Constructing and Evaluating Sensor-Based Statistical Models of Human Interruptibility

Dissertation Proposal

James A. Fogarty
Human Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
jfogarty@cs.cmu.edu

Committee:
Scott E. Hudson (Chair)
Christopher G. Atkeson
Robert E. Kraut
Eric J. Horvitz, Microsoft Research

Abstract

A person seeking a colleague's attention is normally able to quickly assess the colleague's current interruptibility. This ability allows behavior that people consider natural, socially appropriate, or simply polite. In contrast, current computer and communication systems inappropriately interrupt or unduly demand attention because they have no way to determine whether an interruption is appropriate. As a result, people who design or use systems are forced to choose systems that are passive, waiting for the user to initiate action. If reliable models of human interruptibility were available, systems could use this information to negotiate interruptions at appropriate times, thus offering the potential for significant advances in human computer interaction.

In this proposal, I present work to construct and evaluate sensor-based statistical models of the interruptibility of office workers in their normal work environments based on interruptibility observations collected in a minimally intrusive manner. Using an experience sampling technique, my feasibility work shows that analyses of collected data can reliably support the creation of sensor-based statistical models of human interruptibility that are as accurate as or more accurate than estimates provided by human observers. I then discuss `AmIBusy`, a system to create sensor-based statistical models of human interruptibility by combining data collected from many people, thus reducing the disruption experienced by any one person contributing the necessary interruptibility observations. Next are my plans to explore a less intrusive method of collecting interruptibility observations, wherein I will use `AmIBusy` to build a notification application that creates sensor-based statistical models of human interruptibility from the responses of people to the delivery of its notifications. My evaluation will examine how well models based on this feedback can predict the receptiveness of a person to a notification. I will also examine how a model built from this feedback relates to the more generic notion of interruptibility measured by self-reports in my feasibility work. Finally, I present two planned enhancements to my initial implementation of `AmIBusy`: a computational framework for describing interruptions and a distributed method for building sensor-based statistical models of human interruptibility.

Introduction

People have developed a variety of conventions that define behavior that is socially appropriate in different situations [2]. In office working environments, social conventions around interruptions dictate when it is appropriate for one person to interrupt another. These conventions, together with the reaction of the person who has been interrupted, allow an evaluation of whether or not a particular interruption was appropriate. Social conventions around interruptions also allow the development of an *a priori* expectation of whether or not an interruption would be appropriate [13].

In contrast, current computer and communication systems are largely unaware of the social conventions defining appropriate behavior, of the social situations surrounding their usage, and the impact that their actions have on people and social situations. At any given point in time, a person might be notified of the arrival of a new email, receive an instant message from a colleague, be reminded by a handheld computer of an upcoming appointment, receive a phone call on their office or mobile phone, and be involved in a face-to-face interaction with a colleague. Any one of these demands for attention can be addressed relatively easily, but the sum of repeated demands can be disruptive. Without a model of whether it is appropriate to interrupt, it is impossible for current systems to develop informed *a priori* expectations about the impact their interruptions will have on people and their social situations.

People who design or use computer and communication systems can currently adopt one of two strategies for managing the damage caused by inappropriate interruptions. One strategy is to avoid building or using proactive systems, forcing systems to be silent and wait passively until a user initiates interaction. Although this approach is reasonable for many applications in a desktop computing environment, many applications in intelligent spaces and other mobile or ubiquitous computing environments could benefit from a system being able to initiate interactions [15]. A second strategy is to design and use systems that can be temporarily disabled during potentially inappropriate time intervals. However, this approach can be self-defeating. Turning off a mobile phone prevents unimportant interruptions, but it also prevents interruptions that could have conveyed critical information that was needed at that moment. Because current systems generally do not have a mechanism for weighing the importance of information against the appropriateness of an interruption, people are forced into extremes of either allowing all interruptions or forbidding all interruptions. This problem is amplified because people often forget to re-enable systems after the potentially inappropriate time interval has passed [25, 29].

If relatively robust estimators of human interruptibility were available, they might support a variety of significant advances in human computer interaction and computer mediated communication. These advances might allow the users of these systems to be more efficient and more comfortable with the systems. Note that the development of systems capable of modeling human interruptibility does not imply that those systems will deprive people of control. For example, mobile phones could inform a caller that the person being called appears to be busy, allowing the caller to consider the importance of the call in deciding whether to interrupt the apparently busy person or instead leave a message [29]. Email and messaging applications could delay auditory notification of less important messages when the recipient appears to be having a meeting or conversation in

his or her office, but would never prevent delivery of the information. Information displays might choose between several methods of conveying information according to the importance of the information and the current appropriateness of each method of communication. Many specific applications might also be designed for different domains. For example, information about interruptibility might be combined with information on expertise and other relevant factors to automatically route incoming technical support requests to the most appropriate member of a technical support staff.

In this proposal, I present work to construct and evaluate sensor-based statistical models of the interruptibility of office workers in their normal work environments based on interruptibility observations collected in a minimally intrusive manner. I start by reviewing work related to my research on sensor-based statistical models of human interruptibility. I then present a series of studies to examine the feasibility of creating reliable sensor-based statistical models of human interruptibility. Using an experience sampling technique, my feasibility work shows that analyses of collected data can reliably support the creation of sensor-based statistical models of human interruptibility that are as accurate as or more accurate than estimates provided by human observers. I then discuss `AmIBusy`, a system to create sensor-based statistical models of human interruptibility by combining data collected from many people, thus reducing the disruption experienced by any one person contributing the necessary interruptibility observations. Next are my plans to explore a less intrusive method of collecting interruptibility observations, wherein I will use `AmIBusy` to build a notification application that creates sensor-based statistical models of human interruptibility from the responses of people to the delivery of its notifications. My evaluation will examine how well models based on this feedback can predict the receptiveness of a person to a notification. I will also examine how a model built from this feedback relates to the more generic notion of interruptibility measured by self-reports in my feasibility work. I finish my discussion of proposed work by presenting two planned enhancements to my initial implementation: a computational framework for describing interruptions and a distributed method for building sensor-based statistical models of human interruptibility. I then present my contributions, my proposed schedule, and my acknowledgements.

Related Work

Related work can be broadly categorized into five approaches: laboratory studies of interruptions in controlled environments, field studies of how people currently perceive and manage interruptions, models based on activity recognition and an assumed relationship to interruptibility, models built from data sets that have been manually labeled with interruptibility information, and work to use automatically collected data to model presence, the likelihood that a person is reachable via a given communication medium, or some other concept related to interruptibility.

McFarlane tested four known methods for deciding when to interrupt people, measuring the effects of different interruption coordination techniques on performance in a laboratory task [22, 23]. McFarlane's results have implications for structuring appropriate interactions, but no single method emerged as best across all performance measures. Gillie and Broadbent studied the impact of mental arithmetic and free recall interruptions on a task that required participants to remember what parts of the task had been accomplished before the interruption [11]. They found that the complexity of an

interruption and its similarity to the interrupted task affected participant resumption of the task, but neither the length of the interruption nor the opportunity for the participant to mentally rehearse their current state before handling the interruption had an effect. Czerwinski *et al.* studied instant messaging interruptions during participant completion of list searching tasks and tasks in common office productivity software [5, 6, 7]. Importantly, these studies found that even an ignored instant messaging notification was still disruptive to task performance. Adamczyk and Bailey studied interruptions at different points in a document editing task, a video summarization task, and a web search task [1]. They found evidence that interruptions at task breakpoints were less annoying to participants and were described by participants as more respectful of their primary task.

Field studies of how people currently perceive and manage interruptions can provide insight into the problems with current systems and how models of interruptibility might benefit human computer interaction. Hudson *et al.* used an experience sampling technique to explore the perceptions that managers in a research environment had about interruptions [19]. They found a tension between the desire for uninterrupted time to complete tasks and the desire for the helpful information sometimes conveyed by an interruption. In a result similar to that discussed by Perlow [28], Hudson *et al.* found that people consider inappropriate interruptions such a problem that they sometimes isolate themselves by physically moving away from their computer or even away from their office. I point out that this demonstrates the problem discussed in our introduction, people choosing to forbid all interruptions because the systems that they use cannot determine whether a potential interruption is appropriate.

Some systems approach interruptibility by recognizing activities and assuming a relationship between those activities and interruptibility. Oliver *et al.* built SEER, which uses computer activity and analyses of audio and video streams to recognize a set of office activities [27]. The recognized activities are a phone conversation, giving a presentation, a face-to-face conversation, engagement in some other activity, conversation outside the field of view of the camera, and not present. They evaluate the recognition of these activities in a controlled environment, but do not directly measure interruptibility. Kern and Schiele discuss two axes of interruptibility, personal and social, and present a system for using wearable sensors to estimate these axes of interruptibility [21]. Their evaluation examines the system's ability to categorize data that they have labeled according to these axes, but they do not examine whether their results correspond to how people actually feel about their own interruptibility.

Work to directly model interruptibility has been based on collecting data sets that are paired with interruptibility information. My feasibility work, discussed in detail later in this proposal, has used an experience sampling technique to collect interruptibility information [9, 10, 20]. Studying people in their natural working environments, I have recorded information that might be useful as an indicator of a person's interruptibility and collected interruptibility labels once or twice an hour by prompting the participant to report their current interruptibility level on a five-point scale. Horvitz and Apacible haven taken a retrospective labeling approach to this problem, collecting a total of 15 hours of audio and video recordings from the offices of 3 participants who then viewed the recordings and labeled themselves as in a state of low, medium, or high interruptibility [16]. While these approaches allow the direct modeling of interruptibility,

collecting the necessary interruptibility information requires the time and attention of the person being modeled.

Automatically collected data sets have been used to model concepts related to interruptibility, such as presence or the likelihood that a person is reachable via a given communication medium. Begole *et al.* have analyzed activity logs of an awareness client to extract temporal patterns of presence [3, 4]. For example, their system can detect that a person is typically not online for approximately one hour starting at 1:00 PM on Thursdays, which might be due to a recurring meeting. Mynatt and Tullio use electronic calendars and patterns of attendance to infer the likelihood of a person attending a future calendar item [26]. Though the work was based on a manually-maintained diary of what calendar items a person attended, Mynatt and Tullio point out that such information could reasonably be collected by a location tracking system. Horvitz *et al.* have used patterns of presence, together with calendar information, to model when a person is likely to be available on a device and consider forwarding messages to mobile devices [17, 18].

Feasibility Studies

I have conducted a series of feasibility studies to examine the possibility of creating robust sensor-based statistical models of human interruptibility. I first collected audio and video recordings of office workers and used a Wizard of Oz technique to examine the utility of many different simulated sensors, some of which might have been quite difficult to actually build. I then used the collected audio and video recordings to examine the performance of human observers, establishing an important point of comparison for the performance of sensor-based statistical models of human interruptibility. I most recently used real sensors deployed in the offices of a larger and more diverse group of office workers to examine the robustness of the sensors and modeling techniques that I developed based on the results of my initial study.

Wizard of Oz Simulated Sensor Study

Many different sensors seem like they might relate to interruptibility, but the uncertainty surrounding their actual usefulness makes it very likely that implementing them and then evaluating them would result in significant time and resources being spent on sensors that are ill-suited or sub-optimal for predicting human interruptibility. My initial feasibility work instead uses more than 600 hours of audio and video recordings collected from the normal working environments of four office workers. While collecting these recordings, I prompted subjects at random time intervals approximately once every thirty minutes to provide self-reports of their interruptibility on a five-point scale.

I then used a Wizard of Oz technique to simulate the output of many potential sensors. Paid employees viewed the collected recordings and used the custom software shown in Figure 1 to simulate the output of such potential sensors as the presence of a guest, whether the subject was sitting or standing, whether the subject was speaking, interaction with a desk or file cabinet, the handling of papers, and many other sensors that seem like they might relate to interruptibility and could probably be implemented. Note that I included potential sensors that might be very difficult to actually implement, because knowing that they would be useful could motivate their construction.

Using the simulated sensor output, I built statistical models of the collected interruptibility self-reports. The models constructed from the simulated sensors

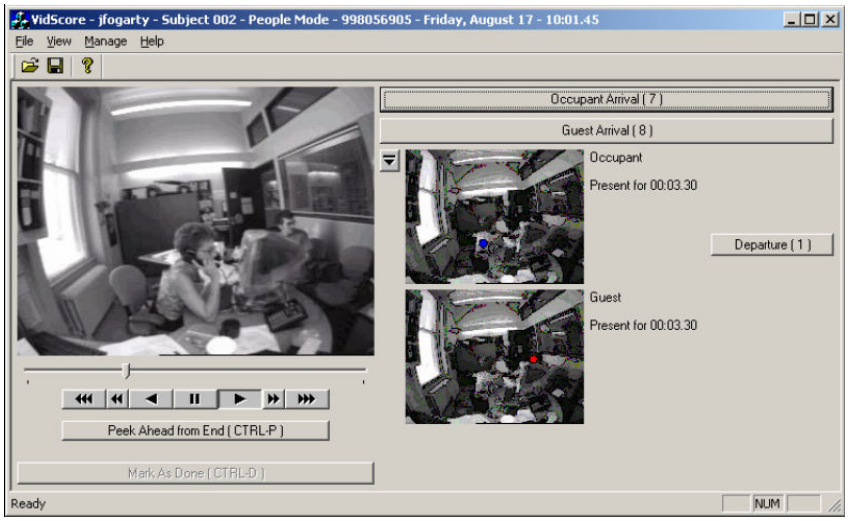

Figure 1. Custom Interface Used to Simulate Sensors.

distinguished situations reported as "highly non-interruptible" from other situations with an accuracy of 82.4%, significantly better than a chance accuracy of 68% that could be obtained by always estimating that a situation was not "highly non-interruptible" ($\chi^2(1, 1344) = 37.56$, $p < .001$). It is worth noting that current systems generally perform at this base level of accuracy, because they have no model of human interruptibility. I created these models using established machine learning techniques, including decision trees and naïve Bayes models. I also examined the use of support vector machines and AdaBoost with decision stumps, finding their performance comparable on this problem.

Beyond their performance, the models I created in this study had several good properties. First, they required relatively few sensors. For example, the most predictive individual sensor, detecting whether anybody in the room is talking at the time of a potential interruption, by itself has an accuracy of 75.9%. Second, the sensors used to create the models are very practical. I constructed a model with an accuracy of 79.2% from sensors indicating whether anybody is talking in the office, whether the phone is off its hook, whether a person is generating mouse or keyboard activity, and the time of day. These properties of the models indicate that it is not only possible to create sensor-based statistical models of human interruptibility, but that they can also be created from inexpensive and non-intrusive sensors.

Human Observer Performance Study

While my initial Wizard of Oz simulated sensor study showed that simulated sensors could support the construction of statistical models of human interruptibility that perform significantly better than chance, it was not clear what level of performance should be considered good. People use a complex system of subtle cues to negotiate entry into conversation [12], and it seems that models of human interruptibility should be part of a negotiation, rather than treated as if they provide absolute guidance. In order to determine what level of accuracy might be appropriate for a system making an initial judgment of interruptibility prior to negotiating an interruption, I studied the accuracy of human observers in estimating a person's interruptibility.

Using the audio and video recordings collected as a part of my Wizard of Oz study, I showed forty subjects short segments of audio and video recordings from the normal working environments of four office workers. After viewing each segment of the recordings, the human observers estimated the interruptibility of the person in the recordings. I compared these estimates with the response given by the person in the recordings when that person was prompted to describe their interruptibility. The human observers were able to distinguish situations reported as "highly non-interruptible" from other situations with an accuracy of 76.9%.

The 82.4% accuracy of the statistical models of human interruptibility created from my simulated sensor data is significantly better than the 76.9% accuracy of human observers established in this study ($\chi^2(1, 3072) = 9.51$, $p < .01$). This finding indicates that sensor-based statistical models of human interruptibility can perform well enough to inform negotiated approaches to interruptions.

Sensor and Model Robustness Study

My first two studies showed the possibility of creating sensor-based statistical models of human interruptibility that performed as well as or better than human observers, but left open several questions. The four people studied all had similar job responsibilities, and it was unclear if the approach would work for more diverse groups of office workers. There was also a question of whether real sensors could be implemented well enough to support models of human interruptibility that would perform as well as models based on the simulated sensors.

In order to examine the robustness of the sensing and modeling techniques explored in my Wizard of Oz study, I deployed real sensors into the normal working environments of ten office workers with relatively diverse job responsibilities. I logged the level of audio activity, as heard by a microphone placed on a shelf in the office and as heard by the built-in microphone found on many laptop computers. I also logged the amount of mouse and keyboard activity, what applications were in use, whether the phone was off its hook, whether the door was open, closed, or cracked, and the activation of two motion sensors in the office. While collecting this sensor data, I prompted the subjects for self-reports of their interruptibility at random intervals approximately once per hour.

A sensor-based statistical model of the interruptibility of all ten subjects, ignoring the differences in their job responsibilities and working environments, had an accuracy of 79.5%, better but not significantly different than my previously established measure of human performance ($\Delta z = 1.34$, $p \approx .18$). More accurate models resulted when I examined subsets of the data from subjects with similar job responsibilities and working environments. A model of two first-line manager subjects had an accuracy of 87.7% ($\Delta z = 1.34$, $p < .001$), a model of five researcher subjects had an accuracy of 81.1% ($\Delta z = 0.89$, $p \approx .37$), and a model of three interns working in offices shared with another intern had an accuracy of 80.1% ($\Delta z = 0.17$, $p \approx .86$). These results indicate that statistical models of human interruptibility created from real sensors can perform as well as or better than human observers for a variety of office workers.

Interestingly, I found that sensor-based statistical models of human interruptibility can be created from a variety of different sensors, as long as those sensors can detect indications of social engagement and task engagement. For example, using only the sensing mechanisms available in a normal laptop computer, I built a model of the two

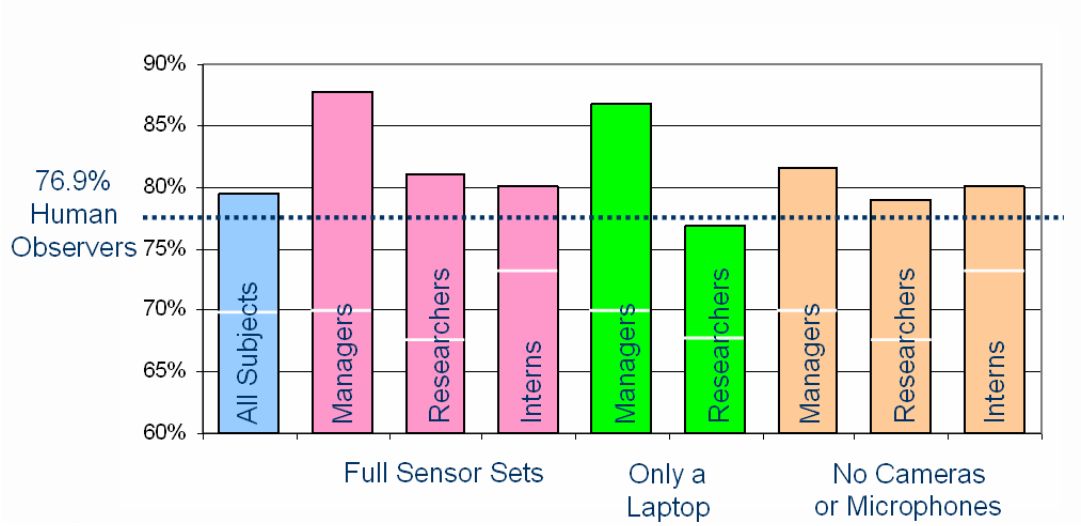

Figure 2. Summary of Robustness Study.

first-line manager subjects with an accuracy of 86.8% ($\Delta z = 3.55$, $p < .001$) and a model of four of the researcher subjects with an accuracy of 76.8% ($\Delta z = 0.91$, $p \approx .36$). I also built models that used neither cameras nor microphones, thus avoiding some of the associated privacy concerns. Those models performed at 81.6% for the first-line manager subjects ($\Delta z = 1.48$, $p \approx .14$), 78.9% for the researcher subjects ($\Delta z = 0.20$, $p \approx .83$), and 80.1% for the intern subjects ($\Delta z = 0.17$, $p \approx .86$).

The accuracies of different sensor combinations for subsets of the collected data are summarized in Figure 2, with the white lines across each bar representing the chance performance for that problem. While some models performed better than others, they all performed at least as well as the human observers in my previous study. These findings indicate that real sensors can support reliable statistical models of human interruptibility for office workers with diverse responsibilities and working environments.

System Support for Combining Data from Many People

While my feasibility studies have demonstrated the potential for creating sensor-based statistical models of human interruptibility, it remains difficult to deploy applications based on this approach. A significant number of interruptibility observations are required to learn the statistical models, and the prompts requesting self-reports can be disruptive. In order to reduce the disruption caused by this data collection, I will build a system that creates models of human interruptibility by combining data collected from many people. This will reduce the number of observations that need to be collected from any one person, thus reducing the disruption experienced by any one person. It will also allow applications to use data collected from other people to provide estimates of a person's interruptibility even before that person has provided any interruptibility observations.

`AmIBusy` (*Automatically Modeling Interruptibility by Unobtrusively Sensing You*) will use a client-server approach to collecting data from many people to create models of human interruptibility. `AmIBusy` clients will upload their collected data to a server, which will use the data from clients to create a sensor-based statistical model of human interruptibility, which the clients will then download and use. As a system, `AmIBusy` will also provide a research contribution by making it easier to build applications that use

interruptibility estimates. Providing a set of mechanisms for logging sensor data, collecting observations of a person's interruptibility, and using the collected logs and observations to create sensor-based statistical models of human interruptibility, AmIBusy will allow application developers to use models of human interruptibility without requiring that they acquire specialized knowledge of machine learning. The defining characteristics of AmIBusy are:

> AmIBusy will make it easier for application programmers to create and use sensor-based statistical models of human interruptibility.
>
> AmIBusy will be extensible to support a wide variety of sensors.
>
> AmIBusy will analyze data from many people by extracting situations and buildings models of the interruptibility of people in those situations.

This subsection discusses each of these characteristics of AmIBusy, presenting my rationale and my plans for implementation.

Making it Easier for Applications to Model Human Interruptibility

By encapsulating mechanisms for logging sensor data, collecting interruptibility observations, and using the collected logs and observations to create sensor-based statistical models of human interruptibility, AmIBusy will make it easier for applications to use models of human interruptibility. Application programmers will be able to focus on their application and how it uses a model of human interruptibility, rather than on implementing logging and aggregation mechanisms or understanding the machine learning techniques used by AmIBusy. There are two primary types of applications that AmIBusy will support.

The first type of application is interested in only a generic notion of interruptibility. AmIBusy will allow such applications to obtain estimates of a person's interruptibility, without requiring that the applications provide interruptibility observations. For this type of application, the necessary interruptibility observations will instead be collected by a standard mechanism, such as once-per-day random prompts for a self-report. Because models based on these observations are generic, many different applications can be based on the same model. For example, researchers familiar with my work have proposed that application programmers might use database queries for a person's interruptibility to inform the delivery of messages [14].

The second type of application is interested in a specialized notion of interruptibility. For example, one application might be interested in whether it is currently appropriate to present a dialog requesting a person provide information needed to continue an automated process. Another application might be interested in whether a person is likely to be available to read an instant message. AmIBusy will support these applications by allowing them to provide observations of the specialized notion of interruptibility they are interested in, such as whether a persons selects a "Now is a Bad Time" option in response to a dialog or whether a person ignores a notification for a new instant message. AmIBusy will then use these observations to create a sensor-based statistical model of the desired notion of interruptibility.

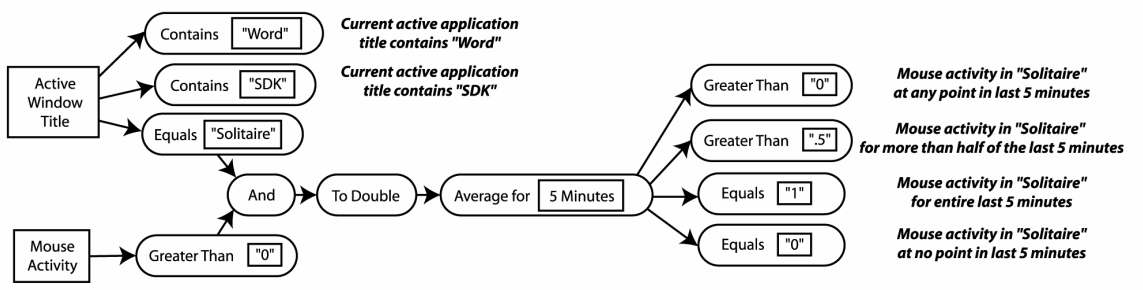

Figure 3. Applying Operators to Create Features from Raw Sensors.

Supporting a Wide Variety of Sensors

In order to support the consideration of whatever sensors are most relevant to a model of human interruptibility, `AmIBusy` will be extensible to support a wide variety of sensors. `AmIBusy` will support extensions in the form of sensors that run asynchronously or in a child thread of `AmIBusy`, but it will also support the addition of complex data sources like those that might be provided by the Context Toolkit [8]. Supporting the integration of such complex data sources will allow `AmIBusy` to take advantage of existing sensing mechanisms, instead of requiring that they be separately implemented for `AmIBusy`.

In my feasibility studies, I found that raw sensor data is typically insufficient for creating statistical models of interruptibility, and that it is instead necessary to extract features from the raw sensor data. For example, one of my studies found that the activation of a talking sensor in the moment before a potential interruption was less predictive than whether the sensor had been activated for 60% of the last five minutes. The same study found that a high level of mouse activity was not by itself predictive, but that it was predictive when it occurred in a particular application. Rather than requiring that programmers explicitly implement each such potential feature, `AmIBusy` will allow programmers to focus on implementing the core functionality of sensors. `AmIBusy` will then automatically explore features based on combinations of sensors and temporal sequences of sensor readings. Figure 3 presents an illustration of how these features will be constructed. Starting with raw sensors on the left, operators are applied to create features of increasing complexity. An exploration of potential features will occur during the offline model creation process, and only the features that prove useful will be computed when the model is in use.

Modeling Interruptibility in Different Situations

My feasibility studies found that different sensors were most predictive for people with different job responsibilities. However, I also found that data from one type of office worker was still useful to models being created for a different type of office worker. I believe this is in part because some basic social behaviors and cues are valid across many types of people. I also believe that the studied office workers share some responsibilities, such as participating in small meetings, and that these common responsibilities allow useful statistical relationships to be extracted despite individual differences.

My approach to aggregating data from many people will be based on the idea that different sensors are predictive for different people in part because they spend their time

doing different tasks. `AmIBusy` will cluster interruptibility observations according to the similarity of their corresponding sensor readings, forming unlabeled groups that I will refer to as *situations*. I might, for example, expect situations to emerge that represent "alone in a quite office," "in office with another person," or "working in programming environment." Note that I can use the interruptibility observations to supervise the clustering process and determine how many clusters produce the best model for the collected data. When using a model constructed in this manner, current sensor data provides a likelihood that a person is in each situation contained in the model. A model of each situation uses the sensor data to estimate the person's current state of interruptibility, and the estimates are combined using the likelihood that the person is in each situation.

I have chosen this approach because I believe it will allow models to very quickly begin to provide reliable estimates of a person's interruptibility, without requiring that a person spend attention choosing a model type or specifying model parameters. For example, a model would likely include a situation representing "working in programming environment," and the model would already have a notion of what sensors are usually predictive in that situation. Such a model can be effective immediately, as opposed to a model that adapted from a null initial model or required some amount of time to decide that a person was a "programmer" and then began to use an appropriate model.

Collecting Less Intrusive Observations of Interruptibility

As another approach to minimizing the disruption caused by collection of interruptibility observations needed to build sensor-based statistical models of human interruptibility, I will examine the use of observations that can be collected less intrusively. I will create an application that allows people to subscribe to notification streams, such as the community information presented by "What's Happening?" [30] or updates generated by news websites. This application is interesting in the context of interruptibility models because the individual notifications provided by such streams are typically of low value, even if the overall information content of the stream is of high value. Presenting low-value notifications at inappropriate times can be perceived as disruptive, causing people to abandon or disable an application. This has generated a significant amount of interest in using peripheral displays to present this type of information, but I am also interested in directly addressing the problem of delivering the notifications at appropriate times.

The notification application I create will provide a mechanism for a person to indicate that a notification was delivered at an inappropriate time and will note when notifications are clicked in order to view a website or otherwise display more information. It will use these observations of the receptiveness of a person to a notification to build a sensor-based statistical model of human interruptibility, which it will use to inform the presentation of its notifications. My evaluation, discussed in the next section, will examine the performance of models built from these less intrusive indications of interruptibility versus models built from self-reports. Although `AmIBusy` will support a wide variety of sensors, this application will focus on computer activity and other sensors that can be deployed in software, such as analyses of the audio from the built-in microphone of laptop computers.

Evaluation of Model Construction

My evaluation of the models constructed by AmIBusy in the context of my notification application will be two-fold. I will examine the ability of the feedback collected by the notification application to predict the receptiveness of a person to the notifications delivered by the application. I will also examine the relationship of models built from this feedback with the interruptibility self-reports collected in my feasibility work.

In order to evaluate models created from user feedback to my notification application, I will deliver some notifications at random times. I expect that these randomly-timed notifications will be better received when the model indicates that a person is interruptible, as measured by a lower likelihood of the person indicating that the notification was delivered at an inappropriate time and a higher likelihood of the person clicking to obtain additional information.

I will also examine the relationship between models built from user feedback to my notification application and the more generic notion of interruptibility measured by self-reports in my feasibility work. To this end, I plan to ask some people who are using my notification application to provide occasional self-reports of interruptibility, using a technique like that used in my feasibility studies. Given the combination of interruptibility self-reports and feedback to the notification application, there are several interesting questions to consider. It would be very interesting to find that models built from only the feedback to the notification application could perform well at estimating the interruptibility self-reports. This finding would indicate that more subtle forms of model feedback can succeed, possibly removing the need for explicit prompts for self-reports. A more likely result is that the feedback given to the notification application is not by itself sufficient for building generic models of interruptibility, but may provide enough additional information to allow equally reliable models to be created from fewer self-reports. Showing the ability of models based on self-reports to predict the receptiveness of people to notifications would demonstrate some external validity the self-report measures of human interruptibility.

Further System Development

While I will initially focus on developing AmIBusy to the point that I can deploy my notification application and begin my evaluation, I will then further develop AmIBusy to better support additional applications. Two major parts of this further development will be a computational framework for describing interruptions and a distributed method for building sensor-based statistical models of human interruptibility.

A Computational Framework for Describing Interruptions

My initial development of AmIBusy will focus on supporting one type of interruptibility observation at a time. If two or more applications are collecting different types of interruptibility observations, they will need to use separate instances of AmIBusy. But different types of interruptibility observations, such as the feedback to my notification application and the self-reports I will collect in my evaluation, have different properties and it seems important to support these differences.

I will design and implement a computational framework for describing interruptions. When an application reports an interruptibility observation, it will use this framework to describe the nature of the observation. When requesting an estimate of a person's

interruptibility, an application will have the option of using the framework to describe a potential interruption. If a description of a potential interruption is provided, AmIBusy will use the description to provide a more specific estimate of a person's interruptibility.

My current plans are to base this framework on the taxonomy of human interruption presented by McFarlane [24]. This taxonomy presents several dimensions along which interruptions can be classified, but is not currently appropriate for use as a computational framework. I envision describing interruptions according to a set of properties that correspond to the dimensions of McFarlane's taxonomy. For each property, I will create a hierarchy of possible values. Application programmers can then assign one or more values from this hierarchy to the corresponding property. They can also insert new values into the hierarchy and assign them to the property. When asked to reason about a person's interruptibility, AmIBusy will use the properties of a potential interruption to determine what prior interruptibility observations are related to the query. If the query is for a node in the hierarchy about which AmIBusy has very little knowledge, it will instead use its knowledge of parent nodes to provide a best estimate.

Including a framework for describing interruptions in AmIBusy will allow the exploration of several possibilities. This framework will allow different types of interruptibility observations, such as the feedback provided to my notification application and the more intrusive prompts for self-reports, to be analyzed together, potentially allowing a variety of less intrusive interruptibility observations to remove or significantly reduce the need for more intrusive observations. There is also a possibility that properties describing who is initiating an interruption might allow AmIBusy to discover that a person is typically more receptive of interruptions initiated by "Person / Colleague / James" than of interruptions initiated by "Computer / Microsoft Office Assistant." AmIBusy might also discover that a person prefers that interruptions of content type "Notification / Calendar / Meeting Reminder" be delivered to the channel "Mobile Phone Notification" when away from their office, but to the channel "Wall Projector Notification" when in their office.

A Distributed Method for Building Models of Human Interruptibility

In my initial implementation of AmIBusy, clients will upload their collected sensors logs and interruptibility observations to a server, which will then use the aggregated logs to create a model of interruptibility, which the clients will then download from the server. While there are concerns related to the computational expense of analyzing large amount of data on a server, there are also concerns about what types of sensor logs are appropriate to collect on a central server. For example, it seems interesting to pursue analyses of whether social chit-chat in instant messaging clients relates to interruptibility differently than work-related conversation, but it also seems potentially inappropriate for a widely deployed system to collect logs of instant messaging chats on a central server.

I will design and implement a distributed method for building sensor-based statistical models of human interruptibility. AmIBusy clients will be responsible for analyzing their own data, rather than uploading the raw sensor logs and interruptibility observations to a central server. Based on their analyses, clients will upload an intermediate representation that contains much less potentially sensitive information than is contained in the raw sensor logs. The server will be responsible for combining these intermediate representations to produce a model of interruptibility. This approach will address both

the computational concerns related to conducting large analyses on the server and help to protect the potentially sensitive information contained in raw sensor logs.

It is not yet clear what intermediate representation will be appropriate for this problem, and it may be that several rounds of data exchange between clients and the server are necessary to distribute the model building process. The Bayesian models I have used in my feasibility work are created from counts of how often a sensor was active when a person was interruptible versus how often the sensor was active when the person was not interruptible. This seems to be a good starting point for the representation, as clients that upload only this data are releasing relatively little information.

Contributions

Current systems often interrupt at inappropriate times or unduly demand attention because they have no way to model interruptibility and cannot make an informed *a priori* decision about the appropriateness of an interruption. This proposal presents several contributions towards addressing this problem.

I contribute a series of studies to examine the feasibility of creating sensor-based statistical models of human interruptibility. My studies present a novel use of a Wizard of Oz technique to simulate a wide variety of potentially useful sensors. My studies also demonstrate that sensor-based statistical models can distinguish situations reported as "highly non-interruptible" from other situations with accuracies as good as or better than human observers.

I contribute a system that reduces the intrusiveness of collecting the observations necessary for sensor-based statistical models of human interruptibility by combining data collected from many people. Because data is combined from many people, fewer observations need to be collected from any one person. This system will also make it easier for application developers to use models of human interruptibility, will be extensible to support a wide variety of sensors, and will model the interruptibility of people in different situations.

I contribute an example of using a less intrusive method to collect interruptibility observations, using feedback provided to a notification application to build sensor-based statistical models of human interruptibility. This type of interruptibility observation can be collected without the disruption associated with a prompt for a self-report. Notification applications are also interesting because individually low-value notifications delivered at inappropriate times can be considered disruptive, leading people to abandon a notification application. By using a model of interruptibility to inform the delivery of notifications, this application will minimize the disruption caused by its notifications.

I contribute evaluations of sensor-based statistical models of human interruptibility built by combining data from many people and by collecting less intrusive interruptibility observations. While my feasibility studies have shown that reliable sensor-based statistical models can be created from interruptibility self-reports, these evaluations will examine models created using observations collected from many people and observations based on feedback to my notification application. Both of these approaches offer to reduce the disruption caused by collecting the interruptibility observations necessary to build sensor-based statistical models of human interruptibility, and so they offer significant additional potential.

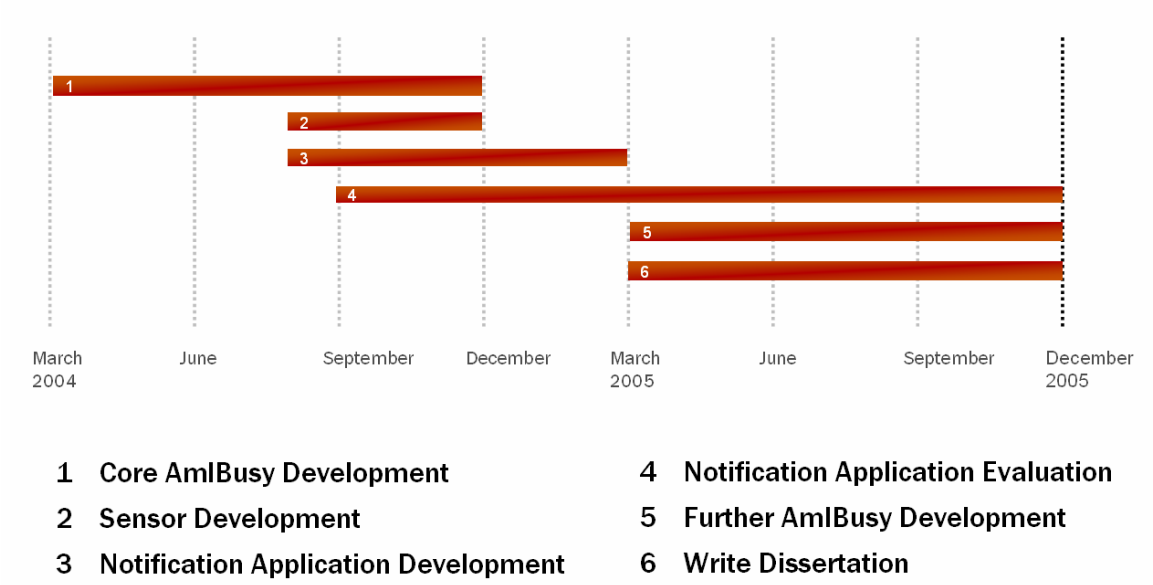

Figure 4. Proposed Schedule.

Schedule

My proposed schedule is presented in Figure 4. I intend to spend the remainder of the summer implementing and refining core components of the `AmIBusy` architecture. I will then begin parallel development of my notification application and the sensors necessary for the notification application. This development will probably lead me to further refine some aspects of the core `AmIBusy` implementation. As soon as feasible, I will begin to collect training data by deploying the notification application with a small group of early adopters. I will conduct a broader evaluation of the notification application once I feel that I have collected enough training data from the early adopters to provide a reasonable initial experience to a larger user group. Once the notification application is broadly deployed, I will focus on further development of `AmIBusy`.

Acknowledgements

I would like to thank Scott Hudson for advising me in this work. This work has been conducted as a part of research projects on Situationally Appropriate Interaction and Managing Human Attention. It has benefited greatly from the contributions of the other members of these research projects, including Christopher Atkeson, Daniel Avrahami, Laura Dabbish, Jodi Forlizzi, Sara Kiesler, Robert Kraut, Johnny Lee, and Jie Yang. My work has also benefited from discussions with and knowledge shared by many other people, including Sonya Allin, Bo Begole, Ryan Baker, Jim Christensen, Al Corbett, Darren Gergle, Eric Horvitz, Wendy Kellogg, Jennifer Lai, Joonhwan Lee, Jeffrey Nichols, Fleming Seay, Desney Tan, and David Tolliver. I have undoubtedly forgotten to include other people who have provided important contributions. The data for my robustness study was collected during an internship at IBM Research. This work was funded in part by DARPA and by the National Science Foundation through my Graduate Research Fellowship and through grants IIS-01215603, IIS-0205219, IIS-9980013, and IIS-0325351.

References

1. Adamczyk, P.D. and Bailey, B.P. (2004) If Not Now, When?: The Effects of Interruption at Different Moments Within Task Execution. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2004)*, 271-278.
2. Barker, R.G. (1968) *Ecological Psychology*. Stanford University Press.
3. Begole, J.B., Tang, J.C. and Hill, R. (2003) Rhythm Modeling, Visualizations, and Applications. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2003)*, 11-20.
4. Begole, J.B., Tang, J.C., Smith, R.B. and Yankelovich, N. (2002) Work Rhythms: Analyzing Visualizations of Awareness Histories of Distributed Groups. *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2002)*, 334-343.
5. Cutrell, E., Czerwinski, M. and Horvitz, E. (2001) Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance. *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT 2001)*, 263-269.
6. Czerwinski, M., Cutrell, E. and Horvitz, E. (2000) Instant Messaging and Interruptions: Influence of Task Type on Performance. *Proceedings of the Australian Conference on Computer-Human Interaction (OZCHI 2000)*, 356-361.
7. Czerwinski, M., Cutrell, E. and Horvitz, E. (2000) Instant Messaging: Effects of Relevance and Time. *Proceedings of the British HCI Group Annual Conference (HCI 2000)*, 71-76.
8. Dey, A.K., Salber, D. and Abowd, G.D. (2001) A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. *Human-Computer Interaction (HCI) Journal*, *16* (2-4). 97-166.
9. Fogarty, J., Hudson, S., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. and Yang, J. (2004) Predicting Human Interruptibility with Sensors. *To Appear, ACM Transactions on Computer-Human Interaction (TOCHI)*.
10. Fogarty, J., Hudson, S. and Lai, J. (2004) Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2004)*, 207-214.
11. Gillie, T. and Broadbent, D. (1989) What Makes Interruptions Disruptive? A Study of Length, Similarity, and Complexity. *Psychological Research*, *50*. 243-250.
12. Goffmann, E. On Facework. In Goffmann, E. ed. *Interaction Ritual*, Random House, New York, 1982, 5-45.
13. Hatch, M.J. (1987) Physical Barriers, Task Characteristics, and Interaction Activity in Research and Development Firms. *Administrative Science Quarterly*, *32*. 387-399.
14. Heer, J., Newberger, A., Beckmann, C. and Hong, J.I. liquid: Context-Aware Distributed Queries. In Dey, A.K., Schmidt, A. and McCarthy, J.F. eds. *"UbiComp 2003: Ubiquitous Computing" ACM UbiComp Proceedings*, Springer-Verlag, 2003, 140-148.
15. Horvitz, E. (1999) Principles of Mixed-Initiative User Interfaces. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 1999)*.
16. Horvitz, E. and Apacible, J. (2003) Learning and Reasoning about Interruption. *Proceedings of the International Conference on Multimodal Interfaces (ICMI 2003)*, 20-27.
17. Horvitz, E., Jacobs, A. and Hovel, D. (1999) Attention-Sensitive Alerting. *Proceeding of the Conference on Uncertainty and Artificial Intelligence (UAI 1999)*, 305-313.
18. Horvitz, E., Koch, P., Kadie, C.M. and Jacobs, A. (2002) Coordinate: Probabilistic Forecasting of Presence and Availability. *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 2002)*, 224-233.
19. Hudson, J.M., Christensen, J., Kellogg, W.A. and Erickson, T. (2002) "I'd be overwhelmed, but it's just one more thing to do": Availability and Interruption in

Research Management. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2002)*, 97-104.

20. Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J. and Yang, J. (2003) Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2003)*, 257-264.
21. Kern, N. and Schiele, B. (2003) Context-Aware Notification for Wearable Computing. *Proceedings of the IEEE International Symposium on Wearable Computing (ISWC 2003)*.
22. McFarlane, D.C. (2002) Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction. *Human-Computer Interaction*, *17* (1). 63-139.
23. McFarlane, D.C. (1999) Coordinating the Interruption of People in Human-Computer Interaction. *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT 1999)*.
24. McFarlane, D.C. and Latorella, K.A. (2002) The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-Computer Interaction*, *17* (1). 1-61.
25. Milewski, A.E. and Smith, T.M. (2000) Providing Presence Cues to Telephone Users. *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2000)*, 89-96.
26. Mynatt, E. and Tullio, J. (2001) Inferring Calendar Event Attendance. *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2001)*, 121-128.
27. Oliver, N., Horvitz, E. and Garg, A. (2002) Layered Representations for Recognizing Office Activity. *Proceedings of the International Conference on Multimodal Interaction (ICMI 2002)*, 3-8.
28. Perlow, L.A. (1999) The Time Famine: Toward a Sociology of Work Time. *Administrative Science Quarterly*, *44* (1). 57-81.
29. Schmidt, A., Takaluoma, A. and Mäntyjärvi, J. (2000) Context-Aware Telephony Over WAP. *Personal and Ubiquitous Computing*, *4* (4). 225-229.
30. Zhao, Q.A. and Stasko, J.T. (2002) What's Happening? Promoting Community Awareness Through Opportunistic, Peripheral Interfaces. *Proceedings of the Conference on Advanced Visual Interfaces (AVI 2002)*, 69-74.