Towards an Artificial Receptionist: Anticipating a Persons Phone Behavior

Peter Vorburger, Abraham Bernstein
Department of Informatics, University of Zurich, Zurich, Switzerland
vorburger, bernstein@ifi.unizh.ch

***Abstract*— People are subjected to a multitude of interruptions, which in some situations are detrimental to their work performance. Consequently, the capability to predict a person's degree of interruptability (i.e., a measure of detrimental an interruption would be to her current work) can provide a basis for a filtering mechanism. This paper introduces a novel approach to predict a person's presence and interruptability in an office-like environment based on audio, multi-sector motion detection using video, and the time of the day collected as sensor data. Conducting an experiment in a real office environment over the length of more than 40 work days we show that the multi-sector motion detection data, which to our knowledge has been used for the first time to this end, outperforms audio data both in presence and interruptability. We, furthermore, show, that the combination of all three data-streams improves the interruptability prediction accuracy and robustness. Finally, we use these data to predict a subject's phone behavior (ignore or accept the incoming phone call) by combining interruptability and the estimated importance of call. We call such an application an artificial receptionist. Our analysis also show that the results improve when taking the temporal aspect of the context into account.**

***Index Terms*— Interruptability, motion detection, multi-sector, efficiency, phone, receptionist, context-awareness, user behavior**

I. Introduction and Motivation - Interruptability and Work Efficiency

An ordinary office day is often disrupted by interruptions such as phone calls, email, people coming by to ask questions, etc. Some of them might provide some benefit to an office worker by, for example, providing some important piece of information relevant for the current task or bringing an urgent issue to attention [1], [2]. But others are annoying, interrupt the current stream of thought, require an oftentimes time-consuming task switch, and are, thus, detrimental to the overall work performance [3], [4]. Oftentimes, the gravity of the detrimental effect is not only dependent on the actual interruption [5], [6] but the current context (or state) of the interrupted person [7]: Some activities require utmost concentration to maintain the train of thought (e.g., deriving a mathematical proof, complex decision tasks) and any interruption will lead to a very costly task-switch, whereas others are not as susceptible to interruption costs (e.g., reading one's email, simple tasks) [8]. In other words, the current task of a person determines his *interruptability* - the amenability to interruption his/her current working state has. Consequently, the *decision (usually made by an administrative assistant) of whether an interruption should be allowed to proceed to its intended recipient or whether it should be held back for later processing is usually made based on both the nature of the interruption and the interruptability of the person*. Some interruptions, such as phone-calls, don't carry any descriptive element of their content (beyond possibly a caller-id which might indicate a prior probability of the call importance). And *the less we know about the content of an interruption the more the person's interruptability determines what to do about an interruption.*

This paper investigates a novel way of predicting a person's interruptability in an office based setup. More specifically, we gather information about the subject's context by considering audio and motion detection sensors as well as the time of the day. Using data collected in a real office environment over the length of over 40 work days we, first, show how well we can predict the subject's presence in the office from our observations, where motion detection outperforms the other two data streams. Second, we show that we can predict his/her interruptability with good accuracy. Specifically, we demonstrate that multi-sector motion detection is superior to audio (which has been known as best interruptability indicator so far [9]) and that a combination of all data streams reaches an even higher prediction accuracy. We also show that dividing the motion detection information into different sectors representing different activity regions of the subject improves the prediction power. Finally, we use the inferred degree of the person's interruptability and the estimated importance of call (based on the phone number) to predict the person's attitude towards the incoming phone call. More specifically, we predict whether the user would like to ignore or accept the incoming phone call. The presented analysis also takes the temporal aspect of office work into account.

The remainder of the paper is organized as follows. First, we introduce the experiment we conducted to support our claims, which involves presenting the methodology and the technical setup. Then we evaluate the gathered data and analyze the results. Finally, after comparing the results with related work, we close with a discussion and the prospect of future work.

II. Experiment

Inferring a person's interruptability can be a central element in helping to manage interruptions, such as incoming phone calls. But can we design a device that automatically predicts

a person's context and, hence, interruptability? The studies of Hudson and colleagues [9], [7] have shown, such a prediction seems feasible in general. We assert that the predictive power of such a device can be significantly improved if it uses motion detection in addition to audio (and other the typically used devices). To that end we designed a long-term experiment, which would allow us to collect the necessary sensor streams to test our assertion with real-world data. This section describes this experiment starting with the requirements to the experimental setup and continuing with a description of the data collection. The next section then discusses the results.

A. Requirements to the Collected Data

In order to be able to make the desired predictions we needed to collect sensor data containing sufficient information about the subject's context in its environment, i.e., his/her office. We, therefore, decided to record both audio and video as well as self-reports provided by the subject.
For the motion detection recording we used a camera reporting changes in different sectors of the office as dynamics might be a significant indicator of someone's context or context changes. For simplicity, we did not consider face recognition or any other high level image recognition techniques in this work. The camera's microphone recorded the auditory surrounding of the person in the office. To support our approach for the artificial receptionist we logged the phone number of incoming calls and a self-report by the user. The phone call content itself was not recorded. The self report was structured along four dimensions:

1) His/her level of interruptability (How disrupting was the phone call?). The level of interruption has been broken down to five classes in a range from "ok, I don't care" to "do not disturb".
2) The level of importance of call as expected to be before answering only knowing the caller's number and/or name as it shows up on the phone display. The spectrum of the level ranges from "unimportant" to "highly important" broken down to five classes.
3) The level of the importance of call after hanging up knowing how important the call really was. Here the report included the same five classes as in the estimated importance of call before answering the call.
4) The action (accept or ignore) that the subject would have executed in advance if he/she had known the content the actual phone call. There are two classes to choose between: "would have better ignored" and "good that I answered".

Due to the limited number of (expected) phone calls to the subject taking part in the experiment, we decided to conduct another closely related experiment in parallel. As explained in more detail in the next section we prompted the subject on a regular basis to report his/her level of interruptability, similar to the interruptability report on incoming phone calls. There are two advantages on this procedure. First, we can collect a larger data collection about interruptability without depending on phone calls, which leads to more precise statements about the interruptability. And second, we can compare these self-reports with the self-reports on the phone calls, which elevates the significance of the latter predictions as well, if similar.

B. Method

According to Feldmann-Barrett and Barrett [10], there are three ways to conduct experience sampling:

1) *Interval contingent*: Sampling occurs at regular intervals.
2) *Event contingent*: Events of interest trigger the sampling procedure.
3) *Signal contingent*: Sampling is performed randomly over a period of time.

The annotation of incoming phone calls corresponds to event contingent experience sampling. Our concept of the interruptability self-report on a regular basis corresponds to a mixture of interval and signal contingent experience sampling. To ensure an upper and lower limit of the number of annotations we generate acoustic signal (or "beep") every 15 minutes. We also used a variance of 10 minutes on the signal to avoid "training" the users to expect the signal and thus altering their behavior. According to the subject the frequency of "beeps" turned out not to be too disruptive after some days of experimentation but their occurrence was still frequent enough to collect a significant number of self-reports.
The subject was asked to adhere to the following directions during the experiment. When a "beep" occurs the subject has to perform the self-report (assuming that the person is present in the office). This instruction allowed us to also gather information about the subject's presence in the office. Furthermore, the subject was asked to answer all incoming phone calls even if he/she preferred to ignore a call allowing a correct ex-post specification of the importance of call.

C. Data Collection Setup

The environment of the experiment is an office with three work places/locations (Figure 1). The office is typically used by one person only. The two remaining seats are used sporadically by other people as well as by the subject. The subject corresponds to the researcher profile of [9].
The audio and video data were recorded by an off-the-shelf webcam (Logitech QuickCam Pro 4000). We added a wide angle lens widening the aperture from 45 to 75, such that the entire office could be overviewed as shown in Figure 1. The audio recording was set to CD quality but mono instead of stereo (i.e., the settings are 16bit, 44.1 kHz, mono). The recorded video had 320*240 pixels (in color) at 25 frames per second. We compressed the video stream using the XviD codec setting I420 to ensure that one day of recording would fit on one DVD, while ensuring good recording quality. The recorded files were saved in "avi"-format for further processing, keeping the audio and video streams synchronized.
For the self-report we used a modified keyboard. All information sources were collected on a single PC on working days from 8.15am to 6.15pm.

Fig. 1. On the picture on the left, the office is seen through the webcam. The picture on the right, shows the camera mounted in the corner of the office.

Fig. 2. Sources of video inferences. On the left, another person than the subject is in the office. On the right an open window covers the office partly.

D. Sources of Interference

The experiment was conducted in a real-life environment. As a consequence, much interference influenced the gathered data. In this section we provide a list of possible interferences. The video used as motion detector was sensitive to all kinds of movements in the office. Therefore, the quality of collected data suffered from the presence of people other than the subject in the office - especially, when the subject was not present as seen in Figure 2 on the left. Furthermore, disturbances such as objects (like an open window) covering part of camera's view or changing brightness influenced the recording quality. Background noise interfered with the audio recordings. Sources of such background noise originate from outside the office (e.g., people chatting on the corridor, or the neighing horse on the paddock next to the university) or from inside the office (e.g., computer ventilators).

Error sources in the annotation procedure stem from the subject ignoring "beeps" or phone calls as well as inadvertent annotation mistakes. Addressing this risk we implemented a control mechanism using a feedback message for impossible annotation sequences.

Finally, as a matter of course this experiment was influenced by the experiment itself. The "beeps" prompting for a report on the subject's interruptability were disruptive for the subject.

E. Preprocessing

Beside the synchronization of all data streams we had to preprocess the raw audio and video data to get the most appropriate features for our problem. First, we extracted the features from audio (spectral center of gravity, temporal fluctuations of spectral center of gravity, tonality, mean amplitude onsets, common onsets across frequency bands, histogram

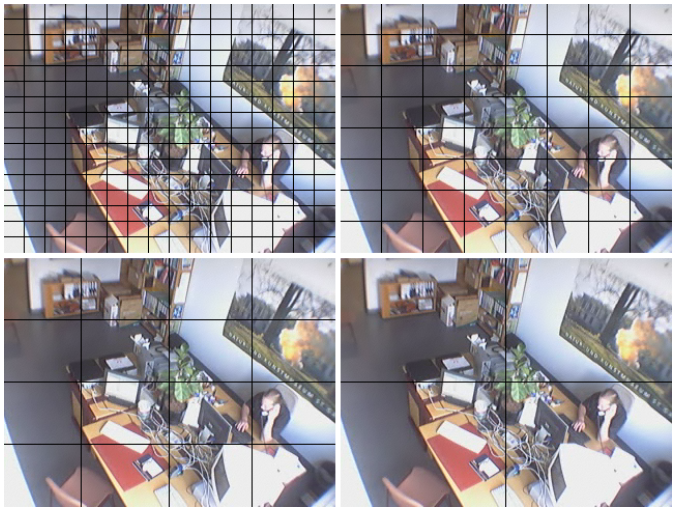


Fig. 3. Illustration of the motion detection features obtained by equidistant splitting

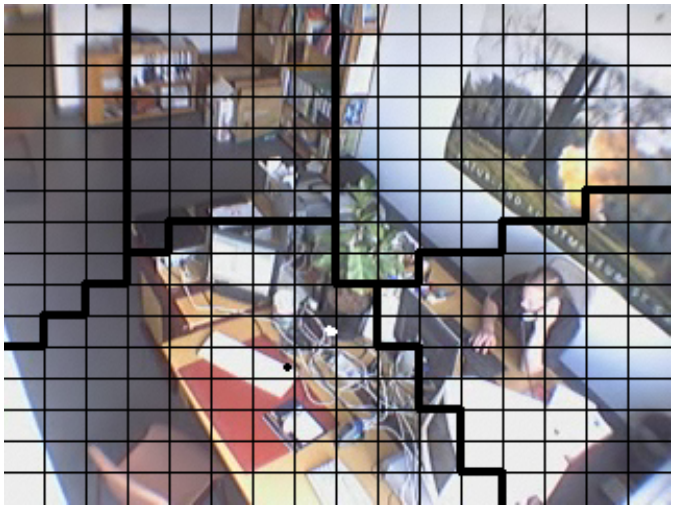


Fig. 4. Illustration of the motion detection features obtained by assigning five activity regions.

width, variance, mean level fluctuations strength, zero crossing rate, total power, and the 10 first cepstral coefficients) as described in [11]. This resulted in audio feature vectors of 20 features for every second of the recording.

As we used the video stream as a motion detector we calculated the changes between two frames separated by one second in the video stream. We divided each frame into rectangles of the size of 15x20 pixels resulting in 256 (16x16) distinct fields. To measure the motion in the office we summed the number of changed pixels between the two frames. Hence, for every second of the recording we obtained feature vectors with 256 features. Based on this large feature set we calculated smaller sets by summarizing the values of adjacent rectangles such that we got feature sets of the size of 64 (8x8), 16 (4x4), 4 (2x2), and 1 as depicted in Figure 3. Additionally, we created another feature set by dividing the room into five sectors as shown in Figure 4. The borders of the five sectors separate different activity regions. Three of the regions are located at

Fig. 5. Histogram of the movements recorded by the video camera on the left and the corresponding camera view on the right side.

the three work places; the other two are only active when people walk around. Thus, our motion detector is a little more sophisticated than the usually used motion detectors because it distinguishes between different sectors, except where we employed the feature set of size 1.
Finally, we constructed a two-dimensional feature vector representing the time of the day by taking the hour and distinguishing between am and pm.

III. Results

This section presents the results obtained after conducting the experiment. The experiment lasted 41 (working) days. During this time the two data sets were recorded. The data set generated by the ”beeps” consists of 1349 self-reports. In the following we refer to this data set as the ”beep” data set. The data set concerning the phone calls consists of 98 self-reports to which we refer as the ”phone” data set.

A. Data Overview (Descriptive Statistics)

1) Motion Detection: The motion detection data shows patterns as depicted in Figure 5. The usual location of the subject can easily be identified as the bright area. There are other lighter regions near the door and around the second work place. The Figure additionally shows the sectors of the five features we have chosen as activity regions. The borders of the particular sections overlap partially with the motion pattern.

2) Presence: The larger data set generated by the ”beeps” contains information about both the subject’s interruptability and his/her presence in the office. Figure 6 shows the overall presence of the subject illustrating that the subject is in his/her office about 45.1% of the time. The histogram on the right graphs the presence depending on the time of the day. The lunch break manifests itself as a dip at noon. The (average) presence decreases at both ends of the day (note that the distinctive decrease at 8am and 6pm are mainly due to the partial recording).

3) Interruptability: When present, the subject self-reported his/her degree of interruptability on a scale from 1 ”easily interruptible” to 5 ”not at all interruptible”. Figure 6 on the left shows the distribution of the interruptability in the ”beep” data. Class 2 ”quite interruptable” is dominant with a prior probability of 29.3% followed by class 5 with 25.2%. Figure 6 on the right is the corresponding distribution for the phone data. The two distributions are of similar shapes.

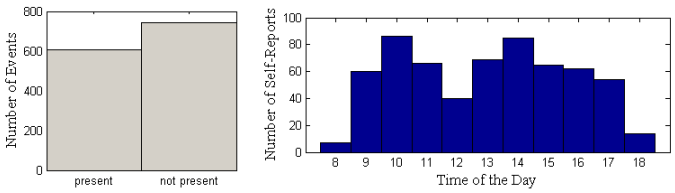


Fig. 6. The histogram on the left shows the overall presence of the subject in the office; the histogram on the right shows the presence vs. the time of the day.

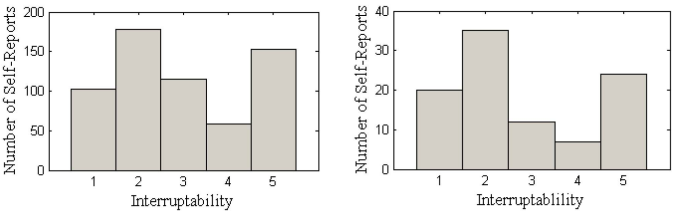


Fig. 7. Interruptability self-reports on a scale of five classes. On the left are the self-reports derived from the ”beep” data set and on the right the equivalent histogram but from the telephone data set.

4) Phone Calls / Phone Call Data Set: The phone call data set contains only annotations to answered phone calls but no information about missed phone calls, thus there is no information about presence in it. The histogram in Figure 8 shows the distribution of the (anonymized) phone numbers. Note that 9 out of 22 phone numbers where nonrecurring. This data set is about one sixth as large as the presence part of the ”beep” data set but contains more self-report dimensions. The first dimension of the self-reports is the interruptability as introduced in the larger data set. The histogram on the right of Figure 6 shows a similar distribution as seen in the larger data set. The dominant class 2 (degree of interruptability) has a prior probability of 35.7%. The biggest difference between the two histograms in Figure 6 is the decrease of class 3. Hence, we infer that the smaller data set is nearly as significant as the larger dataset because of the similarity of these two distributions.
The next two self-report dimensions are the importance of the incoming phone calls, before and after answering the call. Figure 9 shows both histograms. The figure on the right shows that most of the phone calls turned out to be important. The figure on the left illustrates the estimated importance of call before answering it. By comparing two histograms it is interesting to note that the estimations made by the subject ex-ante are higher than the importance of the phone call actually

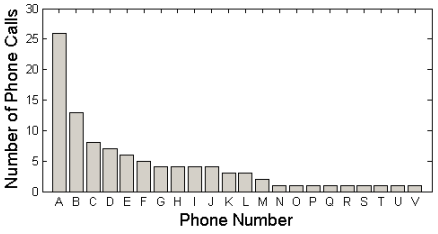


Fig. 8. Distribution of incoming phone calls. Each character represents a unique phone number.

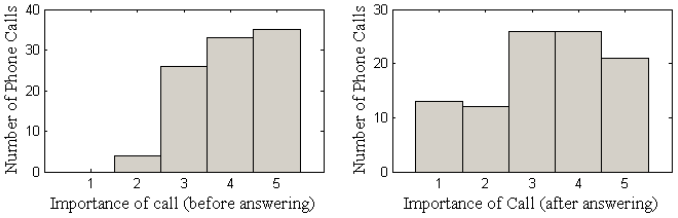


Fig. 9. The histogram on the left shows the distribution of the self-reports on the estimated importance of call, where ”5” corresponds to ”very important”. On the right is the distribution of the importance of call as it turned out after answering the call.

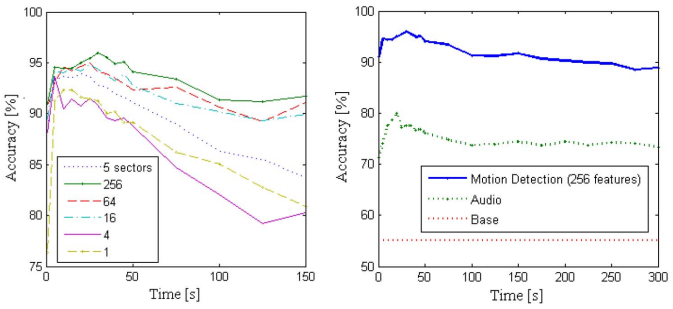


Fig. 10. Accuracies of presence prediction dependent on the past time. On the left, we see the motion detection with all six features for feature selection, on the right, the presence prediction from audio data compared with the prediction of from motion detection.

was.

5) Behavior/Action: The fourth and last dimension of the phone call self-reports is the behavior of the subject. The two considered behavior patterns are ”ignore” or ”accept”. The two requested actions are reported after answering a phone call in the experiment; however the goal of this dimension is that an artificial receptionist anticipates these actions. The class ”accept” dominates with 77.6% of all calls.

B. Prediction Quality

In this section, we present the prediction results of the subject’s self-reports from the sensor streams. First, we explain the prediction methods followed by the results.

1) Applied Classification Methods: We used the Weka 3 machine learning software package [12] to predict the subject’s self-reports. For all classification tasks we tested the data with two standard learning algorithms: naïve Bayes and the ”J48” decision tree learner. We preprocessed the data by normalizing and discretizing it with the standard Weka algorithms for better predictions. For the predictions, we took data up to 5 minutes prior to the event into account. We incorporated this information by an additional processing of the data by averaging the data (with equal weight) for each self-report. The depth of this averaging defines how much of the information about the past is incorporated. For each original feature the resulting new feature vector then contains the mean and standard deviation[1]. All results reported below are based on a 10 fold cross-validation.

2) Presence: For a future artificial receptionist application it is important to determine if the person is present in the office or not. Both graphs in Figure 10 show the prediction accuracy versus past time. The graph on the left shows the prediction from motion detection evaluating all six possible feature-combinations. The largest feature set (the most finely grained with 256 rectangles per frame) turns out to be the most predictive. The graph on the right, compares the best motion detector prediction with audio. Both audio and motion detection show a distinct maximum at about 20 seconds. Motion detection reaches an accuracy of 96% at 20 seconds using the J48 decision tree classifier outperforming audio that reaches 89.9% using naïve Bayes. This shows that interruptability prediction contains an important temporal aspect. Furthermore, the alignment of the best predictions accuracies of both models when taking the last 20 seconds into account could indicate that the time window to be considered of 20 seconds is inherent to the contextual phenomena investigated (i.e., presence) rather than a sensory artifact. This makes particularly sense with presence detection, as just less than 20 seconds is about the time a person might stay still (motion detection) and be quiet (i.e., not type, cough, etc.) during typical work.

TABLE I
CONFUSION MATRICES FOR PRESENCE PREDICTION BASED ON 20 SECONDS OF PAST DATA. (CLASS 1 CORRESPONDS TO ”NOT PRESENT” AND CLASS 2 TO ”PRESENT”).

Presence, audio

		Model prediction	
		1	2
Self-Report	1	582	159
	2	98	510
Accuracy: 80.9% Base: 54.9%			

Presence, motion

		Model prediction	
		1	2
Self-Report	1	719	22
	2	32	576
Accuracy: 96.0% Base: 54.9%			

Presence, time of day

		Model prediction	
		1	2
Self-Report	1	570	171
	2	365	243
Accuracy: 60.3% Base: 54.9%			

When taking only the time of the day into consideration to infer the presence we reach an accuracy of 60.3% which is still better than the prior annotation distribution of 54.9% (see Table I for the detailed confusion matrices). We combined the three classifications by meta-classifiers on their class prediction probabilities but the results were not better than the prediction from motion detection. Thus, audio and the time of the day do not contribute any new information to achieve better accuracies but might contribute to higher robustness.

3) Interruptability: We have two data sets to infer the degree of the subject’s interruptability. This prediction task is a 5-class classification prediction with a base prior of 29.3% for the larger and 35.7% for the smaller data set. Figure 11 shows that the 5-sector feature of the motion detector is the most predictive. Specifically, note that the incorporation of domain knowledge into the model through the informed choice of the 5 sectors pays off, as it leads to an improved prediction accuracy[2]. Figure 12 shows the

[1]We also tried Markov chains and hidden Markov models. However, they were outperformed by our coarse approach.

[2]We intend to investigate whether the sectors could also determined automatically through analyzing the histogram of the movement data (c.f. Figure 5)

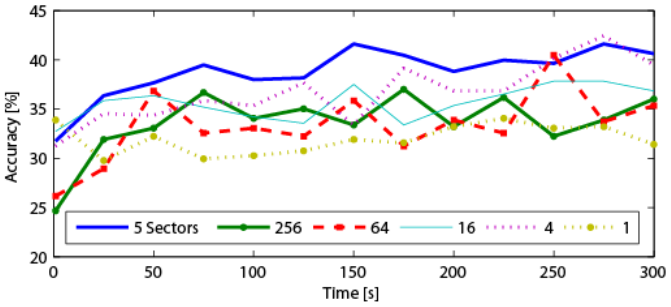


Fig. 11. Interruptability detection from all features setups of the motion detector using J48.

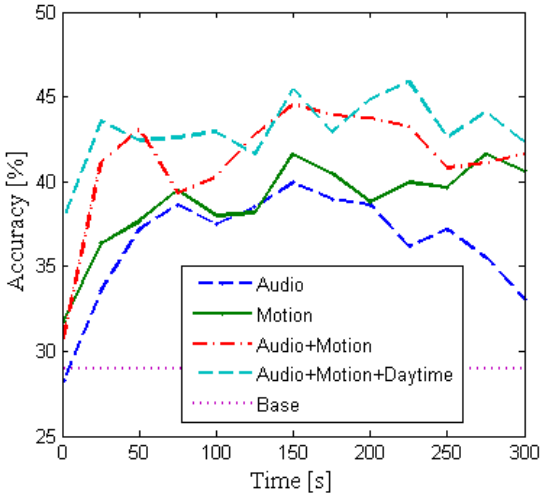


Fig. 12. 5-class interruptability prediction vs. time to past from the larger "beep" data set. The figure shows the audio and motion detection predictions and the combinations.

prediction accuracies of audio and motion detection vs. the time into the past. Both audio and motion detection show very good results with a maximum at 150s. Motion detection is superior (41.6%) to audio (40%) and, furthermore, motion detection seems to be more robust to variations on the time dimension. Predicting the interruptability from the time of the day using naïve Bayes results in an accuracy of 35.9%. Combining audio and motion detection by a naïve Bayes meta-classifier results in a remarkably better prediction result (maximum at 150s: 44.6%) indicating that both sensor inputs provide partly independent information. Combining all three information sources (audio, motion detection, and time of the day) results in an even better result with a maximum accuracy of 45.4% at 150s. Furthermore, the combination of the three sources results in a much more robust result in terms of time dependency. Again we see that that the temporal aspect of the contextual situation is very prominent. Specifically, we see that the prediction quality of the models rises until at least about 20 seconds have been taken into account.

Table II shows the performance of the different calculations on the basis of confusion matrices. The graph on the right of Figure 13 shows the same predictions based on the smaller telephone data set. The curve based on the smaller data set show a much weaker performance than the results of the larger data set. On one hand this indicates that the model based on

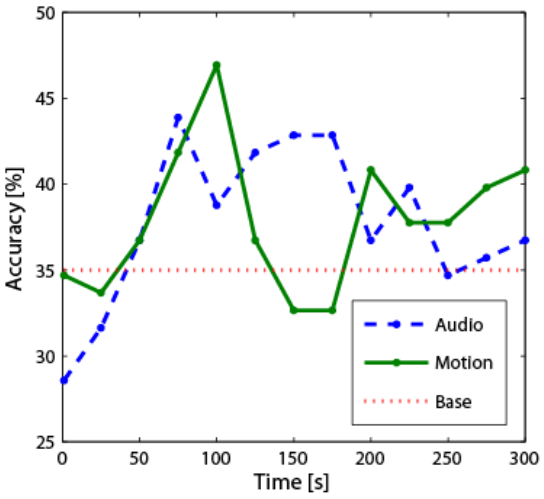


Fig. 13. 5-class interruptability prediction vs. time to past from the smaller phone data set.

motion detection needs more training instances than the model based on audio data. On the other hand the subsequeent results based on this interruptability calculation will suffer from this effect.

4) Importance of Call: We estimate the importance of an incoming call based on the caller's phone number. Of course, the target label is the importance of call after answering (ex-post), as it is closer the actual importance of call as opposed to the (ex-ante) estimation of the subject before answering the call. Our prediction task takes the subject's self-report about the estimated importance of call as benchmark for our classification based on the phone number. The confusion matrix on the left in Table III shows that the subject often overestimated the importance of the incoming phone calls as seen in the overview of the collected data. The subject reached an accuracy of 37.8% compared to the base of 26.5%. Applying a naïve Bayes classifier on the phone numbers

TABLE II
CONFUSION MATRICES FOR THE 5-CLASS INTERRUPTABILITY DETECTION ON THE "BEEP" DATA SET.

Interruptability, audio

Self-Report \ Model prediction (naïve Bayes)	1	2	3	4	5
1	12	51	11	5	24
2	13	106	17	8	34
3	8	65	14	5	23
4	3	15	2	6	33
5	8	25	8	7	105

Accuracy: 40.0% (at 150s)
Base: 29.3%

Interruptability, motion

Self-Report \ Model prediction (J48)	1	2	3	4	5
1	27	45	6	0	25
2	22	125	3	3	25
3	15	75	4	1	20
4	9	25	0	1	24
5	12	33	5	7	96

Accuracy: 41.6% (at 150s)
Base: 29.3%

Interruptability, time of the day

Self-Report \ Model prediction (naïve Bayes)	1	2	3	4	5
1	1	40	1	0	61
2	5	104	4	0	65
3	4	46	0	0	65
4	1	19	0	0	39
5	4	36	0	0	113

Accuracy: 35.9%
Base: 29.3%

Interruptability, all combined

Self-Report \ Model prediction (naïve Bayes)	1	2	3	4	5
1	35	24	18	8	18
2	15	103	38	4	18
3	7	53	34	3	18
4	6	11	9	2	31
5	8	11	22	10	102

Accuracy: 45.4% (at 150s)
Base: 29.3%

TABLE III

CONFUSION MATRICES OF THE PREDICTION OF THE ACTUAL AND THE ESTIMATED IMPORTANCE OF CALL.

Importance of call, phone number

Self-Report \ Subject's prediction	1	2	3	4	5
1	0	3	4	5	1
2	0	1	6	3	2
3	0	0	10	10	6
4	0	0	4	11	11
5	0	0	2	4	15

Accuracy: 37.8%
Base: 26.5%

Importance of call, phone number

Self-Report \ Model prediction (naïve Bayes)	1	2	3	4	5
1	0	0	11	2	0
2	0	0	10	1	1
3	2	0	14	7	3
4	0	0	19	4	3
5	0	0	6	3	12

Accuracy: 30.6%
Base: 26.5%

TABLE IV

ACTION PREDICTION BASED ON THE SELF-REPORTS.

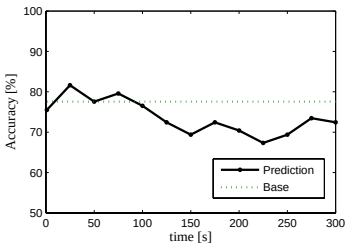
Interruptability	Estimated importance of call before answering	Actual importance of call after answering	Prediction of behavior
X			77.6%
	X		81.6%
		X	88.8%
X	X		79.6%
X		X	94.9%

(phone numbers are treated as independent classes) we reach a prediction accuracy of 30.6% as shown in detail in the confusion matrix on the right in Table III. It is not surprising that the subject outperforms the naïve Bayes algorithm because the subject already had prior knowledge about the phone numbers. This advantage is intensified by the fact that 9 of 22 phone numbers were non-recurring and hence the algorithm was not able to include them in the model (c.f. phone number histogram in Figure 8).

5) Subject's Phone Behavior: The subject's phone behavior is the last prediction category of our study. The goal is to predict the appropriate action of the subject, i.e. ignoring the phone call or accepting it. The following predictions show how precise the action can be determined from the two dimensions "interruptability" and "importance of call". This provides a benchmark for the following estimations under the assumption that the two used dimensions have been determined with an perfect accuracy of 100%.

First we examine how precise the action can be determined directly from the subject's self-reports disregarding the sensor data. Table IV shows the prediction of the action using a naïve Bayes classifier.

The "X" marks which self-report have been taken into account. It turns out that the combination of knowing the degree of interruptability and knowing the actual importance of call best determines the most appropriate action. Obviously, the predictions involving the estimated importance of call do not perform as well as the predictions based on the actual importance of call due to the bias in the subject's estimation. It is remarkable that the actual importance of call contributes more to the action prediction than the interruptability (the interruptability alone only reaches the prior probability of the action distribution of 77.6%). Nevertheless, the combination of those two dimensions shows that both of them are required for reliable predictions. These values are used as benchmark



Subject's Behavior

Self-Report \ Model prediction	ign.	acc.
ign.	13	9
acc.	9	67

Accuracy: 81.6% (at 25s)
Base: 77.6%

Fig. 14. On the left the prediction of the action vs. time. On the left the confusion matrix for the maximal accuracy at 25 seconds.

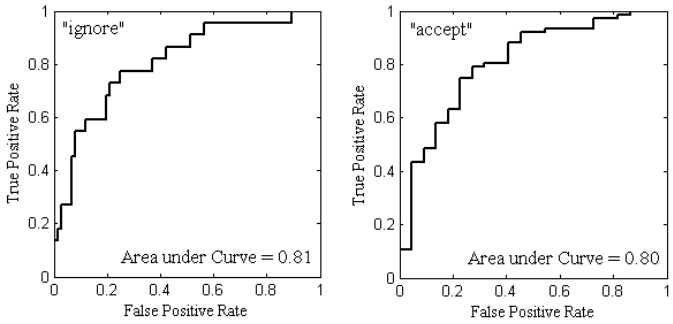


Fig. 15. ROC curves for the two classes of the subject's behavior. On the left the curve for "ignore", on the right the curve for "accept".

for the following predictions based on the sensor data streams. For the prediction of the subject's phone behavior we determined the interruptability from the audio, motion detection, and the time of the day data. These three predictions as well as the prediction of the importance of call were combined using a naïve Bayes meta-classifier. Figure 14 shows the accuracy vs. time to past. At a first look this result does not seem to be overwhelming except for the region around 25s with an accuracy of 81.6% compared to the base accuracy of 77.6%. But at a second look this value of 81.6% is *better than the prediction from the benchmark calculation based on the self-reports of interruptability and the estimated importance of call* showing that the algorithm classifies as good as the subject would. Even more, from the benchmark calculation in Table IV we can induce that improving the model on the importance of call could result in a much better accuracies up to 94.9%. Usually, it is very difficult to evaluate classifiers in unequal distributions like in this problem using classification accuracy. Therefore, we use the alternative approach of receiver operating characteristics (ROC) curves - a measure from signal theory that has gained high acceptability in machine learning [13]. The ROC curve graphs the true positive rate versus the false positive rate. The random classifier corresponds to the diagonal, the closer the curve to the upper left corner the better it is. As the curves are always monotonically increasing the area under the curve can be used as a measure for the classifier performance. Figure 15 shows the two ROC curves for the two classes "ignore" and "accept" for the model generated at 25 seconds. The curve on the left for "ignore" has an area under curve of 0.81 and the curve on the right for "accept" has an area of 0.8, which is good.

ROC curves allow a cost independent investigation of clas-

sifiers. Even more we have the possibility to assign a cost-based cutoff for classifications using probability estimators. Reconsidering our desired phone application a subject can individually assign cost to the actions. E.g., a subject is rather willing to accept phone calls that are misclassified as ”accept” than to miss a really important call - so the subject could weight the costs of a misclassification of an ”ignored” call higher.

IV. Comparison to Related Work

Our work can be best compared to the studies by Forgarty and colleagues [7], [9], as their setup is the most comparable to ours. In [7] they present a so-called ”Wizard of Oz” feasibility study to predict people's interruptability. They simulate a sensor-equipped office using a video and audio recording of the office, which are then hand-coded by people determining features such as the number of people currently in the office, who is speaking, what task objects are being manipulated, whether the phone is on or off the hook, and other similar facts about the environment. In a follow-up study [9], [14] they equipped an office with real physical sensors. They placed microphones in the office, logged the beginning and end of each non-silent interval, after applying a speech recognition tool that detected conversations. Additionally, two magnetic switches, one near each side of the top of the door frame, allowed them to sense whether the door was open, cracked, or closed. Two one-sector motion sensors were put in each office but not used in any of the predictions. Another magnetic switch was used to determine whether a person's phone was physically off its hook. 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. The interruptability annotation was done the same way as introduced in the preceding study by audibly prompting the subjects (i.e., self-reporting). All this information together with the interruptability annotation was used to build interruptability predictors. Forgarty and colleagues found a prediction accuracy of 51.5% for the 5 class prediction problem using all of their sensors. At the first sight this outperforms our predictions. Note, however, that their data was clearly different, in that they reported a higher prior of up to 40.9% (for one of his subjects) with an average of 31.9% whereas we had a prior of 29.2% lowering the overall accuracy. Furthermore, their study found that the 1-sector motion detector didn't add any prediction accuracy to the model. Our study shows that *multi-sector motion detection* outperforms pure audio predictions (c.f. Figure 10), which were the most predictive in their study. We also show that *multi-sector motion detection is complimentary to audio* (c.f. Figure 12). Consequently, it is reasonable to assume that multi-sensor motion detection extracts useful information about the context in a non-obtrusive way. Hence, we believe that augmenting the setup by Fogarty and colleagues with motion detection would either improve their overall prediction accuracy or make some of the other sensors they used obsolete. Alternatively, we can assume that our prediction accuracy could be further improved with the addition of some of the sensors proposed by them.
Even though there are other studies to predict interruptabiliy in both the office setting [15], [16] and the mobile setting [17], [18], [19] these are extremely difficult to compare to our setup. [15], [16] uses information stored on the PC as well as a key-logger to predict accuracy. Consequently, those studies only use the ”virtual” context rather than the physical environment making them complimentary to our investigation. [17], [18], [19] all present wearable devices to predict a person's interruptability exploring a different style of work rather than the office-based setup.
At this time we have no knowledge about studies examining the prediction of a person's phone behavior.

V. Limitations and Future Work

The major drawback of this study is that the experimental setup is restricted to only one single subject. To strengthen the external validity of the experiment we intend to conduct this experiment with a broad range of different people. These measurements will also allow comparing different user profiles and the adaptability of models between people. Furthermore, we also plan to consider multi-person offices to examine the influence of multiple persons on the prediction quality as well as the presence of additional persons on someone's interruptability - a subject that has not yet been investigated in any study we know.
Furthermore, we plan to complement the camera simulating a motion detector by infrared cameras. These offer the ability to detect objects that are warmer than background, which allows identifying persons or active devices (like screens) without taking motion (and time) into account and could lead to better prediction accuracies. Consequently, we are expecting stronger predictive power from infrared motion detectors. We also plan to apply our approach to other areas such as instant messaging systems or phone applications in vehicles (since a vehicle can be viewed as a room on wheels). Last but not least, we are investigating the use of more sophisticated prediction algorithms.

VI. Conclusions

In this study we successfully introduced multi-sector motion detection to augment context-awareness in office-like setups. We found that we can predict whether a person is present in the office or not, based on motion detection, audio, and daytime data, where multi-sector motion detection clearly outperforms the others. We also found that we can predict the person's degree of interruptability from these three information sources, where multi-sector motion detection again turned out to be the most reliable sensor-stream, which could, however, be improved when complemented with audio. Our study shows clearly that the interruptability of people can be predicted with a highly simple and non-obtrusive sensory setup (of only one small web-cam). Combined with a prediction of the nature of the interruption this simple setup can provide a major building block for a cheap and reliable device to manage interruptions and, thus, improve overall work performance.
Summarizing, this study represents a first step towards the

design and implementation of an artificial receptionist. Its findings strongly indicate that such a receptionist can be implemented with simple sensors.

VII. Acknowledgements

We would like to thank Alen Zurfluh for his substancial support in the initial stage of this project. Furthermore, we would like to thank B. Schiele and N. Kern for sharing their audio preprocessing code.

References

[1] E. B. Cutrell, M. Czerwinski, and E. Horvitz, "Effects of instant messaging interruptions on computing tasks," in *CHI '00: CHI '00 extended abstracts on Human factors in computing systems*. New York, NY, USA: ACM Press, 2000, pp. 99–100.

[2] B. O'Conaill and D. Frohlich, "Timespace in the workplace: dealing with interruptions," in *CHI '95: Conference companion on Human factors in computing systems*. New York, NY, USA: ACM Press, 1995, pp. 262–263.

[3] T. Gillie and D. Broadbent, "What makes interruptions disruptive? a study of length, similarity, and complexity," in *Psychological Research 50*, 1989, pp. 243–250.

[4] J. M. Hudson, J. Christensen, W. A. Kellogg, and T. Erickson, "ïd be overwhelmed, but it's just one more thing to do: availability and interruption in research management," in *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM Press, 2002, pp. 97–104.

[5] D. C. McFarlane, "Interruption of people in human– computer interaction: A general unifying definition of human interruption and taxonomy," in *Technical Report NRL/FR/5510-97-9870*, 1997.

[6] ——, "Coordinating the interruption of people in human-computer interaction," in *Proceedings of Human-Computer Interaction (INTERACT'99)*. IOS Press, The Netherlands, 1999, pp. 295–303.

[7] S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee, and J. Yang, "Predicting human interruptibility with sensors: a wizard of oz feasibility study," in *Proceedings of the conference on Human factors in computing systems*. ACM Press, 2003, pp. 257–264.

[8] C. Speier, J. S. Valacich, and I. Vessey, "The influence of task interruption on individual decision making: An information overload perspective," in *Decision Sciences, 30 (2)*, 1999, pp. 337–360.

[9] J. Fogarty, S. E. Hudson, and J. Lai, "Examining the robustness of sensor-based statistical models of human interruptibility," in *Proceedings of the 2004 conference on Human factors in computing systems*. ACM Press, 2004, pp. 207–214.

[10] L. F. Barrett and D. J. Barrett, "An introduction to computerized experience sampling in psychology," *Social Science Computer Review*, vol. 19, no. 2, pp. 175–185, Summer 2001.

[11] J. Syrjälä, "Context classification using audio data for wearable computer," Master's thesis, Swiss Federal Institute of Technology (ETH), 2003.

[12] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[13] F. J. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine Learning*, vol. 42, no. 3, pp. 203–231, 2001.

[14] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, "Predicting human interruptibility with sensors," *ACM Trans. Comput.-Hum. Interact.*, vol. 12, no. 1, pp. 119–146, 2005.

[15] E. Horvitz, P. Koch, C. M. Kadie, and A. Jacobs, "Coordinate: Probabilistic forecasting of presence and availability," in *Proceedings of the Eighteenth Conference on Uncertainty and Artificial Intelligence (UAI '02)*, 2002, pp. 224–233.

[16] E. Horvitz and J. Apacible, "Learning and reasoning about interruption," in *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*. New York, NY, USA: ACM Press, 2003, pp. 20–27.

[17] D. Siewiorek, A. Smailagic, J. Furukawa, A. Krause, N. Moraveji, K. Reiger, J. Shaffer, and F. L. Wong, "Sensay: A context-aware mobile phone," in *ISWC '03: Proceedings of the 7th IEEE International Symposium on Wearable Computers*. Washington, DC, USA: IEEE Computer Society, 2003, p. 248.

[18] N. Kern and B. Schiele, "Context-aware notification for wearable computing," in *Proceedings of the 7th International Symposium on Wearable Computing*, New York, USA, October 2003, pp. 223–230.

[19] N. Sawhney and C. Schmandt, "Nomadic radio: scaleable and contextual notification for wearable audio messaging," in *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM Press, 1999, pp. 96–103.

Peter Vorburger holds a master's degree in physics obtained at the ETH Zurich, Switzerland. He is PhD. candidate at the Dynamic and Distributed Information Systems Group in the Department of Informatics at the University of Zurich, Switzerland. Before joining the University of Zurich he worked for four years as physicist in industry and as a CRM freelance consultant in the financial services sector. His research interests include context-awarenss and machine learning theory.

Abraham Bernstein is Associate Professor and heads the Dynamic and Distributed Information Systems Group in the Department of Informatics at the University of Zurich, Switzerland. Before joining the University of Zurich he was Assistant Professor in the Department of Information, Operations and Management Sciences at New York Universitys Leonard N. Stern School of Business and received a Ph.D. from MIT's Sloan School of Management. His research interests include the various aspects of supporting dynamic (intra- and inter-) organizational processes with a special focus on machine learning, the semantic web, and pervasive computing.