Predicting Human Interruptibility with Sensors

JAMES FOGARTY, SCOTT E. HUDSON, CHRISTOPHER G. ATKESON, DANIEL AVRAHAMI, JODI FORLIZZI, SARA KIESLER, JOHNNY C. LEE, and JIE YANG
Carnegie Mellon University

A person seeking another person's attention is normally able to quickly assess how interruptible the other person currently is. Such assessments allow behavior that we consider natural, socially appropriate, or simply polite. This is in sharp contrast to current computer and communication systems, which are largely unaware of the social situations surrounding their usage and the impact that their actions have on these situations. If systems could model human interruptibility, they could use this information to negotiate interruptions at appropriate times, thus improving human computer interaction.

This article presents a series of studies that quantitatively demonstrate that simple sensors can support the construction of models that estimate human interruptibility as well as people do. These models can be constructed without using complex sensors, such as vision-based techniques, and therefore their use in everyday office environments is both practical and affordable. Although currently based on a demographically limited sample, our results indicate a substantial opportunity for future research to validate these results over larger groups of office workers. Our results also motivate the development of systems that use these models to negotiate interruptions at socially appropriate times.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Collaborative computing*; H.1.2 [**Models and Principles**]: User/Machine Systems; I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Design, Measurement, Experimentation, Human Factors

Additional Key Words and Phrases: Situationally appropriate interaction, managing human attention, context-aware computing, sensor-based interfaces, machine learning

1. INTRODUCTION

People have developed a variety of conventions that define what behavior is socially appropriate in different situations [Barker 1968]. In office working

This work was funded in part by DARPA, by the National Science Foundation under Grants IIS-01215603, IIS-0205219, IIS-9980013, and by J. Fogarty's NSF Graduate Research Fellowship.
Author's address: J. Fogarty, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburg, PA 15213-3891; email: jfogarty@cs.cmu.edu.

environments, social conventions 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 an interruption is appropriate. Social conventions around interruptions also allow the development of an *a priori* expectation of whether or not an interruption would be appropriate [Hatch 1987].

Current computer and communication systems are largely unaware of the social conventions defining appropriate behavior, of the social situations surrounding them, and the impact that their actions have on social situations. Whether a mobile phone rings while its owner is in a meeting with a supervisor or a laptop interrupts an important presentation to announce that the battery is fully charged, current computer and communication systems frequently create socially awkward interruptions or unduly demand attention because they have no way to determine whether it is appropriate to interrupt. It is impossible for these systems to develop informed *a priori* expectations about the impact their interruptions will have on users and the social situations surrounding usage. As computing and telecommunications systems have become more ubiquitous and more portable, the problem has become more troublesome.

People who design or use computer and communication systems can currently adopt 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, applications in intelligent spaces and other mobile or ubiquitous computing environments could benefit from a system being able to initiate interactions [Horvitch 1999]. 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 convey critically important information. Because systems 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 forget to re-enable systems after a potentially inappropriate time interval has passed [Milewski and Smith 2000].

If we could develop relatively robust models of human interruptibility, they might support a variety of significant advances in human computer interaction and computer-mediated communication. Such models do not need to deprive people of control. For example, mobile phones could automatically 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 to leave a message instead [Schmidt et al. 2000]. Email and messaging applications might delay potentially disruptive auditory notifications for less important messages, but never prevent delivery of the information. Information displays might choose between several methods of conveying information according to the current appropriateness of each method of communication. Many specific applications could 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.

McFarlane [1999, 2002] tested four known methods for deciding when to interrupt people. Although his results have implications for structuring appropriate interactions, no single method emerged as best across all performance measures. Czerwinski et al. [2000a, 2000b] and Cutrell et al. [2001] studied interruptions created by instant messages and the effect of these interruptions on different computer tasks. Importantly, they found that an instant messaging notification is disruptive to task performance even when it is ignored. These studies focused on very specific computer tasks and leave open questions related to the effect of interruptions on the social situations surrounding computer usage. Voida et al. [2002] discuss such social situations while analyzing tensions in instant messaging related to uncertainty about the level of attention being given by a remote person. They suggest that instant messaging applications might benefit from providing better indications of the availability of a remote person. Begole et al. [2002, 2003] present temporal analyses of activity logs from an awareness application for distributed workgroups. They find that certain patterns may indicate when a person will become available for communication, but note that only information related to computer usage is available for their analyses.

Horvitz et al. [1998] have shown that models can be used to infer goals and provide appropriate assistance. Observing low-level mouse and keyboard events, their Lumière prototype modeled tasks that a person might be performing and used its interpretation to provide assistance. Oliver et al.'s [2002] SEER system uses models to recognize a set of human activities from computer activity, ambient audio, and a video stream. These activities are a phone conversation, a presentation, a face-to-face conversation, engagement in some other activity, conversation outside the field of view of the camera, and not present. The activities SEER models may relate to interruptibility, but they are examined only in a controlled environment and cannot directly estimate interruptibility.

Horvitz et al. [1999] present methods for estimating the importance of a potential interruption in their discussion of the Priorities prototype. Although they focus on using a text classification strategy to identify important emails, they note that the methods they present can apply to other classes of notifications. These types of methods will be significant in creating systems that balance interruptibility against the importance of potential interruptions.

Hudson et al. [2002] used an experience sampling technique to explore the perceptions that managers in a research environment had about interruptions. They found that there was a tension between desiring uninterrupted working time and the helpful information sometimes obtained from an interruption. In a result similar to that discussed by Perlow [1999], Hudson et al. found that people sometimes isolate themselves from potential interruptions by ignoring notifications or moving to a different physical location. We point out that this strategy demonstrates the problem we previously discussed, that people forbid

all interruptions because the systems they use cannot determine whether a potential interruption is appropriate. Hudson et al. propose that researchers focus on making interruptions more effective and suggests socially translucent systems [Erickson and Kellogg 2000] as an approach. Bellotti and Edwards [2001] express a similar concern that context-aware systems will not always get it right, and the systems need to be designed so that they defer to people in an accessible and useful manner.

This article describes work to develop and quantitatively evaluate sensor-based statistical models of human interruptibility. Because people use social conventions and externally visible cues to estimate interruptibility rather than relying on invisible internal phenomena like a cognitive state, it should be possible to develop such models empirically. One approach would be the top-down creation, deployment, and evaluation of various combinations of models and sensors. However, the uncertainty surrounding the usefulness of various sensors makes it very likely that significant time and resources would be spent building and evaluating sensors ill-suited or suboptimal for the task. This work is instead based on a bottom-up approach, in which we collected and analyzed more than 600 hours of audio and video recordings from the actual working environments of four subjects with no prior relationship to our research group. We simultaneously collected self-reports of the interruptibility of these subjects. Using these recordings, we have examined human estimates of the interruptibility of the people in the recordings. We have also created models of interruptibility based on the assumption that changes in behavior or context are indicative of interruptibility. These models use sensor values that were manually simulated by human coding from the recordings, using a Wizard of Oz technique [Dahilbäck et al. 1993; Maulsby et al. 1993].

This article shows that models of interruptibility based on simple sensors can provide estimates of interruptibility that are as good as or better than the estimates provided by people watching audio and video recordings of an environment. More specifically, we present a study demonstrating that people viewing the audio and video recordings can distinguish between "Highly Non-interruptible" situations and other situations with an accuracy of 76.9%. A model based on manually simulated sensors makes this same distinction with an accuracy of 82.4%. Both of these accuracies are relative to a chance accuracy of 68% that could be obtained by always estimating that a situation was not "Highly Non-interruptible." These types of models can be built using only a handful of very simple sensors. While the study is based on a limited demographic and will need to be validated for different groups of office workers, the result is still very promising. The favorable comparison between human judgment and our models indicates an opportunity for using interruptibility estimates in computer and communication systems.

In the following section, we introduce our subjects, the collection of audio and video recordings in their work environments, and the specifics of their interruptibility self-reports. Then we present an overview of this collected data, as described by the interruptibility self-reports and our set of manually simulated sensors. This is followed by a presentation of our first study, examining human estimates of interruptibility based on the recordings. We then move

Fig. 1. Representative frames from the recordings.

to our second study, discussing models of interruptibility based on manually simulated sensors, including an analysis of the usefulness of various sensors and a comparison of these models to human estimates. We next present models based on limited automated analyses of the recordings. Finally, we offer a short conclusion and discuss opportunities for future work.

2. DATA COLLECTION

The recordings discussed in this article were collected in the actual working environments of four subjects with no prior relationship to our research group. To increase uniformity for this exploratory work, we selected four subjects with similar working environments and tasks. Each subject serves in a high-level staff position in our university with significant responsibilities for day-to-day administration of a large university department and/or graduate program. The subjects have private offices with closable doors, but their responsibilities require them to interact with many different people and they generally do not have full control over their time. They usually work with their doors open and responded to a variety of "walk in" requests. Because they almost never close their office doors, it is likely that the absence of this explicit indication of non-interruptibility makes it more difficult to estimate their interruptibility.

Recordings were collected using a computer with an 80GB disk and an audio/video capture card connected to a small camera and microphone. Subjects could disable recording for thirty minutes by pressing the space bar. The computers had speakers used for informing subjects that recording had been disabled, to advise them recording was about to resume, and to request interruptibility self-reports. They did not have displays. Signs were posted to alert guests to the presence of a recording device, and the subjects were encouraged to disable recording if they or a guest was uncomfortable. We also provided subjects with a mechanism for retroactively requesting that recordings be destroyed.

Grayscale cameras with wide-angle lenses were mounted in the office so that both the primary working area and the door were visible. Figure 1 shows images from two of the cameras. Video was captured at approximately 6 frames per second, at a resolution of 320×240. Audio was captured at 11KHz, with 8-bit

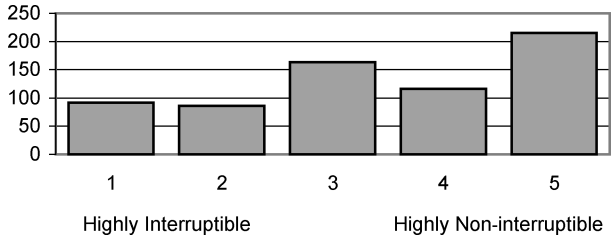


Fig. 2. Interruptibility self-report distribution.

Table I. Individual Subject Self-Report Distributions

	Highly Interruptible				Highly Non-Interruptible
	1	2	3	4	5
Subject 1	9 6.6%	14 10.2%	40 29.2%	18 13.1%	56 40.9%
Subject 2	17 10.2%	21 12.7%	58 34.9%	27 16.3%	43 25.9%
Subject 3	52 31.5%	26 15.8%	20 12.1%	10 6.1%	57 34.5%
Subject 4	14 6.9%	25 12.3%	45 22.1%	61 29.9%	59 28.9%
All	92 13.7%	86 12.8%	163 24.3%	116 17.3%	215 32.0%

samples. The machines were deployed for between 14 and 22 workdays for each subject, recording from 7am to 6pm on workdays. Our setup worked well except in one case where a week of data was lost because an undetected improper compression setting caused the disk to fill prematurely. For this subject, we collected an additional 10 days of data at a later date. A total of 602 hours of recordings was collected from the offices of these four subjects.

Subjects were prompted for interruptibility self-reports at random, but controlled, intervals, averaging two prompts per hour. This is an experience-sampling technique, or alternatively a beeper study [Feldman-Barrett and Barrett 2001]. To minimize compliance problems, we asked a single question rated on a five-point scale. Subjects could answer verbally or by holding up fingers on one hand, but almost all responses were verbal. Subjects were asked to "rate your current interruptibility" on a five-point scale, with 1 corresponding to "Highly Interruptible" and 5 to "Highly Non-interruptible." A sign on the recording machine reminded the subject which value corresponded to which end of the scale. Subjects were present for a total of 672 of these prompts.

3. DATA OVERVIEW

This section characterizes the data collected from our subjects. The overall distribution of interruptibility self-reports is shown in Figure 2. The distributions for individual subjects are shown in Table I. For 54 of these 672 samples, the subject was present and clearly heard the prompt, but did not respond within 30 seconds. We examined these individually and determined that the subject was either on the phone or with a guest for the vast majority of the 54 cases.

Table II. Frequency of Events During Times When the Office Occupant was Present

Door Open	98.6%	Door Close	0.7%
Occupant Sit	88.9%	Occupant Stand	13.1%
Occupant at Desk	74.0%	Occupant at Table	21.2%
Occupant Keyboard	22.6%	Occupant Mouse	19.6%
Occupant Monitor	46.8%	Occupant File Cabinet	1.0%
Occupant Papers	28.0%	Occupant Write	5.5%
Occupant Drink	1.0%	Occupant Food	1.4%
Occupant Talk	32.6%	Occupant on Telephone	12.7%
One or More Guests Present	24.1%	Two or More Guests Present	3.0%
One or More Guests Sit	9.3%	Two or More Guests Sit	1.5%
One or More Guests Stand	14.2%	Two or More Guests Stand	0.8%
One or More Guests Talk	20.7%	Two or More Guests Talk	1.7%
One or More Guests Touch	0.5%	Two or More Guests Touch	0.0%

Results in the literature suggest that these activities are highly correlated with non-interruptibility, and this expectation is validated in the remainder of our data. To simplify analysis and model building, we have placed these 54 cases in the "Highly Non-interruptible" category.

While there are clearly differences in the self-report distributions for the individual subjects, it is especially important to note that subjects self-reported "Highly Non-interruptible" for 215 prompts, or approximately 32% of the data. An informal inspection found that responses of "Highly Non-interruptible" were sometimes given calmly and other times curtly by agitated subjects. For many of the analyses in this article, we will examine this distinction and evaluate the ability of estimators to distinguish "Highly Non-interruptible" situations from other situations.

Table II presents how often particular events occur in the recordings. These values are based on manually simulated sensors that will be discussed later in this article. They are also based on the periods for which the subject was present, as opposed to the entirety of the recordings. As previously mentioned, these subjects almost always had their doors open. The lack of the explicit non-interruptibility cue provided by a closed door probably makes it more difficult to estimate their interruptibility. The subjects spent most of the day sitting, and most of that time sitting at their desks. A guest was present approximately 25% of the time when the subjects were present, but there was very rarely more than one guest present. While subjects frequently interacted with a computer, they also spent a significant amount of time handling papers or talking.

4. HUMAN ESTIMATION

In order to evaluate the difficulty of estimating interruptibility and establish an important comparison point for our models, we conducted an experiment examining the human estimation of interruptibility. Subjects that we will refer to as *estimator subjects* were shown portions of the recordings collected from the original subjects which we will refer to as *video subjects*. Using the same scale as the video subjects, the estimator subjects estimated the interruptibility of the video subjects. The estimator subjects distinguished "Highly Non-interruptible" situations from other situations with an accuracy of 76.9%.

Fig. 3. The interface used by estimator subjects for human estimation.

4.1 Methodology

Using a website that advertises experiments conducted at our university, we recruited 40 estimator subjects, each of whom was paid for a session that was scheduled for one hour. A majority of our estimator subjects were students at our university or at another university within walking distance. To protect the video subjects, the estimator subjects were shown still images of the video subjects and asked if they recognized any of the video subjects. They were only shown recordings of video subjects they did not recognize.

Each session started with an explanation of the task. Estimator subjects were told to evaluate the recordings as if they were walking into that situation and needed to decide how interruptible the video subject was prior to deciding whether to interrupt the video subject. A practice portion was started, and the experimenter introduced the estimator subject to the interface in Figure 3. The interface presented five initially unchecked radio buttons for each estimate. Estimator subjects were told that they could watch the video more than once, and they were advised that they should be as accurate as possible without worrying about speed. The estimator subject then used the interface to estimate the interruptibility of a video subject for 6 randomly selected practice self-reports. This was followed by the main portion in which the estimator subject estimated the interruptibility of video subjects for 60 self-reports. The main portion self-reports were selected randomly without replacement between estimator subjects, ensuring that every self-report would be used once before any self-report was used twice. After the main portion was completed, estimator subjects provided information about their general strategies during the main portion and their specific strategies for making estimates from particular recordings. We will not further discuss their strategies, but informally note that

Table III. Confusion Matrix for Human Estimates of Interruptibility

Video Subject Value	Estimator Subject Value: Highly Interruptible 1	2	3	4	5: Highly Non-Interruptible
1	172 7.2%	92 3.8%	41 1.7%	31 1.3%	10 0.4%
2	94 3.9%	110 4.6%	72 3.0%	36 1.5%	5 0.2%
3	150 6.3%	204 8.5%	133 5.5%	79 3.3%	45 1.9%
4	82 3.4%	110 4.6%	116 4.8%	73 3.0%	39 1.6%
5	89 3.7%	121 5.0%	101 4.2%	145 6.0%	250 10.4%
Overall Accuracy: 30.7% Accuracy Within 1: 65.8%					

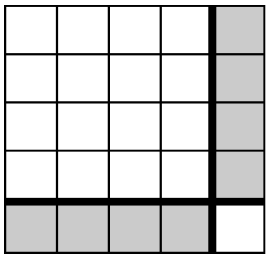
subjects reported strategies consistent with our intuition and the available literature indicating that social and task engagement are important [Seshadri and Shapira 2001]. We finally collected answers to two seven-point Likert scales discussed later in this section. The sessions were not timed, but none lasted longer than the scheduled hour.

During both the practice and main portions, the interface alternated between showing 15 or 30 seconds of the recordings from immediately before a self-report. Half of the estimator subjects started with 15 seconds, and half started with 30 seconds. We chose to use 15 seconds of the recordings because people naturally make these estimates very quickly. A person glancing in an open office door can usually decide whether it is appropriate to interrupt. We felt that showing too much of the recordings for each estimate might affect how the estimator subjects made their decisions. While it would normally be considered inappropriate to look in an open office door for 15 seconds, we felt that the additional temporal information presented in 15 seconds should help to correct for differences between normal circumstances and our recordings. The 30-second condition was included to determine whether additional time improved accuracy. As we will discuss later in this section, our estimator subjects felt 15 seconds was sufficient and their performance did not improve with the longer recordings.

Of the original 672 interruptibility self-reports, recordings for 587 self-reports were used with the estimator subjects. The others were not used because they were potentially sensitive or because a technological artifact, such as a gap in the video shortly before a prompt, might have been distracting to the estimator subject. As 40 subjects provided estimates for 60 self-reports selected randomly without replacement, each of the 587 self-reports had four or five estimates generated for it, including at least two based on 15 seconds of the recordings and at least two based on 30 seconds.

4.2 Experiment Results

Table III presents the human estimates in the form of a confusion matrix. Rows correspond to the values reported by the video subjects, and columns correspond



Video Subject	Estimator Subject: Other Values	Estimator Subject: Highly Non
Other Values	1595 66.5%	99 4.1%
Highly Non	456 19.0%	250 10.4%
	Accuracy: 76.9%	

Fig. 4. Transforming the 5-choice problem into 2-choice problem.

to the values from the estimator subjects. The unshaded diagonal represents instances when the estimator subject correctly estimated the same value given by the video subject. Summing the diagonal, we can see that estimator subjects were correct for 738 instances, or approximately 30.7% of the data. Because "Highly Non-interruptible" is the most common value, always estimating that value establishes a baseline chance accuracy of 706 correct, or 29.4%. Our estimator subjects performed only slightly better than chance, a difference which is not significant ($\chi^2(1, 4800) = 1.01$, $p > .31$). This indicates that interruptibility estimation, as posed, is difficult.

We note that the mistakes made by the estimator subjects appear to include a certain amount of bias, perhaps related to self-interest. If the mistakes were random, we might expect approximately the same number of entries in the upper-right half of the confusion matrix as in the lower-left half. This would mean estimator subjects were equally likely to confuse video subjects for being more interruptible as they were to confuse video subjects for being less interruptible. Instead, there are 450 entries in the upper-right half, approximately 18.7% of the data, and 1212 entries in the lower-left half, approximately 50.5% of the data. Aggregating for each estimator subject, estimator subjects reported significantly lower values than the video subjects ($t(39) = -8.79$, $p < .001$). This may imply a systematic bias towards viewing another person as interruptible when we are interested in making an interruption.

Figure 4 illustrates a transformation that reduces the problem to distinguishing between "Highly Non-interruptible" responses and other responses. Because this reduced form will be used throughout this article, it is worth clarifying that the bottom-right cell represents instances when both the video subject and the estimator subject responded with "Highly Non-interruptible." The upper-left cell represents instances in which both the video subject and the estimator subject responded with any other value. The other two cells represent instances when either the video subject or the estimator subject responded with "Highly Non-interruptible," but the other did not. For this problem, the estimator subjects have an overall accuracy of 76.9%, significantly better than a chance performance of 70.6% ($\chi^2(1, 4800) = 24.5$, $p < .001$).

While an accuracy of 76.9% may seem low for a task very similar to everyday tasks, we find this level of accuracy believable because of the context in which people normally make interruptibility estimates. People do not typically make an initial estimate and then blindly proceed. Instead, the evaluation of

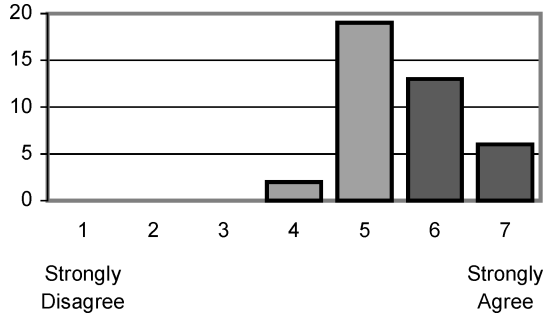


Fig. 5. "I am confident in the accuracy of my judgments."

interruptibility is an early step in a negotiated process [Goffmann 1982]. An initial determination that a person is not interruptible allows an early exit from negotiation, but other cues allow a person to decide against interrupting despite an initial evaluation that they could. Other cues can include eye contact avoidance and the continuation of the task that would be interrupted. In designing systems to use interruptibility estimates, it will be important to support a negotiated entry, rather than assuming that interruptibility estimates provides absolute guidance.

4.3 Estimator Subject Confidence

The validity of our human estimation results is strengthened by confidence data collected from the estimator subjects. The first Likert scale in the experiment stated "I am confident in the accuracy of my judgments." Each estimator subject responded on a seven-point scale ranging from "Strongly Disagree," which we will refer to as 1, to "Strongly Agree," which we will refer to as 7. Given the results for this scale, as shown in Figure 5, it is clear that our estimator subjects were confident in the accuracy of their estimates. We believe these confidence levels indicate the recordings provided enough information for estimator subjects to make estimates with which they were comfortable.

Interestingly, the subjects who were most confident in their estimates did not perform better. In the 5-choice problem, subjects responding with a 6 or 7 actually did slightly worse than subjects responding with a 4 or 5, though this difference is not significant ($\chi^2(1, 2400) = 1.94$, $p > .15$). They also performed slightly worse in the 2-choice problem, but this difference was also not significant ($\chi^2(1, 2400) = 0.83$, $p > .36$).

4.4 Recording Duration

As discussed in introducing this experiment, we felt 15 seconds of the recordings would be sufficient for estimating interruptibility, and we included cases with 30 seconds to determine whether the additional time was helpful. This section presents evidence supporting our initial belief that 15 seconds of the recordings was sufficient.

The second Likert scale in the experiment stated "The 15 second videos were long enough for making judgments." Figure 6 shows the estimator subject

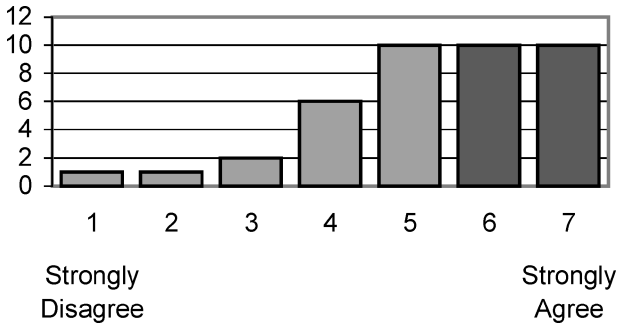


Fig. 6. "The 15-second videos were long enough for making judgments."

responses which indicate that the estimator subjects generally found 15 seconds to be sufficient. Subjects who indicated a confidence level of 6 or 7 using 15 seconds of the recordings did slightly worse in both the 5-choice problem and the 2-choice problem than subjects who indicated a lower confidence level, though these differences are not significant ($\chi^2(1, 1200) = 2.59$, $p > .10$, $\chi^2(1, 1200) = 0.07$, $p > .78$). These results show that estimator subjects generally felt 15 seconds of the recordings was sufficient and that the estimator subjects who desired more information did not do any worse than estimator subjects who were comfortable with the amount of information available.

Further evidence that 15 seconds of the recordings was sufficient is seen in the lack of an improvement when 30 seconds were available. In the 5-choice problem, the overall accuracy of estimates based on 30 seconds of the recordings is slightly worse than that of estimates based on 15 seconds, but this difference is not significant ($\chi^2(1, 2400) = 1.76$, $p > .18$). In the 2-choice problem, estimates based on 30 seconds of the recordings were better than estimates based on 15 seconds, but not significantly better ($\chi^2(1, 2400) = 0.06$, $p > .80$). These results indicate that the extra information available in 30 seconds of the recordings did not improve accuracy, which is consistent with the human ability to make these decisions very quickly in everyday environments.

4.5 Discussion

This section has presented an experiment to explore human estimation of interruptibility. The experiment showed that human estimators performed only slightly better than chance when asked to estimate interruptibility on a 5-point scale from "Highly Interruptible" to "Highly Non-interruptible". These human estimators appear to have systematically interpreted the video subjects as being more interruptible than the video subjects reported. By reducing the problem to distinguishing between "Highly Non-interruptible" conditions and other conditions, we establish a human estimator accuracy of 76.9%.

Taken as a whole, these results seem to indicate that automatic estimates of human interruptibility can be based on short periods of time immediately preceding a potential interruption. Because human estimators had difficulty accurately estimating the interruptibility of a video subject on a 5-point scale, it seems that it might be reasonable for automatic estimators to focus on

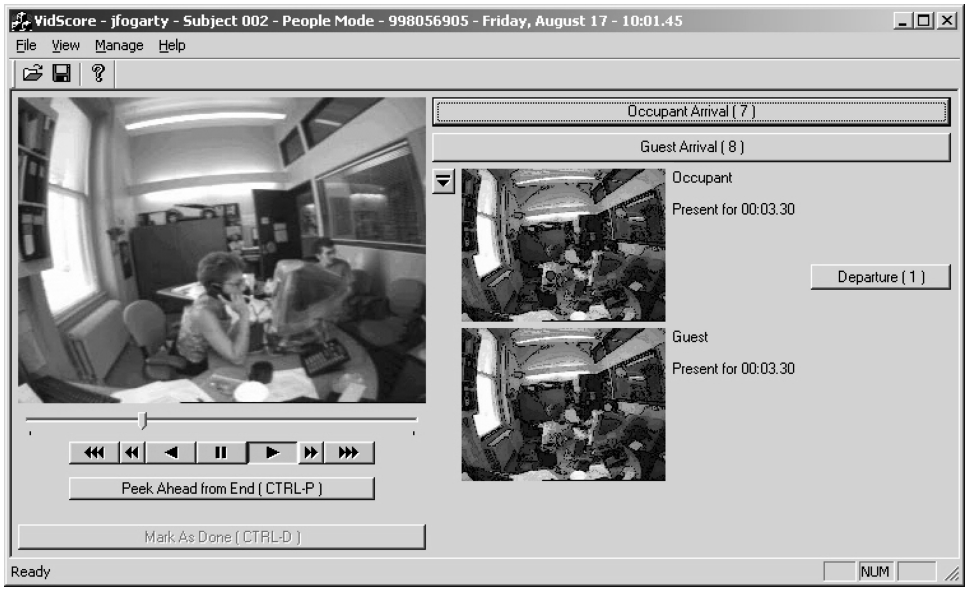


Fig. 7. Custom interface used for Wizard of Oz sensor simulation.

recognizing "Highly Non-interruptible" conditions. Automatic estimators could identify extremely inappropriate times for interruptions and allow a system to avoid them while using negotiated approaches during other times. This strategy appears to work well in human interaction [Goffmann 1982] and also seems worth pursuing as an approach to human computer interaction.

5. MODELS BASED ON WIZARD OF OZ SIMULATED SENSORS

While people regularly estimate interruptibility during everyday tasks, we are interested in whether models based on practical sensors can automatically provide these estimates. This section presents sensors simulated using a Wizard of Oz technique [Dahlbäck et al. 1993; Maulsby et al. 1993]. As discussed in our introduction, the decision to use simulated sensors allows us to consider a variety of sensors without requiring that we build them first. We can thus limit the time and resources spent on sensors that are ill-suited or suboptimal for predicting interruptibility. After discussing our simulated sensors, this section presents and analyzes models based on these simulated sensors. This section partially duplicates preliminary results discussed in a previous paper [Hudson et al. 2003], but significantly adds to the sensors, models, and analyses presented in that paper.

5.1 Manual Sensor Simulation

The sensors discussed in this section were manually simulated using a custom interface shown in Figure 7. The interface presents recordings in 15-second segments. A coder could playback the recordings at normal speed or double speed, at their option. At the end of each segment, a coder could go to the next segment

Table IV. Wizard of Oz Simulated Sensors for Each 15-Second Segment

Occupant Related	• Occupant presence. • Speaking, writing, sitting, standing, or on the phone. • Touch of, or interaction with: desk (primary work surface), table (large flat surface other than the primary work surface), file cabinet, food, drink, keyboard, mouse, monitor (gaze at), and papers (including books, newspapers, and loose paper).
Guest Related	• Number of guests present. • For each guest: sitting, standing, talking, or touching (any physical contact or very close physical proximity with occupant, including handing occupant an object).
Environment	• Time of day (hour only). • Door open, closed.
Aggregate	• Anybody talk (combines occupant and guest talk values).

or watch the current segment again. This interface, and the set of sensors it is used to simulate, was developed after an initial exploratory coding of data from our first subject. Data from all four subjects was coded after the procedures were finalized. Coders began their work training for consistency. We evaluated agreement among coders by recoding a randomly selected 5% of the recordings and found 93.4% agreement at a granularity of 15 second intervals. In order to minimize coding time, and because we believe information in close temporal proximity will be most useful in predicting interruptibility, we have only coded the 5 minutes preceding each self-report, for a total of 56 hours of coded recordings.

Using a total of four passes, our coding of the recordings identified the 24 events or situations included in Table IV. This set of manually simulated sensors was chosen because we had an *a priori* belief that they might relate to interruptibility, because we believed that a sensor could plausibly be built to detect them, and because they could be observed in our recordings. While we believe that information like what applications are running on a computer could be useful, we could not directly observe such information in our recordings. Some sensors would be easier to build than others, and we have included sensors that would be difficult to build because knowing they are useful might justify the effort necessary to develop them.

Using these simulated sensor values, we computed a number of *derivative sensors* to capture recency, density, and change effects. These are shown in Table V, and were computed for time intervals of 30 seconds, 1 minute, 2 minutes, and 5 minutes. We will use the names in the left column to refer to derivatives of sensors, and so "Occupant Talk (Any-300)" refers to the Any derivative of the Occupant Talk sensor over a 5 minute interval.

5.2 Predictiveness of Individual Features

Based on the literature and our own intuitions, we expect that the strongest indicators of non-interruptibility would be related to task engagement and social engagement [Seshadri and Shapira 2001]. We informally note that it is almost always considered rude to interrupt a person who is talking. It is also

Table V. Derivations Applied to Manually Computed Sensors

Imm	Whether the event occurred in the 15 second interval containing the self-report sample.
All-N	Whether event occurred in *every* 15 second interval during N seconds prior to the sample.
Any-N	Whether event occurred in *any* 15 second interval during N seconds prior to the sample.
Count-N	The number of times the event occurred during intervals in N seconds prior to the sample.
Change-N	The number of consecutive intervals for which the event occurred in one and did not occur in the other during N seconds prior to the sample.
Net-N	The difference in the sensor between the first interval in N seconds prior to the sample and the sensor in the interval containing the sample.

Table VI. Information Gain Ordering of the 30 Most Predictive Individual Features

1	Any Talk (Count-30)	11	Telephone (Count-30)	21	Telephone (All-60)
2	Any Talk (Imm)	12	Occupant Talk (Count-120)	22	Telephone (Count-120)
3	Occupant Talk (Imm)	13	Occupant Talk (Any-60)	23	Telephone (Count-300)
4	Occupant Talk (Count-30)	14	Occupant Talk (Change-60)	24	Any Talk (Count-300)
5	Any Talk (Count-60)	16	Telephone (Imm)	25	Occupant Talk (Count-300)
6	Any Talk (Any-30)	15	Any Talk (Any-60)	26	Any Talk (All-60)
7	Occupant Talk (Any-30)	17	Telephone (All-30)	27	Telephone (Change-60)
8	Occupant Talk (Change-30)	18	Telephone (Count-60)	28	Telephone (Any-30)
9	Occupant Talk (Count-60)	19	Any Talk (All-30)	29	Telephone (Change-30)
10	Any Talk (Count-120)	20	Occupant Talk (All-30)	30	Occupant Talk (Change-120)

particularly inappropriate to interrupt a person who is speaking on a telephone, perhaps because the remote party cannot participate in the subtle nonverbal negotiation of the interruption.

While we felt that these types of activities would need to be detected to produce good estimates of interruptibility, it was not clear exactly which sensors would be the most helpful. It was also not clear which easily-built sensors might work almost as well as sensors that would be very difficult to build. To gain some insight into these issues, we examined the predictive power of individual features using an *information gain* metric [Mitchell 1997].

Described simply, information gain is based on sorting a set of observations according to the value of a feature associated with each observation. The sorting removes the entropy associated with variations in that feature. This reduction in entropy provides an estimate of the predictiveness of that feature. The absolute value of this difference is not particularly interesting, only the relative values for the features. Further, information gain only indicates potential usefulness in prediction and cannot, by itself, indicate whether a feature indicates interruptibility or non-interruptibility. Finally, the notion of predictiveness measured by information gain includes sensitivity to frequency, and so an event that always indicates interruptibility, but almost never occurs, would not be highly ranked.

Table VI presents an ordered list of the 30 most predictive individual features, as indicated by information gain when distinguishing between "Highly Non-interruptible" self-reports and other self-reports. This number of features

Table VII. Features Selected with a Correlation-Based Feature Selection Technique

1	Telephone (Count-30)	9	Monitor (Count-300)	17	Any Talk (Net-300)
2	Any Talk (Imm)	10	Telephone (All-300)	18	Telephone (All-30)
3	Any Talk (Count-60)	11	Guests Sit (Net-60)	19	Mouse (Count-120)
4	Telephone (Imm)	12	Telephone (Net-120)	20	Any Talk (All-120)
5	Mouse (Count-60)	13	Telephone (Count-300)	21	Food (Count-300)
6	Any Talk (Count-300)	14	Any Talk (Count-30)	22	Table (Change-30)
7	Telephone (All-60)	15	Writing (Change-30)	23	Guests Sit (All-300)
8	Occupant Talk (Imm)	16	Stand (Change-300)	24	Table (Count-300)

was selected arbitrarily and is only intended to allow an examination of the most predictive individual features. Although we had expected talking and the telephone to be important indicators, it is very interesting to note that *all* 30 of the top individual features are related to either the telephone or talking. This metric does not consider the redundancy between the features in the chart. While sensors for talking and the telephone will be important throughout this article, the models discussed in the rest of the article will also examine what additional features can complement the information gained from talking and telephone sensors. This metric shows that, if allowed to use only one sensor, a sensor related to talking or the telephone is the most useful.

5.3 Correlation-Based Feature Selection

As we begin to examine multiple features, we note that the combination of manually simulated sensors and sensor derivations yields a very large number of possible features. Using all of these features to build models could have very negative effects. In a phenomenon known as *overfitting*, a model mistakenly interprets minor details or quirks in data as representative of data it will be asked to evaluate in the future. The overall accuracy of its future estimates is then lower than it should be, because it is confused by differences in the minor details that it previously mistook for important. Overfitting is very similar to degree-of-freedom problems found in models with excessive parameters.

In order to prevent overfitting, we applied a correlation-based feature selection technique [Hall 2000] as implemented in the Weka machine learning software package [Witten and Frank 1999]. This technique uses correlations between different features and the value that will be estimated to select a set of features according to the criterion that "Good feature subsets contain features highly correlated with the (value to be estimated), yet uncorrelated with each other" [Hall 2000]. Table VII lists the 24 features selected for distinguishing between "Highly Non-interruptible" conditions and other conditions, in the order of their selection. Unlike Table VI, the number of features selected here is not arbitrary. The correlation-based feature selection technique indicates the point at which it believes additional features are redundant and may lead to overfitting which, in this case, is after the (Count-300) derivative of the Table feature.

In the next section, we will create models of human interruptibility based on the features selected in this section. While we will revisit feature selection in a later section, the feature selection technique used here has some good

Table VIII. Accuracy of Models Built from the Correlation-Based Features in Table VII

Video Subject	Naïve Bayes: Other Values	Naïve Bayes: Highly Non	Decision Tree: Other Values	Decision Tree: Highly Non
Other Values	380 56.5%	77 11.5%	415 61.8%	42 6.3%
Highly Non	72 10.7%	143 21.3%	115 17.1%	100 14.9%
	Accuracy: 77.8%		Accuracy: 76.6%	

qualities. First, this technique is computationally very cheap compared to the feature selection techniques we use later. In a deployed system, the feature selection techniques used here could regularly examine a huge number of possibly interesting features and quickly select an appropriate subset. Second, this technique is independent of the models that will be created from the selected features. As such, the selected features are appropriate for use with a variety of modeling techniques.

5.4 Initial Model Construction

This section presents models constructed using several standard machine learning techniques. Specifically, we will be using decision trees [Quinlan 1993] and naïve Bayes predictors [Duda and Hart 1973; Langley and Sage 1994]. We have obtained similar results with support vector machines [Burges 1998] and AdaBoost with decision stumps [Freund and Schapire 1997], but will not discuss them here for the sake of brevity. We will also not attempt to fully describe each of these techniques. Instead, interested readers are encouraged to consult the original references or a machine learning text, such as Mitchell [1997]. All of our models were constructed using the Weka machine learning software package [Witten and Frank 1999], a widely available open source software package.

Confusion matrices for models constructed from the features in Table VII are presented in Table VIII. Remember that chance is an accuracy of 68.0%, which could be obtained by always predicting “Other Values”. The results in this section have all been obtained using a standard cross-validation approach involving multiple trials of model construction. In each of 10 trials, 90% of the data is used to train, and the remaining 10% is used for testing. Each instance is used to train 9 trials and to test 1 trial. The values reported are sums from the 10 trials.

These results show that models based on manually simulated sensors with features selected according to a correlation-based feature selection technique can estimate human interruptibility as well as our estimator subjects. Both models perform significantly better than the 68.0% chance (Naïve Bayes: $\chi^2(1, 1344) = 16.41$, $p < .001$, Decision Tree: $\chi^2(1, 1344) = 12.50$, $p < .001$), and neither is significantly different than the 76.9% performance of our estimator subjects (Naïve Bayes: $\chi^2(1, 3072) = 0.27$, $p > .60$, Decision Tree: $\chi^2(1, 3072) = 0.02$, $p > .89$). The difference between the models is also not significant ($\chi^2(1, 1344) = 0.27$, $p > .60$). Given that we used a feature selection technique that is independent of the modeling technique and reproduced the

Table IX. Results Using Wrapper-Based Feature Selection with a Naïve Bayes Classifier

1	Any Talk (Imm)
2	Drink (Any-30)
3	Desk (Change-300)
4	Telephone (Imm)
5	Time of Day (Hour Only)
6	Stand (Any-120)
7	Stand (Net-120)
8	Guests Sit (Net-30)
9	Desk (Net-60)
10	Drink (Count-300)

	Naïve Bayes	
Video Subject	Other Values	Highly Non
Other Values	411 61.2%	46 6.8%
Highly Non	80 11.9%	135 20.1%
	Accuracy: 81.25%	

results with distinct learning techniques, these results make us quite hopeful that models with accuracies in the range of 75% to 80% can be driven by sensors.

5.5 Wrapper-Based Feature Selection and Model Construction

While the correlation-based feature selection technique used earlier has several good properties, it is a heuristic and we cannot be sure the features it selects are optimal. This section presents an alternative feature selection technique that chooses features according to their usefulness in a particular model. It is based on slowly adding features to a model until additional features do not improve accuracy, and is known as a wrapper technique because it can theoretically be wrapped around any model [Kohavi and John 1997]. Because this technique requires the repeated application of a machine learning technique, it is computationally much more expensive than techniques like correlation-based feature selection. The results presented were obtained in conjunction with a feature search strategy that starts with an empty set of features and adds or removes features from the set until there is no change that results in an improvement. This approach is limited by the fact that it selects features appropriate to the particular model used during feature selection. Used with a naïve Bayes model, for example, this method will not select locally predictive features that could be useful to a decision tree model.

Table IX presents the results of applying a wrapper-based feature selection technique with a naïve Bayes classifier. The 10 features shown here were selected as good features for the naïve Bayes classifier. They yield a model with an accuracy of 81.25%, significantly better than the 68.0% chance ($\chi^2(1, 1344) = 31.13$, $p < .001$), significantly better than the estimator subjects ($\chi^2(1, 3072) = 5.82$, $p < .05$), and better than the naïve Bayes classifier built with the correlation-based feature selection, though this difference is not significant ($\chi^2(1, 1344) = 2.42$, $p > .11$). Table X presents similar results obtained with a decision tree classifier. Coincidentally, 10 features are also selected in this case, though they are different than the features selected for use with the naïve Bayes classifier. The selected features yield a decision tree classifier with an accuracy of 82.4%, significantly better than chance ($\chi^2(1, 1344) = 37.56$, $p < .001$), significantly better than the estimator subjects ($\chi^2(1, 3072) = 9.51$, $p < .01$), and significantly better than the decision tree classifier built with the correlation-based feature selection ($\chi^2(1, 3072) = 9.51$, $p < .01$). The difference

Table X. Results Using Wrapper-Based Feature Selection with a Decision Tree

1	Any Talk (Imm)
2	Telephone (Count-30)
3	Time of Day (Hour Only)
4	Desk (Change-120)
5	Monitor (Any-300)
6	Occupant Talk (Net-120)
7	Writing (Count-30)
8	Writing (Count-60)
9	Papers (Count-300)
10	Mouse (All-120)

	Decision Tree	
Video Subject	Other Values	Highly Non
Other Values	413 61.5%	44 6.5%
Highly Non	74 11.0%	141 21.0%
	Accuracy: 82.4%	

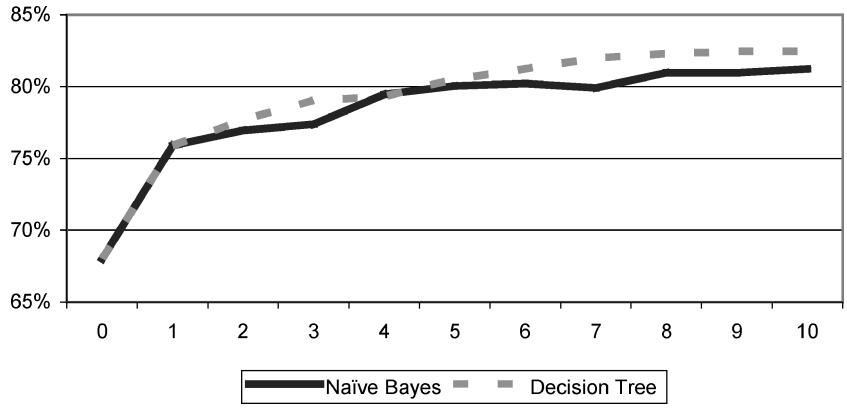


Fig. 8. Classifier accuracy versus number of features.

between the decision tree model and the naïve Bayes model built here is not significant ($\chi^2(1, 1344) = 0.32$, $p > .57$).

The models presented in this section both distinguish "Highly Non-interruptible" situations from other situations significantly better than the 76.9% accuracy of our estimator subjects. The tradeoff for obtaining these better results is that we have expended many more computational resources during model creation and we have selected features that may be appropriate only with the modeling techniques we used when selecting them. These results, taken with the results in the previous section, support the view that it should be possible to create robust models of human interruptibility. Because the estimates given by our models match and even surpass the accuracy of estimates given by our estimator subjects, it should be possible to design systems that effectively use these estimates as part of a negotiated interruption process.

5.6 Model Accuracy and Number of Features

Given that wrapper-based feature selection chose only 10 features from a possible set of almost 500 features, it is interesting to examine how the accuracy of the models is improved by each additional feature. Figure 8 plots the accuracy of the two wrapper-based models presented in the previous section as a function of the number of features. Both models start at a baseline accuracy of 68% for no features. They then have a very sharp improvement in accuracy when the

Table XI. Results with "Easy to Build" Features and a Naïve Bayes Model

1	Any Talk (Imm)
2	Any Talk (Count-30)
3	Telephone (All-30)
4	Mouse (Imm)
5	Mouse (Change-30)
6	Telephone (All-120)
7	Mouse (All-120)
8	Mouse (Net-60)

	Naïve Bayes	
Video Subject	Other Values	Highly Non
Other Values	396 58.9%	61 9.1%
Highly Non	81 12.1%	134 19.9%
	Accuracy: 78.9%	

first feature is added. In both cases, this is the Any Talk (Imm) feature. The next handful of features yields a small, but noticeable, improvement. After this, very little improvement is associated with each feature added, and the feature selection terminates after 10 because no additional feature improves accuracy.

This relationship between the features and the accuracy of the models has important implications. Our data indicates that a single sensor to detect whether anybody in the office is currently speaking can by itself yield an accuracy of 75.9%. While this is worse than the performance of our estimator subjects, the difference is not significant ($\chi^2(1, 3072) = 0.28$, $p > .59$). This might seem too simple to be reasonable, but we point out that speaking correlates with many other activities that one might wish to recognize when estimating interruptibility. For example, people normally speak when on the telephone. It is also generally expected that people speak to a guest who is currently in their office. This result suggests that it may not be necessary to use expensive sensor networks or vision-based systems to estimate interruptibility, but that we might instead build much less expensive systems that perform nearly as well as more expensive alternatives.

5.7 An "Easy to Build" Feature Set

Given the results of the previous section, we now consider models using only sensors that are readily available or could be easily constructed. In fact, we originally created the Any Talk simulated sensor because it would be easier to build than a sensor that differentiated between the occupant of an office talking and guests talking. This proposed sensor could be combined with simple software that detects mouse and keyboard activity. Inexpensive hardware placed between the telephone and the wall can sense whether the phone is currently off the hook. Finally, the time of day is readily available. Throughout this section, we will refer to this set of 5 sensors from our manually simulated data as "Easy to Build" features.

Table XI and Table XII present the features and models resulting from a wrapper-based feature selection with the "Easy to Build" features. The naïve Bayes result of 78.9% overall accuracy is better than the 76.9% accuracy of our estimator subjects, though not significantly ($\chi^2(1, 3072) = 1.19$, $p > .27$), and worse than the 81.25% accuracy of the model in Table IX that was built from the full set of sensors, but not significantly ($\chi^2(1, 1344) = 1.19$, $p > .27$). The decision tree model accuracy of 79.2% is also better than our estimator subject accuracy, but the difference is not significant ($\chi^2(1, 3072) = 1.58$, $p > .20$). It is

Table XII. Results with "Easy to Build" Features and a Decision Tree Model

1	Any Talk (Imm)
2	Telephone (Count-30)
3	Time of Day (Hour Only)
4	Any Talk (Net-120)

	Decision Tree	
Video Subject	Other Values	Highly Non
Other Values	396 58.9%	61 9.1%
Highly Non	79 11.8%	136 20.2%
	Accuracy: 79.2%	

Table XIII. Results of Wrapper-Based Feature Selection with a Naïve Bayes Classifier

1	Occupant Talk (Any-30)
2	Sit (All-30)
3	Table (Count-60)
4	Food (Any-120)
5	File Cabinet (Any-300)
6	Door Close (Any-300)

		Naïve Bayes				
		Highly Interruptible				Highly Non-Interruptible
		1	2	3	4	5
Video Subject Value	1	33 4.9%	6 0.9%	27 4.0%	0 0.0%	26 3.9%
	2	12 1.8%	12 1.8%	39 5.8%	4 0.6%	19 2.8%
	3	11 1.6%	2 0.3%	106 15.8	7 1.0%	37 5.5%
	4	7 1.0%	1 0.1%	55 8.2%	23 3.4%	30 4.5%
	5	11 1.6%	5 0.7%	42 6.3%	11 1.6%	146 21.7%
		Overall Accuracy: 47.6% Accuracy Within 1: 71.7%				

worse than the 82.4% accuracy of the model in Table X that was built from the full set of sensors, but not significantly ($\chi^2(1, 1344) = 2.32$, $p > .12$).

These results for the "Easy to Build" sensors are very promising because they indicate that models of human interruptibility can be based on technology that is already available or easily built. This implies that we do not need to solve hard computer vision problems or hard artificial intelligence problems before proceeding with creating systems that use models of human interruptibility.

5.8 Models of the 5-Choice Problem

Up until this point, we have focused on models to distinguish "Highly Non-interruptible" situations from other situations. This section presents models of the full 5-point scale and discusses how these models can support a level of flexibility that is not available with models of the 2-choice problem. It is important to note that the techniques used here do not have any notion that our five possible values represent a scale. As far as the techniques are concerned, the five values are completely unrelated. While there are techniques that support values in a scale, informal experimentation with some of these techniques did not yield an improvement over the results presented here.

Table XIII presents the results of wrapper-based feature selection for the 5-choice problem with a naïve Bayes classifier. The 47.6% overall accuracy of this model is significantly better than our estimator subjects 30.7% performance

Table XIV. Results of Wrapper-Based Feature Selection with a Decision Tree

1	Occupant Talk (Any-30)
2	Sit (All-30)
3	Desk (Any-120)
4	Telephone (Imm)
5	Occupant Talk (Net-30)
6	Food (Any-300)
7	Drink (Any-300)
8	Stand (All-120)
9	Any Talk (Any-30)
10	Present (Count-120)
11	Papers (Net-30)
12	Present (All-120)
13	Door Open (All-60)

		Decision Tree				
		Highly Interruptible				Highly Non-Interruptible
		1	2	3	4	5
Video Subject Value	1	40 6.0%	4 0.6%	25 3.7%	2 0.3%	21 3.1%
	2	18 2.7%	13 1.9%	39 5.8%	1 0.1%	15 2.2%
	3	9 1.3%	4 0.6%	113 16.8%	8 1.2%	29 4.3%
	4	9 1.3%	2 0.3%	53 7.9%	28 4.2%	24 3.6%
	5	12 1.8%	5 0.7%	37 5.5%	9 1.3%	152 22.6%
Overall Accuracy: 51.5% Accuracy Within 1: 75.1%						

($\chi^2(1, 3072) = 66.17$, $p < .001$). Table XIV presents the results from a decision tree model. Its 51.5% overall accuracy is significantly better than the estimator subjects ($\chi^2(1, 3072) = 98.88$, $p < .001$) and better than the naïve Bayes model, though this difference is not significant ($\chi^2(1, 1344) = 2.01$, $p > .15$).

Models of the 5-choice problem allow systems to provide an additional level of flexibility. People who feel they are being interrupted too often could use the system's interface to request that they be interrupted less frequently. Instead of initiating a negotiated interruption for a value of 4 or lower, the system could then only negotiate interruptions when its model estimates a value of 3 or lower. Alternatively, systems could use the value of the estimate to decide how subtly to initiate an interruption. Estimates of 3 or 4 could be used by a system to decide when to initiate a negotiated interruption with an ambient information display [Fogarty et al. 2001; Heiner et al. 1999; Redström et al. 2000], while estimates of 1 or 2 could be used by the system to decide when to initiate with a more direct method.

5.9 Discussion

This section has presented a variety of statistical models of human interruptibility. We first demonstrated that models based on manually simulated sensors can differentiate "Highly Non-interruptible" situations from other situations with an accuracy as high as 82.4%, significantly better than the 76.9% performance of our human estimator subjects. This initial result is made more interesting by the observation that the Any Talk simulated sensor alone can provide an accuracy of 75.9% and that a set of sensors we consider easy to build can provide an accuracy as high as 79.2%. This set of sensors does not require any vision-based techniques and could be built and used for a very low cost.

If used in conjunction with models of the importance of different interruptions and systems designed to allow negotiated entry into an interruption, the models presented in this section could support significant advances in human computer interaction and computer mediated communication. While this work has not attempted to solve the hard artificial intelligence problems related to

truly understanding human behavior, we have quantitatively demonstrated that simple sensors can effectively estimate human interruptibility. By using passive sensors instead of requiring that people create and maintain calendars or otherwise explicitly indicate their interruptibility, our approach helps to make interruptibility estimation practical for use in everyday systems.

6. AUTOMATED ANALYSIS OF THE RECORDINGS

While we did not initially intend to automatically analyze our recordings, the results of our manually simulated sensor analysis made the possibility interesting. Specifically, the significance of the Any Talk simulated sensor makes it worth examining whether the audio we collected from a single microphone placed in the corner of an office allows us to approximate the Any Talk simulated sensor sufficiently well to support models of human interruptibility.

Because we recorded audio with a microphone placed beside the computer used for recording, our recordings include a significant amount of fan noise from the recording computer. There are many situations where the combined audio and video recordings make it clear that a person is talking and the manually simulated Any Talk sensor has a value of true, but only a faint murmur is actually audible over the fan noise in the audio. It is much more difficult to identify these instances without video, and we would expect automated techniques to encounter difficulties.

6.1 Silence Detection

As an approximation of the Any Talk manually simulated sensor, we decided to use the silence segmentation functionality of the Sphinx speech recognition package [CMU Sphinx]. For each recording configuration, the silence segmentation software was calibrated with a short bit of "silent" audio. For these calibrations, we used recordings from early in the morning before the subject arrived. These recordings contained fan noise created by our recording machine, but did not contain any other activity. After calibrating, we used the silence segmentation with 4 different threshold configurations, designed at one extreme to identify only the loudest activity, and at the other extreme to identify activity even slightly above the silence calibration. For each threshold, we built a set of features representing how much of a time interval was not silent.

To determine if these features could reasonably approximate our Any Talk simulated sensor, we used the features from the 15 seconds before each interruption to attempt to predict the value of the Any Talk (Imm) simulated sensor value. This is intended only as a rough estimate of the usefulness of these features as there are some problems related to using the 15 seconds before the interruption versus the 15 seconds that were the basis for the manually simulated sensor value. Given this qualification, we built a naïve Bayes model that predicted our Any Talk (Imm) simulated sensor with an accuracy of 79.2% and a decision tree with an accuracy of 80.1%, both significantly better than the 70.4% chance accuracy that could be obtained by always predicting "Not Talking" (Naïve Bayes: $\chi^2(1, 1344) = 13.73$, $p < .001$, Decision Tree:

Table XV. Results Using a Naïve Bayes Model with Silence Detector Features

1	Telephone (Imm)
2	Silence Detector (Medium Thres-10 Sec Interval)
3	Telephone (All-30)
4	Keyboard (Change-60)

	Naïve Bayes	
Video Subject	Other Values	Highly Non
Other Values	429 63.8%	28 4.2%
Highly Non	131 19.5%	84 12.5%
	Accuracy: 76.3%	

Table XVI. Results Using a Decision Tree with Silence Detector Features

1	Telephone (Count-30)
2	Silence Detector (High Thres-300 Sec Interval)
3	Silence Detector (Medium Thres-5 Sec Interval)
4	Keyboard (Any-30)
5	Telephone (Count-300)
6	Keyboard (Change-120)
7	Telephone (Any-300)
8	Keyboard (Change-300)
9	Keyboard (Net-60)
10	Silence Detector (Highest Thres-300 Sec Interval)
11	Silence Detector (Low Thres-300 Sec Interval)
12	Mouse (All-120)
13	Telephone (Change-300)

	Decision Tree	
Video Subject	Other Values	Highly Non
Other Values	435 64.7%	22 3.3%
Highly Non	133 19.8%	82 12.2%
	Accuracy: 76.9%	

$\chi^2(1, 1344) = 16.87$, $p < .001$). This indicates that our silence detection features have predictive value despite difficulties with the fan noise.

6.2 Hybrid Models

To further evaluate our implementation of the Any Talk sensor, we combined it with time of day and our manually simulated sensors for the telephone, keyboard, and mouse. As discussed in our "Easy to Build" section of the manually simulated sensor discussion, these sensors are already available or very easily built. They can also be expected to produce very reliable results.

Table XV shows a naïve Bayes model built using wrapper-based feature selection. Its overall accuracy of 76.3% is not significantly different from the 76.9% accuracy of our human estimator subjects ($\chi^2(1, 3072) = 0.08$, $p > .77$). The decision tree model shown in Table XVI has an overall accuracy of 76.9%, which is equivalent to our human estimator subjects ($\chi^2(1, 3072) = 0.001$, $p > .97$). The difference between these two models is not significant ($\chi^2(1, 1344) = 0.07$, $p > .79$).

This shows that a single microphone in the corner of an office, when combined with the time of day, a sensor for whether the phone is in use, and activity information for the mouse and keyboard, can provide enough information to estimate human interruptibility as well as our human estimators. The result does not require expensive infrastructure, and so it seems very practical for use in everyday systems. The result also shows that the implementation of an Any

Talk sensor does not need to be perfect, as our silence detector features only predict our Any Talk sensor with an accuracy of 80%, but are still useful for interruptibility estimation.

7. DISCUSSION AND FUTURE WORK

Given the results in this article, there is room for substantial work to validate and build upon our results with larger groups of people in a wider range of environments. There are also a variety of issues to consider in other environments such as the additional noise of open-plan offices. Mobile workers pose a different set of challenges. One issue of particular interest is development of an appropriate Any Talk sensor. The silence detector used here adapts to background noise well enough to work in the office environments of our video subjects, but it is not clear whether it is sophisticated enough to identify talking in noisier environments. A substantial body of research on segmenting and classifying audio [Lu et al. 2002] can be applied to this problem.

The estimator subjects in our study were not personally familiar with the video subjects, and it is possible they might have performed better if they were. However, many of the cues that people might use such as learned patterns of availability can be modeled [Begole et al. 2002, 2003]. There is room to improve our models by examining the strategies people use to estimate the interruptibility of colleagues. We are also interested in the bias our estimator subjects had in estimating that video subjects were more interruptible than the video subjects reported. Additional studies might examine whether this bias would be removed or reversed if they were told to act as an assistant regulating access.

In more recent work, we have used the results of this work to support the deployment of real sensors into the offices of ten office workers [Fogarty et al. 2004a]. We logged the output of these sensors and collected interruptibility self-reports. Analyses of the collected data support the results presented in this article, demonstrate models for a wider variety of office workers than was studied in this article, examine some questions regarding the amount of training data required for these models, and explore the potential of different combinations of sensors. Recent work by Horvitz and Apacible [2003] examined models of interruptibility based on calendar information, computer activity, and real-time analyses of audio and video streams. They collected a total of 15 hours of audio and video recordings from three office workers. The office workers then viewed the recordings and annotated them with a description of their interruptibility. This work is complimentary to ours, but the differences in our data and the data collected by Horvitz and Apacible make it inappropriate to directly compare model performance.

We intend to build systems that use the types of models presented in this article. Functional systems will allow us to continue to evaluate and improve upon these models, including examining models that learn the individual nuances of people over time. Building systems will also allow us to explore many issues related to application use of these models. These issues include balancing the importance of a piece of information with the cost of the interruption required to deliver it. We are also interested in estimates of human interruptibility as

one part of a multi-stage negotiation of an interruption. There are also a variety of issues to consider relating to, use of models in awareness and communication applications, some of which we have recently examined by building a communication client that shares automatically sensed information about a person's context and interruptibility [Fogarty et al. 2004b].

We have presented studies that quantitatively demonstrate that models created from simple sensors can estimate human interruptibility as well as our human estimator subjects could from the recordings. Because anyone talking in a room is the most predictive feature we examined, our models do not require complex sensors such as vision-based techniques, and can instead be built from a single microphone in an office and very simple sensors for telephone, mouse, and keyboard activity. By using a passive approach, instead of requiring people to explicitly indicate interruptibility, or create and maintain calendars, our approach makes interruptibility, estimation feasible for use in everyday systems. Used with models of the importance of potential interruptions and system designs that support negotiated interruptions, our models offer to support significant advances in human computer interaction.

ACKNOWLEDGMENTS

We would like to thank everyone who has contributed to Weka and Sphinx. We would like to acknowledge all the members of our Situationally Appropriate Computing research group. We thank Darren Gergle and Ryan Baker for cheerfully answering our statistical questions, though any mistakes should not be blamed on them. We would like to acknowledge our video coders: Ben Davies, Rick Ebert, Rucha Humnabadkar, Becky Kaplan, Matt Mowczko, and Long Pan.

REFERENCES

BARKER, R. G. 1968. *Ecological Psychology*. Stanford University Press.

BEGOLE, J. B., TANG, J. C., AND HILL, R. 2003. Rhythm modeling, visualizations, and applications. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2003)*. 11–20.

BEGOLE, J. B., TANG, J. C., SMITH, R. B., AND YANKELOVICH, N. 2002. Work rhythms: Analyzing visualizations of awareness histories of distributed groups. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2002)*. 334–343.

BELLOTTI, V. AND EDWARDS, K. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Hum.-Comput. Interact. 16*, 2-4, 193–212.

BURGES, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Disc. 2*, 2, 121–167.

CMU Sphinx: Open Source Speech Recognition. http://www.speech.cs.cmu.edu/sphinx/.

CUTRELL, E., CZERWINSKI, M., AND HORVITZ, E. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT 2001)*. 263–269.

CZERWINSKI, M., CUTRELL, E., AND HORVITZ, E. 2000a. Instant messaging and interruptions: Influence of task type on performance. In *Proceedings of the Australian Conference on Computer-Human Interaction (OZCHI 2000)*. 356–361.

CZERWINSKI, M., CUTRELL, E., AND HORVITZ, E. 2000b. Instant messaging: Effects of relevance and time. In *Proceedings of the British HCI Group Annual Conference (HCI 2000)*. 71–76.

DAHLBÄCK, N., JÖNSSON, A., AND AHRENBERG, L. 1993. Wizard of Oz studies—Why and how. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI 1993)*. 193–200.

DUDA, R. O. AND HART, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley and Sons.

ERICKSON, T. AND KELLOGG, W. A. 2000. Social translucence: An approach to designing systems that support social processes. *ACM Trans. Comput.-Hum. Interact. (TOCHI) 7*, 1, 59–83.

Feldman-Barrett, L. and Barrett, D. J. 2001. Computerized experience-sampling: How technology facilitates the study of conscious experience. *Soc. Sci. Comput. Rev.* 19, 175–185.

FOGARTY, J., FORLIZZI, J., AND HUDSON, S. E. 2001. Aesthetic information collages: Generating decorative displays that contain information. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2001)*. 141–150.

FOGARTY, J., HUDSON, S., AND LAI, J. 2004a. Examining the robustness of sensor-based statistical models of human interruptibility. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2004)*. 207–214.

FOGARTY, J., LAI, J., AND CHRISTENSEN, J. 2004b. Presence versus availability: The design and evaluation of a context-aware communication client. *Int. J. Hum.-Comput. Stud. (IJHCS) 61*, 3.

FREUND, Y. AND SCHAPIRE, R. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci. 55*, 1, 119–139.

GOFFMANN, E. 1982. On Facework. *Interaction Ritual*, E. Goffmann, Ed. Random House, New York, 5–45.

HALL, M. A. 2000. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the International Conference on Machine Learning (ICML 2000)*. 359–366.

HATCH, M. J. 1987. Physical barriers, task characteristics, and interaction activity in research and development firms. *Admin. Sci. Quart. 32*, 387–399.

HEINER, J. M., HUDSON, S. E., AND TANAKA, K. 1999. The information percolator: Ambient information display in a decorative object. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 1999)*. 141–148.

HORVITZ, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 1999)*.

HORVITZ, E. AND APACIBLE, J. 2003. Learning and reasoning about interruption. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI 2003)*. 20–27.

HORVITZ, E., BREESE, J., HECKERMAN, D., HOVEL, D., AND ROMMELSE, K. 1998. The lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Conference on Uncertainty and Artificial Intelligence (UAI 1998)*. 256–265.

HORVITZ, E., JACOBS, A., AND HOVEL, D. 1999. Attention-sensitive alerting. In *Proceeding of the Conference on Uncertainty and Artificial Intelligence (UAI 1999)*. 305–313.

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. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2002)*. 97–104.

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. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2003)*. 257–264.

KOHAVI, R. AND JOHN, G. H. 1997. Wrappers for feature subset selection. *Artif. Intel. 97*, 1–2, 273–324.

LANGLEY, P. AND SAGE, S. 1994. Induction of selected Bayesian classifiers. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI 1994)*. 399–406.

LU, L., ZHANG, H., AND JIANG, H. 2002. Content analysis for audio classification and segmentation. *IEEE Trans. Speech Audio Process. 10*, 7, 504–516.

MAULSBY, D., GREENBERG, S., AND MANDER, R. 1993. Prototyping an intelligent agent through wizard of Oz. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 1993)*.

MCFARLANE, D. C. 2002. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Hum.-Comput. Interact. 17*, 1, 63–139.

MCFARLANE, D. C. 1999. Coordinating the interruption of people in human-computer interaction. In *Proceedings of the IFIP Conference on Human-Computer Interaction (INTERACT 1999)*.

MILEWSKI, A. E. AND SMITH, T. M. 2000. Providing presence cues to telephone users. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2000)*. 89–96.

MITCHELL, T. M. 1997. *Machine Learning*. McGraw-Hill.

OLIVER, N., HORVITZ, E., AND GARG, A. 2002. Layered representations for recognizing office activity. In *Proceedings of the International Conference on Multimodal Interaction (ICMI 2002)*. 3–8.

PERLOW, L. A. 1999. The time famine: Toward a sociology of work time. In *Admin. Sci. Quart. 44*, 1, 57–81.

QUINLAN, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

REDSTRÖM, J., SKOG, T., AND HALLNÄS, L. 2000. Informative art: Using amplified artworks as information displays. In *Proceedings of Designing Augmented Reality Environments*.

SCHMIDT, A., TAKALUOMA, A., AND MÄNTYJÄRVI, J. 2000. Context-aware telephony over WAP. *Pers. Ubiquit. Comput. 4*, 4, 225–229.

SESHADRI, S. AND SHAPIRA, Z. 2001. Managerial allocation of time and effort: The effects of interruptions. *Manage. Sci. 47*, 5, 647–662.

VOIDA, A., NEWSTETTER, W. C., AND MYNATT, E. D. 2002. When conventions collide: The tensions of instant messaging attributed. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2002)*. 187–194.

WITTEN, I. H. AND FRANK, E. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Received January 2003; revised August 2003, February 2004; accepted February 2004 by Shumin Zhai and Victoria Bellotti