Large-Scale Evaluation of Call-Availability Prediction

Martin Pielot
Telefonica Research, Barcelona, Spain
martin.pielot@telefonica.com

ABSTRACT

We contribute evidence to which extent sensor- and contextual information available on mobile phones allow to predict whether a user would pick up a call or not. Using an app publicly available for Android phones, we logged anonymous data from 31311 calls of 418 different users. The data shows that information easily available in mobile phones, such as the time since the last call, the time since the last ringer mode change, or the device posture, can predict call availability with an accuracy of 83.2% (Kappa = .646). Personalized models can increase the accuracy to 87% on average. Features related to when the user was last active turned out to be strong predictors. This shows that simple contextual cues approximating user activity are worthwhile investigating when designing context-aware ubiquitous communication systems.

Author Keywords

Availability; Phone Calls; Mobile Phones; Prediction; Interruptibility; Attentiveness

ACM Classification Keywords

H.5.m Information interfaces and presentation: misc.

INTRODUCTION

Incoming mobile phone calls aren't always on time. Previous work shows that about one third of all calls are missed [2, 18]. Prominent reasons are that users don't hear the phone, cannot pick up the call for social reasons (*e.g.*, being in a meeting), or prefer to focus on other activities, such as sleeping or playing games [18].

In our recent work [13] on monitoring the effect of mobile phone, we argued that providing better management of the expectations towards responsiveness by communicating recipient (non-)availability is one of the most important strategies in lowering the severity of interruptions in computer-mediated communication. However, how to automatically communicate (non)availability has not yet been shown.

Previous research has explored the concept of availability in two ways: first, learning through subjective feedback, what affects a user's availability to phone calls [4, 5, 11, 20] – with the goal of understanding which contextual factors are strong predictors for call availability; and second, predicting interruptibility through sensors and other information availability on ubiquitous devices in the context of interruptions by people, emails, messengers, or phone notifications [1, 6, 9, 10, 14, 15].

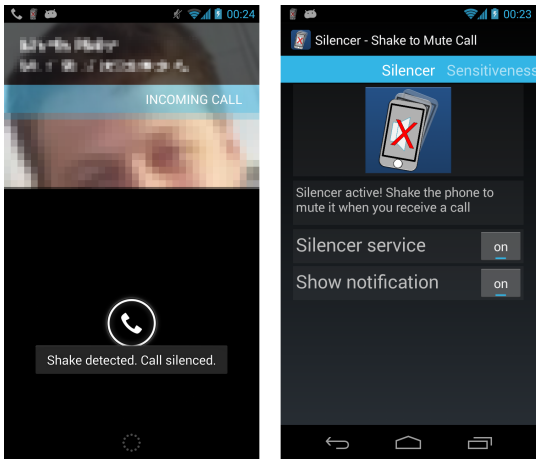

Figure 1. The Silencer's main view, and showing a "toast" message right after having silenced the call.

The aim of this work is to investigate to what extent sensor- and contextual information available on mobile and ubiquitous devices can predict availability to phone calls. We report from a large-scale study, where we published a phone-call handing app called *Silencer* on Google Play. Via this app, we collected data on how 418 users handled 31311 calls, alongside contextual information available through the phone itself. The contribution of this note is three-fold:

1. the first validation of previous work on call availability prediction in the large, *i.e.*, in a natural setting with hundreds of users from a diverse sample.
2. evidence that features extracted from contextual information and common sensors in mobile phones achieve respectable accuracy (accuracy = 83.2%, *Kappa* = .646) in predicting call availability.
3. evidence that personalized models can increase accuracy to beyond 87.0% (*Kappa* = .640).

RELATED WORK

Interruptibility has been extensively studied for personal interruptions [9], emails [10], phone notifications [6, 15], IM

UbiComp '14, September 13 - 17, 2014, Seattle, WA, USA

http://dx.doi.org/10.1145/2632048.2632060

[1], and mobile messaging [14]. These studies show that it is possible to use sensors and contextual information to estimate whether a user will be available for communication attempts through these channels.

With respect to phone calls, De Guzman *et al.* [5] found in a 4-week diary study with 13 participants that, before making a call, people desire to know detailed contextual information, including location, time, physical, social, and emotional availability, and current activity. Knittel *et al.* [11] surveyed 132 phones users and found that people are only willing to share some contextual information, such as current location, current activity, or presence of appointments.

What contextual factors matter for call receivers in deciding whether to take a call has been studied as well:

Danninger *et al.* [4] conducted a pilot study, where 4 lab colleagues frequently responded to experience-sampling questionnaires administrated via their PC. They found that that location, time, activity, presence of others, level of engagement and importance & urgency of the current activity are amongst the features that matter most.

Ter Hofte [20] conducted a one-week experience-sampling study with 10 colleagues. The questions focused on current activity and presence of other people, and whether the participants would be taking a call in this situation. Trained on the basis of 750 samples, a model could predict with 63.9% accuracy if a person would report to be available for a call.

Horvitz *et al.* [8] explored to what extent meeting details, such as meeting duration, subject, location, or organizer, can predict the cost of interruptions from phone calls. Informal tests with 2 colleagues over 4 months indicate that the system performed well.

One of the most closely related works is by Rosenthal *et al.* [16]. On the Android phones of 20 students, they logged features identifying context (*e.g.*, location, time of day, ...) and importance of the alert (*e.g.*, whether the contacting person is in the favorites list or how often this person had contacted the user in the past), and used experience sampling to learn the user's preferences regarding whether the phone should have been muted. In a two-week follow-up study, they tested the model and automatically muted the phone. 13 of 19 users reported to be satisfied with the classification.

Recent work by Boehmer *et al.* [2] presents a novel UI for handling interrupting calls more smoothly when the user is active in another app: instead of switching to the full-screen call UI, call-handling buttons are displayed on the screen corners as an overlay, which allows people to finish their tasks in the current app.

In summary, most studies on predicting availability were done with only 2 - 20 participants – often recruited amongst university students or co-researchers. Further, contextual information and ground truth were mostly established from subjective feedback, which humans cannot always assess properly. What is missing is a study with a larger and more representative sample, that uses objective data as ground truth and for approximating the user context.

METHOD

To investigate predicting availability to phone calls on mobile phones, we conducted a large-scale in-the-wild study [7, 12, 17]. That is, we published a call-handling application called *Silencer* on Google Play and studied actual phone calls. For each call, we logged how the user handled it, and the current context of the user, approximated by information and sensor data available to the mobile phone.

The Silencer Application

The Silencer is available for free on Google Play for Android phones. It allows users to temporarily mute the phone ringer on incoming calls by simply shaking the phone. Figure 1 shows the main configuration view and a message telling that the call has been silenced.

The app was published end of 2012 and, by the time of writing (May 2014), has been downloaded 14849 times from Google Play. The study logger software was added in March 2014 with a major update. We added a paragraph informing users that anonymous data will be collected to investigate the call handing of the Silencer. Users had to manually agree to the added permissions in order to update the application.

To avoid silencing the phone unintentionally, *e.g.* when running or driving in a car, the Silencer uses call-based shake-sensitivity adaptation. During the first two seconds of each call, the accelerometers simply record the level of acceleration, but the call cannot yet be muted. The maximum acceleration observed during this period multiplied by 1.5 serves as threshold. Only if the acceleration exceeds this threshold after the calibration period, the phone is muted.

Data Collection

For each call, the app logged the following 15 features:

- the ringer mode and when it last changed,
- charging state and when the phone was last (un)plugged,
- the screen state (on/off) and when it last changed,
- the day of the week,
- the hour of the day,
- the proximity sensor (display covered/not covered),
- the pitch angle of the display,
- the level of acceleration right before the call,
- how often the same caller had called before,
- the time since the last call,
- whether the last call was picked up,
- whether the last call was silenced,
- and whether the user took this call (ground truth).

These features comprise those sensors and contextual information related to calls and user activity that are easy to access on Android phones without requiring excessive permissions or computation.

To save battery, the application only monitors changes to the ringer mode and the screen in the absence of a call, as it is

important to know the status of these sensors prior to the call. All other used sensors are activated only on incoming calls.

Participants

Within a time-frame two months, we collected phone calls from the phones of 418 distinct users. The phones reported timezones from all around the world. 48 different locales were reported, the five most frequent being: en-US (28.8%), en-GB (12.9%), ja-JP (12.6%), de-DE (4.7%), and ar-AE (4.4%). Hence, the participants represent a diverse sample of Android users from a wide range of places around the world.

RESULTS

To clean the data, we initially removed those calls where it was likely that the user simply had tested the application. Thus, we excluded all calls that were received less than 3 hours after installation, and where strong acceleration forces larger than 10G were observed in the 2-sec calibration phase.

The resulting data set contains 31311 calls. 19175 (61.2%) of the calls were picked up, 8816 (28.2%) were missed, and 3320 (10.8%) were muted by shaking the Silencer and then left ringing. These numbers are in line with previous findings that about one third of all calls are missed [2, 18].

The median time interval between the first ring and picking up the call was 8.0 seconds.

Classes and Classifier Selection

We consider the user *available* if the user picked up the incoming call, and *unavailable* if the user missed the call or muted it. This allowed us to treat the problem as a simple classification task with 19175 (61.2%) instances of *available* and 12136 (38.8%) of *unavailable*.

We tested and empirically compared the performance of a wide range of well-known classifiers that are available in Weka[1]. For all tests, we used 80% of the data as training set and 20% as test set. Thus, our results show how well the model can predict availability from context that has not yet been seen by the classifier. We obtained the best performance with Random Forests [3], and thus used them throughout the remaining analysis.

Classification Accuracy

With standard configuration $seed = 1$ and 10 trees, the Random Forest model achieved an accuracy of 83.2%. As suggested by Strobl *et al.* [19], we confirmed that varying the configuration achieves similar results: between 82.77% and 84.03%. The Kappa statistic of .646 indicates that the classifier performs 64.6% better than a random guess. As comparison, using Naive Bayes as classifier achieves an accuracy of 69.2%. Table 1 shows the confusion matrix.

available	unavailable	← classified as
3364	480	available
569	1849	unavailable

Table 1. Confusion Matrix for the classification.

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

Table 2 shows detailed accuracy by class. Precision and recall indicate that the algorithm performs better for identifying those instances where the model predicts *available*. From an application perspective, this is desirable, as we consider it worse if the system tells a caller that the receiver will pick up the call, but in fact does not. With a precision of 85.5%, the model will be correct roughly 5 out 6 tries when predicting that a user will pick up a phone call with a given context.

Class	Precision	Recall	ROC
available	.855	.875	.899
unavailable	.794	.765	.898

Table 2. Detailed accuracy by class.

We empirically tested randomly-selected subsets of the ground truth data. With roughly 500 calls the model starts to approach the 80% accuracy mark. However, our tests did not reveal that the accuracy flattens out beyond any points.

Feature Ranking

To understand the importance of the individual features, we ranked them using Weka's "select attributes" facility. As feature evaluator, we used *ClassifierSubsetEval*, which "evaluates attribute subsets on training data or a separate hold out testing set. Uses a classifier to estimate the 'merit' of a set of attributes." As search method, we used *GreedyStepwise*, which works by "traversing the space from one side to the other and recording the order that attributes are selected." The feature evaluation used 10-fold cross validation.

Table 3 shows the generated average ranking of the features. Average merit indicates the average loss of accuracy when removing the given feature. Average rank indicates at what rank the feature evaluation determined for the given feature during each of the 10 folds.

Feature	avg. rank	avg. merit
Last ringer change (time)	1.0	−.074
Last screen change (time)	2.0	−.026
Screen status	3.6	−.018
Last (un)plugged (time)	5.4	−.014
Last call (time)	6.8	−.013
Activity / Acceleration	7.3	−.013
How often called by caller	7.6	−.017
Day of the week	9.4	−.012
Charger (un)plugged	10.0	−.012
Hour of the day	10.1	−.012
Ringer mode	11.4	−.011
Last call silenced	12.4	−.011
Pitch of phone	12.5	−.011
Screen (not) covered	13.0	−.011
Last call picked	14.1	−.011

Table 3. Ranking of the tested features.

All top-five features are an indication whether the user is or has just been active with the phone. Ranks 6 to 10 further introduce the relation to the caller and proxies to the current activity. Thus, current activity and relationship to the caller appear to be the strongest factors that can be best inferred from standard phone sensors.

Since Random Forests sometimes performs better when reducing the number of features, we tested removing low-ranking features. A model built from the top-10 / top-5 features achieved an accuracy of 81.41% / 79.62% respectively. This indicates that even the bottom-5 features are still notably contributing to the overall accuracy.

Personalized Models

Next, we explored the potential gain from personalizing the model. We computed models for each users built from 10, 20, 30, ... 190 calls. If a user did not receive that many calls during the study period, the data for this user was discarded. For each of the personal models, we computed its accuracy and the kappa value via 10-fold cross validation. With 50 calls, the personalized models, on average, started to outperform the generic model with an accuracy of 84.0%. With 120 calls, the personalized models exceeded the generic models Kappa (.64) and clearly outperformed the generic model in terms of accuracy 87.02%. This indicates that personalizing the model is a vital approach to further increase accuracy.

DISCUSSION

Our data set shows that predicting availability from sensors and device information is not only possible for personal interruptions [9], instant messaging [1], and phone notifications [15], but also for phone calls.

With an accuracy of 83.2% for the generic model and over 87.02% for the personalized models, our approach outperforms previous work [5], which reports 63.9% accuracy from a model made from subjective *in-situ* feedback. Personal models matching the generic model's accuracy can be generated from as little as 50 calls. This shows that automatically communicating non-availability, as proposed before to manage expectations [13], is feasible.

Previous work [4, 5, 20] has suggested several hard-to-measure features, such as the nature of the current activity, for predicting availability to mobile phone calls. Our work highlights that availability-prediction models with satisfying accuracy can be built by relying on much simpler features.

The strongest predictors were those that approximate user activity, either physical or with the phone, and approximation of daily routines. This is in line with previous work [1, 5, 20], which found that features related to user activity are strong predictors of availability. Advancing prior work, we show that phone sensors can be viable proxies for user activity.

Previous work also has shown that people desire to know a wide range of contextual information before making a phone call [4, 11]. However, sharing this information can raise privacy concerns and is not always desired [11]. Our findings show that a more privacy-protected approach is feasible as well: instead of sharing detailed information, such as the current location, a service could simply share the prediction and optionally the confidence of the prediction.

One open question is how to implement this service: continuous prediction would currently be too battery intensive. Obtaining a prediction during a call or on request might be more feasible, but would require an explicit trigger by the caller.

As for all large-scale studies, one limitation of the presented work is the lack of control over the tasks [2]. Neither researchers nor users control, *e.g.*, how often and in what situations users received calls, and whether the measured reactions align with the actual intent. For example, our data set may be sparse on certain situations and users may at times have reacted differently from how they felt. Nevertheless, by running the study in a naturalistic setting, we obtain data from calls when they actually occur, and we forgo any bias arising from inaccurate self-judgment.

The findings advance prior work in the following ways: compared to Horvitz *et al.* [8], who explored to mute phones according in meetings according to the details in the calendar, our work covers a wider range of features and situations. Compared to Rosenthal *et al.* [16], we did not measure whether users do not want to be interrupted, but whether they would pick up a call. With respect to both works, our study comprises a much larger and more diverse sample, which yields to insights with comparably high ecological validity and generalizability.

CONCLUSIONS

On the basis of a large-scale study with 418 phone users, we show that monitoring 15 features on a mobile phone, which approximate user activity and state, allow to predict with 83.2% accuracy if a person will pick up a call or not. When computing individual models for users with more than 120 calls, those individual models achieved an accuracy of 87%. It shows that the previously proposed approaches to automatically predict availability for personal interaction, emails, notifications, and messaging, can be applied to phone calls too.

The findings presented in this paper demonstrate the viability of automated, lightweight call availability predictors for mobile phones. In contrast to previously proposed solutions of sharing detailed contextual information, this allows to share availability information with potential callers in a privacy-preserving way, as no detailed context information, such as current location or recent call activity, has to be shared.

By restricting our investigation to simple features, we leave room for future improvements, such as testing more complex features, or investigating how to approximate hard-to-measure features, such as "the importance of the current activity", as proposed in more human-centered approaches, in automated ways.

ACKNOWLEDGMENTS

We sincerely thank our participants for their contributions! Special thanks goes to Aleksandar Matic who encouraged the submission of the work.

REFERENCES

1. Avrahami, D., and Hudson, S. E. Responsiveness in instant messaging: predictive models supporting inter-personal communication. In *Proc. CHI '06*, ACM (2006).
2. Boehmer, M., Lander, C., Gehring, S., Brumby, D., and Krueger, A. Interrupted by a phone call: Exploring designs for lowering the impact of call notifications for smartphone users. In *Proc CHI '14*, ACM (2014).
3. Breiman, L. Random forests. *Machine Learning 45*, 1 (2001), 5–32.
4. Danninger, M., Kluge, T., and Stiefelhagen, R. Myconnector: Analysis of context cues to predict human availability for communication. In *Proc ICMI '06*, ACM (2006).
5. De Guzman, E. S., Sharmin, M., and Bailey, B. P. Should i call now? understanding what context is considered when deciding whether to initiate remote communication via mobile devices. In *Proc GI '07*, ACM (2007).
6. Fischer, J. E., Greenhalgh, C., and Benford, S. Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proc. MobileHCI '11*, ACM (2011).
7. Henze, N., and Pielot, M. App stores: External validity for mobile hci. *interactions 20*, 2 (Mar. 2013), 33–38.
8. Horvitz, E., Koch, P., Sarin, R., Apacible, J., and Subramani1, M. Bayesphone: Precomputation of context-sensitive policies for inquiry and action in mobile devices. In *Proc UM '05* (2005).
9. Hudson, S., Fogarty, J., Atkeson, C., Avrahami, D., Forlizzi, J., Kiesler, S., Lee, J., and Yang, J. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proc. CHI '03*, ACM (2003).
10. Iqbal, S. T., and Bailey, B. P. Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Trans. Comput.-Hum. Interact. 17*, 4 (Dec 2010), 15:1–15:28.
11. Knittel, J., Sahami Shirazi, A., Henze, N., and Schmidt, A. Utilizing contextual information for mobile communication. In *Proc CHI EA '13*, ACM (2013).
12. McMillan, D., Morrison, A., Brown, O., Hall, M., and Chalmers, M. Further into the wild: Running worldwide trials of mobile systems. In *Proc Pervasive '10*, Springer-Verlag (2010).
13. Pielot, M., Church, K., and de Oliveira, R. An in-situ study of mobile phone notifications. In *Proc. MobileHCI '14* (2014).
14. Pielot, M., de Oliveira, R., Kwak, H., and Oliver, N. Didn't you see my message? predicting attentiveness to mobile instant messages. In *Proc. CHI '14*, ACM (2014).
15. Poppinga, B., Heuten, W., and Boll, S. Sensor-based identification of opportune moments for triggering notifications. *Pervasive Computing 13*, 1 (2014), 22–29.
16. Rosenthal, S., Dey, A. K., and Veloso, M. Using decision-theoretic experience sampling to build personalized mobile phone interruption models. In *Proc. Pervasive '11*, Springer-Verlag (2011).
17. Sahami, A., Henze, N., Dingler, T., Pielot, M., Weber, D., and Schmidt, A. Large-scale assessment of mobile notifications. In *Proc. CHI '14* (2014).
18. Salovaara, A., Lindqvist, A., Hasu, T., and Häkkilä, J. The phone rings but the user doesn't answer: Unavailability in mobile communication. In *Proc MobileHCI '11*, ACM (2011).
19. Strobl, C., Malley, J., and Tutz, G. An introduction to recursive partitioning: Rational, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods 14*, 4 (2009), 323–348.
20. ter Hofte, G. H. H. Xensible interruptions from your mobile phone. In *Proc MobileHCI '07*, ACM (2007).