Now where was I? Physiologically-Triggered Bookmarking

Matthew K.X.J. Pan, Jih-Shiang Chang, Gokhan H. Himmetoglu, AJung Moon, Thomas W. Hazelton, Karon E. MacLean, Elizabeth A. Croft
The University of British Columbia
Vancouver BC V6T 1Z4
maclean@cs.ubc.ca

ABSTRACT
This work explores a novel interaction paradigm driven by implicit, low-attention user control, accomplished by monitoring a user's physiological state. We have designed and prototyped this interaction for a first use case of bookmarking an audio stream, to holistically explore the implicit interaction concept. Here, a user's galvanic skin conductance (GSR) is monitored for orienting responses (ORs) to external interruptions; our prototype automatically bookmarks the media such that the user can attend to the interruption, then resume listening from the point he/she is interrupted. To test this approach's viability, we addressed questions such as: *does GSR exhibit a detectable response to interruptions,* and *how should the interaction utilize this information?* In evaluating this system in a controlled environment, we found an OR detection accuracy of 84%; users provided subjective feedback on its accuracy and utility.

Author Keywords
Physiological signals, human-computer interaction, interruption, orienting response, galvanic skin response.

ACM Classification Keywords
H.5.2 [**User Interfaces**]: Input devices and strategies, Haptic IO; I.5.5 [**Pattern Recognition Implementation**]: Interactive Systems

General Terms
Measurement, design, experimentation, human factors.

INTRODUCTION
Most modern user interfaces developed for consumer electronic products rely on explicit channels of command and control. That is, information is most often exchanged between users and devices through explicit gestures like key strokes, verbal commands and finger swipes, and received through visual or auditory channels. When well designed, this clear, controlled form of interaction has the important benefit of leaving little room for misinterpretation.

However, there are circumstances in which the demands and costs of an explicit interface encumber its utility and efficiency, most typically in multitasking scenarios [1]. In a time- and safety-critical example, steering and handling a vehicle often requires most of the driver's attention; yet a variety of secondary vehicle controls (e.g., adjusting audio system settings) compete for the same attentional resources. The resulting diversion can impair safety. In other cases, the cost may be in convenience, privacy or attentional fragmentation: a user may wish to alter a device's behavior, but cannot or prefers not to use explicit commands (e.g., answering the phone while washing dishes). A badly timed device notification may temporarily dip in priority and then be forgotten (e.g., a pop-up notification).

Theory
Foreseeing the rise of ubiquitous computing, Weiser and Brown argued that rather than abruptly demanding a user's focus, interactions should slide transparently between the attentional periphery and center [2], "calmly" providing context and orientation.

In this vein, we propose a fundamentally different approach to device interaction, which employs implicit control channels to reduce an interface's attentional demand. Building on the knowledge that some human affective (emotional) states can be estimated from physiological signals captured with off-the-shelf biometric sensors [3], our implicit channel is a user's voluntary and involuntary physiological responses to a situation. We incorporate this biometric response directly into the application's interaction loop, such that it responds to changes implied by the user's affective state (Figure 1). The loop is closed with an immediate but unintrusive indication of system recognition and response to the user's estimated state; in ongoing work, we are exploring haptic feedback for this role [4,5].

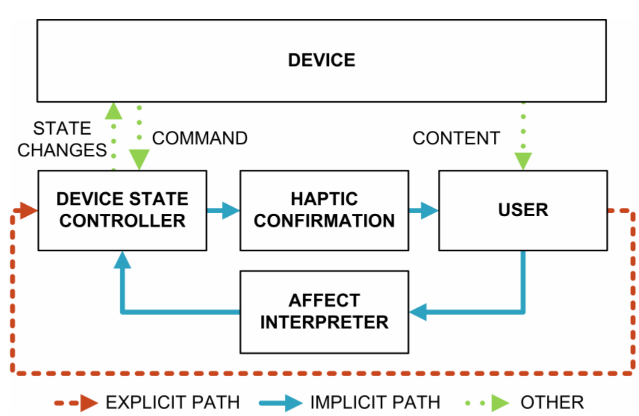

Figure 1. Proposed implicit interaction loop structure.

Approach and Use Case Description

We are interested in the utility and effectiveness of this implicit control paradigm in the context of a simple but representative use case, illustrated by this scenario:

A woman listens to an audiobook on her portable media player in a dentist's waiting room. Her name is called; her focus shifts from the audiobook to the hygienist. She fumbles in her bag to pull out the device, and eventually finds the "pause" button. The hygienist waits.

Or,

… her focus shifts. Reluctant to make the hygienist wait, she pulls out her earbuds and gets up, leaving her player running until the book ends. Later, she iteratively scrolls and listens backwards through the audio stream, trying to find the last familiar point and estimating forwarding intervals when she overshoots.

Both trajectories highlight how an explicit interface can 'demand' a user's attention, with negative outcomes if attention is not given. But what if instead:

… her focus shifts. An electrode slipped over her finger detects this and marks the audio stream, which continues to play. She feels a small tap on her finger that signals bookmark placement. She pulls out the earbuds and follows the hygienist. Later when alone again, she jumps back through two or three placed marks, locates the one placed just before the interruption where she stopped listening, and continues playback from there.

This is the implicit interaction method: an 'auto-bookmarking' system detects and marks a user's orienting response (OR) to an interruption by monitoring galvanic skin response (GSR) in real time. Bookmarks placed near the estimated point of interruption can be accessed later via a graphical or haptic tool. This use case highlights the example of a mobile application used in a predominantly hands-free mode. At this stage, we do not seek to change the underlying nature of existing explicit control or replace its channels; but rather to bypass points of dysfunction with a new, lower-effort channel when appropriate.

Contributions

This research example of attentional bookmarking, applied here to the initial use case of interrupted audiobook listening, defines a model for a unique user-device interface driven by implicit, low-attention control. With this platform, we have assessed the viability of a holistic interaction model that may be appropriate for a wide range of applications. We have focused on two questions pertaining to our implicit interaction model, with respect to the audiobook-listening use case:

1. *Does GSR exhibit a consistent and detectable response to interruptions in realistic contexts?*
2. *How should the system respond to this information?*

In this paper, we demonstrate empirically that GSR can be used to detect orienting responses to interruptions under controlled conditions, and have data suggesting that this ability can be extended to noisier, chaotic environments. We built an OR detection system based on this result, and to explore Question 2, made quantitative and qualitative observations of its use in a setup simulating audiobook listening scenarios. The latter effort gave us specific insights towards how best to refine our implicit interaction model in both this and more general cases. We explored the interaction paradigm without the haptic feedback component shown in Figure 1, as the feedback dynamics warrant a more in-depth examination than can be offered here.

In the remainder of this article, we summarize related work (`Related Work`), describe our hardware/software setup for detecting and displaying ORs (`Use Case System Description`), and present the experimental methods and results by which we quantified GSR-based ORs to interruptions and applied it to our interaction paradigm use-case (`Experiments`). We conclude with discussion and future work (`Discussion` and `Conclusions & Future Work`).

RELATED WORK

Affective State Detection and Implicit Interaction

Past work has related a human's measured physiology to affective state. This includes characterization of vision-based facial expression and auditory speech [6]; however, external manifestations (i.e., mouth and eye motions, voice pitch, etc.) can vary and be voluntarily suppressed. Physiological signal analysis is a popular alternative. Researchers have processed physiological sensor data and used filtering, statistical analysis and machine learning classifiers to derive models of affective user state [7-11]. Here, the difficulty lies in signal quality that differs by individual and day or hour, non-specific responses and other sensor noise.

In their early work in affective computing, Picard et al. classified emotions using physiological sensors [3,7,8]. With algorithms such as Fisher Projection and Sequential Floating Forward Search, they obtained emotional state recognition accuracy of up to 81% after 2-4 min of algorithm training per subject, for eight categories. Kim et al. aimed to reduce signal monitoring times and system training requirements through feature extraction and pattern classification of data (electrocardiography, skin temperature variation and electrodermal activity) pooled from multiple subjects [9]. A recognition rate of 78% for three emotion categories was obtained after 50s of monitoring.

There have also been initial attempts to use physiological classification to augment traditional interaction techniques. For example, Healey and Picard's StartleCam - a wearable video camera - monitors a user's GSR to detect when a user is startled [10]. When GSR indicates a user's heightened arousal, a time series of digital images are saved to mimic the user's "flashbulb" memory of the event causing the startle, autonomously generating an image-based diary for offline examination and memory assistance.

Our interaction model builds on these results with the aim of providing transition support between primary and secondary tasks. Beyond offline review, we seek a fluid continuous interaction where marked moments are used to propel the user towards a goal with minimal disruption.

GSR and the Orienting Response

The OR is an immediate reaction to the perception of a novel element or stimulus that is not sudden or intrusive enough to elicit a startle reflex. ORs are often examined to gather insights on human attention shifts and information processing [12]. An OR can be detected in many ways, including heart rate and electroencephalography (EEG), but the simplicity of measuring GSR and its strong relationship to OR is attractive.

The GSR (i.e., electrodermal activity) has been studied since the late 1800s. It is currently believed that the GSR is caused by the electrical activity of the sweat glands, and is connected to the sympathetic nervous system; it has been linked with the physiological instantiations of emotion, arousal and attention [13-16]. Firth and Allen showed that short-term changes in GSR reflect ORs [12]. Previous research has shown that within normal ranges of ambient room temperature and controlled subject state and motion, there is a high correlation between OR and GSR [14]. Other literature suggests that GSR measurements can be more easily discriminated than other physiological measures such as heart rate and EEG, since they can be detected quickly and without complex analysis [15].

GSR measurements are most sensitive on volar surfaces, suggesting a future possibility for sensors that can be worn unintrusively as a ring or shoe insert during daily activities. GSR sensors are inexpensive, wearable and pose no risks for the user. Disadvantages include latency in signal detection (i.e., 1-4s lag periods are common), significant response variance between subject groups (e.g., gender, age), possible habituation over time, and non-specificity [16].

We chose GSR as, on balance, the most promising physiological input to our control loop - keeping its limitations in mind. Since we require only OR information, with GSR we can bypass the complexity of a full affective model.

USE CASE DESCRIPTION

In our use case, marks are placed automatically during interrupted listening to an informative audio stream.

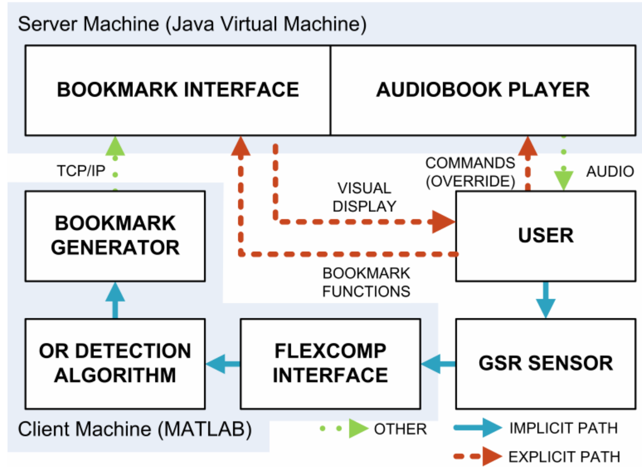

Figure 2. Diagram of bookmarking system architecture.

Bookmarks are placed near GSR-detected ORs, without explicit user direction. The marking system used here is intended as a platform to demonstrate and study the broader implicit interaction paradigm.

In this section, we describe our physical implementation. The "loop" is illustrated in Figure 2: the user's GSR was sensed, processed and sent by network to a control computer which analyzed the GSR stream, placed bookmarks, and made the bookmarks accessible to the user via a graphical list. A user can control the audio stream directly and/or via bookmark selection. Users could also place bookmarks manually ("Bookmark Functions" in Figure 2). Some parts of the system were developed specifically for this use case (e.g., the audiobook player and bookmark manager). Others, in particular the interruption detection algorithm and software, are of more general applicability.

GSR Apparatus and System Architecture

Our GSR measurements were obtained with Thought Technology's ProComp Infiniti® physiology-measurement hardware system [17]. The ProComp encoder reads data from a GSR sensor which uses dry electrodes attached to the index and middle fingers of the non-dominant hand. The encoder transmits the filtered, digitized signal to a notebook computer via USB. Skin conductance was measured in micro-Siemens (μS) and recorded at 256 Hz.

To test various feature-based OR detection approaches, we developed a distributed system based on a TCP/IP client-server architecture (see Figure 2). The client CPU received GSR data via a custom MATLAB™ program which performed OR detection and bookmark generation. Bookmarks are sent via TCP/IP to a custom MP3 audiobook player running on the server notebook. This architecture was designed for flexibility and future use in implicitly-controlled media players in mobile or distributed environments, with the wired USB connection replaced with a Bluetooth link from a wearable sensor.

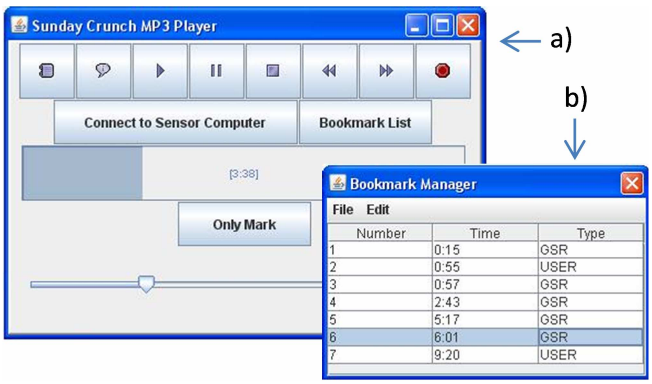

Figure 3. a) Audiobook player; b) Bookmark manager.

Audiobook Player and Bookmark Manager

Users interacted with a custom Java audiobook player (Figure 3a), on which the user could place bookmarks explicitly using a custom bookmark-manager graphical interface (Figure 3b). Participants used the bookmark GUI to navigate through existing marks labeled by type (GSR-derived and explicitly user-created) and time. In the future, we plan to render device-to-user communication and bookmark placing through other channels (e.g., haptics) to further reduce demand on a user's visual faculties.

Interruption Detection Software

Figure 4a shows a process flow diagram outlining the detection process and subsequent auto-bookmarking behavior, which was implemented in MATLAB™. The GSR was measured in real-time, then down-sampled from 256 to 32 Hz to permit online processing (literature and our own results have shown that GSR rise times for ORs are in the range of 1-3s [15]; the Nyquist-Shannon theorem translates this to sampling requirement of >5 Hz). The resampled signal is smoothed by convolution with a 32-point Bartlett window (Figure 4b), and then differentiated (Figure 4c) similar to the procedure used by Kim et al. [9].

To detect an OR, we first identified zero-crossings in the first derivative of the smoothed GSR, noting the signal's direction at these points ('-' to '+' or '+' to '-'). Consecutive zero-crossings, from negative to positive then back to negative, signify peaks in the raw GSR signal. A bookmark is placed if the following inequality holds true:

$$Threshold \leq GSR_{peak} - SMA_{640}(GSR)$$

where *Threshold* is a pre-determined value established in `Experiment 1`, GSR_{peak} is the maximum value of a peak, and $SMA_{640}(GSR)$ represents the moving average of the last 640 samples (20s). To avoid the placement of extraneous bookmarks, new bookmarks were suppressed for a 20s "blackout" window following the most recent placed bookmark. This procedure incorporates contextual information (GSR amplitude) for the signal under investigation, which can vary greatly across individuals and time.

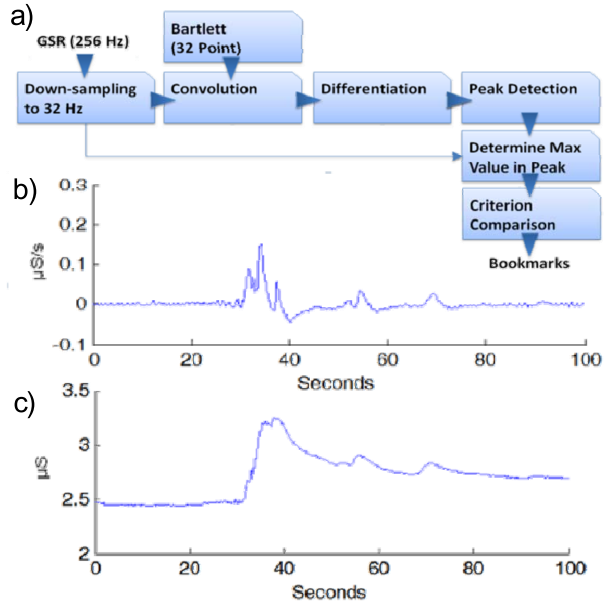

Figure 4. a) Flowchart of OR detection process. b) Typical smoothed GSR waveform (3200 samples). c) Smoothed first derivative GSR waveform (3200 samples).

Latency and Window Length: GSR measurements experience a natural latency of 1-4s [15]. The auto-bookmarking system introduces a delay due to differencing, smoothing and other operations, measured at 100ms. Computational latency associated with OR detection was experimentally assessed to range from 0.5-2s due to differences in GSR signal rise time after different stimuli. When summed, there is a 1.5-6s lag between the actual occurrence of a stimulus and the GSR-based recognition of the OR.

We have hypothesized that users could more easily reorient after jumping to a few seconds prior to a perceived interruption, than to the exact point of interruption. Experimentally, we found that users tended to rewind audiobooks to approximately 10s prior to the start of an interruption*. We therefore placed bookmarks 15s prior to the detection of an OR, corresponding to 9-13s before the actual interruption based on our estimate of recognition latency. Figure 5 shows a timeline of these events.

The choice of averaging window length was the result of a compromise. If the averaging window is too long, false detections arise from large-amplitude, low-frequency GSR changes that are not characteristic of ORs. If the window is too short, the analysis will be too localized and the algorithm will fail to identify ORs. Through pilot studies, we found that a 20s window provided the most stable algorithm sensitivity; hence this value was used in our experiments*.

EXPERIMENTS

Our main objective was to support development of and then validate the general implicit interaction paradigm, within the context of the initial audiobook marking use case. We

* See `Experiment 3`.

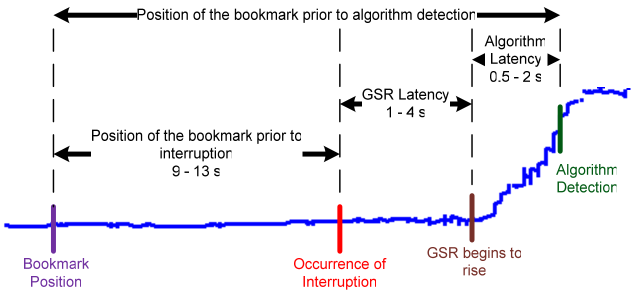

Figure 5. Approximate timeline of the auto-bookmarking system. Horizontal bars indicate timeline variability, black vertical lines show an example timeline for a participant introduced to vocal stimulus.

structured our research questions to be broad, but answerable and relevant from this use case perspective. Experiment 1 (E1) examined the usability of GSR as an indicator of true interruption. We used Experiment 2 (E2) to refine parameters of the bookmark algorithm; e.g., the distance by which bookmarks should be advanced from the point of measured interruption. In Experiment 3 (E3) we showed our bookmarking implementation to users and invited their qualitative feedback. Finally, in a preliminary pilot experiment (E4), we logged and examined additional GSR data in an uncontrolled, noisy environment to assess the algorithm's viability in more realistic situations.

Experiment 1: Evaluation of GSR Utility

Our first research question addressed physiological responses to interruptions - specifically:

Q1: *Does GSR exhibit a consistent and detectable change in response to interruptions of a level that distract users?*

Bookmarks should be placed at or near the point where the user is interrupted. To do this, we need to know what GSR looks like when the user has really been interrupted. Only some events at some times are interruptive enough to distract users who are focused on a primary task (here, audiobook listening). For example, one may be able to mentally block a nearby conversation in a café, but will be distracted by the sound of a door opening in a quiet room. We would like to differentiate between these two situations using only features extracted from GSR measurements. E1 was conducted to answer Q1 by specifically addressing the following hypotheses:

H1: *The GSR signal amplitude exhibits a range that corresponds to ranges of interruption levels; and*

H2: *Disruptive interruptions exhibit a characteristic GSR profile and amplitude range that can be detected with usably high (>80%) true-positive rates.*

This would provide the empirical grounding necessary for development of an OR detection system. An 80% accuracy rate was chosen as baseline. This was comparable to results obtained by Kim et al. [9] for the recognition of three emotion types and determined to be adequate for practical applications. Usability of detection rates is examined in the `Discussion`.

Design

Four female and seven male subjects (n = 11) aged 23-30 participated in E1. Subjects were pre-screened for potential confounds including diagnosed attention disorders (obsessive-compulsive or attention-deficit) and familiarity with the audiobook used. Data from one male subject was discarded due to sensor malfunction.

At the beginning of the experiment, half of the subjects were asked to put in their pocket a cell phone provided by the experimenter that was programmed to emit an auditory ringtone triggered by an experimenter. The other half were asked for their own cell phone numbers, then asked to set their cell phones to vibrate mode and place them in a pocket. This was done to determine if there is a net noticeable difference in GSR when a familiar stimulus from the participant's own cell phone is presented verses an unfamiliar one from cell phone provided by experimenters. We observed no discrepancy between the two cases.

During the experiment, subjects were asked to sit in a silent experiment room facing a wall and to don a GSR sensor and pair of headphones. To reduce noise in the GSR, subjects were asked to avoid making large physical motions. The entire experiment was video-recorded for post-hoc comparison with GSR data. The experimenters sat on the other side of a visual divider from the subject, as shown in Figure 6, to avoid unintentional distraction or anticipation thereof. To encourage interruption responses typical of focused listening, subjects were instructed to listen to the audiobook carefully as they might be tested on concepts described in the audiobook; no test was actually administered.

After collecting two minutes of baseline GSR data where the participants were asked to silently relax, experimenters started playback of the first chapter of the audiobook "Free" by Chris Anderson [18]. The chapter, which describes the invention and marketing of Jell-O and Gillette disposable razors, was verified to be of neutral arousal level in a pilot study; that is, listening to the audiobook did not impact the user's GSR signal except at the start or end of play. The content was considered boring by most subjects and was reused in E1, E2 and E3.

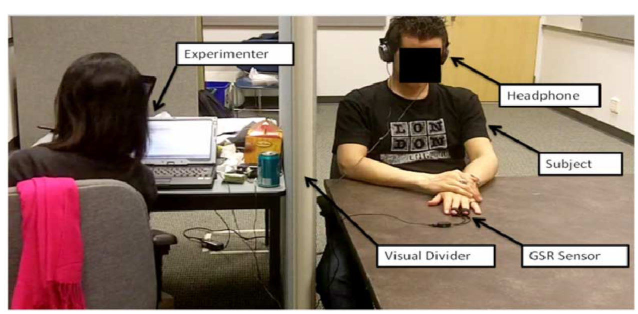

Figure 6. Experiment 1 set-up.

After approximately a minute of playback, the experimenter began to cause interruptions without pausing or stopping the player. Four different interruptions were used: knocking on the experimenter's desk three times {K}, tapping the subject on the shoulder twice from behind {T}, calling the cell phone in the subject's pocket {C}, and verbally activating the subject {V}. These interruptions were chosen as they were perceived to be common in daily-life settings within the context of an audio-listening task. Each interruption was used twice. Intervals of at least 1 min between interruptions allowed the subject's GSR to settle.

Eight interruptions were presented to all subjects in the following order: {K}-{T}-{C}-{V}-{V}-{C}-{T}-{K}. During piloting, we did not discover any difference in any measured quantities as a result of stimulus order; therefore, to standardize the experience of the interruption order across users we did not randomize this sequence. Subjects were not informed of the order of interruptions. The same experimenter introduced all interruptions throughout all sessions with consistent volume and tone. The duration of phone rings was constant, and the two verbal instructions in the experiment were scripted as:

V1: "[Subject Name], I just want to let you know that you are doing great."

V2: "[Subject Name], we are going to continue to take more measurements, and let you know when the experiment is over, OK?"

The experiment was followed by a post-hoc questionnaire and a semi-structured interview. The questionnaire aimed to discover which interruptions were most disruptive. For each type of interruption, the subjects were asked whether they: a) were distracted, b) found it hard to refocus on the book after the interruption, and c) would have liked to rewind the book to a time just before the occurrence of interruption. In the interview, we collected general feedback on the experiment with a focus on each interruption. Qualitative data from the questionnaires and interview was statistically analyzed to test H1. Quantitative data from the recorded GSR was analyzed with pattern detection algorithms to test H2. The length of E1 excluding preparation and administration of the questionnaire/interview was 20 min.

Results

Questionnaires & Interviews: Post-experiment questionnaires revealed that 81% of participants agreed or strongly agreed that they were engaged by the audiobook. 73% of participants agreed that they were interrupted during the experiment. From a set of yes/no questions for each type of interruptions, 91% of participants agreed that they were interrupted by verbal interaction {V}. 55% of subjects stated that {K}, {T} and {C} were interruptive. Qualitative analysis of the post-experiment interview showed that verbal interruption was found to be the most disruptive; it was also found that {K} did not register as either annoying or disruptive to most participants. Three subjects reported that they did not notice a knock at all. Users found the level of disruption by tapping or cell phone to be in between that of knocking and verbal interruption. Hereafter, we refer to {K} as non-disruptive and {V}, {T}, and {C} as disruptive interruptions.

Name	Definition
True positive (TP)	A detection occurred within a 15s time window from the start of a labeled interruption (-4 to +11s).
False negative (FN)	No detection occurred within a 15s time window from the start of a labeled interruption.
False positive (FP)	A detection occurred without a corresponding a labeled interruption within a 15s period.
True negative (TN)	No detection occurred when there was no labeled interruption within a 15s period.
POWER	Proportion of interruptions that are detected by the algorithm (N_{TP} / N_{actual}).
PRECISION	Proportion of detections are interruptions ($N_{TP}/N_{detected}$).

Table 1. Definitions of test outcomes, conditions and measures.

GSR Signal Analysis: The algorithm's performance was evaluated by power and precision metrics as defined in Table 1. We considered an interruption "detected" if an OR occurred within -4 to +11s from the presentation of an interruptive stimulus; this range was derived from the sum of the 1–6s detection latency (see `Interruption Detection Software`) and ±5s which was allocated to account for variation of when ORs were detected. We tested both the power and precision of the algorithm's interruption detection. Power reflects the frequency of true positives, i.e. detection of when an interruption occurred. For example, if 4 out of 6 disruptive interruptions were detected in a single trial, power = 67% (N_{TP}=4, N_{FN}=2). Precision indicates robustness to false positives. If 4 out of 10 detections correspond to disruptive interruptions, precision = 40% (N_{TP}=4, N_{FP}=6). A false positive proved difficult to define because nonspecific responses – those that occur in the absence of an identifiable stimulus – could be caused by valid internal or otherwise unobservable stimuli.

Figure 7 shows the results of our OR detection algorithm for a typical trial, illustrating disruptive interruptions and detected ORs. Supporting our previous approximation of 1.5 to 6s, the average interval between an interruption's actual occurrence and "true" OR capture was 2.52s (std 2.35s).

We examined the algorithm's sensitivity to detection threshold by computing the average power and precision for threshold values ranging from 0.01 to 0.70μ*S* as shown in Figure 8. Power and precision are inversely and directly proportional, respectively, to threshold in a roughly linear fashion. A threshold of 0.07μ*S* (shown at the arrow in Figure 8) was found to be optimal, with maximum power of 84% and a precision of 32%.

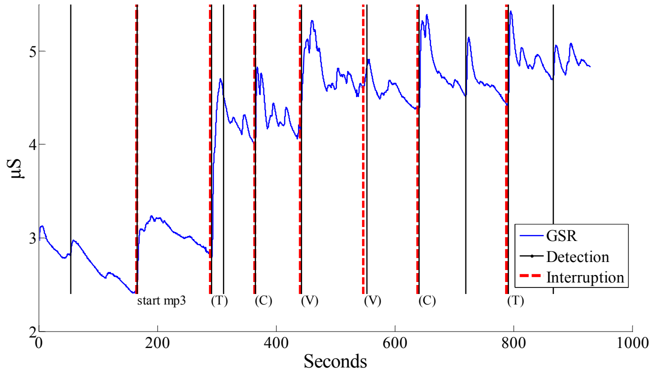

Figure 7. GSR signals overlaid with interruption and detection lines for one representative trial (threshold set at 0.07 µ*S*).

E1 subjects received 20 of each interruption type for a total of 80. The average power of each type of interruption is plotted in Figure 9. Additionally, we calculated the power of detected interruptions deemed to be disruptive, by excluding the apparently non-disruptive {K} types. With the detection threshold set to 0.07µ*S*, the average trial power of disruptive interruptions was 84% (std 15%). A two-sample right-tail T-test verified that the power of disruptive interruptions - {V}, {T} and {C} - is significantly greater than the power of non-disruptive {K} interruptions ($p=0.02$). However, inspection of GSR data from E1 does not show notable qualitative differences between interruption types.

Experiment 2: Qualitative/Quantitative Observation of Desired System Behavior

In E2, we sought data that would help build the auto-bookmarking application for an average user. Specifically,

Q2a: *Where should a mark be placed for greatest utility?*

Q2b: *Is there a correlation between interrupt duration and how far back participants rewind the audio stream?*

According to E1, verbal interruptions (conversation, verbal instructions and questions) were the most disruptive of those tested; verbal interruptions are also controllable in duration. Thus, to maximize the number of rewind events, we used only verbal interruptions in E2.

Design

Four females and seven males (n=11) aged 23 to 30 participated in E2. The same audiobook used in E1 was reused in E2. We required that subjects chosen for E2 had not previously listened to nor read the book prior to the experiment; hence, no participants from E1 participated in E2.

The set-up for E2 is shown in Figure 10. The visual divider used in E1 was omitted, as we felt verbal interruptions lacking eye contact was unnatural. A second experimenter recorded interrupt durations. Both experimenters were out of the participant's immediate sight to prevent anticipation of stimuli. Subjects used the MP3 player shown in Figure 3a; the pause function and 'track time' display were disabled to simulate the scenarios described in `Approach and Use Case Description`. After a brief training session on

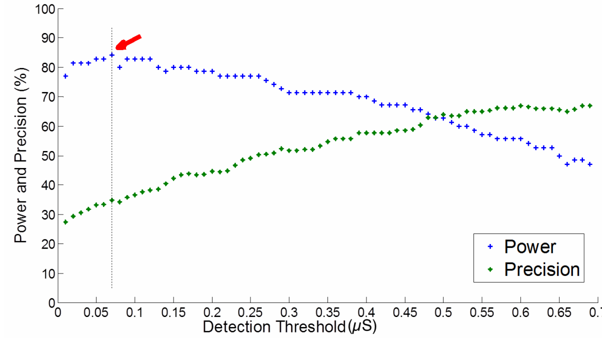

Figure 8. Average power and precision versus detection threshold. The arrow points to the maximum power when the threshold is set to 0.07µ*S*.

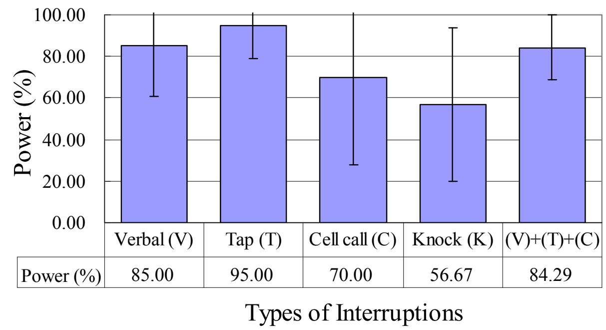

Figure 9. Average algorithm power (proportion of interruptions detected of total) for interruption types. Error bars represent std variability of power across all subjects.

the customized player, subjects were instructed to pay close attention to the content of the audiobook, to rewind whenever desired, and that there would be content-related questions following the listening session. The player automatically logged timestamps of user rewind actions.

Subjects were asked to relax for two minutes at the outset of the experiment to collect baseline GSR data. Experimenter 1 initiated verbal interruptions periodically after the subjects' GSR had settled, with short (2-13s), medium (14-28s), and long (29+s) conversations. A total of six verbal interruptions (two short, two medium and two long) were used for each subject. One medium and one long question were related to the content of the audiobook. Finally, the video recording of the experiment was played back to the subject in a semi-structured interview, to understand the rationale for various rewinding patterns.

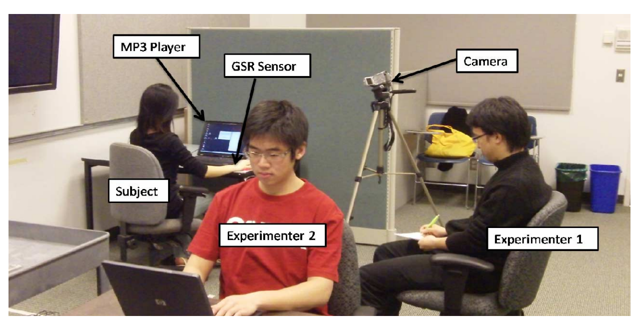

Figure 10. E2 set-up. A similar set-up was also used for E3.

Results

We expected all 66 stimuli to be interruptive - six per subject across the 11 subjects. 19 of these generated user-driven rewind events that corresponded to an interruption. 33 of the remaining stimuli did not cause any rewind event, and the other 14 were discarded either because they caused unrelated rewinding (due to mind wandering etc.), or subjects were not able to remember their reasons for rewinding the audiobook at that particular incident during the interview. It was found that users tended to rewind multiple times for a short duration after an interruption for reasons such as seeking for the 'best' location. This was still counted as a single valid rewind event.

We measured how many seconds the subject rewound prior to the onset of the interruption. For example, we recorded a value of -2 if a subject rewound the audiobook to a position 2s prior to the start of an interruption. The average rewind period during the 19 valid interruptions was found to be -9.37s (std: 7.65s), i.e. on average, a subject rewound the audiobook to 9.37s prior to the occurrence of the disruptive interruption event. We also recorded the duration of each interruption. The duration of the 19 valid interruptions ranged from 7 to 56s. The correlation coefficient between interruption length and extent of the rewind event was r=0.03, which suggests that the rewind distance was unrelated to length of interruption.

Experiment 3: Interaction Utility Assessment

In E3, we tested the utility of our tuned GSR-based auto-bookmarking algorithm. E3 re-used E2's procedure and set-up (Figure 10) but employed the bookmarking algorithm in real-time. Data from E3 provided insight into user reactions to the bookmarking prototype.

Design

One female and four males (n=5) aged 23-29 participated in E3. The system used in E3 is described with its parameterizations in the `Use Case System Description`. While the previous experiments provided us with information on how to detect disruptive interruptions and how to use that information, E3 specifically addresses the question:

Q3: *Do our system-generated bookmarks consistently provide utility after interruptions?*

We used the same pre-screening procedure and setup as in E2. In addition, E3 subjects were asked to use the custom MP3 player and associated custom bookmark manager interface (Figure 3b). Subjects were instructed to use this manager as their first method of rewinding when rewinding was desired, reverting to manually navigating through the audiobook only if the bookmarks were unsatisfactory. Subjects could rewind from any bookmark, but were advised to try the most recent one first. Interactions with the MP3 player and bookmark manager were logged, and video and audio recordings of the experiment were saved.

To better understand their strategies for interrupt recovery, we asked subjects to view the recorded video stream while commenting on each interruption in a post-experiment questionnaire. Finally, we conducted an interview to understand user satisfaction and usefulness of each bookmark.

Results

Each subject in E3 was exposed to six interruptive stimuli, for a total of 30 events. 26 of these stimuli were reported as disruptive, valid interruptions. 21 of these caused subjects to rewind immediately following the interruption. 18 of the 21 rewinds utilized only system-generated bookmarks; in 13 of these, subjects used only the latest system-generated bookmark, whereas subjects in the other five instances continued to try older bookmarks. In the remaining three rewind events, subjects manually navigated through the audiobook. Subjects rated 76% of the automatically-placed bookmarks that they used as 'appropriately positioned'.

The semi-structured interviews provided constructive feedback: users were surprised to see such a system, and were interested in seeing further application areas of such interaction. In general, subjects provided positive feedback on the system's usefulness and utility. Few suggestions on bookmark follow-up were raised. Two subjects suggested the system should pause when the user is interrupted, and to play again by user's manual control. One subject reported that the system's visualization and/or interaction quality could be improved.

Experiment 4: GSR Field Assessment (Pilot)

As a final early validation step, we assessed the potential effectiveness of this implicit interaction implementation based on GSR informativeness in a less controlled environment. We were keen to preview through a pilot what would be involved as we moved to the field to guide our own future work. Specifically, we wanted an initial indication of:

Q4: *Are there quantifiable differences between GSR measurements obtained from a controlled vs. uncontrolled, possibly noisy environment which may affect utility or effectiveness of the interaction?*

Design

Two male subjects (n=2) aged 23 participated in the E4 pilot. During rush hour at a busy outdoor bus terminal, each participant was seated (alone) on a bus bench for the duration of the study and asked to attend to an audiobook played through headphones. Subjects wore a GSR sensor and were instructed to use a pushbutton marking device to signal moments where they felt interrupted enough by an event to wish the audiobook to be rewound. An experimenter (out of the participant's view) also recorded major events that could have potentially caused ORs.

Results

An informal analysis of data from two male participants showed little qualitative (wave shape, event responsiveness) difference between raw GSR measurements obtained from our controlled tests (E1 and E2 results) and the bus terminal. The primary distinguishing characteristic was absolute GSR amplitude: subjects in E1 and E2 tended to exhibit baseline and OR GSRs in the range of 3-6µS, whereas GSR measurements in E4 were 6-15µS; an example comparison is shown in Figure 11. As the software algorithm designed for the auto-bookmarking use case does not utilize absolute signal amplitude, but rather amplitude of peaks relative to the moving average, this preliminary result suggests that system operation should be minimally affected by a chaotic, noisy environment.

Although E4 results are informal and preliminary, this pilot study suggests (though not proves) to us that at the cost of more sophisticated signal processing, information regarding ORs will still be present in GSR data collected from less controlled environments.

DISCUSSION

Q1: Does GSR indicate disruptive interruptions?

E1 showed that GSR can be used to detect disruptive interruptions in an audiobook listening context. According to subject interviews and self-reports, our {V}, {T} and {C} interruptions did disrupt subjects. E1 results (Figure 9) confirmed that different types of interruptions correspond to different levels of disruption as indicated by GSR. Thus, H1 is confirmed for the range of interruptions that we tested. Moreover, for GSR data obtained in E1, we achieved an 84% recognition rate in detecting disruptive interruptions, (i.e., true positives of type {V}, {T} and {C}). This supports H2, where disruptive interruptions produce salient GSR signal features, detectable by our algorithm with an acceptable true-positive rate; false positives were difficult to detect, as explained in `Experiment 1`.

Only four types of interruption stimuli were tested in these experiments; the design space of possible interruption stimuli and contexts is much larger. As such, validity of our results is coupled to the primary task and interruption stimuli used here. For example, while a gentle knock was not disruptive in this context, it certainly could be in other situations. Nevertheless, our approach - evaluating the GSR signal in context and using this to tune an algorithm variant to that context - is general enough to be applied to other primary tasks such as watching TV, housecleaning, gardening, working at a computer and perhaps driving.

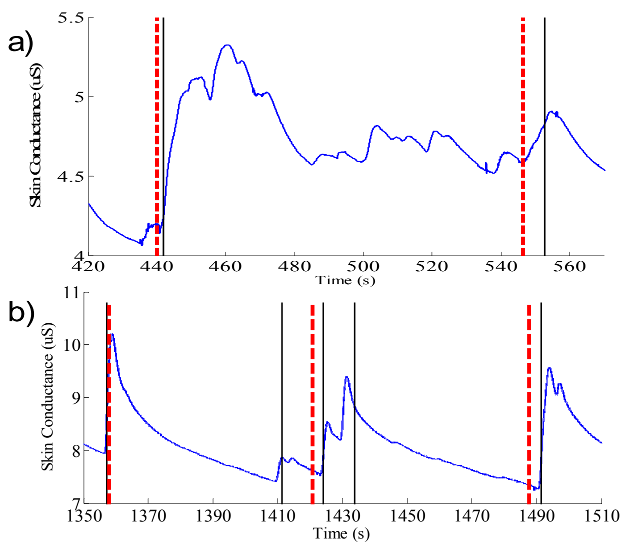

Figure 11. Raw GSR data collected from a participant in a) a controlled test environment and b) a bus terminal at rush hour. Thick, dotted lines signifies interruptions; thin, solid lines represent bookmarking system OR detections.

As seen in Figure 8, there is a tradeoff in power versus precision as a function of detection threshold, i.e. between false negatives and false positives. Based on subject interviews, we believe that higher power with slightly lower precision is a reasonable tradeoff. However, the optimal balance will depend on the environment and individual responsiveness to ORs.

Q2: Where should a mark be placed relative to interruption?

E2 provided data for designing the bookmark placement system. Its results indicated that users preferred to rewind the audiobook to a location approximately 10s prior to the onset of an interruption, independently of interruption length. This implication for mark placement advance may be applicable to interruptions in other use contexts. Conveniently, the results also imply that it is not crucial to measure or estimate the length of interruption, a potentially challenging task in the field. However, we believe this finding only applies to interruptions under a minute. Longer interruption may affect user's behavior and the desire of bookmark location; for instance, a 10-minute interruption may cause the user to replay the entire chapter of the audiobook, because the content has slipped out of his short-term memory. Likewise, our observation of mark-advance independence from interruption length may apply to other use cases, but this will need to be verified.

Q3-Q4: Do our system-generated bookmarks consistently provide utility after interruptions? Is GSR informative in uncontrolled environments?

GSR signals are known to be susceptible to noise from sources such as bodily motion and cognitive workload [15]. We generally found (E1-E3) that our algorithm was quite robust to small hand or arm movements and subtle posture changes, but we had false-positive detections in the absence of an identifiable stimulus. Although it is not feasible to explain each one, we expect that some occur through internal distractions, such as when a phrase or a word in the media being perused initiates mind wandering. E4 results suggest that GSR is robust in a chaotic environment.

Subjects did not always find rewinding the audiobook necessary despite being interrupted – particularly for short interruptions (<11s – as determined by the length of the longest 'short' interruptive conversation in E2 without causing users to rewind). We found that no rewinding occurred for 81% of the short interruptions presented to subjects in the experiments. Subjects reported that even if they stopped listening to the audiobook to attend to short interruptions, they could quickly pick up the gist of what was being said afterwards. We believe this was representative of this use case, and it seems that users sometimes use a ‘multi-tasking’ strategy to cope with interruptions.

E3 and E4 served only to briefly explore the potential of our implicit control algorithm in controlled and uncontrolled setting; the small sample size limits our ability to provide any concrete evidence pertaining to the usefulness our system in the field and thus caution is required when evaluating the significance of these results. They give some indication that our algorithm provided a usable detection rate; further investigations with larger sample sizes are required. Subjects reported appreciation of and surprise at the automatic bookmarking. Perhaps most relevantly, many comments were directed to bookmark follow-up as opposed to the actual bookmark generation, supporting the conclusion that the current bookmarking algorithm is reasonably accurate, our vision of a loop-closing system is of potential utility and our next step is to concentrate on using the bookmarks themselves.

CONCLUSIONS AND FUTURE WORK

In this paper, we have demonstrated a novel approach to human-computer interaction based on an implicit communication paradigm. We have developed and validated the theory that a computer algorithm can, in real-time, detect features within a GSR stream that correspond to ORs caused by external disruptive interruptions. We have exploited these findings with a system that provides OR-based auto-marking for a representative audiobook use case. We have found that with our algorithm, we can achieve a true-positive interruption detection power of 84%.

The work presented here demonstrates the feasibility of an implicitly-controlled interaction model which provides a possible solution for managing the interplay between primary and secondary tasks in various environments. This use case represents a first step in a larger effort in which we aim to develop an implicit interaction loop driven by physiological parameters such as GSR.

We anticipate that the false-positive rate found in this use case may be addressable by dynamically tailoring the OR detection algorithm to individuals, through adjusting sensitivity, bookmarking delay and dead-zone period until another bookmark can be placed. Improvements to the present system’s usability in the field (e.g., further reducing noise in the GSR introduced by bodily movements) may be assisted by integrating contextual information such as accelerometer and location data from a worn or carried device.

Our next step is to provide users with a low-attention notification that the application has acted in response to the user’s physiology (for the use case presented here, that a mark has been placed). To further test and validate this model, we plan to study the closed-loop dynamics of other use cases such as streaming media, internet browsing and vehicle operation. We would also like to explore how to discriminate ORs to the media itself from responses to external events.

REFERENCES

1. Wickens CD. *Engineering Psychology and Human Performance*. Merrill; 1984:513.
2. Weiser M, Brown JS. Designing Calm Technology. *Powergrid Journal.* 1996;1.01(July):94-110.
3. Picard RW. *Affective Computing*. Cambridge, MA: MIT Press; 1998.
4. MacLean K. Putting Haptics into the Ambience. *IEEE Trans. on Haptics*. 2009;2(3):123-135.
5. Baumann MA, Maclean KE, Hazelton TW, McKay A. Emulating Human Attention-Getting Practices with Wearable Haptics. In: *IEEE Haptics Symp.* ; (2010).
6. Cowie R, Douglas-Cowie E, Tsapatsoulis N, et al. Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing*. 2001;26(1):32-80.
7. Picard RW, Vyzas E, Healey J. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Trans. on Pattern Analysis and Machine Intelligence*. 2001;23(10).
8. Picard RW. Toward Agents that Recognize Emotion. *MIT Media Lab Perceptual Computing Section Technical Report*. 1998;(515).
9. Kim, K.H., Bang, S.W., Kim, S.R. Emotion recognition system using short-term monitoring of physiological signals. *Med. & Bio. Eng. & Computing.* 42, 3 (2004).
10. Healey J, Picard RW. StartleCam: a cybernetic wearable camera. *Intl. Symp. on Wearable Comp.*, IEEE Comp. Soc. (1998), 42-49.
11. Conati, C. Probabilistic Assessment of User's Emotions in Educational Games. *Applied Artificial Intelligence 16*, 7-8 (2002), 555-575.
12. Frith, CD, Allen, HA. The skin conductance OR as an index of attention. *Biological Psychology*. 1983;17(1).
13. Boucsein, W. *Electrodermal Activity*. Plenem Press, New York, New York, USA, 1992.
14. Cacioppo JT, Tassinary LG. Inferring psychological significance from physiological signals. *The American Psychologist*. 1990;45(1):16-28.
15. Bernstein, A. An analysis of the skin conductance OR in samples of American, British, and German schizophrenics. *Biological Psych*. 1982;14(3-4):155-211.
16. Cacioppo JT, Tassinary LG, Berntson GG. The Electrodermal System. In: *Handbook of Psychophysiology*. 3rd ed. New York, New York, USA: Cambridge University Press; 2007:159-181.
17. www.thoughttechnology.com
18. Anderson, C. *Free: The Future of a Radical Price*. HarperCollins, New York, New York, USA.