# Constructing and Evaluating Sensor-Based Statistical Models of Human Interruptibility

**James Fogarty**
Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jfogarty@cs.cmu.edu

**ABSTRACT**
A person seeking a colleague's attention is normally able to quickly assess the colleague's interruptibility. In contrast, current computer and communication systems interrupt at inappropriate times or unduly demand attention because they have no way to consider human interruptibility. If reliable models of human interruptibility were available, they might support a variety of advances in human computer interaction.

In this summary of my research, I first present a series of studies that we have conducted to examine the feasibility of creating sensor-based statistical models of human interruptibility. I then present my plans to develop a system to support applications based on such models. Finally, I present my plans to use this system to examine two approaches to reducing the disruption associated with collecting the observations of human interruptibility needed to build statistical models: combining data collected from many people and collecting less intrusive types of interruptibility observations.

**Categories and Subject Descriptors:**
H5.2. Information interfaces and presentation: User Interfaces; H1.2. Models and Principles: User/Machine Systems.

**Author Keywords**
Interruptibility, context-aware computing, sensor-based interfaces, situationally appropriate interaction, managing human attention, machine learning.

**INTRODUCTION**
Modern office workers increasingly find computing and communication systems to be at the core of their everyday work experience. 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 or simultaneous demands can be disruptive. Current systems interrupt inappropriately or unduly demand attention at least in part because they have no way of determining when it is appropriate to interrupt. A colleague preparing to call a person typically has no way to know that the person is in the middle of a face-to-face meeting, and an email client about to announce the arrival of a new message cannot determine whether an obvious or subtle notification is currently more appropriate. If reliable models of human interruptibility were available, they might support a variety of advances in human computer interaction.

This paper summarizes my research on constructing and evaluating sensor-based statistical models of human interruptibility. I first contribute a series of studies to examine the feasibility of creating sensor-based statistical models of human interruptibility. I then contribute a system to support the use of sensor-based statistical models of human interruptibility in a variety of applications. Finally, I contribute the development and evaluation of two approaches to reducing the disruption associated with collecting the observations of human interruptibility that are needed to build statistical modes: combing data collected from many people and collecting less intrusive types of interruptibility observations.

**FEASIBILITY STUDIES**
Many different sensors seem like they might relate to interruptibility, but the uncertainty surrounding their actual usefulness make it very likely that implementing and then evaluating them would result in significant time and resources being spent on sensors that are ill-suited or sub-optimal to predict interruptibility. Our initial feasibility work instead collected 600 hours of audio and video recordings from the normal environments of four office workers and used a Wizard of Oz technique to simulate the presence of a variety of potentially useful sensors [2, 5]. While these recordings were being collected, the office workers were prompted to provide self-reports of their interruptibility at random but controlled time intervals averaging one prompt every thirty minutes. Statistical models based on the simulated sensor data distinguished situations that the subjects self-reported as "highly non-interruptible" from other situations with an accuracy as high as 82.4%, significantly better than a chance accuracy of 68.0% that could be obtained by always predicting that people were not "highly non-interruptible" ($\chi^2(1, 1344) = 31.13$, $p < .001$). Note that this base level of performance


*UIST '04*, October 24–27, 2004, Santa Fe, New Mexico, USA
ACM 1-58113-962-4/04/0010

characterizes current systems, which typically act as if a person is always interruptible.

To evaluate the performance of these statistical models, the collected audio and video recordings were shown to human observers who estimated the interruptibility of the people in the recordings [2]. These human observers distinguished situations reported as "highly non-interruptible" from other situations with an accuracy of 76.9%. This study shows that our sensor-based statistical models created from simulated sensors perform significantly better than the human observers ($\chi^2(1, 3072) = 5.82$, $p < .05$). While an accuracy of 76.9% may seem low for human performance of a task very similar to everyday tasks, people do not typically make an initial estimate of interruptibility and then blindly proceed according to this initial estimate. Instead, the evaluation of interruptibility is an early step in a negotiated process. An initial estimate indicating that a person is not interruptible allows an early exit from a negotiation, but other cues, such as eye contact avoidance or the continuation of a task that would be interrupted, allow a person to determine that they should not pursue an interruption, despite an initial evaluation indicating that they could. In designing systems to use interruptibility estimates, it will be important to support a negotiated approach, rather than assuming that an interruptibility estimate provides absolute guidance.

My most recent feasibility study extended these results by deploying real sensors into the normal environments of ten office workers with more diverse job responsibilities [3]. We deployed sensors to detect motion, whether the phone was off its hook, whether the door was open, closed, or cracked, the audio level in the office, the active computer application, and the level of mouse and keyboard activity. The office workers were prompted to provide self-reports of their interruptibility at random but controlled intervals averaging once every forty to sixty minutes. A model of all ten workers, ignoring the differences in their job responsibilities and working environments, had an accuracy of 79.5%, better but not significantly different from human observers ($\Delta z = 1.34$, $p \approx .18$). More accurate models resulted when I examined subsets of the data for 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.

## SYSTEM SUPPORT FOR INTERRUPTIBILITY MODELS

While my feasibility studies have shown the potential for creating sensor-based statistical models of human interruptibility, significant obstacles remain to deploying applications based on this approach. To address some of these obstacles, I am building a system for *Automatically Modeling Interruptibility by Unobtrusively Sensing You*, or AmIBusy. AmIBusy provides mechanisms for logging sensor data, collecting observations of interruptibility, and automatically analyzing the collected sensor logs and interruptibility observations to create statistical models of human interruptibility.

AmIBusy will support two different types of applications. The first type is those applications that desire a generic estimate of interruptibility. Such applications can use standard mechanisms, such as once per day random prompts for a self-report, to collect the interruptibility observations needed to build a statistical model. Because these types of observations measure a generic notion of interruptibility, many different applications can share the same model. The second type of application is interested in a more specialized estimate of interruptibility, such as whether a person will be receptive to a notification on a mobile device. These applications can build specialized models by providing observations of the desired notion of interruptibility, such as when a person reads a notification on a mobile device or indicates that a notification was inappropriate.

AmIBusy will be extensible to support a wide variety of sensors, including simple sensors running as a child thread, sensors running asynchronously, and complex data sources like those that might be provided by the Context Toolkit [1]. My feasibility studies found that raw sensor data at the moment of an interruptibility observation is typically inadequate for creating statistical models, and that we instead need to consider different combinations of sensors and their values in the time leading up to an interruptibility observation. AmIBusy will allow developers to focus on implementing the core functionality of sensors, rather than requiring them to explicitly address the different ways that a sensor might interact with other sensors or the temporal relationship that a sensor might have with interruptibility. This is accomplished by automatically applying sets of operators to create features from raw sensor readings.

A major difference between AmIBusy and previous systems is that AmIBusy will automatically build models of human interruptibility using data collected from many people. Clients will regularly connect to an AmIBusy server, contributing interruptibility data and obtaining an updated model based on the data contributed by many clients. I will leave a discussion of the benefits of this approach to the next section, but point out that my decision to combine data collected from many people is informed by findings in my feasibility work that individual differences in what sensors predict interruptibility are not so large that they prevent models from being predictive over different kinds of office workers. Because this approach introduces the possibility of privacy concerns related to a client uploading detailed sensor logs to a server, I will also develop a distributed method for building interruptibility models. Each client will analyze its own detailed sensor

logs, uploading only a minimal representation needed by the server to create a model.

## MINIMIZING THE DISRUPTION OF BUILDING MODELS

Statistical models are built by extracting relationships between an independent variable, interruptibility in my work, and dependent variables, which in my work are features derived from the raw sensor data collected by AmIBusy. Existing approaches to collecting the necessary observations of the independent variable, interruptibility, are rather disruptive. My feasibility work is based on collecting prompted self-reports more than once per hour. Horvitz and Apacible have taken a retrospective labeling approach, recording several hours of activity in a person's office and asking the person to later watch the recordings and provide labels of their interruptibility [4]. While these two approaches have their differences, they both require significant time and attention from the person whose interruptibility is being modeled.

I will examine two approaches to minimizing the disruption associated with collecting the interruptibility observations needed to build sensor-based statistical models of human interruptibility. The first approach, alluded to in the previous section, is to combine data collected from many people. The second approach is to collect less intrusive types of interruptibility observations. This section discusses my plans to implement these approaches and then evaluate their effectiveness.

### Combining Data Collected From Many People

Combining data collected from many people provides two primary ways to reduce the disruption associated with building statistical models of human interruptibility. First, fewer observations need to be collected from any one person. In the case of self-reports, this might mean that people are prompted at most once per day, instead of the more than once per hour used in my feasibility work. Because fewer observations are collected from any one person, any one person will experience less disruption associated with providing the necessary observations of interruptibility. Second, there is no need for an initial training period in which a model learns a person's interruptibility and performs very poorly. Systems that require significant attention but provide little value are likely to be abandoned by office workers who are already busy with their work responsibilities. Instead, AmIBusy can provide an initial model based on data collected from other people. The person's own data contributes to the model over time, and there is no initial period in which the model performs very poorly.

### Less Intrusive Types of Interruptibility Observations

I will examine less intrusive types of interruptibility observations by developing and deploying a notification application, using the "What's Happening?" system as a starting point for my design [6]. Notification applications are interesting in the context of interruptibility because, even though notification streams can be of high value in the aggregate, individual notifications are often of low value and can be considered disruptive when delivered at inappropriate times. People may thus become frustrated with a notification application and disable it, even though it provides value. This has generated interest in peripheral displays for such notifications, but I am also interested in directly addressing the problem of delivering notifications at appropriate times.

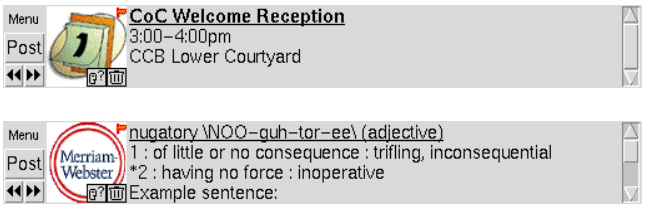


**Figure 1. Two notifications by "What's Happening?" [6].**

I will build a model of human interruptibility based on how people respond to the delivery of notifications. An explicit indication that a notification was delivered at a bad time or an ignored notification both provide some evidence that a person is not interruptible. Mousing over a notification for its text or clicking through the notification to obtain additional information both provide some evidence that a person is interruptible. Using this feedback, my notification application will build a model of the receptiveness of people to its notifications and deliver its notifications according to that model. I will implement this application using sensors that can be deployed in software to a typical laptop, such as analyses of computer activity, audio as heard by a laptop's built-in microphone, and estimates of location based on network connectivity. While AmIBusy will support a wide variety of sensors, focusing on software-deployable sensors will allow a larger evaluation than would be feasible if I needed to deploy hardware.

### Evaluation Data Collection

I will deploy my notification application in stages, with data collected from early adopters providing a better initial model to later adopters. I will log the content of each notification and how people respond to it. While every such observation of a response to a notification is appropriate for use in training my models of interruptibility, I will need to collect an unbiased sample for use in evaluating the performance of my models. I will collect this unbiased sample by delivering a small percentage of notifications at random times.

I will also ask a subset of the people who are using my notification application to provide occasional prompted self-reports of their interruptibility. My current plan is to prompt for such self-reports no more than once per day. Collecting this data will allow me to examine the relationship between responses to my notifications and the more generic notion of interruptibility measured by self-reports, as discussed in the next subsection.

### Evaluation Analyses

Having collected a dataset containing observations of how people respond to the delivery of notifications and a dataset containing interruptibility self-reports, I plan to conduct four primary analyses. Presented in the order from those that I have the most reason to believe will be successful to those that it is less clear will succeed, these are:

A. Training models from self-reports and predicting other self-reports.

B. Training models from notification responses and predicting responses to other notifications.

C. Training models from self-reports and predicting responses to notifications.

D. Training models from notification responses and predicting self-reports.

Analysis *A*, using a cross-validation approach to train models from self-reports and then predict other self-reports, is the most similar to the analyses I conducted in my feasibility work. Given the results of my prior analyses, I expect that models will perform as well as or better than the measure of human performance that I established in my feasibility work. However, this analysis will improve upon the results of my feasibility work because it will be based on data collected from many more people than were involved in my feasibility work. This will validate my approach of combining data from many people to minimize the disruption experienced by any one person.

Analysis *B*, using a cross-validation approach to train models from responses to my notification application and then predict responses to other notifications, will explore the effectiveness of sensor-based statistical models in a real-world application. I expect that models will be able to determine when it is appropriate to deliver a notification, as measured by a lower likelihood of a person indicating that a notification was delivered at an inappropriate time and a higher likelihood of a person clicking through for additional information about the notification, at least well enough to prevent notifications at the most inappropriate times. This will validate the effectiveness of my approach with a specialized notion of interruptibility, as opposed to the very generic notion of interruptibility measured by self-reports.

Analysis *C*, using models trained from self-reports to predict the receptiveness of people to the notification application, will examine the external validity of interruptibility self-reports. I again expect that these models will be able to determine when it is appropriate to deliver a notification at least well enough to prevent notifications at the most inappropriate times. This will demonstrate that models of interruptibility based on generic measures like self-reports can be used to predict more specialized notions of interruptibility. This is important because it would indicate that applications can focus on building a good general model of interruptibility, rather than needing to build many specialized models.

Analysis *D*, using models trained from responses to the notification application to predict the self-reports, examines the possibility of creating models of a generic notion of interruptibility from observations that can be collected very non-intrusively but also measure a very specialized notion of interruptibility. If successful, this would indicate that we might be able to completely do away with prompts for self-prompts and other explicit interruptibility measures. If responses to the notification application are not themselves sufficient for building reliable models of the self-reports, I am also interested in hybrid approaches. For example, if a given level of reliability can be reached by training with 100 self-reports, it might be possible to reach the same level of reliability by training with 25 self-reports and 200 responses to my notification application. This would indicate an opportunity for using less intrusive types of interruptibility observations to reduce the number of more explicit observations that need to be collected.

## CONCLUSION

Sensor-based statistical models of human interruptibility offer the potential for significant advances in human computer interaction. My dissertation will pursue this potential by contributing a series of feasibility studies, a system to support applications using sensor-based statistical models of human interruptibility, and the development and evaluation of two techniques to reduce the disruption associated with collecting the interruptibility observations needed to build statistical models.

## REFERENCES


1. 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.
2. 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)*.
3. 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.
4. Horvitz, E. and Apacible, J. (2003) Learning and Reasoning about Interruption. *Proceedings of the International Conference on Multimodal Interfaces (ICMI 2003)*, 20-27.
5. 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.
6. 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.