment in which the interface is situated (office vs lobby).
- What are the objects in this space? Knowledge of key objects in the space, (monitors, chairs, white-boards, etc).
- Are there people in the space? Presence Detection is used to signal potential command input and to activate monitoring processes.
- Where are the people? Location Determination is used to perform automatic device selection, i.e. which cameras to use for tracking, which microphones to use for audio input.
- Who are the people in the space? Automatic Identification is used to authenticate the user to the system. This could be achieved through a combination of face and speaker recognition technologies.
- What are they paying attention to? Attention Tracking is used to detect when a user issues a speech command as opposed to talking to another person.
- Is anyone talking? Speech Detection used to prime the recognition engine.
- Where is the speaker? Speaker Location used to direct attention of the computer.
- What are they saying? Speech Recognition from a set of distant microphones.
- Are they happy with my performance? Interaction Quality Assessment is used to automatically obtain feedback about the users' satisfaction with the computers performance through facial expression and voice analysis.
- What do they mean? Natural Language Understanding to interpret the spoken commands.

A computer thus empowered with context through the use of multimodal inputs will have to process the user's service request and respond to him in an appropriate manner. Effective communication of information from a computer to the user is in itself a very involved process which will be only lightly touched upon in this paper.

Autonomic interfaces have much in common with the multimodal user interfaces that have been written about in the past, Weiser's Computer of 21st Century [51], Smart Rooms [36] at MIT, the Future Computing Environments effort at Georgia Tech[17], EasyLiving at Microsoft [2], and several other efforts[39, 26]. They allow interaction with the user through a variety of modes of communication- primarily visual, auditory and tactile. In using this array of options, particularly for input, the aim is to achieve a natural user interface that is intuitive and easy to use for the user, whether a novice or an expert. However, by using the term autonomic we wish to stress particular attributes that can be designed into the system, and go beyond the aims of mere multi-modality.

Autonomic user interfaces do not simply allow the intuitive and transparent switching between modalities expected in multi-modal interfaces, but they can exploit the redundancy in modalities to achieve greater accuracy and greater ability to detect, diagnose and be resilient to, equipment failures and unforeseen circumstances. A system equipped with cameras and microphones for instance, can use both face recognition and speaker identification to identify users with enhanced accuracy, and robustness to both background noise and visual occlusion. Very high accuracy can be obtained by integrating even more identification modalities, such as gait, lip motion, typing and 'badging-in' by linking together multiple identification instances across time with visual tracking, and exploiting the continuity of identity of a person across a visual track. Furthermore, such a system can verify that the speech and vision are correctly correlated, and from the same source, thus preventing 'replay' attacks that attempt to defeat the system by playing back recorded biometric data.

Cross-modal resilience is also seen when, for instance, a user continues to give voice commands when there is an unusually large amount of background noise. Here, the system can gracefully begin to rely more heavily on visual speech (lip-reading) than is normally useful in a quiet environment. If a device, such as a keyboard or mouse, fails, the system can tell (from camera observation) that the user is attempting to use it, and infer that the device has failed, and inform the user, possibly while simulating the device's behavior by interpreting the visual input. Such autonomic interfaces are also have capabilities of self-design and self-configuration. A system should be able to correlate inputs from multiple cameras and microphones and infer the relationships between them, without an extensive calibration procedure, and also warn, as the system is being installed, of areas that are not observed by the current camera configuration, suggesting changes in the current set-up, or predicting the potential benefits of adding more input devices.

Section 2 presents an operating scenario which provides a context for further discussions. Since autonomic user interfaces are based on pervasive sensing of the users, privacy is a critical issue, this is discussed in section 3. A number of advanced technologies like speech recognition and visual tracking are necessary to enable an AUI, all the component technologies for an AUI are discussed in section 4. We propose a black board architecture which coordinates the diverse set of component technologies, this is discussed in section 5. Section 6 discusses how an AUI can be made self-aware and self-healing. Thoughts on evaluating an AUI are presented in 7 followed by the conclusions.

2 Operating scenario

An autonomic interface can be deployed in a variety of spaces, some of which are discussed later. The following scenario of a work environment with an autonomic interface is used as a running example through the document.

Figure 1 shows a plan view of a typical office environment. Once such an environment is equipped with an autonomic interface, here are some of the possible ways in which the autonomic interface could assist the user who could choose to interact with the system using voice commands.

1. S1: Computer, where did I park my car? The system tracks cars as they arrive in the parking lot and associates the parking space to the particular user upon entering the building.

2. S2: Computer,is John Doe in today? The system can identify people as they enter the building.

3. S3: Computer, set privacy level to "high". The privacy level can be configured by each user to suit their requirements, for instance user identities can be masked by the system.

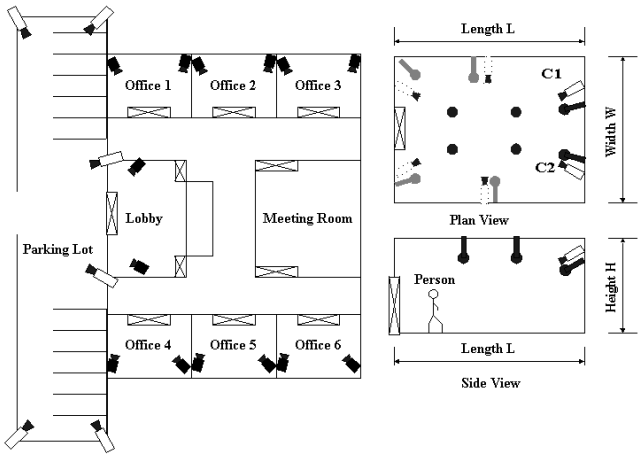

Figure 1: **Left:** An Office Environment equipped with an autonomic interface. **Right:** Plan and Side views of a room instrumented with multiple cameras and microphones.

4. S4: John Doe is meeting with me in my office, and wants to schedule a follow up next Tuesday at 10 AM. I turn, to the computer and say, Am I free Tuesday at 10? The computer recognizes the command, checks the calendar and responds.

5. S5: My friend calls me at on my office telephone number, and the computer knowing that I am currently in the lab automatically forwards the call to the lab.

Autonomic user interfaces may be situated different spaces and serve a variety of purposes. The office environment discussed in section 2 is one possible place for an autonomic interface. Following are some of the other potential places for an autonomic interface.

Home: The autonomic interface would optimally control several devices at home while learning the personal preferences of the occupants, like controlling lights, music. The AUI would also keep track of activities of people in the home acting like a scratch pad. Several other functions monitoring and providing support to elders have been explored in the Aware Home project [29].

Vehicle: Autonomic interfaces in vehicles are already being implemented for cell phone control and climate control. Additionally, AUI can aid in navigation and in tracking driving behavior for safety purposes.

Public Spaces: The role of autonomic interfaces in public spaces is more observation than interaction. However an autonomic interface based solely on cameras can keep track of traffic, space usage [20], measure the effectiveness of displays and perform several other functions. Consider the following example.

I am walking around a mall and pause by a product display. The display sensing my presence says, “Hello, welcome to Spectrum Sports.” Tracking my gaze, the system identifies that I am looking at the tennis racquet and says “Could I give you more information about our line of tennis equipment?” I say, “Yes,” and the system starts a promotional video.

3 Privacy issues in natural interfaces

In order that interfaces exhibit meaningful autonomic behaviors, it may often be necessary that they acquire, store and learn a significant amount of personal user information. The subjects of this observation, whether users of the system or individuals who are unknowingly exposed to the system's sensing mechanisms, must have the rights and mechanisms to know what information is being gathered, how it is used, and there must be mechanisms for assuring that the information is used only in ways permitted by law and agreed to by the users.

Users will be concerned about how the stored information could be abused for purposes other than those for which it was originally gathered from the users, whether the novel purpose was merely unforeseen, or the result of a new technology enabling a novel use of thought to be innocuous when previously gathered. In either situation, the users would like to know their choices in correcting the existing unsatisfactory situation. Clearly, the scope of privacy issues is very broad and these issues are just beginning to be addressed, mostly in the restrictive context of the voluntarily disclosed information and much remains to be done to formulate clear guidelines for more diverse content. The purpose of this section is to discuss these privacy issues within the context of the autonomic interfaces.

Autonomic interfaces will be equipped with an array of sensors, some of which provide very rich information (e.g. cameras) and others could be potentially very intrusive (e.g. mm radar). Further, the inputs to autonomic interfaces are often implicit, the interfaces could be physically pervasive, disembodied and in a position to observe the users over extended periods of time, in many situations and with perfect recall. Consequently the users of an autonomic system may be largely unaware of what information is being gleaned by the system, what is being stored and how is it being used.

Given that autonomic systems could be using extended sophisticated learning capabilities, it will be difficult to identify and understand what information is being used by the system in its modelling of and reaction to the users. In particular, it may be difficult to prevent a system from capturing subtle, extraneous, objectionable, or even illegal correlating experience (e.g. shabbily clad customer may imply poor returns on service), or subsequently prove that this was taking place.

On the other hand, the existing manual/semi-automatic systems are scarcely above reproach when it comes to use of fair use of personal information. If existing laws and lawsuits are any guidance, we cannot help but reach the conclusion that these systems are ridden with many problems involving (un)conscious, unconscionable and unfair biases of people.

The advantage of having a concrete automated system is that the system behavior can be modelled deterministically, the (inter)actions of the systems can be meticulously logged, and thus provide an objective basis for investigation. In an increasingly automated world, then, the issues are more concerned with whether all the concerned parties have fair access to the information and designs which ensure how these properties of the system can be effectively guaranteed. While we are very far away from accomplishing these lofty goals, right now it appears that closed system designs supporting fool-proof access accountability could provide a possible interim solution.

More positively, given our preference for the interface being sensitive to human issues and that some of the communication media are secure and others less so, the progressive interfaces are expected to automatically determine the best possibly methods communicating different messages conveying information of different human nature. With access to multiple modalities of communication and to the user model, autonomous interfaces are in a position to be more sensitive to its users' privacy needs.

4 Component technologies

In this section we review the major technological components that would provide information to the autonomic user interface. The primary sensing modalities, echoing the primary human senses, are acoustic and visual. These provide the richest sources of data, but will be supplemented in many cases by other modalities.

4.1 Detecting and locating people

Detecting and locating people are of fundamental importance to the activity of autonomic interfaces. Knowing that people are present is a minimum requirement for useful interaction, and knowing how many people are present may be required by a variety of processes, from usage monitoring to directing fire rescue.

Knowing peoples' locations enables the interface to focus attention on the users, both for input and output. For example, a pan-tilt-zoom camera can be directed towards the speaker to capture a better image of his/her face (for face recognition, facial expression analysis, lip tracking, or simply for transmission); microphone arrays can be steered toward a speaker to pick up his/her speech signal with high fidelity for subsequent analysis by a speech recognition system or simply for transmission; directional loudspeakers can be used to spatially direct sound output towards the user [49] ; visual output can also be directed towards the user (by projecting the display at the user's location [30], or rendering graphical objects on a monitor with the user's location as the focus, or turning a synthetic avatar towards the user).

Both detection and location of people can be carried out by a number of technologies. The most direct and reliable of these are explicit presence tags such as active badges [50] or active bats [21] which broadcast their presence or location using infra red or radio. A survey of location technologies can be found in [22]. An obvious disadvantage of such possession-based systems is that they entail the inconvenience of having users remember to these contraptions and misplaced tags result in failure of the system.

The second category of the technologies used for human detection are based on what humans do: Humans are vocal and mobile; they can be detected based upon when they speak, move or act. Some specific acts, such as swiping a badge through a reader or hitting a key on a keyboard give clear indication of the presence of a user, but are clearly limited in scope. Detecting speech or motion are do not necessarily indicate human presence, but are useful cues that can be exploited in constrained circumstances, or in conjunction with other cues.

A number of human physiological characteristics (such as pulse, breathing, heat, exhalation of carbon dioxide) have been used to detect humans in adversarial situations, but are not generally applicable to the user interface scenarios. Sensing of human presence using near and far infra-red has received some attention lately [53, 40]. The near infra-red sensing of humans is an emerging technology which is mostly restricted to low resolution imaging of humans (e.g. 1 sq. ft. pixels) and therefore, it may be difficult to differentiate humans from other warm bodies using this technology. The far infra-red imaging systems offer detailed imagery but is prohibitively expensive for widespread deployment. Both technologies suffer from interference from other sources of infra-red irradiation. One of the strengths of these technologies is that no illumination is required, and combining these sensors with other modalities broadens the range of conditions that can be handled, and can make person detection and location more robust.

Vision is the sense that provides us with our richest source of information about the world. Many autonomic user interfaces will need to have computer vision not only because it is a practical way of acquiring the information that they need, but to achieve natural (i.e. human-like) abilities we would wish them to have the same sensing capabilities as a human, and see the world the way we do.

Human detection using visible light cameras has some obvious strengths. First and foremost is the richness of information available from the sensor. Not only can people be detected and located, but details of their appearance can be measured, and inanimate objects can be sensed in the same way. At a basic level, computer vision can be used to detect the presence of people. The popular technique of background subtraction [25], by modelling the appearance of the static objects in a scene, is able to segment out the moving or previously unseen parts of a scene. Since people are constantly moving, this or the simpler frame differencing, can signal the presence of people in the scene, and their location within the acquired image.

If a person is within the field of view of more than one camera, stereo or triangulation can be used to accurately determine the person's position in a three-dimensional coordinate system, assuming that the cameras are calibrated. Even without multiple views, knowledge of the geometry and the assumption that the person is standing on the ground plane gives a good idea of the person's location.

There are some obvious disadvantages to using vision: (i) occlusion and camouflage Occlusion from other objects or humans can often become problem for effective detection. Similarly, insufficient contrast with the background may also cause failures in detection. (ii) complexity: human form unfortunately comes in all shapes and forms. Our understanding of independent visual detection of humans with different clothing, hair styles, jewelry, disfigurements is at best limited. (iii) illumination: sufficient ambient illumination is necessary for sensing of human beings; (iv) privacy concerns: In addition to presence/absence information, the visual appearance provides many other pieces of information and may potentially be objectionable to the subjects.

Given that any particular modality for human detection may be (occasionally) prone to failure, one of the favorite methods of human detection uses both the human form appearance as well as their motion.

4.1.1 Speaker Location

Knowing the location of the speakers in a space may be more important than knowing where all the people are, and is clearly a valuable source of information for such general person location. Acoustic source localization methods can be broadly classified as steered beam-former based, spectral estimation-based, and time delay estimation-based. Of these, the time delay estimation-based locator is by far the most popular. Acoustic source localization techniques have the advantage of operating in poor lighting conditions, performing omni-directional sensing, and ability to localize in spite of visual occlusion.

Time delay estimation-based locators use the difference in time of arrival of a speech signal at different microphones, along with knowledge of geometry of the microphones, to estimate the location of the speech signal. With two microphones, the time delay in arrival of speech defines a three-dimensional hyperbolic surface on which the source can lie, commonly referred to as the cone of confusion. Three or more microphones can be used to localize the sound source in 3D. The biggest challenge in time delay estimation is reverberation in real world environments. As a result, numerous techniques have been proposed in the literature for optimal estimation of time delay. One of the earliest and popular techniques is generalized cross correlation [31]. The principle was extended in [18] to better separate direct sound from reverberant sound and use the characteristics of speech to avoid picking up other sounds. Others have employed cross-power spectrum phase techniques with large arrays of microphones [41]. A closed-form solution for quadruples of microphones was presented in [5] as an alternative to compute-intensive optimization techniques needed with microphone arrays. Another approach for dealing with reverberation is the adaptive eigenvalue decomposition algorithm [3]. Here, instead of correlating the two signals, the technique tries to directly determine the relative delay between the direct paths of two estimated channel impulse responses.

Attempts at combining visual and acoustic processing to estimate speaker location have started ap-

pearing only in recent years [38, 39], although there has been a longer history in multimodal systems that combine speech recognition with visual lip-reading [37]. Combining acoustic and visual processing for speaker localization can result in a more robust and more accurate system. This is particularly important in an autonomic interface. An autonomic interface should have both visual and acoustic localization systems, and the ability to integrate these systems, or use them individually. The autonomic system should continually associate *confidence factors* with acoustic and visual modalities. The system should be able to switch between modalities or integrate modalities based on these confidence factors.

4.2 Determining the user's pose

Within an autonomic user interface, it is important in many circumstances for the interface to be able to determine the focus of the user's attention. In human interaction, our gaze, head pose and gestures all transfer important information about the person being spoken to, the object referred to, a spatial location and so on. The user interface must be able to pick up these cues and interpret them correctly.

In case the user, is reading something off a monitor, change in attention could be detected by change in eye gaze [56]. In the case of autonomic interfaces one of the critical requirements is to allow distant interaction with devices, in such a case, head orientation is the most significant cue for attention tracking. Specifically, for the task of attention tracking the most critical head parameters are the azimuth or pan followed by the elevation or tilt.

Estimating the head orientation of a person can be done by various means and has been used mainly in VR systems[1]. However most of these require the user to wear a special head tracking device which is not desirable in the majority of interfaces. Tracking head orientation based on camera input provides the advantage that it does not encumber the user. There have been several research systems for head tracking[48]. The techniques that can track head azimuth and elevation rely on three distinct techniques, namely feature-, appearance- and template-based. Feature-based techniques rely primarily on locating known landmarks on the face and using their geometry to determine pose. Appearance-based techniques typically learn the relationship between the pixel appearance of the face and the orientation using a technique such as neural networks or principal component analysis. Template-based methods rely on doing a pixel by pixel match between a known template and the current image to determine the pose. Most of the techniques in literature operate on fairly high resolution face images, typically where the user is 2–8 feet from the camera. Also, most techniques can measure face orientation only in a limited range of azimuth, rarely more than $\pm 90^o$. In the case of autonomic interfaces the the face orientation can vary through 360 degrees and the subject is likely to be much farther away from the camera, resulting in much lower resolution head images. Detecting features in low resolution images is very challenging, hence an attention tracking system for distant interaction will be appearance-based and use additional cues from the body shape and motion to estimate head orientation.

4.3 Acoustic signal acquisition

Speech is of fundamental importance for person-to-person communication, and as such people have long aspired to making speech a mode of interaction with computing devices. Research over the last thirty years has begun to make this dream a reality. In an autonomic interface, the system should be able to react to spoken information in a natural manner, with understanding comparable to that of another human. To make this possible, a number of problems must be solved by the autonomic infrastructure. In the first place, simply detecting whether someone is talking is a difficult problem. The traditional method of detecting acoustic energy at a microphone, works only for microphones close to the speaker's mouth and in quiet environments.

To work in more challenging conditions with background noise, several speakers and with microphones built into the infrastructure, rather than held close to the speaker's mouth, then the problem becomes much harder, and we look to humans for ways to find who is talking. Firstly, people are good at recognizing types of noises, and research has been carried out on, for example, distinguishing music from speech [44]. In addition to using multiple microphones to limit spatially the signal acquisition, blind source separation and auditory scene analysis can be used to tune the signal to a particular sound source.

Stationary background noise can also be modelled (e.g. Parallel Model Combination [19]) or subtracted from the signal (cepstral means subtraction [52]) to leave more interesting parts of the signal present, but further research is necessary to endow the autonomic interface with the human powers of distinguishing a human voice from other sounds, including, say, knowing the difference between a voice on a radio and a live voice. Clearly, though, humans use visual cues to solve this problem. We know someone is talking, and which person, by seeing their lips move. A system that determines if a person is speaking is described in [12]. This system combines person detection, head pose (for focus of attention) as well as lip motion and acoustic energy to determine if a user is speaking. Thus far, such a system works only under controlled circumstances, such as a desktop computer user, or [42] the user of a kiosk, but with the techniques described in section 4.1 for acquiring high resolution face and far-talking speech signals, it should be possible to extend this to allow an autonomic interface acquire such information throughout its domain.

4.4 Person identification

Knowing who the people in the space are is of fundamental importance in many applications. We can define two types of identity— relative and absolute. Absolute identity is an identification of a person tying them to a known individual record in a database, perhaps containing name, employee number and an index of security clearances. Usually the user has been pre-enrolled and some biometric data (such as a photograph) captured at enrollment so that identification can be carried out. Relative identity merely consists of maintaining *continuity of identity* — tracking a person so we know that the person tracked at time t' must be the same person tracked at time t because they've been well tracked throughout the interval. This requires continuous tracking, or methods (as simple as recognizing someone's clothing) to associate tracks at different times with one another. Similarly, some biometric data (e.g. face or voice as discussed below) could be acquired at time t and verified later.

Absolute identity requires acquiring enough information about the user from the sensors in the environment to identify him or her unambiguously. The information can be solicited from the user (for instance requesting a user name at a terminal, or a badge-swipe at a doorway) or derived from measured data (sensing an active badge, recognizing a face). To prevent users from trivially impersonating one another, such identification should involve a unique personal possession (such as an employee badge or credit card); proof of secret information (a password or PIN); or biometric verification (face, voice, fingerprint recognition).

In many autonomic installations, where we wish the interface to be minimally intrusive, the system should acquire reliable knowledge of the identity without making demands on the user. In this case acquiring biometric data at a distance is highly desirable. Here there are two biometrics that stand out as being particularly applicable— face recognition and speaker identification. A space equipped with cameras can passively acquire face images from the people in the space as described above, and these can be identified using a face recognition system. Similarly, if the user is speaking, or can be requested to speak, the user's voice can be captured with microphones, and recognized with a speaker ID system.

Clearly the problem is not so simple. Acquiring facial images or speech with sufficient quality to determine identity is no trivial matter in an unconstrained situation. Acquiring such signals is dealt with in

sections 4.3 speech. However, it is worth mentioning here that most face recognition systems are designed to work from frontal images, with performance degrading as the subject turns away from the frontal view. Some research has investigated the recognition of profile images, or recognition from 3-dimensional data. Also there has been some research on integrating information from many frames of video to give a better determination of identity.

Another biometric of relevance for passive acquisition is gait recognition. This is still in its infancy, and has a relatively low discriminative power, but it may help distinguish among a small group of subjects, or provide additional information for maintaining continuity of identity. Lip motion identification — recognizing the unique way an individual's lips move while speaking — has also been used for identification, though principally this is used in combination with speaker identification and face recognition. This *multi-biometric* system uses three different biometrics to determine an individual's identity, providing a lower error rate and much greater resistance to replay attacks by impostors.

An autonomic interface that determines the user's identity will use just such a combination of methods to determine and corroborate the identification. A single biometric may be susceptible to noise, failure or deliberate deceit, but having multiple modes of identification makes the system much more reliable, accurate and trustworthy. The corroboration process, where multiple sources of biometric information are *fused* together has received some interest in recent years [24, 33] though the adaptive reweighting according to dynamic reevaluation of the perceived reliability of each of the information sources is still an open problem.

Since it may only be possible to identify individuals at certain times (such as when the pass in front of a camera, speak or swipe a badge), *continuity of identity* is important even in absolute identity systems. Having identified a person, we must track them to know their identity at a later stage when the identity can not be verified directly. The tracking can, of course, precede the identification operation, so that, for instance, we track a person leaving a car and crossing the car park before they badge into a secure building, or can be seen by a face recognition system. At that point they can be identified, and we can retrospectively associate the car with the individual. Likewise shopping habits (Which products were browsed but not bought? How long did the shopper spend in the store?) can be constantly monitored, and finally associated with an individual's record when the person shows their loyalty card at the checkout.

Further, continuity of identity allows the fusion of evidence across time. We may see a user's face, and then track them, later hearing them say something, all the while acquiring gait information and successively building up a more confident estimation of their identity, until they log in at a computer with a password and fingerprint, putting their identity beyond doubt. Such evidence accumulation must, of course, take into account any inaccuracy of the tracking, to recognize that there are circumstances when continuity of identity cannot be assured, as when two people go out of range or sight of the tracking sensor.

4.5 Speech recognition and understanding

Hitherto, speech recognition has largely been confined to users in quiet rooms speaking directly into microphones. Recently it has become possible to voice dial a cell phone in a car, though on a vocabulary larger than ten to fifty words such a system would be unusable, and the performance relies to some extent on the regularity of the noise in a car.

In an unconstrained, challenging environment the autonomic interface of the future will marshal a wide range of complimentary techniques to provide good, and potentially superhuman speech recognition capabilities. As with the other tasks, described in this paper, one of the keys to robust, reliable speech recognition is the combination of different modalities, and in this we can mimic the well-known human ability [46] to use visual information to enhance the comprehension of heard speech. Lip images can be acquired by a

telephoto lens steered onto the user's lips, and used to improve the decoding accuracy through joint modelling – ideally adapting the relative reliance on the two signals according to their perceived reliability, just as 2001's HAL is able to lipread when the acoustic signal is unavailable.

Another powerful way of improving the usability of a speech system is by controlling the language that the system is able to understand. In a difficult, noisy environment, limiting the vocabulary to digits, results in acceptable performance, but in very good conditions, the user should be able to converse naturally with the machine, using spontaneous (and thus poorly pronounced and ungrammatical) natural language. Strong modelling and continuous re-estimation of the domain can simplify the speech task (by reducing the perplexity, a measure of the branching factor or uncertainty in a conversation) without imposing artificial constraints on the user.

Recognizing speech is, for some applications such as dictation, an end in itself, but most of the time, the autonomic interface needs to interpret the speech, and act appropriately according to what was said. Some of the speech that the autonomous interface hears, must simply be transcribed (say to record the minutes of a meeting), but some of the speech consists of commands directed at the computing infrastructure. Distinguishing the two situations is of crucial importance, and depends largely on the speech content, but this can be combined with other cues (eye contact with an embodiment of the interface, change in tone of voice) to determine the user's intention underlying the speech act.

Interpreting the commands involves the understanding of natural language, itself an extremely difficult task. So far practical systems with humans interrogating a machine system have restricted themselves to limited domains, such as a virtual travel agent [27] although recent years have seen these systems progress from constrained grammar, limited vocabulary systems to systems capable of understanding natural language within the chosen domain. For the foreseeable future, any autonomous interface is liable to have a restricted domain, or at best be able to model a small number of domains, perhaps switching between them based on explicit or implicit cues. Further into the future, systems should be able to understand anything that we would expect a human to, using domain constraints to understand otherwise ambiguous sentences, but able to switch domains abruptly as the speaker changes topic.

5 Operation of a natural interface

An autonomic interface involves the interaction of multiple software processes — some which observe the user and others which respond to user requests. These processes are inherently distributed, for example, the speech recognition will run independently of the visual person tracking, but both of them need to interact in the context of the autonomic interface. There are several mechanisms which facilitate such distributed process interaction. Basically we require an architecture containing a set of autonomous processes operating independently on their own domain of expertise and making available results for use by other experts.

At a high level, we could consider this like a blackboard systems [8], with all experts' results available to all other experts though in practice, looked at at a finer scale, the communication can be much more directed and local, and is more likely to be implemented with message-passing between components that discover each other through a hierarchically organized registry scheme.

Figure 2 shows a global architecture for an autonomic interface. It consists of a shared solution space which is accessible to all the knowledge sources for both reading and writing. Each knowledge source can post a set of known events and reacts to a events from some of the other knowledge sources. The events posted vary in complexity ranging from simple events like *person detected* which could be a boolean variable through *person track state* which could be a complex measurement of various aspects of a person's motion or a speech transcript. Several of the knowledge sources shown in figure 2 have already been discussed in

the previous sections, details of other knowledge sources are presented below. In fact several of these KSs are themselves complex systems consisting of a number of separately operating knowledge sources.

Person Tracker KS: This knowledge source performs the tracking of multiple people in the AUI space using typical multi camera tracking techniques [45]. The trackers used here typically estimate the parameters of a full body person model. This typically includes the position of the person in the space and various body parameters like head orientation, joint angles for the limbs, etc.

Gesture Recognition KS: Depending on the application the AUI may use several visual gestures as input. For example, the head orientation may be used to determine the users attention and his hand pose may be used for pointing gestures. The gesture recognition knowledge source uses the output of the person tracking KS in conjunction with the world model 6.1 to recognize the users gestures.

Command KS: This knowledge source uses the outputs of the speech recognition knowledge source in conjunction with the gesture recognition knowledge source, natural language understanding and the command vocabulary to interpret the user input to generate commands to the system.

Output Generation KS: This knowledge source selects the appropriate modality and device for responding to the user interactions with the AUI. It uses information from the world model, user preferences and context information in communicating the system output to the user. For example, if the output is response to a the query, "How are the traffic conditions on I287", the system may respond with audio output if the user is not near a monitor. In the case of displaying an email, the system may project the email onto a projection display.

Autonomic Behavior KS: This knowledge source is responsible for managing the autonomic behaviors of the system, which include error detection, adaptation and error recovery. A detailed discussion of the autonomic behaviors is presented in section 6

Implicit Communications KS: One of the unique aspects of autonomic interfaces is their ability for implicit communication. Implicit communication is the ability of the computer "to know" things which the user has not explicitly communicated to the system. For example, screen savers would be an implicit communication in today's systems, i.e, the system observes the lack of keyboard activity for a period if time and turns on the screen saver based on this observation. In autonomic interfaces this capability is more advanced and uses cues of human behavior to drawing inferences based on the users action.

Figure 3 shows an example of the interactions between the knowledge sources and the blackboard during a single user interaction. When a person first enters the scene the person detection KS detects and posts a *new person event* onto the blackboard. The *new person event* is observed by the person tracker and person id KS's. The person tracker KS tracks various body parts of the person and estimates a set of body pose parameters. The person ID source uses a combination of face and voice (if the person is speaking) to identify the person. This results in the blackboard being updated by a the *person ID* and the *person track state*. The user preference KS observes the *person id* posted and posts the users *personalization data* onto the blackboard. The posting of the *personalization data* is used by the person tracker KS to update its privacy settings, the output generation KS updates its personalization data (for example, a particular user may not want to use audio outputs). The speech recognition KS loads a personalized vocabulary and training data. The posting of the *speech event* by the speech detection KS causes the speech recognition engine to

start the recognition process which posts the speech transcript onto the blackboard. Simultaneously, the gesture recognition KS and the world model KS are observing the person track state, resulting in the world model posting a *output device proximity event* as the user approached a output device. In the meantime, the command KS acts on the speech transcript to parse it for commands and posts a *command event* onto the blackboard. The output generation KS observes the command and acts on it to send the response to the appropriate output device. This interaction clearly illustrates the parallel distributed processing that is necessary for an autonomic interface which is provided by the blackboard architecture.

6 Autonomic behaviors

Explicitly managing/maintaining large complex hetereogenous interfaces is very cumbersome and labor intensive. Systems capable of exhibiting autonomic behaviors – those which are aware of their environment, their own state and are able to adapt to changes and heal themselves are desirable from the perspective of effective management as well as transparency. The following are the most important autonomic behaviors.

Environment Awareness: In the case of a AUI, this entails the system having knowledge of the environment in which it is installed. For example, in order to automatically forward a phone call, the AUI must be aware of the location of the user in the building and the telephone closest to the user. Such environment aware services require the system to have a model of the world with geometric reasoning capacities. Section 6.1 presents the details of a world model.

Self Awareness: Self awareness is the foundation for a error detection and self healing. In addition, self awareness is essential in various functions such as tracking across multiple camera's, automatic switching between audio and visual modalities, etc. Self awareness is achieved through a combination of the world and system models (section 6.2).

Error Detection and Correction: Error detection can be classified into two types, namely component failures (like camera, microphone, server or software failures) and performance degradation (like reduced confidence in person identification or speech recognition, etc). Component failures will involve using alternative sensors, servers etc, while performance degradation will involve system tuning. Section 6.3 discusses the various aspects error detection and correction.

Learning and Adaptation: The autonomic interfaces should have ability to adapt to different users and different usage situations by learning differences in the environment, and knowledge, style, and preferences of their users. There are three central features for an adaptive interface. First, the interface needs to maintain a user model which can be easily inspected and modified. Second, the interface can learn the environment and users' behaviors to improve its performance in some task domain based on partial experience with the domain. Finally, the interface can adapt its behavior by using the learned knowledge and make recommendations to the user. Details are discussed in section 6.4.

6.1 The world model

The world model is a representation of the "world" in which the autonomic interface is situated. The model includes a representation of space, objects in the space, function etc. The model should support queries of the following kind.

Proximity queries: Given a spatial location, find all objects in the world with a certain distance, given an object find where it is.

Visibility queries: Given a spatial location and gaze direction and field of view, find all visible objects.

Attribute queries: Find all objects that can be used as a projection surface. Find all exits from the space.

Constraint queries: Is it OK to perform function X at a given location. Can I forward a telephone call to a meeting room?

The requirements of the world model can be satisfied by an object-oriented geometric representation. Figure 4 shows a geometric representation of the world. The world is represented as a set of objects, each object has a base coordinate system which is defined with reference to its parent object. Ultimately each object in the model can be expressed in the world coordinate system. The following shows an example of the structure of an object.

```
Object( \\
Transform:                      ; Transform with reference to
parent\\
Link to Parent                  ; A link to parent object\\
Object ID:                      ; Unique Identifier\\
Dimension: WxLxH                ; Bounding volume\\
Object Count: N                 ; Number of contained objects\\
Contained Object Pointers[N]    ; Pointers to contained objects\\
Connection Count: C             ; Number of entry-exits     \\
Connection Object Pointers[C]   ; Pointers to connected objects\\
Renderable Description:         ; Facet model of object\\
Mobility:                       ; Fixed, Semi-permanent\\
Functions:                      ; Enumerated Type\\
Audio Visual Properties:        ; Enumerated Type\\
Interaction Status              ; Enumerated Type\\
 )

Functions (\\
MONITOR, PROJECTION SURFACE, TELEPHONE, CAMERA, MICROPHONE, UNKNOWN) \\
```

In the proposed model, the object oriented representation allows for extensibility while the geometric basis allows for geometric reasoning. In order to increase the efficiency of access, the bounding volumes of the objects can be indexed using a spatial indexing structure like the R-Tree [43]. The issue of acquiring a model for a large space can be fairly expensive. A method of bootstrapping is described in section 6.1.1. Several of the research efforts in context aware computing have proposed the use of similar models [6, 23, 2].

6.1.1 Model acquisition

Model acquisition can be a time-consuming step in the installation of any user interface of the kind we are describing. In the past the creation of the world model has been a laborious manual operation requiring extensive measurement, calibration, labelling and meticulous organization.

One of the advantages of the future autonomic user interface is, however, the relative ease with which the world model can be constructed, much of it being acquired through an automatic process, with minimal supervision and external input.

While the component technologies are not yet fully developed, and integration is even further into the future, the following is the scenario that we envisage for a self-installing autonomic user interface.

- The system designer chooses the areas that need to be covered by cameras, microphones or other sensing modalities
- The cameras and microphones are temporarily installed in the actual space, or in a computer model of the space, so that the system can analyze the actual coverage achieved and determine where sensors need to be moved or added to achieve specified performance requirements.
- Cameras and microphones with pre-calibrated intrinsic parameters are installed in the space. Overlapping fields of view for the cameras are hypothesized based on background appearance, and confirmed by observing people moving around the space. In a densely instrumented space, the spatial geometry can be learned in this way, but in sparser installations cameras' interrelationships may need to be acquired from another source, such as GPS in the cameras or a human designed CAD model.
- Microphones are likewise calibrated, semi-automatically with reference to audio-visual events.
- Fine calibration can be carried out by a human or robotic agent traversing the space with an object of known size and a device for emitting noise pulses.
- Objects in the world model may be identified by sophisticated object recognition technologies, or manually labelled by the designer. Self-labelling will involve techniques such as sending known patterns to a specific display and observing where those patterns appear, dialling particular telephones etc.
- The system may have to learn much other information either explicitly or experientially over a period of time. For instance learning that a particular room is known as "the lab" can either be explicitly taught; can be learnt by association; or the information can be explained in response to the system asking for an explaination of an unknown term.

Clearly much of this automation lies in the far future of artificial intelligence, so in the short term, much more manual intervention will be required. Dimensions and layout of rooms must be acquired from digitizing building plans; furniture and objects must also be measured and inserted into the CAD model, and labelled explicitly. Cameras must be individually and collectively calibrated and registered with the model. [9]

6.2 The system model

The goal of the system model is to provide the necessary representation which allows the system to be "self-aware" and "self-healing". Automatic reconfiguration for both hardware and software components and the ability to localize faults and isolate affected system components are essential prerequisites to self awareness and healing.

Plug and Play: The AUI should have the architecture to support resource discovery and management for both hardware and software components. This can be accomplished by augmenting the blackboard architecture with plug and play capabilities. Plug and Play architecture has been used to set up, configure and add peripherals to a PC's [54]. A similar approach can be used to provide dynamic reconfiguration capabilities to the AUI. Figure 5 shows the device abstraction for a camera, server and the person id service. The abstractions provide a set of standard services which provide information about the device and also allow for controlling the device.

For example, an AUI installation has been using face recognition technology to achieve person identification. A new speaker recognition module (SRM) is added to the AUI. The SRM starts to add new speaker identification results into the blackboard. This information is automatically recognized by the person ID module which uses the information to improve person identification accuracy.

Fault Localization: Given that an error is detected in the functioning of the AUI, the diagnosis procedure will be required to localize the error to a particular component. This requires a representation of the information flow within the AUI. Figure 6 shows the system configuration graph for a portion of the AUI. As discussed in the camera failure scenario, this graph is used to isolated dependent components. The system configuration graph is a dynamic structure which is updated as the configuration changes due to various autonomic behaviors.

6.3 Error Detection and Correction

In this section we discuss the error detection and correction processes in an AUI. The errors in AUI's can be classified into *component failures* and *performance degradation*.

6.3.1 Component Failures

Component failures can be detected using a query-response mechanism in combination with an interrupt on error mechanism. Each component (both hardware and software) in the AUI will have an addressable software module called the component monitoring module (CMM) which responds to a central monitoring module called the autonomic behavior manager (ABM). For example, associated with each camera is a CMM which monitors the image stream to detect camera failures. If a camera failure is detected, the CMM sends a message to the ABM. The ABM can also query the CMM for status, this allows for the detection of catastrophic failures (like the crashing of the machine running the CMM).

Once a malfunction in one of the cameras is detected by the Autonomic Behavior Manager. The ABM uses the the system configuration graph to isolate all he KSs that receive input from the camera and shuts them down. The ABM initiates a resource discovery call to locate another camera with similar space coverage as the failed camera. Once the camera is located, the ABM restarts all the affected processes on a machine which can receive the camera input.

6.3.2 Performance Degradation

Unlike traditional error detection and correction techniques in which the transmitted signal is coded to achieve error detection and correction, in the case of signal interpretation (speech, face reco, visual tracking etc) the challenge is to measure the error in the "estimation process". Most signal interpretation algorithms have model parameter estimation as one of their steps. The parameter estimation process typically produces

a residual error which is an indication of how well the data fits the model. Algorithms can use residual errors along with several other internal parameters to produce a confidence measure for the results they produce. A threshold on the confidence can be used to detect errors.

Performance monitoring is achieved through a mechanism similar to component failures. Each signal interpretation module has a attached to it a performance monitoring module (PMM) which reports a normalized performance metric to the ABM. Once the ABM detects a performance problem with a particular component it can take one of take one of the following actions through the PMM.

Adapt Algorithm Parameters: For example, in case the visual tracking module has a low performance due to changing lighting, the background estimation rate can be increased to make the tracking adapt faster to the changing lighting.

Adapt Modality Weights: In the case, that speech recognition performance drops due to background noise, the visual features in the speech recognition can be weighted more than the audio features.

Adapt Signal Acquisition: An alternative way to address the drop in speech recognition performance could be through switching to the closest microphone, or be steering a microphone towards the location of the speaker.

6.4 The user model

Naturally, since the purpose of the autonomic interface is to make the user's life simpler, the interface must have a model of the user in addition to a model of the inanimate objects that make up the space in which the user operates. There are several components to a world model: generic user traits; specific user traits; and current user state. We address each of these in turn.

The generic user traits capture general information that can be used as priors to guide estimation of specific user traits, and which can be assumed in the absence of data. Such a model will comprise such anthropometric data as the size, shape and articulation of people, how fast people move. It may also comprise assumptions of preferences: what tasks are likely to be performed in which locations in the space, how is a user most likely to configure the lighting in the space etc. Further, it can be considered to comprise more specialized data such as the models for speaker-independent speech recognition.

As a system observes a particular user, and tracks that user over a period of time, through *continuity of identity*, the system can build up a considerable knowledge base about the person- the person's appearance, specifically clothing appearance and face appearance for identification, and specific values for the anthropometric data such as height and limb length; the person's voice- the system can adapt individual models to give enhanced speech recognition, and create a speaker identification model. As the user controls the interface- changing the lighting, carrying out specific tasks in specific places etc., the interface can learn this information and passively acquire preference data to be used in making future decisions.

6.4.1 Affective state

One particular aspect of user state that it is desirable for the user model to contain is the user's affective state. Being aware of how the user is receiving a piece of information provided is very valuable. Is the user satisfied, confused, frustrated, or simply sleepy? Being able to know when the user needs more feedback, by not only keeping track of the user's actions, but also by observing cues about the user's emotional experience also has advantages. Based on the user's reaction, the autonomic interface can be adapted by the user or self-adapted to the user's preferences. To achieve this, ideally an integration of multiple cues will be used. For

example, the interface can recognize the user's emotion through body language, voice and facial expression. Also the user can give explicit information through speech.

Although voice can express some human emotional states [35, 28], acoustic information does not provide as much information as vision for emotion analysis. Recent progress in computer vision brings the possibility of using facial and hand information to find if the user is satisfied. Since faces are at the center of human-human communication, conveying a person's internal emotional state, intentions or social communications, it would seem natural and desirable to give faces an important position at the center of human-computer interaction. The face can express emotion sooner than people verbalize or even realize their feelings. Facial expression analysis has been being an active research topic for psychologists since 1872 [11]. Within the past decade, significant efforts have been made in developing methods of automatic facial expression analysis [4, 13, 16, 32, 34, 47, 55]. For automatic facial expression recognition, having located the user's face, individual facial features need to be tracked and their motion measured. Finally the expressions represented must be recognized.

There are two big challenges need to be stressed. One of the important challenges for the autonomic interface is whether the facial expression reveals the true emotions of the user, which may require higher level knowledge. The interpretation of facial expression is dependent upon contextual information such as the user's culture, the setting and accompanying body gesture or speech. Another important challenge is that emotion is often communicated by subtle changes in one or a few discrete facial features, such as a tightening of the lips in anger or obliquely lowering the lip corners in sadness [7]. Change in isolated features, especially in the area of the eyebrows or eyelids, is typical of paralinguistic displays; for instance, raising the brows signals greeting [14]. To capture such subtlety of human emotion and paralinguistic communication, automated recognition of fine-grained changes in facial expression is needed. Several systems were developed to recognize subtle facial expressions [10, 13, 47].

Hand gesture is another important information of the autonomic interface to find if the user is satisfied. To achieve a natural and intelligent interaction, there are three important issues: gesture modelling, gesture tracking, and gesture recognition.

In order to evaluate if the user is satisfied, all or part of the voice, facial, and hand gesture information can be used. For example in S4, when the computer checks the calendar and displays the response on the computer screen, if I frown with displeasure and point to the wall, the computer might then project the response on the wall.

In addition to speech and vision, there is a wide selection of other ways of measuring affective state (including such indicators as heart rate, skin resistance, eye movemements) though in general measuring these requires special sensors attached to the user, making them inappropriate for most user interface situations.

7 Performance evaluation

An interface is designed to achieve effective communication/interaction between the system and its users. The performance evaluation of a generic interface is based on how effectively the user can interact with the system for performing the specified tasks and how efficiently system utilizes its resources [15]. Effectiveness of an interface can quantified using metrics such as appropriate invocation of outputs associated the given inputs, speed of execution, ease of use, ease of user learning, system adaptation to user etc. Efficiency metrics may include cost of the system, percentage utilization of individual components, etc. Relative significance of each these individual performance components to overall performance obviously depends on the specific tasks user interface is performing and on the costs of achieving user acceptable system behavior in each of the individual performance dimensions. For instance, interface in a volatile workspace may

place more importance on ease of learning where as a healthcare situations may emphasize more on system adaptation to users. Derivation of the appropriate quantitative performance models of the interface typically relies on user inputs and the reader referred to literature on user studies [15] related to user interfaces for further reading. The specific issues related to the autonomic interface evaluation arise primarily because of the imperfect behaviors of underlying emerging technologies.

An autonomous interface relies on many pattern recognition technologies mentioned elsewhere in this work. Performance of the individual modules depends upon the inherent discriminative information available in the input signal presented to the module and world/user/signal models of the module. It is well known that error rates of existing practical user identification systems, speech recognition systems, user localization systems, gesture recognition systems are not insignificant and could themselves be a direct cause of the unacceptable overall performance of the system. Similarly, the existing state-of-the-art integration technologies (e.g., audio-visual, speaker-speech) also suffer from imperfect behaviors because of the limited models or limited information in the overall signal presented to the system.

Building the models of the individual components and modelling the system performance in terms of the component models is intractable for even most simple pattern recognition systems and the evaluation studies have predominantly relied on empirical data for estimating the system performance. For any empirical performance evaluation experiment to be able to precisely generalize to the entire population of interest, the test data needs (i) to be *representative* of the population and (ii) contain enough samples from each category of the population. In the context of the autonomous systems both components present a huge problem. Often, the system measurements themselves may in-deterministically change the user behavior and hence capture of realistic inputs may not be trivial. For instance, the Further, due to implicit nature of the inputs, capturing realistic set of samples (e.g., asking user to frown may be significantly different from the realistic user frown) may need thoughtful experimentation. The issues of realistic capture is further exacerbated in case evaluation of the security performance of system under adversarial attacks with unknown threat model. Because of the rich and multidimensional nature of inputs, capture of sufficient number of samples may be very expensive and an intractable problem in itself. The complex interaction between system adaptation to user and user adaptation to the user may imply different system operating points for different users. Issues related to user habituation and system event histories may pose further challenges to sample size estimation.

Emergence of technology depends upon a number of related entities: need for a functionality which provides a value; the application of a technology to obtain the value; and the perceived performance of the technology to provide that value. Bother-less human computer communication provides a genuine value to a user. With increasingly inexpensive computing power and sensors, audio-visual technologies appear appear to be the only candidates for providing a bother-less, convenient, pervasive, and natural interface. Whether this will indeed happen, will depend upon how quickly the overall performance natural interfaces are characterized and compared against other alternatives.

8 Conclusions

In this paper we presented a vision for an autonomic user interface which provides a natural human computer interaction. The interface relies on advanced speech recognition, visual tracking and multi-modal technologies operating in a blackboard like architecture. The combination of speech and visual technologies provides the interface with both human awareness and context awareness both of which are essential for natural human computer interaction. In addition to providing natural human computer interaction, the AUI also uses a combination of the world model and system model to achieve self awareness, and self healing properties.

The vision of an autonomic user interface presented in this paper can be realized today in many limited situations. For example in a well defined space like an automobile, a combination of limited vocabulary speech recognition and visual tracking using stereo could be integrated into the automobile control system to generate fairly advanced autonomic behaviors. However the more general case of instrumenting a larger space like an office environment (where the range of activities is much larger) with a AUI is further off into the future.

Acknowledgements

The authors wish to thank the Adventurous Research program of the Services & Software Department of the IBM Research Division for its support.

References

[1] Azuma and Ronald T. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, 1997.

[2] Brumitt B, Meyers B, Krumm J, Kern A, , and Shafer S. Easyliving: Technologies for intelligent environments. In *Handheld and Ubiquitous Computing*, 2000.

[3] J. Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization. *Journal of the Acoustical Society of America*, 107(1):384–391, October 2000.

[4] M. J. Black and Y. Yacoob. Trcking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proc. Of International conference on Computer Vision*, pages 374–381, 1995.

[5] M. Brandstein, J. Adcock, and H. Silverman. A closed-form method for finding source locations from microphone-array time-delay estimates. In *Proc. Of International conference on Acoustics Speech and Signal Processing*, pages 3019–3022, 1995.

[6] Mark Burnett, Paul Prekop, and Chris Rainsford. Intimate location modeling for context aware computing. In *Proceedings of the Workshop on Location Modeling for Ubiquitous Computing, Atlanta, Georgia*, 2001.

[7] J. M. Carroll and J.A. Russell. Do facial expression signal specific emotions? *Journal of Personality and Social Psychology.*, 70:205–218, 1996.

[8] N. Carver and V. Lesser. The evolution of blackboard control architectures. In *Expert Systems with Applications, Special Issue on The Blackboard Paradigm and Its Applications*, 1994.

[9] Xing Chen, james Davis, and Philips Slusallek. Wide area camera calibration using virtual calibration objects. In *Proceedings of IEEE CVPR 2000*, 2000.

[10] J. F. Cohn, A. J. Zlochower, J. Lien, and T. Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, 36:35–43, 1999.

[11] C. Darwin. *The Expression of Emotions in Man and Animals*. John Murray, reprinted by University of Chicago Press, 1965, 1872.

[12] P. de Cuetos, C. Neti, and A. Senior. Audio-visual intent to speak detection for human-computer interaction. In *International Conf. on Acoustics Speech and Signal Processing*, 2000.

[13] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Classifying facial actions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 21(10):974–989, October 1999.

[14] Eihl-Eihesfeldt. *Human ethology*. NY: Aldine de Gruvter, 1989.

[15] ERGOWEB. Ergonomics standards and guidelines - iso 9241, 2002.

[16] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.

[17] Abowd et al. Living laboratories: The future computing environments group at the georgia institute of technology. In *Proceedings of the 2000 Conference on Human Factors in Computing Systems (CHI 2000)*, The Hague, Netherlands, April 2000.

[18] D. R. Fischell and C. R. Coker. A speech direction finder. In *Proc. Of International conference on Acoustics Speech and Signal Processing*, pages 19.8.1–19.8.4, 1984.

[19] M.J.F. Gales and S.J. Young. HMM recognition in noise using parallel model combination. In *Eurospeech*, pages 837–840, Sept 1993.

[20] Haritaoglu and Flickner. Detection and tracking of shopping groups in stores. In *CVPR*, 2001.

[21] Jeffrey Hightower and Gaetano Borriella. Location systems for ubiquitous computing. *IEEE Computer*, 34(8):57–66, 2001.

[22] Jeffrey Hightower and Gaetano Borriello. A survey and taxonomy of location systems for ubiquitous computing.

[23] Fritz Hohl. Next century challenges: Nexus - an open global infrastructure for spatial-aware applications. In *Proceedings of the Fifth Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'99), Seattle, Washington, USA*, pages 15–20, 1999.

[24] L. Hong and A.K. Jain. Integrating faces and fingerprints for personal identification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(12):1295–1307, 1998.

[25] Harwood Horprasert and Davis. A statistical approach for real-time robust background subtraction and shadow detection. In *ICCV FRAME-RATE Workshop*, 1999.

[26] http://www.research.ibm.com/ecvg/PeopleIntr.html. Peoplevision: Human aware environments.

[27] L. Julia, L. Neumeyer, M. Charafeddine, A. Cheyer, and J. Dowding. HTTP://WWW.SPEECH.SRI.COM/DEMOS/ATIS.HTML. Presented at AAAI '97 Spring Symposium, 1997.

[28] A. KAPPAS, U. Hess, and K.R. Scherer. *Voice and Emotion*, pages 200–238. Cambridge University Press, New York, 1991.

[29] et al Kidd, Cory D. The aware home: A living laboratory for ubiquitous computing research. In *Proceedings of the Second International Workshop on Cooperative Buildings - CoBuild'99*, October 1999.

[30] Rick Kjeldsen, Claudio Pinhanez, Gopal Pingali, Jacob Hartman, Tony Levas, and Mark Podlaseck. Interacting with steerable projected displays. In *Proc. of the 5th International Conference on Automatic Face and Gesture Recognition (FG'02), Washington (DC), May 20-21*, 2002.

[31] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics Speech and Signal Processing*, ASSP-24(4):320–327, October 1976.

[32] Jenn-Jier James Lien, Takeo Kanade, Jeffrey F. Cohn, and C. C. Li. Detection, tracking, and classification of action units in facial expression. *Journal of Robotics and Autonomous System*, 31:131–146, 2000.

[33] B. Maison, C. Neti, and A. Senior. Audio-visual speaker recognition for video broadcast news: Some fusion techniques. In *Multi-media Signal Processing*, 1999.

[34] Kenji Mase. Recognition of facial expression from optical flow. *IEICE Transactions*, E. 74(10):3474–3483, October 1991.

[35] I. Murray and J. Arnott. Toward the simulation of emotion in synthetic speech: A review of the litterature on human vocal emotion. *Journal of the Acoustical Society of America*, 93(2):1097–1108, October 1993.

[36] Alex Pentland. Smart rooms. In *Scientific American*, www.sciam.com/ 0496issue/ 0496pentland.html, April 1996.

[37] E. D. Petajan, B. J. Bischoff, D. A. Bodoff, and N. M. Brooke. An improved automatic lipreading system to enhance speech recognition. In *Proc. of ACM Conference on Human Factors in Computing Systems*, 1988.

[38] G. Pingali. Integrated audio-visual processing for object localization and tracking. In *Proc. of SPIE Conference on Multimedia Computing and Networking*, volume SPIE Vol. 3310, pages 206–213, San Jose, CA, 1998.

[39] G.S. Pingali, G Tunali, and I Carlbom. Audio-visual tracking for natural interactivity. In *Proceedings of ACM Multimedia*, 1999.

[40] F. J. Prokoski. Security and non-diagnostic medical uses for thermal imaging,. In *Proc. American Academy of Thermology, 1997 Annual Meeting, Pittsburgh*, U.S. patent 5,163,094, April 1997.

[41] D. Rabinkin, R. Renomeron, J. French, and J. Flanagan. Estimation of wavefront arrival delay using the cross-power spectrum phase technique. *Journal of Acoustic Society of America*, 4(2):2697, October 1996.

[42] James M. Rehg, Kevin P. Murphy, and Paul W. Fieguth. Vision-based speaker detection using Bayesian networks. In *Conference on Computer Vision and Pattern Recognition*, volume II, pages 110–116. IEEE CS, 1999.

[43] Hanan Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1994.

[44] J. Saunders. Real-time discrimination of broadcast speech/music. In *International Conf. on Acoustics Speech and Signal Processing*, pages 993–996, 1996.

[45] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling. In *Second International workshop on Performance Evaluation of Tracking and Surveillance systems*, 2001. CVPR workshop.

[46] A.Q. Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36:314–331, 1979.

[47] Y.L. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):1–19, February 2001.

[48] Kentaro Toyama. Prolegomena for robust face tracking. *Microsoft Research Technical Report*, MAR-TR-98-65, 1998.

[49] Thomas Funkhouserand Nicolas Tsingos and Jean-Marc Jot. Computational sound for graphics, virtual reality, and interactive systems. In *Course Notes 45, SIGGRAPH 2002, San Antonio, Texas (To appear)*, `http://www.cs.princeton.edu/ funk/course02.pdf`, July, 2002.

[50] Roy Want, Andy Hopper, Veronica Falcao, and Jon Gibbons. The active badge location system. Technical Report 92.1, ORL, 24a Trumpington Street, Cambridge CB2 1QA, 1992.

[51] Mark Weiser. The computer for the 21st century. *Scientific American*, 9 1991.

[52] Martin Westphal. The use of cepstral means in conversational speech recognition. In *Proc. Eurospeech '97*, pages 1143–1146, Rhodes, Greece, 1997.

[53] R. H. Wolfe, P. Hobbs, and S. Pankanti. Footprints: an ir approach to human detection and tracking. In *SPIE Conf. Multispectral Image Processing and Pattern Recognition, Object detection, classification, and tracking technologies, Proc. SPIE, Vol. 4554, Eds. J. Shen, S. Pankanti, and R. Wang, 22-24 Oct 2001, Wuhan China*, page ix, 2001.

[54] www.upnp.org. Understanding universal plug and play.

[55] Y. Yacoob and M. J. Black. Parameterized modeling and recognition of activities. In *Proc. 6th IEEE Int. Conf. on Computer Vision*, pages 120–127, Bombay, India, 1998.

[56] Alexander Zelinsky Yoshio Matsumoto. An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In *Proceedings of IEEE Fourth International Conference on Face and Gesture Recognition (FG'2000)*, March 28-30, 2000.

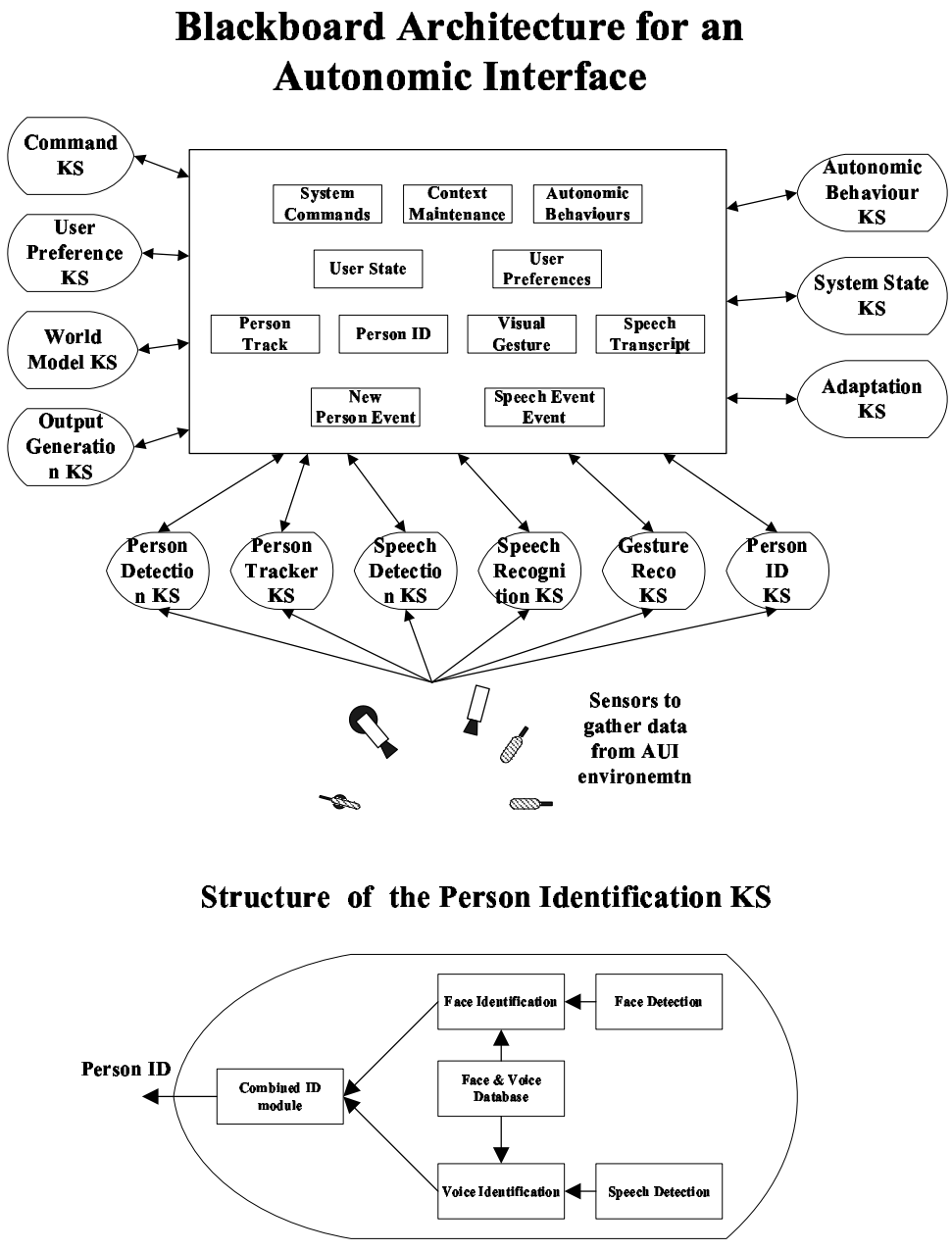

Figure 2: Architecture for an Autonomic Interface

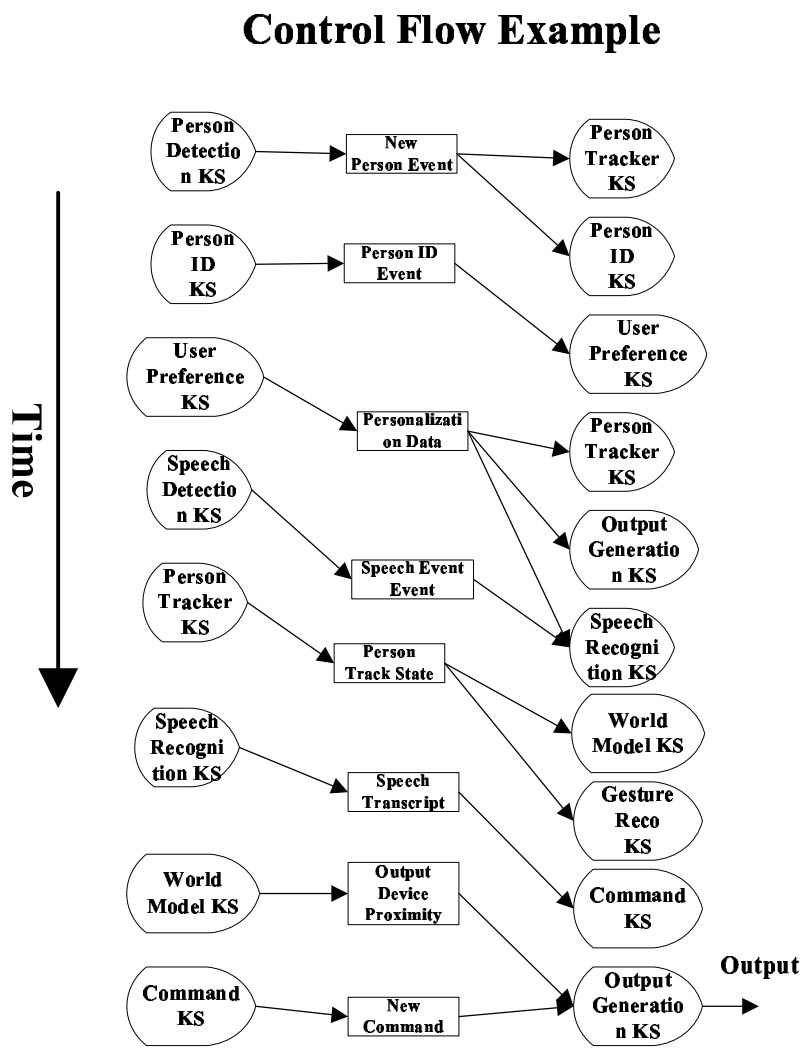

Figure 3: Example Control flow in an Autonomic Interface

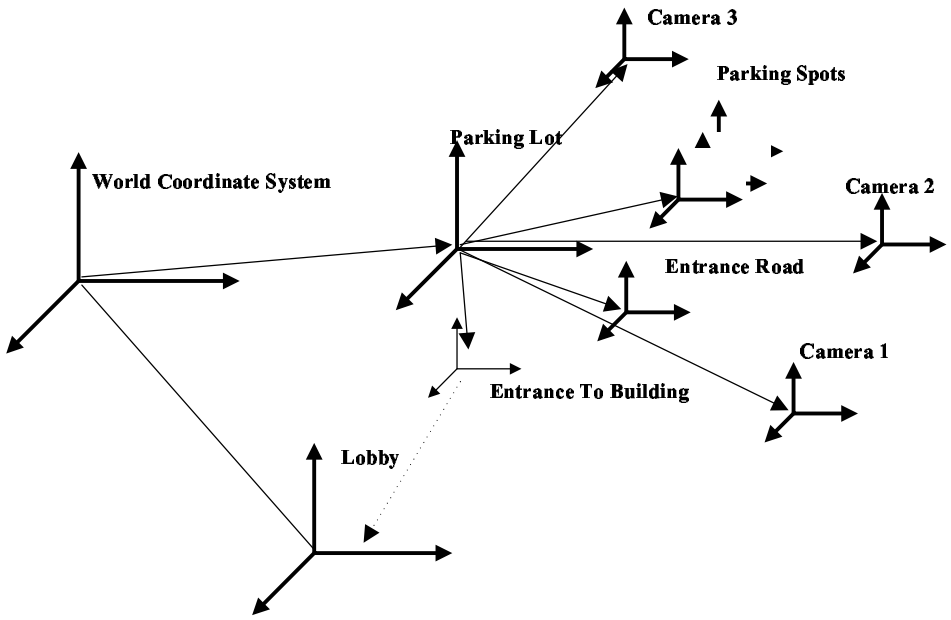

Figure 4: Geometric Representation of the World Model

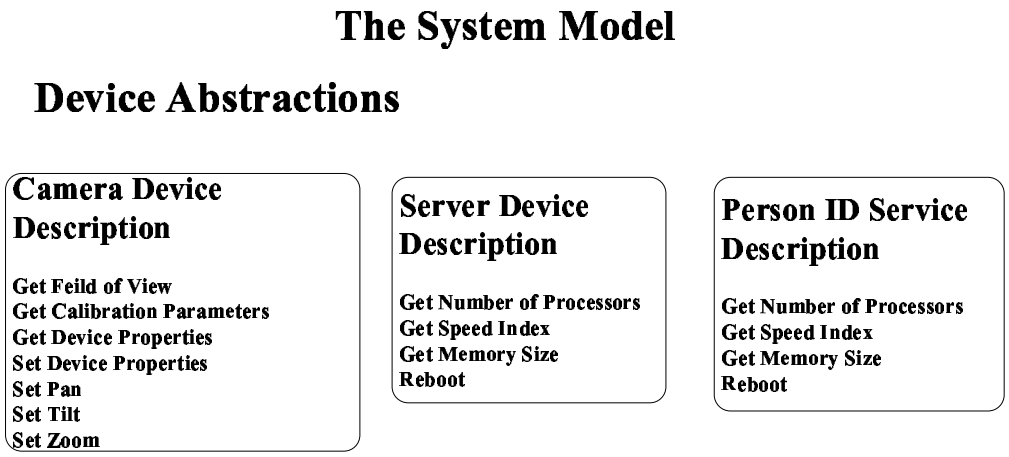

Figure 5: System Model: Device Abstraction

The System Model

Dependency Graph

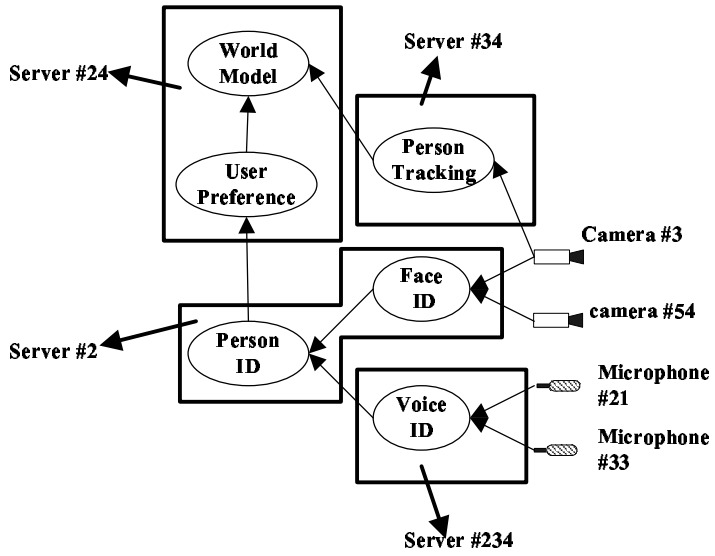

Figure 6: System Model: Component Dependency Graph