Accepted for publication in:
J. PERVASIVE COMPUT. & COMM. 2 (2), MARCH 2006. ©TROUBADOR PUBLISHING LTD


# A Scenario-Based Approach for Direct Interruptablity Prediction on Wearable Devices

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

*Received: January XX 2005; revised: November XX 2005*

***Abstract*— People are subjected to a multitude of interruptions. This situation is likely to get worse as technological devices are making us increasingly reachable. In order to manage the interruptions it is imperative to predict a person's interruptability - his/her current readiness or inclination to be interrupted. In this paper we introduce the approach of direct interruptability inference from sensor streams (accelerometer and audio data) in a ubiquitous computing setup and show that it provides highly accurate and robust predictions. Furthermore, we argue that scenarios are central for evaluating the performance of ubiquitous computing devices (and interruptability predicting devices in particular) and prove it on our setup. We also demonstrate that scenarios provide the foundation for avoiding misleading results, assessing the results' generalizability, and provide the basis for a stratified scenario-based learning model, which greatly speeds-up the training of such devices.**

***Index Terms*— Interruptability, Context Awareness, Wearable Devices, Evaluation, Direct Prediction, Scenario**

## I. Introduction

People are subjected to a multitude of interruptions. This situation is likely to get worse as technological devices are making us increasingly reachable. But the problem is not as simple as it looks like at the surface. As examples, consider the following situations:

- A physician visits the opera. Should she turn off her pager as directed or leave it on? But what if she gets an important page to perform some emergency surgery? What about the daily financial update page from her investment advisor?
- A manager is about to start an important meeting in his office. Should he turn his phone off? What if the manager misses some important breaking news that would significantly change the course of the meeting?

What both these situations have in common is that a person performs some task and can be interrupted by a technological artifact. When interrupting the person (s)he incurs some cost, be it the angry looks of other opera visitors or the interruption of the ongoing meeting. On the other hand, the person might greatly profit from receiving the message. The surgeon might have to save somebody's life or the manager might improve the meetings outcome. The decision on whether to be informed about the interruption or not is highly context dependent. Research in context-aware computing has investigated whether the interruption of a person based on sensory input is possible. The findings are promising. Some studies show a quite acceptable accuracy when attempting to predict a person's interruptability [1], [2], [3], [4].

In most interruptability studies, however, interruptability is predicted indirectly through the use of predefined categories and it was *difficult to assess the reliability/robustness of the results* towards changes in contexts and applications. Indirect prediction requires the prior definition of categories such as location of the subject and/or its actual activity. We show that a direct prediction of interruptability is possible and provides high accuracies without the need for predefined categories. Addressing the difficulty to assess the reliability of such experiments, this paper argues for a *scenario-based evaluation method*, in which the typical activities of the experiment are reported as part of the experimental description. A scenario - a plan of activities and course of events - puts the evaluation results in context and, hence, provides crucial background information to facilitate their assessment. Basing our investigation on data from a wearable interruptability experiment akin to [4], in which we use a wearable computer with accelerometers and audio sensors to predict a subject's interruptability, we show that a scenario-based evaluation provides a better guide for how a learned interruptability detector performs. In particular we show how misleading the typically reported accuracy statistics are when no detailed scenario is provided. We support our assertion - scenarios help to vastly improve an assessment of the robustness of results - with the following claims sustained by our experimental data: First, the situational variance of the data collected should correspond to the one encountered in real-world use. A scenario helps to ensure such an ample variance. Second, we show how accuracies are typically better if the model has to perform in fewer situations, which is misleading as specialized predictors often poorly generalize. Third, we show how a scenario allows assessing the performance of a predictor in situations it didn't encounter in its training phase - an evaluation, which is difficult without a scenario. As a consequence of these three findings we can see that it is imperative to report on the situation in which the accuracy was trained and attained in order to assess the reliability of a prediction method in general. This even argues to provide scenarios for field-experiments, as some situations might be central to a person's activity but so seldom that

they do not arise within the experiments timeframe. Fourth, we show that scenarios can be used to speed-up the learning phase using a stratified scenario-based learning model, similar to the ones used in speech recognition software. Fifth, we discuss the appropriateness of the sensors used, as we find that accelerometers aren't as predictive and are not as robust towards situational variability in our environment as audio recordings. Last, we question the use of accuracies as a sole measure of prediction quality, as it assumes that the cost of miss-prediction is known at design time (or implicitly assumes equal cost for false positive and false negative classifications) - an assumption which is typically wrong when users are involved.

The remainder of this paper is organized as follows. In the next section we introduce our experimental setup including the technical data collection mechanism and the used office-worker scenario. The following section provides the statistical support for each of our claims based on the data gathered in our experiments. We close with a discussion of related and future work.

## II. Our experimental setup

### A. The technical setup

For our experiments we used a technical setup that was based on Kern and Schiele [4]. Using accelerometers and a microphone - both resource-friendly, easily-available, cheap, and simply-usable sensors - they showed very promising results in terms of prediction quality. [5] also show that different actions can be recognized by accelerometers attached to a subject. In contrast to [4] we used both a different annotation strategy and technology: they attached a PDA (iPaq) to their recording Laptop annotating activities such as "sitting," "standing," "walking," etc., and the auditory context (as "street," "restaurant," etc.) inferring the interruptability tendencies using explicit mapping rules. Avoiding the need for constructing mapping rules and the complexities of perpetually selfrecording one's context using a PDA we decided to use a potentiometer (see also Figure 1, right), which was used by the subject himself to report his interruptability by simply turning the knob. This allowed the subject to simply annotate large amounts of data, without averting his gaze from the current activity.

We connected most sensors, i.e., the accelerometers and the potentiometer to a Smart-it and multiplexer device [6], which transform the analog input into a digital, multiplexed data stream. This data stream could then be accessed through a computer's serial port, allowing the analog sensory data to be queried from programs using standard serial port read/write instructions. The microphone (recording at 44 kHz, 16 Bit mono) was connected to the laptop's sound port.

In order to ensure that we didn't miss important information about the subject's motion we decided to use three 3-D accelerometers. These accelerometers allowed us to measure the motion in 3-D space of the subject's shoulder, wrist[1], and leg (see also Figure 1, left). The subject wore the potentiometer for annotation attached to his belt and the microphone attached like a necklace to his neck. The laptop and Smart-it electronics were all carried in a backpack to allow the subject to move around. The software on the laptop recorded the output from each sensor and wrote it to a file. The output from the microphone was recorded in wave files.

Fig. 1. a) Subject wearing data acquisition setup with accelerometers (red arrows, left side of figure), microphone (green arrow, top right side), and potentiometer (blue arrow, bottom rightside). b) Potentiometer only on the left.

### B. The data acquisition scenario

To assure the representativeness of our measures we constructed a scenario. The scenario was based on the idea that every activity typically undertaken by an office worker should be included at least once. Hence, we assembled a list of typically lived-through tasks of an office worker based on our own experience and informal interviews with office workers. Then the subjects were asked to carry out the tasks using appropriate activities (of their choosing). Note that we left the subject unclear about when certain events would occur to avoid setting the subject's pre-conceptions about the possible course of events influencing his behavior. For example, we left the exact timing of some phone-calls to chance by asking some people to call our subject at random times during the day. After conducting the experiment we examined the activities and categorized them into 6 major activity classes, which we called situations. As Table I shows on the left the scenario includes six rough situations (walking, riding the streetcar, working at the office, visiting the cafeteria, and attending a lecture) totaling in about 4 hours of recording time. The table on the right also shows that some activities have a relatively high prior probability of high interruptability (like streetcar) while others have an extremely low prior (like attending a lecture).

[1]The accelerometers on the wrist were mainly used to synchronize the sound recording with the acceleration sensors by hand clapping.

TABLE I
SCENARIO ACTIVITIES, TIMES, & DURATION

| Activity | Start Time [s] | Duration [s] |
|---|---|---|
| Walking | 0 | 547 |
| Streetcar | 547 | 365 |
| Walking | 912 | 217 |
| Office work | 1129 | 2199 |
| Walking | 3328 | 204 |
| Lecture | 3532 | 1781 |
| Walking | 5313 | 164 |
| Office work | 5477 | 1077 |
| Walking | 6554 | 134 |
| Cafeteria | 6688 | 1270 |
| Walking | 7958 | 207 |
| Office work | 8165 | 225 |
| Walking | 8390 | 70 |
| Meeting | 8460 | 2225 |
| Walking | 10685 | 138 |
| Cafeteria | 10823 | 724 |
| Walking | 11547 | 90 |
| Office work | 11637 | 3161 |
| Walking | 14798 - 15381 | 583 |

| Situation | Interruptability | |
|---|---|---|
| | High [%] | Low [%] |
| Walking | 81.1 | 18.9 |
| Streetcar | 100.0 | 0.0 |
| Office work | 0.2 | 99.8 |
| Lecture | 1.9 | 98.1 |
| Cafeteria | 68.8 | 31.2 |
| Meeting | 4.5 | 95.5 |

## III. RESULTS - SUPPORTING THE CLAIMS

To support our claims we first had to construct an interruptability predictor from the data stream. We extracted 17 features from the audio stream using principal component analysis and normalization, which have been shown to be a good basis for context-aware learning [7]. Furthermore, we sampled all the data streams (accelerometers and audio) down to 1 Hz. Then, we clustered each sensor stream using a simple k-means algorithm. This filtered each data stream into k discrete states, which we then used to learn a simple Markov model for the two prediction classes. At any given point in time the class is predicted using Markov chains. This prediction procedure has two parameters: the number of clusters k and the number of steps s into the past as known as the length of the Markov chain.

All of these computations allow us to support our claims from above, which we will now visit in turn with their supporting evidence.

### A. Direct interruptability prediction

The general accuracy results (ten fold cross-validated) can be seen in Table II, which shows the confusion matrixes for the two-class problem (Interruptible/not Interruptible) for each of the three sensors. As we can see the overall accuracy of our experiment is very good: our accuracies are highly competitive with those reported in related work , who reported accuracies of ranging from 80.1% to 87.7% [1], [4]. The table also reports on the parameter settings k and s for the inference procedure, which we are going to maintain for the rest of the paper. Note that we achieved these high accuracies without any prior definition of categories (symbols) except for the target symbol ”interruptability”. According to the literature [8] we should, therefore, expect better generalization behavior by our prediction method to new situations that have not been encountered/defined before. Subsection 3.5 will discuss the predictor’s generalization behavior to new situations.

TABLE II
MODEL CONFUSION MATRIXES. ”I” STANDS FOR ”INTERRUPTABLE”

| | | **Shoulder** | | **Leg** | | **Audio** | |
|---|---|---|---|---|---|---|---|
| | predicted | I | ¬ I | I | ¬ I | I | ¬ I |
| reported | I | 14.1% | 8.7% | 15.2% | 7.5% | 19.3% | 5.4% |
| | ¬ I | 6.8% | 70.4% | 5.0% | 72.3% | 4.1% | 71.3% |
| Accuracy | | **84.5%** | | **87.5%** | | **90.6%** | |
| Parameters | | k = 200 s = 40 | | k = 150 s = 50 | | k = 240 s = 40 | |

### B. The necessity of situation variance

Our first claim concerning the scenario-based evaluation method states that it is important to ensure that the situations encountered within the experiment are as varying as the ones encountered in real-world use. Using a scenario as a basis for the experiment can help such a correspondence of situation. While the claim is intuitive we used our experimental data to support it numerically. The audio prediction accuracies reported in Table III (the accelerometer results are similar) show how a model learned for each of the situations (rows) predicts the interruptability of all the other situations (in columns). Obviously, the *models predict the situations well, in which they were learned*, resulting in the high prediction accuracies in the diagonal. But we can also see that it is *nearly impossible to infer how well the models are going to perform in situations other than the one they were learned in*. The model learned in the cafeteria situation, for example, predicts walking, streetcar, and cafeteria with relative high accuracies, but performs relatively poor for the rest. As a consequence it is imperative to use a set of situations in an experiment that corresponds with situations that typically occur in an application area. Note that field experiments increase the chance of accruing the necessary situations but don’t guarantee it, as some highly important situations may occur infrequently. It would, consequently, be prudent to enrich field-experiments with scenarios to attain full situational variance.

A second interesting observation can be made from Table III: the *results are asymmetric*. The model learned in the cafeteria, for example, is highly predictive for the streetcar (100% accuracy) but the inverse performs poorly (68% accuracy). This indicates that some situations contain more information for the overall model than others. As a consequence we could theorize about a measure of usefulness of a situation. One might be the prior probabilities shown in Table I on the right, where we can see that the streetcar situation contains no information about non-interruptible situations and the cafeteria has a prior 69% - practically the prediction accuracy of the streetcar model on the cafeteria data. Another would be to use the information content (or entropy) as a measure. Again, we see the necessity for identifying the scenario situations as they can guide our search for information rich situations.

TABLE III
ONE BY ONE ACCURACIES FOR AUDIO

| Training \ Test | Walking | Streetcar | Office work | Lecture | Meeting | Cafeteria |
|---|---|---|---|---|---|---|
| **Walking** | 98.7 | 100.0 | 1.5 | 5.1 | 16.8 | 58.9 |
| **Streetcar** | 80.7 | 100.0 | 0.1 | 0.2 | 3.8 | 68.2 |
| **Office work** | 19.3 | 0.0 | 99.9 | 99.7 | 96.2 | 31.8 |
| **Lecture** | 26.6 | 4.0 | 90.2 | 99.9 | 87.6 | 34.3 |
| **Meeting** | 39.3 | 20.2 | 97.1 | 98.7 | 99.9 | 30.2 |
| **Cafeteria** | 68.7 | 100.0 | 8.8 | 6.9 | 18.0 | 99.9 |

### C. Specialized predictors can be misleading

Our second claim is that specialized predictors seem to perform better. In other words, the more homogenous the learning/prediction sets (i.e., the fewer situations it covers) the higher the prediction accuracy is likely to be. To support this claim we evaluated our learning algorithms within all possible subsets of situations. Figure 2 shows the average overall prediction performance of different subset sizes (from the 6 subsets with only one situation to the one subset with all situations).

At first, consider the results of the models based on accelerometers. When the subsets included only one situation then the predictions where highly accurate (at 0.98%). The more situations are added, i.e., the more heterogeneous the model gets, the less predictive the model becomes. This, again, emphasizes the need for a scenario. First, it highlights how *mere reports of accuracies are problematic, as it makes it impossible to assess the variability of the situations underlying the reported results*. The results could be based, for example on two situations, sitting in a movie theater and sitting in a streetcar, both of which, typically, have homogenous interruptability and are simple to differentiate with a light sensor. Second, without any information about the variability within and between the situations of a scenario it is almost impossible for the reader to assess how a given mechanism might fare in another environment. This is especially true, as the practical application of a prediction mechanism is likely to confront it with many heterogeneous situations, for which the accuracies of a predictor learned form specialized data are *misleading*.

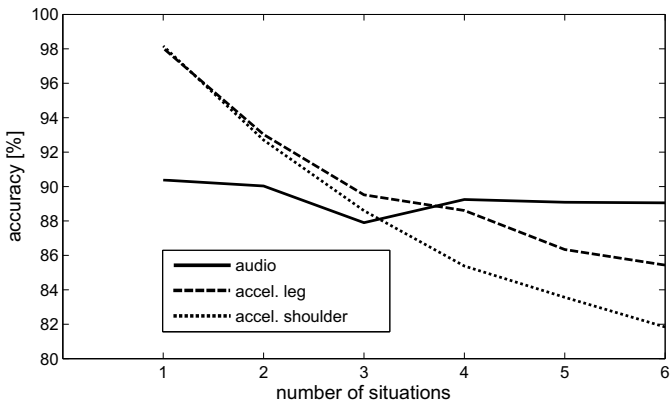


Fig. 2. Specialization accuracies

### D. Sensor robustness towards situational variability

Figure 2 reveals a second interesting finding: in our scenario the predictors based on the audio sensor were much more robust towards situational variability than the ones based on accelerometers. This raises the issue whether the accelerometers are suitable for our task or whether their omission might lead to better results, which has been stated in our fifth claim. This hints at a basic underlying machine learning problem (called over fitting) that sometimes more features for learning might actually misguide the induction algorithm and some sampling of the feature space might improve the algorithm's performance. For our particular problem we can conclude that accelerometers don't seem to provide a highly discriminative measure for interruptability prediction in the used scenario. This shows again how crucial it is to report the application scenario as a part of the results. We will revisit this issue in subsection 3.7 below.

### E. Generalizability of predictors

The ultimate question for any (interruptability) prediction approach is how well it generalizes to situations not encountered in the training phase. We were, therefore, especially interested how our setup would fare in situations not included in the training set. To that end we took our overall dataset and analyzed it as follows. First, we trained the models on one situation and evaluated the resulting model on all the others. Second, we trained on two situations and evaluated the model using the remaining ones, and so forth. The results of this analysis can be found in Figure 3, which shows a similar behavior for all sensor streams. At first, as expected, the results improve the richer the situational variability is in the training set. Interestingly, however, the prediction performance degrades when including a fourth situation. This is due to the symmetry in the experimental setup. In the closed world of the scenario there is only a limited set of situations. Assuming that the situations in the scenario actually provide different sensory readings, the training set starts to be increasingly dissimilar to the test set with every situation that is moved from the test to the training set. With other words, the figure shows two counteracting effects. The more situational variety in the training set the better the model becomes. The less situational variability in the test set the more exceptional its sensory readings become and the more difficult it is to learn a good predictor for it from other situations. Nevertheless, our results are very interesting. First, we see that it is actually possible to *predict never before encountered situations with a fair accuracy of 70%*. Note that this result is better than it looks as the overall performance of the predictor, i.e., the performance including predictions in situations similar to the ones in the training set, is likely to be much better. Second, our evaluation takes place in a somewhat artificially "closed world," where the situational variety is limited. In the real world, the situational diversity is always (much) higher than in the test set. As a consequence, we can expect the improving trend to continue beyond just three situations. Nonetheless, we have to expect that there will always be exceptional situations not covered in any training set, which will lead to miss-

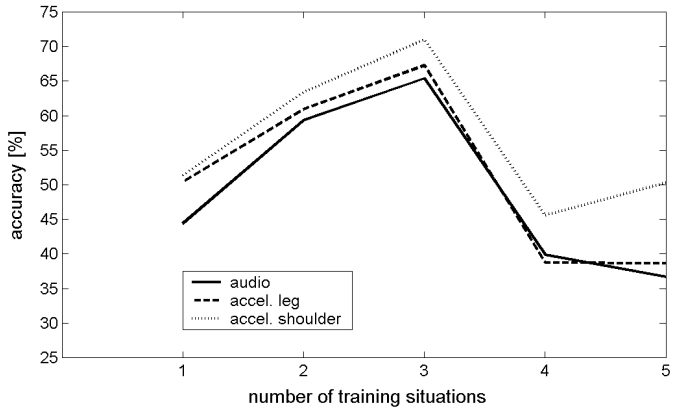


Fig. 3. Generalizability accuracies

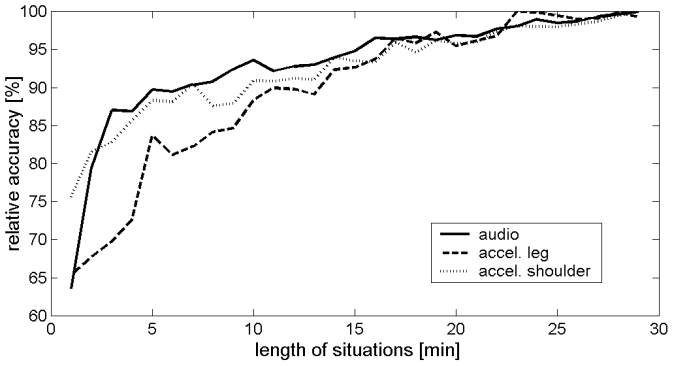


Fig. 4. Relative accuracy gains with scenario-based stratified training

classifications.

Summarizing, we have found that our approach generalizes with acceptable success to situations not encountered in the training set. This analysis shows again how important an underlying scenario is, as it would be nearly impossible to investigate situational generalizability of predictors without a scenario. It also shows that any assessment of generalizability without a scenario is ”doomed” to fail, as it is impossible for the reader to assess the correspondence of situations she might think about with situations in the test and/or training set. Thus, prediction accuracies can only be understood in the context of their evaluatory scenario.

### *F. Scenario-based, stratified training*

The preceding analysis raises the question whether a scenario can inform the selection of data for model induction. In particular, one could hypothesize that if situations *within* a scenario provide a somewhat homogenous data stream then one should be able to plan the initial training data collection using the scenario as a guide to quickly gain good induction performance. In other words, a small sample of each situation within a typical application scenario should be sufficient to learn a well performing model.

To investigate this hypothesis we looked how the prediction accuracy of our model changed when varying the amount of training data available for the scenario situations. We started with randomly chosen one-minute segments from each situation to train a model and determined the accuracy of its performance on the remaining data. We then repeated this evaluation continuously increasing the training segment size by 1 minute until a total of 30 minutes was reached. To avoid non-representative outliers in the trainings set we performed 5 evaluation runs on 5 randomly selected segments from each situation and averaged the reported results. The results of this analysis are shown in Figure 4, which graphs the relative accuracy for each sensor in respect of the best achievable overall accuracy. Thus the figure essentially shows how fast the model climbs the learning curve, i.e., achieves its best attainable model.

The results we found are very promising. Using only a 3 minute audio sample of each situation was enough to attain about 87% of the overall accuracy. The two accelerometer sensors also provide good results, but climb the learning curves slower. A good scenario, therefore, provides the basis for a shortened stratified learning scenario, which is used to learn large percent of the attainable accuracy only with brief situational example segments. Further improvements can then be achieved via relevance feedback within use. Such an approach would significantly lower the barrier for using wearable devices as users oftentimes want to see the benefits of their tools early - otherwise they grow frustrated and don’t bother using it at all [9], [10]. A good scenario could, thus, be the foundation to shorten the deployment time of a wearable interruptability prediction device, akin to the introductory training texts used to train speech recognition software.

The results so far have shown how important a scenario can be for determining the robustness of a wearable interruption prediction mechanism. We have seen how a scenario can assure situational variance, provide context for reported accuracy figures, which ultimately allow evaluating the generalizability of an interruptability prediction approach. Furthermore, we have seen how a scenario can provide the basis for a fast stratified model-learning procedure speeding the learning curve of the prediction models and, thus, lowering the barrier to the approach’s practical use. Our experiment also gave rise to some interesting, scenario-independent findings, which we are going to discuss in the next two sub-sections.

### *G. Are accuracies enough? - The dominance of audio sensors for predicting interruptability*

One issue with interpreting the results of papers about interruptability prediction is that they all report accuracies. Accuracies are problematic: they assume that the cost-tradeoffs between false positives and false negatives (for two class problems) are known at learning time. Most papers we found didn’t report any assumption about the cost of misclassification in which case we have to assume that they implicitly used an equal cost for both false positives and false negatives. Obviously, this doesn’t reflect the application domains. Users typically associate different costs with the various types of misclassification. Wrongly classifying a wanted phone call (such as a surgery request) as unwanted, for example, has a higher cost then occasionally classifying a telemarketer as wanted.

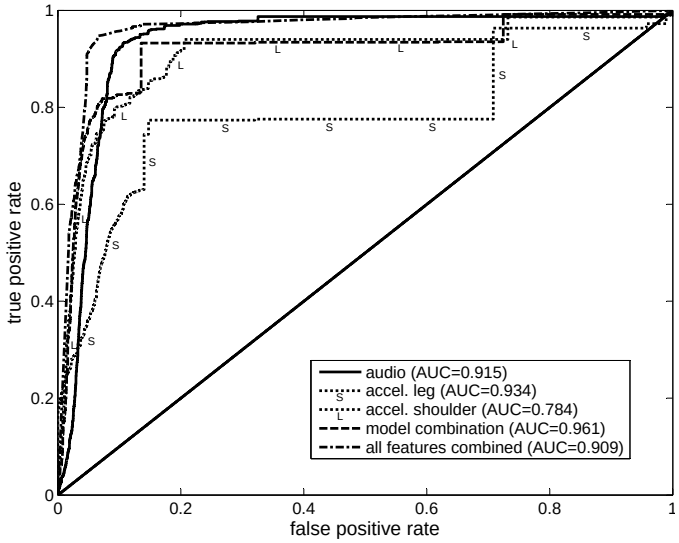


Fig. 5. ROC curves for all sensors

As an alternative to reporting accuracies one could use the receiver operating characteristics (ROC) curve - a measure from signal theory that recently has gained high acceptability in machine learning [11]. 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. Cost is related to the angle of a tangent resulting in a cost-based cutoff for classification using class probability estimators. As a consequence, ROC curves allow comparing two classification approaches independent of any cost assumptions. The area under the ROC curve (AUC) supplies a single number measure for the overall, cost-independent performance of a classifier.

As we can see from Figure 5 the ROC curves differ between the models. Apart from the area left of a false positive rate of 0.08 the audio model clearly dominates both accelerometer models. Only when the cost of a rejected but desired interruption is more than 1.66 times the cost of an accepted, but undesired interruption, then the accelerometer model provides better predictions (for calculation details see [12]). Even when combining the audio with the accelerometers' data the audio model still provides excellent (dominating) results. Figure 5 shows two such combinations: one based on building a classifier based on a combination of all features, another based on combining the three models' prediction using a nave Bayes classifier, which the literature [12] predicts to be superior. But even the feature combined model is outperformed by the audio model, which provides a larger area under the curve. Only the model combination outperforms the audio model, drawing on the accelerometer data to improve the AUC to 0.961 and the accuracy to an excellent 93.7% (see confusion matrix in Table IV).

Summarizing, we can say that reporting ROC curves is superior to accuracies and confusion matrices. We have also seen that under most reasonable cost-assumptions the audio model dominates the accelerometer data stream and only a model combination improves over the audio model's performance. Adding the only somewhat improved accuracy in the combined model and the poor generalizability of this sensor stream (see section 3.3.) raises the question what sensors might be more suitable in our environment for interruptability prediction than accelerometers.

## IV. Related Work

There are essentially two groups of studies that relate to ours: interruptability investigation in the stationary and wearable setting. Belonging to the first group, Horvitz et al. use Bayesian networks based on information stored in a user's calendar including the status of properties of appointments as well as information generated by interactions with computing devices to predict the "relevance" and cost of interruptions [2], [13]. [1] equipped an office with physical sensors such as microphones, magnetic switches determining whether the door is open, motion sensors, etc. Using self-reported interruptability annotations they found that the audio streams where the most predictive sensors with accuracies between 80.1% (for an intern) and 87.7% (for a manager) for a two class prediction (highly non-interruptible and other). These results can be compared with our overall accuracy of 90.6% accuracy for the audio measure. Note that the difference might be both in the choice of features and algorithms. A full comparison would, therefore, require further analysis.

The second group considers wearable setups. The SenSay project [14] connects a notebook to a cell phone, which is connected to a number of sensors like "audio," "accelerometers," "temperature," and "visible light". Another wearable computing platform is the audio-only Nomadic Radio [15] that uses speech recognition, message priority, as well as a contextual notification model to define when a message should be posted on the user's heads-up display. The most similar to our efforts is the project by Kern and Schiele [4]. They connect accelerometers via a sensor module to a laptop. Using these sensory data they predict the activity of the user as defined by the activity classes "sitting," "standing," "walking," "walking up-stairs," "walking down-stairs," and "running" with an accuracy of 86.5%. Furthermore, they classify a person's social context as defined by the states "street," "restaurant," "lecture," and "conversation" using a microphone with an accuracy of 83.17%. Combining these two classifications they infer a tendency for a person's interruptability. A comparison with our approach is difficult, as both our annotation structure and prediction goals (interruptability vs. activity setting/social context) were different.

TABLE IV
Model combination confusion matrix. "I" stands for "interruptable".

| | | Predicted | |
|---|---|---|---|
| | | I | ¬ I |
| Reported | I | 20.8% | 1.2% |
| | ¬ I | 5.1% | 72.9% |

accuracy = **93.7%**

## V. Conclusion and future work

In this paper we introduced a method for accurately predicting a person's interruptability directly from simple sensors without any intermediate steps/symbols. The direct prediction seems to be competitive or even superior to indirect prediction methods and we have not observed any drawbacks yet.

Based on the data gathered we showed that the problem of predicting contextual status from sensory reading can greatly be helped by using an application scenario. It provides a basis to (1) ensure that the *situational diversity* of the training set corresponds to the real-world application, (2) avoid *misleading results of specialized predictors*, (3) assess the *generalizability of approaches* to new situations, and (4) provide the basis for a *stratified scenario-based learning model*, which can greatly speed-up the training time of a wearable predictive device. Finally, using an ROC-analysis, we illustrated the *superiority of the audio sensor stream for our task* raising the question of other suitable sensors.

In the future we intend to focus on two limitations of our current approach: generalization and computational efficiency. The first issue relates to the problem that our current data only stems from one subject. We intend to extend our investigation to multiple subjects over long periods of time to further validate the generalizability of our findings beyond the current single-subject setup. This would, furthermore, allow the investigation of whether models learned for different subjects can be related to each other. Recruiting multiple 'real-world' subject, however, would require a size (and weight) reduction of our data acquisition setup, as only few people would want to wear the (somewhat intrusive) sensors over extended periods of time. The second issue relates to computational efficieny: we intend to investigate whether different algorithms attain similar prediction quality while limiting the computational requirements – a prerequisite for the approach's widespread use and size reduction of the experimental setup.

In closing we would like to highlight that the prediction of a person's interruptability based on contextually collectable information essentially consists of two related sub-problems: the collection of suitable data streams and the efficient inference from these data. While vast advances have been made in the field of machine learning regarding the inference, the problem of which data to collect is still in its infancy and needs further exploration.

## VI. Acknowledgements

We would like to thank Patrice Egger for his substancial support in the initial stage of this project. We would also like to thank B. Schiele and N. Kern for sharing their audio preprocessing code. Furthermore, we would like to thank Haym Hirsh for his feedback.

## References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proceedings of PERVASIVE 2004 (LNCS 3001)*. Berlin Heidelberg: Springer-Verlag, 2004, pp. 1–17.

[6] H. W. Gellersen, A. Schmidt, and M. Beigl, "Multi-sensor context-awareness in mobile devices and smart artifacts," *Mob. Netw. Appl.*, vol. 7, no. 5, pp. 341–351, 2002.

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

[8] R. Pfeifer and C. Scheier, *Understanding intelligence*. MIT Press, Cambridge, Mass., 2000.

[9] J. Grudin, "Groupware and social dynamics: eight challenges for developers," *Commun. ACM*, vol. 37, no. 1, pp. 92–105, 1994.

[10] ——, "Group dynamics and ubiquitous computing," *Commun. ACM*, vol. 45, no. 12, pp. 74–78, 2002.

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

[12] K. M. Ting and B. T. Low, "Model combination in the multiple-data-batches scenario," in *ECML '97: Proceedings of the 9th European Conference on Machine Learning*. London, UK: Springer-Verlag, 1997, pp. 250–265.

[13] 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.

[14] 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.

[15] 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.

**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.

**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.