Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility

James Fogarty, Scott E. Hudson
Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{ jfogarty, scott.hudson }@cs.cmu.edu

Jennifer Lai
IBM T.J. Watson Research Center
Hawthorne, NY 10532
jlai@us.ibm.com

ABSTRACT
Current systems often create socially awkward interruptions or unduly demand attention because they have no way of knowing if a person is busy and should not be interrupted. Previous work has examined the feasibility of using sensors and statistical models to estimate human interruptibility in an office environment, but left open some questions about the robustness of such an approach. This paper examines several dimensions of robustness in sensor-based statistical models of human interruptibility. We show that real sensors can be constructed with sufficient accuracy to drive the predictive models. We also create statistical models for a much broader group of people than was studied in prior work. Finally, we examine the effects of training data quantity on the accuracy of these models and consider tradeoffs associated with different combinations of sensors. As a whole, our analyses demonstrate that sensor-based statistical models of human interruptibility can provide robust estimates for a variety of office workers in a range of circumstances, and can do so with accuracy as good as or better than people. Integrating these models into systems could support a variety of advances in human computer interaction and computer-mediated communication.

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

ACM Classification Keywords
H5.2. Information interfaces and presentation: User Interfaces; H1.2. Models and Principles: User/Machine Systems.

INTRODUCTION

Current computer and communication systems are generally oblivious to the social conventions defining appropriate behavior and the impact that their actions have on social situations. Whether a mobile phone rings in a meeting or a laptop interrupts a presentation to announce that its battery is fully charged, current systems often create socially awkward interruptions or unduly demand attention because they do not have any model of whether it is appropriate to interrupt. As a result, we are forced to design systems that are passive, waiting for a user to initiate action.

Prior work has examined the feasibility of using sensors and statistical models to estimate human interruptibility in an office environment [7, 13]. This work used self-reports of interruptibility and a Wizard of Oz technique to analyze 600 hours of audio and video recordings from the offices of four workers, yielding several results. The first was that human subjects asked to distinguish between "highly non-interruptible" situations and other situations in the recordings had an accuracy of 76.9%. Statistical models based on simulated sensors could make this same distinction with an accuracy as high as 82.4%, significantly better than the human subjects. Interestingly, this prior work showed that much of the accuracy of these models was derived from only a few sensors. By itself, a simulated sensor to determine whether anybody in an office was talking had an accuracy of 75.9%, indicating that social engagement played a major role in the interruptibility self-reports provided by the original subjects. This simulated talking sensor was combined with simulated sensors for keyboard or mouse activity, for using the phone, and for the time of day, resulting in a model with an accuracy of 79.2%. These results showed that statistical models based on simulated sensors can provide useful estimates of the interruptibility of office workers.

While this prior work provided a promising start to this approach to statistical models of human interruptibility, it left several questions unanswered. The four subjects studied in this work had similar jobs as high-level staff in a university, responsible for day-to-day administration of a large university department and/or graduate program. It was unclear whether the results obtained with these subjects would generalize to different types of office workers, such as programmers or others who spend less time interacting with people. There was also a question of whether real sensors could be implemented reliably enough to obtain results as good as those obtained using simulated sensors.

This paper examines several dimensions of robustness in sensor-based statistical models of human interruptibility. First, we develop real sensors based on the simulated sensors used in prior work. Second, we deploy the sensors with a much more diverse group of office workers than was examined in prior work. Results gathered from this deployment demonstrate that this approach to modeling human interruptibility works with a wide variety of people, not just people in the original demographic. Third, we examine the amount of data required to create statistical models that provide reliable estimates of interruptibility. Because it seems that models should adapt to the nuances of an individual person, we use this information to examine the tradeoff between general models and models based on data from a single person. Finally, we examine some of the tradeoffs associated with different sensor combinations. On one hand, models that examine a person's computer activity and use audio from the built-in microphone of a laptop computer can be deployed entirely in software, with no cost for sensing infrastructure. On the other hand, spending resources to instrument an office allows a model of human interruptibility that does not use cameras or microphones.

As a whole, our analyses demonstrate that sensor-based statistical models of human interruptibility can provide robust estimates for a variety of office workers in a variety of circumstances, with accuracy as good as or better than a previously established measure of human performance. The analyses also provide insight into how to include these models in systems. A system might use these models to inform callers that a person appears to be busy, allowing the caller to choose to leave a message or interrupt the apparently busy person. Another system might delay auditory notifications for non-urgent email, still putting new email in the client as it arrives, but preventing the client from playing a potentially disruptive sound when the user appears to be busy. Awareness and communication systems might share information about a colleague's interruptibility, with the goal of encouraging colleagues to not interrupt a person who appears to be busy [8]. It seems that a variety of advances in human computer interaction and computer-mediated communication could be supported by robust statistical models of human interruptibility.

The next section briefly discusses some related work. We then present our data collection method, followed by an overview of the data. Next are our analyses of this data. These start by examining the effectiveness of our sensors and statistical models across the full set of subjects and within interesting subsets of the subjects. We then discuss the effects of training data quantity on model accuracy. Our last analysis considers tradeoffs associated with different combinations of sensors. Finally, we provide a short discussion and some conclusions.

RELATED WORK

Hudson *et al* used an experience sampling technique to explore the perceptions that managers in a research environment had about interruptions [12]. They found a tension between the desire for uninterrupted time and a desire for the helpful information sometimes conveyed by an interruption. Hudson *et al* propose that the focus, therefore, should not be on reducing interruptions, but instead on making them more effective. We agree with this description of the problem, and believe that it is important that people retain control over how they are affected by models of interruptibility. For example, we do not believe that an automated system should inform a caller that the callee appears to be busy and then force the caller to leave a message. Instead, a system should give the caller the option to leave a message or interrupt the apparently busy callee. People can thus consider the importance and timeliness of a call versus the apparent interruptibility of the callee.

Horvitz *et al* have examined a variety of issues related to human attention and computer interfaces [10, 11]. Recent work by Horvitz and Apacible explores models based on calendar information, computer activity, and real-time analyses of audio and video streams [9]. Using a total of 15 hours of observations collected from the working environments of 3 subjects, they evaluate models created from these features and subsets of these features. While this work and our work are complimentary, differences in the data collection methods make it inappropriate to directly compare model performance between this work and our work. Cutrell *et al* examine interruptions created by instant messaging in a relatively specific laboratory study [4].

Begole *et al* examine the automatic extraction of temporal patterns of a person's presence [1, 2]. For example, a regular meeting could be automatically identified by the fact that a person is away from their computer for about an hour at about the same time every Wednesday. However, presence and interruptibility are not necessarily the same. A person could be present, but too busy to be interrupted. Some workers are currently forced to use presence to indicate interruptibility, physically moving away from a computer or office when they do not want to be interrupted [12, 19]. Given these sorts of problems, it seems important to examine models of both presence and interruptibility.

DATA COLLECTION

Our analyses are based on data collected with an experience sampling technique [6], sometimes referred to as a beeper study. After installing sensors in subject offices, we left and subjects went about their normal work activities. At random intervals, our setup played an audio file prompting subjects to report their current level of interruptibility. By simultaneously collecting sensor data, we can later examine which sensors and statistical models would have produced the best estimates of a person's interruptibility.

Our data was collected by a background process running on a subject's primary computer. Because some subjects used laptop computers that they needed to be able to take with them, all of our sensors were attached to the computer by a single connection to a USB hub. Subjects detached this

connection when they took a laptop computer away from an office, and our software gave occasional subtle prompts for subjects to reconnect the hub.

A set of speakers connected to the USB hub was used to prompt subjects to give interruptibility self-reports. The speakers played an audio file asking the subject to "Please give your current interruption level." For the next 10 seconds, our software recorded audio from a microphone attached to the USB hub. During this time, subjects responded orally on a 5-point scale, where a 1 indicated that a subject was highly interruptible and a 5 indicated that subject was highly non-interruptible. A sign was posted in the subject's office to remind them which end of the scale corresponded to which value. At the end of the 10 seconds, a short tone was played to let subjects know that the software was no longer recording audio. Subjects were told that a non-response would be treated as a 5 if they were on the phone and could not answer. Non-responses when the subject was not on the phone were discarded, as there was no way to reliably determine whether the subject was present and had not responded or was just not present. We initially collected this data at random intervals of between 40 and 80 minutes, but later increased the frequency to between 30 and 50 minutes. This increase was because it was very easy for subjects to miss a prompt by stepping out of their office for only a few minutes.

Besides recording subject responses to the interruptibility prompts, the USB microphone was also used as a sensor to determine whether anybody was talking in the office. The microphones were placed on shelves in each office, about 8 feet from the floor and away from computer fans or other noise sources. The audio was analyzed in real-time on the subject's computer, using the silence detector provided by the Sphinx speech recognition package [3]. This software adapts to the relatively constant noise levels generated by fans, air conditioning, or quite background music. It identifies sharp increases in the energy of audio frames collected from the microphone, but does not indicate whether these sharp increases are talking or some other noise. However, conversations tend to go on for many seconds or even many minutes, whereas most other loud noises are relatively short. In our experience, this system works well for detecting extended conversations. We logged the beginning and end times of non-silent intervals, but did not record the audio. We later examine this sensor in a relatively noisy environment, offices in which more than one person normally works.

A custom-built USB sensor board was used to instrument each subject's office. Two magnetic switches, one near each side of the top of the door frame, allowed us to sense whether the door was open, cracked, or closed [17]. Two motion sensors were put in each office, both about 5 feet above the floor, one near the door and one near the subject's desk. Another magnetic switch was used to determine whether a person's phone was physically off its hook. This switch could not detect if the person was using the speaker-phone functionality, but we were not allowed access to the phone systems that would have detected this. In any case, the microphone talking sensor would be likely to detect talking when a subject used the speaker-phone.

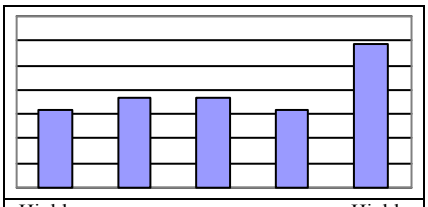
Software on each subject's computer logged, once per second, the number of keyboard, mouse move, and mouse click events in the previous second. It also logged the title, type, and executable name of the active window and each non-active window. We chose to log this information out of the belief that some subjects might be more or less interruptible when working in certain applications. All of the information associated with our study was automatically compressed and uploaded to a local server, so that we could verify that each subject's sensors seemed to be working and so that we could determine when each subject had given the desired number of responses.

DATA OVERVIEW

This section discusses interruptibility self-reports collected from 10 subjects with no prior relationship to this work. The subjects were all employees of a major corporate research laboratory, studied during the course of their normal work. The first two subjects were first-line managers, selected because we felt that their human-centered work was closest to work of the four subjects studied in prior work [7, 13]. Just as social engagement was a key indicator of non-interruptibility in prior work, we expected social engagement to be a key indicator for these two subjects. The next five subjects were researchers who spent a significant amount of their time programming. These subjects were selected because they seem to represent typical knowledge workers, who do interact socially at work but also work on tasks that require focused attention. The last three subjects were summer interns, selected because they shared an office with another summer intern. Given the prior work indicating that a talking sensor is important for estimating interruptibility, these subjects were selected to examine how the regular presence of a second person in the office affected the usefulness of our talk sensor implementation.

We set out to collect 100 interruptibility self-reports from each subject. Figure 1 shows the 975 responses that were actually collected. Data collection for subject 7, one of the researchers, was terminated early because of an external deadline that required the removal of the sensors from his office. Data collection for subject 9, one of the interns, was terminated early because she expressed a feeling that the interruptibility prompts were annoying and asked for the sensors to be removed. Some subjects went slightly over because of the small delay from the subject reaching 100 reports to us taking down the sensors.

While there are individual differences in the distribution of the interruptibility self-reports, note that the most common response was 5, or "highly non-interruptible", accounting for nearly 30% of the data. This distribution is very similar to the distribution found in prior work [7, 13], and seems to



#	Category	Highly Interruptible 1	2	3	4	Highly Non-Interruptible 5
1	Manager	17	31	10	19	37
2	Manager	2	8	36	26	26
3	Researcher	67	7	8	13	10
4	Researcher	23	16	27	3	34
5	Researcher	4	7	24	18	51
6	Researcher	5	34	29	12	27
7	Researcher	0	6	12	16	34
8	Intern	26	28	25	11	13
9	Intern	3	17	12	15	17
10	Intern	11	28	2	25	43
Total		158 16.2%	182 18.7%	185 19.0%	158 16.2%	292 29.9%

Figure 1. Distribution of interruptibility self-reports.

		Model: Other Values	Model: Highly Non
Self-Report	Other Values	640 65.6%	43 4.4%
Self-Report	Highly Non	157 16.1%	135 13.8%
		Accuracy: 79.5% Base: 70.1%	

Figure 2. Performance of model built from all collected data.

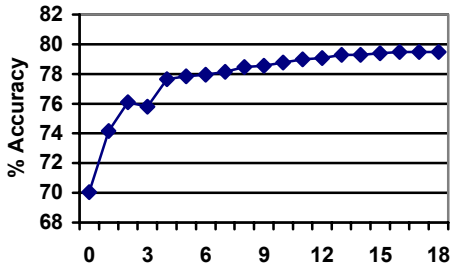


Figure 3. Number of features versus percent accuracy.

indicate that there are circumstances in which people consider themselves to be clearly non-interruptible, as opposed to other times when their interruptibility might be more dependent on the nature of the interruption or some other factor. It therefore seems reasonable to attempt to build models that identify these "highly non-interruptible" circumstances. Our analyses in this paper will focus on models that differentiate between self-reports of 5 versus values between 1 and 4. The base performance for this problem is an accuracy of 70.1%, which could be obtained by always predicting that the person was not "highly non-interruptible". Note that this is what current systems generally do, because they cannot model interruptibility.

MODEL PERFORMANCE

This section presents the performance of statistical models of human interruptibility built from this data. We start by examining models built from all of the data, and then move to considering interesting subsets. All of our models were built in the Weka machine learning environment [20], using a naïve Bayes classifier [5, 15] and wrapper-based feature selection [14]. In a wrapper-based technique, the features used by a model are chosen by adding new features until no potential feature improves the accuracy of the model. Models are evaluated using a standard cross-validation approach, with 10 folds. That is, each model is evaluated in 10 trials, with each trial using 90% of the data to train the model and the other 10% to test the model. The values presented are sums of the 10 trials, so they sum to the total number of self-reports used. Prior work considered several different types of classifiers, but did not find a significant difference in their performance for this problem [7, 13]. We use naïve Bayes classifiers in this work because they are computationally very inexpensive, which is important when using a wrapper-based feature selection technique.

Figure 2 shows the performance of a model built with the full set of self-reports, presented as a confusion matrix. Because we will use several of these matrices in this paper, it is worth clarifying that the upper-left corner indicates that there were 640 cases of the model correctly predicting a self-report value between 1 and 4. The bottom-right corner indicates there were 135 cases where the model correctly predicted a value of 5, or "highly non-interruptible". The upper-right corner shows there were 43 cases where the model predicted a self-report of 5 and the subject had actually responded with a value between 1 and 4. Finally, there were 157 cases where the modeling predicted a value between 1 and 4, but the subject responded with 5. This model has an accuracy of 79.5%, significantly better than the base performance of 70.1% ($\chi^2(1, 1950) = 23.01$, $p < .001$). It is also better than the 76.9% accuracy that prior work found for people estimating another person's interruptibility [7], though this difference is not quite significant ($\Delta z = 1.34$, $p \approx .18$).

The wrapper-based feature selection process used to create our models starts with an empty set, tries every feature to determine which is most predictive, and adds that feature. This process repeats to determine which feature is the best second feature, third feature, etc. until no available feature will improve the accuracy of the model. Figure 3 shows that much of the predictive power of this model comes from the first few features, a finding consistent with prior work [7, 13]. The first feature, whether the phone was off its hook in the last 15 seconds, improves the accuracy of the model from the base of 70.1% to 74.2%. The next three features combine to increase the accuracy to 77.6%. The remaining 14 features contribute very slowly, with the accuracy rising to 79.5% before selection terminates.

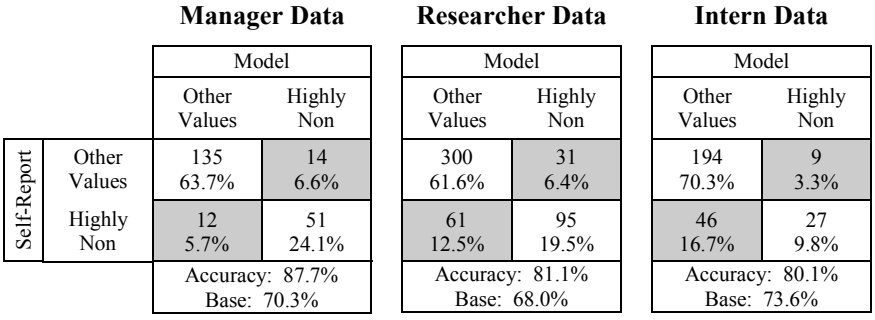


Figure 4. Performance of models built from the manager, researcher, and intern data.

Manager Data

The manager subjects in this study have jobs that seem similar to those of the four subjects studied in prior work, so our expectation was that the sensors found to work well in prior work would also work well for these subjects. The left confusion matrix in Figure 4 shows the result of creating and evaluating a model using only the data from the two manager subjects. The model has an accuracy of 87.7%, significantly better than the 70.3% base for this subset ($\chi^2(1, 424) = 19.47$, $p < .001$) and significantly better than the 76.9% human performance shown in prior work ($\Delta z = 1.34$, $p < .001$). The first feature to be selected, and therefore the most predictive for this data, is whether the phone was off its hook in the last 15 seconds. The third feature selected is whether the talking detector has detected talking for 3 of the last 5 minutes. Both of these indicate social engagement. The second feature selected is whether the subject generated 30 mouse move events in the last 15 seconds. The manager subjects were more interruptible when this feature was true, indicating engagement with the computer. Together, these show that social engagement is the major indicator of non-interruptibility for these subjects.

Researcher Data

Although the researcher subjects interact with colleagues, they also spend a significant amount of time programming, working on papers, or otherwise engaged in individual work. We were interested to see whether the talking sensor would still be useful for these subjects. The middle confusion matrix in Figure 4 shows the result of creating and evaluating a model using only the data from the five researcher subjects. The model has an accuracy of 81.1%, significantly better than the 68.0% base for this subset ($\chi^2(1, 974) = 22.16$, $p < .001$) and better than the 76.9% human performance shown in prior work, but not significantly ($\Delta z = 0.89$, $p \approx .37$). Interestingly, whether or not the phone was off its hook during the last 15 seconds was not selected until the fourth feature. The first feature to be selected, and therefore the most predictive feature, was whether talking had been detected for 30 of the last 60 seconds. The second feature selected, whether the subject had generated 60 mouse move events inside Microsoft Visual Studio in the last 30 seconds, indicated that these subjects were less interruptible when interacting with the programming environment. But that did not mean that they were less interruptible whenever they were active on their computer, as the third feature selected showed that they were more interruptible when they had typed 60 characters in the last 15 seconds. As a whole, the selection of these features seems to show that task engagement was more important for these subjects than it was for the manager subjects. They spent less time on the phone, and so the phone sensor was a less reliable way to detect that they were non-interruptible. But the model was able to handle this difference between the subject groups by learning that the researcher subjects were less interruptible when active in a programming environment.

Intern Data

The intern subjects are interesting because their shared offices meant that noise in the environment was not necessarily associated with the subjects. While it makes sense to expect that a talking sensor might be less effective in such an environment, we included these subjects to see how much of a negative effect the extra noise had on the models. The right confusion matrix in Figure 4 shows the result of creating and evaluating a model using only the data from the three intern subjects. The model has an accuracy of 80.1%, significantly better than the 73.6% base for this subset ($\chi^2(1, 552) = 3.30$, $p \approx .070$) and better than the 76.9% human performance shown in prior work, but not significantly ($\Delta z = 0.17$, $p \approx .86$). As we might have expected, the models compensated for the reduced reliability of the talking sensor by selecting features that consider the talking sensor over a larger time interval. Whereas the model created from the manager data selected the talking sensor at a level of 3 of the last 5 minutes and the model created from the researcher data selected the talking sensor at a level of 30 of the last 60 seconds, the model created from the intern data selected the talking sensor at a level of 15 of the last 30 minutes. This indicates that the model found long conversations more relevant to interruptibility than relatively short activations of the talk sensor. This talk sensor was the third feature selected, while the first feature selected was mouse activity in a window created by javaw.exe. Since all three interns had programming backgrounds, this is probably related to the use of the Eclipse programming environment, which runs as javaw.exe. This shows that this indication of task engagement was important for this group of people, as it was with the researchers. Surprisingly, the second feature

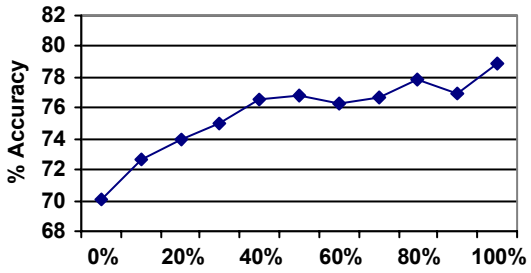


Figure 5. Percent of training data versus percent accuracy on the full data set.

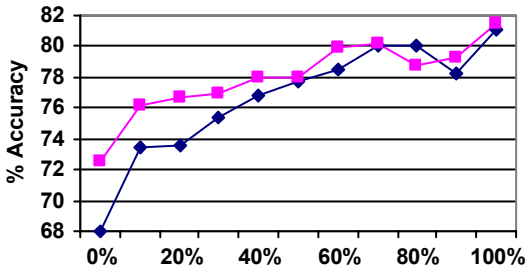


Figure 6. Percent of training data versus percent accuracy on researcher data set, comparing addition of manager data.

selected was whether the motion detectors had been triggered 60 times during the last 30 minutes. The motion detector features were not selected by any of the other models, and certainly not as the second most predictive feature in any of the other models. This seems to indicate that the degradation of the talk sensor associated with the noisier environment made the motion sensors useful for this group of subjects, perhaps because they captured the motion of a guest in the office.

TRAINING DATA AND MODEL PERFORMANCE

The results of the previous section seem to make it clear that statistical models of interruptibility should adapt to the people who are using them. All three models created from a single group of subjects perform better than the model created with the full set of data. While a talking sensor is used for all three groups, the time interval over which to examine the talk sensor is different for the three groups. Different computer applications relate to the interruptibility of the different groups. Furthermore, the usefulness of the motion sensors in the relatively noisy environments of the intern subjects seems like a surprising finding. While a developer might not think to include this particular feature in a model when it is initially deployed, a model should be able to recognize that it is predictive and begin to use it.

If models are to learn which features best predict the interruptibility of people, an important question is how much training data needs to be collected to give reliable estimates. Figure 5 examines this question with the full set of data collected for this work. The values plotted were calculated using a modification of the standard 10 fold cross-validation discussed in the previous section. Recall that each fold in the cross-validation uses 90% of the data for training and 10% for testing. This graph plots the accuracy of models evaluated on the test data as affected by using less than the full 90% of the training data. The value at the 30% mark, for example, is based on using only 30% of the training data. At the 100% mark, the model is trained using the full 90% of the training data, yielding the same accuracy presented in the last section. To ensure that we did not select a particular good or bad subset of the training data, we conducted 10 trials and report the average. Note that the accuracy of the model improves very quickly with the first 10% of the training data. After this initial jump, the accuracy continues to improve at a relatively slow pace.

Given that our models seem to gradually approach their final accuracy, rather than quickly reaching a level of accuracy close to their final accuracy, it seems important to consider how we might improve the accuracy of a model that is still learning the best predictors for the person using it. One approach to this problem would be to use data collected from other people to create an initial model, and then update that model as data becomes available about the person using the system. The idea is that there might be certain indications of interruptibility that are useful across different groups of people. A statistical model might be able to learn this information from data collected for other types of office workers, adapting as more information becomes available about the person using the model.

Figure 6 considers the usefulness of our manager data to a model being trained from and tested against the researcher data. The bottom line in Figure 6 was plotted in the same way as the line in Figure 5, except that it was built using only the data from our five researcher subjects. This shows the same initial jump followed by gradual improvement that we saw in Figure 5. The top line in Figure 6 represents a similar process, except that each training step also included 212 self-reports collected from our manager subjects. Each of the manager self-reports is weighted such that the model treats it with one-fifth of the importance the model gives to each self-report collected from the researchers. So at the 0% mark, the model is training with only the equivalent of 42 manager self-reports, at the 10% mark it is training with the equivalent of 42 manager self-reports and 44 researcher self-reports, and at the 100% mark it is training with the equivalent of 42 manager self-reports and 438 researcher self-reports. The model therefore has the manager data available when the researcher data has not yet provided any information, but the researcher data is treated as much more important when it becomes available. Note that the top line initially performs better when its models are evaluated against the researcher data, and it converges to perform the same as the bottom line when more researcher data becomes available.

The findings in this section seem to indicate that statistical models can adapt to the nuances of interruptibility, and that information collected from other people can help to provide a more accurate initial model. The model can then adapt using a weighted combination of data from other people

		Computer Only Manager Data		Computer Only Researcher Data		No Microphone Manager Data		No Microphone Researcher Data		No Microphone Intern Data	
		Model		Model		Model		Model		Model	
		Other Values	Highly Non	Other Values	Highly Non	Other Values	Highly Non	Other Values	Highly Non	Other Values	Highly Non
Self-Report	Other Values	136 64.2%	13 6.1%	248 61.6%	14 6.4%	145 68.4%	4 1.9%	314 61.6%	17 6.4%	200 72.5%	3 1.1%
	Highly Non	15 7.1%	48 22.6%	75 12.5%	47 19.5%	35 16.5%	28 13.2%	86 12.5%	70 19.5%	52 18.8%	21 7.6%
		Accuracy: 86.8% Base: 70.3%		Accuracy: 76.8% Base: 68.3%		Accuracy: 81.6% Base: 70.3%		Accuracy: 78.9% Base: 68.0%		Accuracy: 80.1% Base: 73.6%	

Figure 7. Performance of models built using only computer activity and the built-in microphone, or built without a microphone.

and data from the person to whom it is adapting. This seems to be a good property, as it means that statistical models of human interruptibility can avoid two pitfalls. The first pitfall would be if a model did not match a person and could not be corrected. As we have shown, this approach seems to adapt well to different types of office workers. The second pitfall would be if a model required extensive training before it could be of any use, which might result in a person giving up on it before it had a chance to be effective. This section shows that models can use data collected from other people to provide better initial performance.

SENSOR COMBINATIONS

This section considers statistical models based on two particularly interesting combinations of sensors. Six of our ten subjects use laptop computers, which typically have a small built-in microphone near the keyboard. Using this microphone and the computer activity, we examine models that could be deployed entirely in software. That is, this sensor combination does not require the installation of any sensing infrastructure and has essentially zero cost. But the extreme proximity of this microphone to the keyboard and the computer fan means that it could be too noisy to be a reliable indicator of interruptibility. Note that while an organization-wide deployment of such an approach could include a phone sensor implemented in software on the organization's phone system, we do not include a phone sensor in this analysis so that the result more accurately reflects what might be expected without such support.

Figure 7 shows the performance of models built using the built-in laptop microphone (instead of the microphone we placed in the office and attached by USB) for our two manager subjects and four of our researcher subjects. These models were not allowed to use our phone, door, or motion sensors. Using just the built-in microphone and computer activity, the manager model has an accuracy of 86.8%, not significantly different from the 87.7% accuracy of the model presented in Figure 4 ($\chi^2(1, 424) = 3.30$, $p \approx .77$) and significantly better than the 76.9% human performance shown in prior work ($\Delta z = 3.55$, $p < .001$). On the other hand, the researcher model, with an accuracy of 76.8%, is not significantly different from the 76.9% human performance shown in prior work ($\Delta z = 0.91$, $p \approx .360$) and worse than the researcher model presented in Figure 4 ($\Delta z = 0.91$, $p \approx .14$). Interestingly, this researcher model does not select any feature related to the laptop microphone.

These models seem to indicate that the laptop microphone was a reliable indicator for the manager subjects, but was not a reliable indicator for the researcher subjects, probably due to the added noise associated with a microphone so close to the computer. Future work would seem to require a more robust implementation of a speech sensor, perhaps using techniques like those presented by Lu *et al* [16]. However, even though the researcher model in Figure 7 was not as reliable as that in Figure 4, it still performed as well as the 76.9% accuracy of human subjects studied in prior work and is still better than the base performance of 68.3% that would be associated with current systems that do not attempt to model interruptibility.

We now turn to the possibility of building models without using a camera or microphone. Such models would require instrumentation of an office, but might be received more positively by a person concerned about the presence of a camera or a microphone in their office. In a discussion of privacy and technology, Palen and Dourish point out that the lack of real-world cues makes it more difficult to know who might have access to information collected by technology and how they might use it [18]. These types of concerns could prevent the deployment of systems using cameras or microphones, so it seems worth considering whether more limited sensors can be useful. For example, a motion sensor clearly can only detect motion and a door sensor can only detect whether the door is open, cracked, or closed. Neither can be abused to record a conversation.

Figure 7 shows the accuracy of models built for each group of subjects without using a microphone sensor. Somewhat surprisingly, the researcher and intern models both have accuracies very close to the accuracy of the models in Figure 4 that did include a microphone ($\chi^2(1, 974) = 0.78$, $p \approx .38$, $\chi^2(1, 552) = 0$, $p \approx 1$). But the manager model accuracy of 81.6% is significantly worse than the 87.7% accuracy of the manager model in Figure 4 that was built with a microphone ($\chi^2(1, 424) = 3.07$, $p \approx .08$). This result seems to reemphasize that social engagement was critical to predicting the interruptibility of our manager subjects, with task engagement being more important for the researcher and intern subjects. Without a microphone, the phone sensor was the only method of detecting social engagement,

and the manager model suffered, though it was still better than the 76.9% human performance found in prior work ($\Delta z = 1.48$, $p \approx .14$).

DISCUSSION AND CONCLUSION

We have shown that statistical models can reliably predict interruptibility as well as, if not better than, a level of human performance established in prior work. While prior work used simulated sensors in the offices of four workers with very similar job responsibilities, these models are based on real sensors deployed with a variety of office workers. The researcher and intern subjects were selected specifically because their work environments are different from those of the subjects studied in prior work. The success of our approach with these subjects gives us reason to believe that reliable statistical models of interruptibility can be created for a variety of office workers in a broad set of circumstances.

Our results also show that models should adjust to the nuances of a person's interruptibility, as doing so seems to allow better performance than is obtained by a general model. But general data is still useful, as it can improve the performance of a model that it is still learning the nuances of a person's interruptibility. It seems that large improvements over the accuracy of the models presented in this paper are likely to come from a more sophisticated model or from considering new features, rather than from collecting more data for use with the same type of classifier used in this work. One fruitful approach might be to consider combining an approach similar to ours with the work on temporal patterns done by Begole *et al* [1, 2].

Finally, we have shown that this approach can succeed even with significant limitations on the sensors that are available. We demonstrated models that could be built without any sensors beyond those already present in a typical laptop computer, and we also showed that models can succeed even if a person will not allow a microphone in their work environment. Even when these limited models did not perform as well as models built from the full set of sensors, they still performed as well as or better than people.

ACKNOWLEDGMENTS

We would like to thank all of the participants in our study. We also thank Ryan Baker and Darren Gergle for cheerfully answering statistical questions and Johnny Lee for his help with the USB sensor board used in our sensing setup. This work was funded in part by DARPA and by the National Science Foundation under Grants IIS-0121560, IIS-0325351, and the first author's Graduate Research Fellowship.

REFERENCES

1. 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.
2. 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.
3. CMU Sphinx: Open Source Speech Recognition. http://www.speech.cs.cmu.edu/sphinx/
4. Cutrell, E., Czerwinski, M. and Horvitz, E. (2001) Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance. *Proceedings of Interact 2001*, 263-269.
5. Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*. John Wiley and Sons.
6. Feldman-Barrett, L. and Barrett, D.J. (2001) Computerized Experience-Sampling: How Technology Facilitates the Study of Conscious Experience. *Social Science Computer Review* (19). 175-185.
7. 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)*.
8. Fogarty, J., Lai, J. and Christensen, J. (2004) Presence versus Availability: The Design and Evaluation of a Context-Aware Communication Client. *To Appear, International Journal of Human-Computer Studies (IJHCS)*.
9. Horvitz, E. and Apacible, J. (2003) Learning and Reasoning about Interruption. *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI 2003)*, 20-27.
10. Horvitz, E., Kadie, C., Paek, T. and Hovel, D. (2003) Models of Attention in Computing and Communication: From Principles to Applications. *Communications of the ACM*, *46* (3). 52-59.
11. 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.
12. 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 SIGCHI Conference on Human Factors in Computing Systems (CHI 2002)*, 97-104.
13. 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 SIGCHI Conference on Human Factors in Computing Systems (CHI 2003)*, 257-264.
14. Kohavi, R. and John, G.H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence* (97). 273-324.
15. Langley, P. and Sage, S. (1994) Induction of Selected Bayesian Classifiers. *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 1994)*, 399-406.
16. Lu, L., Zhang, H. and Jiang, H. (2002) Content Analysis for Audio Classification and Segmentation. *IEEE Transactions on Speech and Audio Processing*, *10* (7). 504-516.
17. Nichols, J., Wobbrock, J.O., Gergle, D. and Forlizzi, J. (2002) Mediator and Medium: Doors as Interruption Gateways and Aesthetic Displays. *Proceedings of the ACM Conference on Designing Interactive Systems (DIS 2002)*, 379-386.
18. Palen, L. and Dourish, P. (2003) Unpacking "Privacy" for a Networked World. *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2003)*, 129-136.
19. Perlow, L.A. (1999) The Time Famine: Toward a Sociology of Work Time. *Administrative Science Quarterly*, *44* (1). 57-81.
20. Witten, I.H. and Frank, E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.