Bringing the High Seas into the Lab to Evaluate Speech Input Feasibility: A Case Study

Joanna Lumsden
Aston University
Birmingham, UK
B4 7ET
+44 (0)121 204 3470
j.lumsden@aston.ac.uk

Nathan Langton
National Research Council of Canada
46 Dineen Dr, Fredericton, N.B.,
Canada, E3B 9W4
nathan.langton@nrc.gc.ca

Irina Kondratova
National Research Council of Canada
46 Dineen Dr, Fredericton, N.B.,
Canada, E3B 9W4
+1 506 444 0489
irina.kondratova@nrc.gc.ca

ABSTRACT
As mobile technologies continue to penetrate increasingly diverse domains of use, we accordingly need to understand the *feasibility* of different interaction technologies across such varied domains. This case study describes an investigation into whether speech-based input is a feasible interaction option for use in a complex, and arguably extreme, environment of use – that is, lobster fishing vessels. We reflect on our approaches to bringing the "high seas" into lab environments for this purpose, comparing the results obtained via our lab and our field studies. Our hope is that the work presented here will go some way to enhancing the literature in terms of approaches to bringing complex real-world contexts into lab environments for the purpose of evaluating the feasibility of specific interaction technologies.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation** (e.g., HCI)]: User Interfaces – *Evaluation/methodology; Voice I/O.*

General Terms
Measurement, Performance, Design, Experimentation.

Keywords
Mobile speech input, evaluation, case study.

1. INTRODUCTION
Mobile technologies are becomingly increasingly pervasive across ever-increasing domains of activity. Accompanying this device penetration is a need to understand the *feasibility* of different interaction technologies and techniques across the expanding range of use case scenarios. It was, in fact, just such a need for understanding that triggered the case study we present in this paper. We were approached by a local entrepreneur (hereinafter, our 'collaborator') who was developing a software data-capture application to run on laptop computers onboard lobster fishing vessels: in essence, he wanted to know whether speech-based data input was a feasible interaction option for use with his software given the complexity and extreme environmental conditions associated with his target context of use.

MobileHCI 2010 September 7 - 10, 2010, Lisboa, Portugal.
ACM 978-1-60558-835-3.

Compared with other interaction techniques, speech has been shown to enhance mobile users' cognizance of their physical environment while interacting with mobile devices [12]. This makes it a strong candidate interaction mechanism for use on a lobster fishing vessel where users are mobile and multitasking. Unfortunately, however, it is estimated that a 20%-50% drop in recognition accuracy can occur when speech is used outside of an office setting; in natural settings, speech recognition accuracy is degraded by people's tendency to speak differently in noisy environments (the Lombard Effect [16]) and by contamination of the speech signal by background noise. Our collaborator could see the potential benefits of using speech-based input with mobile technologies on lobster fishing boats but, for obvious reasons, was concerned as to its feasibility in terms of the attainable accuracy rates and the impact of using speech-based technology on fishermen's environmental cognizance and workload.

To determine how best to acquire the understanding necessary to address these concerns, we looked to the literature on approaches to evaluation of mobile technologies. Given momentum in large part by Kjeldskov *et al's* suggestion that conducting field evaluations is not worth the "hassle" [8], the benefit of undertaking lab v. field evaluations for mobile technologies continues to be the subject of considerable debate [e.g., 7, 15]; as yet, there is no agreed consensus on how best to evaluate mobile technologies. Researchers are only beginning to explore the pros and cons of lab v. field evaluation techniques. The relative infancy of study in this area means that there is meager literature reporting the results of experimental comparisons of field v. lab approaches to mobile evaluations; as such, this debate is often viewed as a matter of opinion [15]. We found little research reporting evaluations of mobile technologies in settings as novel or challenging as a lobster fishing vessel. Furthermore, we found little evidence of comparisons of evaluation approaches where the focus of the evaluation was on a simple assessment of the feasibility of an interaction technique as opposed to more holistic usability or user behavior evaluations. Hence, we found little to guide us in terms of whether to opt for a lab or field study and, in the case of the former, how to approach the design of a lab study such that key aspects of a complex physical environment were adequately 'reproduced' within the lab.

In light of the scarcity of directly applicable literature, the most immediately obvious approach to eliciting the data necessary to

answer our collaborator's concerns was to simply run a field trial on lobster fishing vessels (which, in fact, we did – as previously reported in [13]). Our literature review had, however, highlighted scope for us to contribute to the ongoing debate – albeit within the confines established by our given context and evaluation agenda. Thus, it emerged that our case study offered two opportunities to contribute to knowledge in the mobile HCI community: (1) to demonstrate whether speech-based input is a feasible interaction mechanism for use on a lobster fishing vessel (representing a novel context of use for the community) and to determine the impact of using such a technology on the environmental cognizance and workload of fishermen (as already mentioned, discussion specific to the field study conducted to achieve this goal can be found in [13]); and (2) to reflect on *how* to bring the "high seas" into a lab environment without compromising *relevant* ecological validity as well as to compare the results returned when adopting different approaches to this. Our case study was, therefore, designed to accommodate both agendas.

In this paper, we focus primarily on the second of the opportunities afforded by our case study: that is, we reflect on *how* we went about establishing three evaluation environments in order to examine the feasibility of speech as an input mechanism for use with data-acquisition software on lobster fishing vessels. Our three environments comprise: (1) our original field study on lobster fishing vessels; (2) a lab-based study in which the context of a lobster fishing vessel was abstractly represented (our 'dry-ground' study); and (3) a 'middle-ground' study conducted in an offshore engineering basin (OEB) or wave tank (this representing a middle ground of abstraction between (1) and (2)). We compare the results we obtained after running our study in all three environments.

We hope that a combination of the results previously discussed in our MobileHCI'2008 paper ([13]) and the reflective discussion presented here will go some way to enhancing the literature in terms of approaches to bringing complex real-world contexts into lab environments for the purpose of evaluating the feasibility of specific interaction technologies.

In reviewing the relevant literature, we note that much of the work conducted to date focuses on evaluating the use of mobile phone (or similar) technologies, an underlying assumption being that *mobility* equates to *user* mobility. In this discussion, we adopt a more encompassing definition of mobility to additionally include physical environment-induced user mobility – that is, situations in which a user is physically unstable as a result of his/her physical environment (e.g., in a moving vehicle) whilst using technology. Such environmentally-induced motion impacts on users' ability to accurately interact with 'mobile' technologies – whether it be technology integrated within the fabric of the physical environment (e.g., in-car systems) or portable equipment used within the physical environment (e.g., as in our case, a laptop being used within a moving fishing vessel).

In this paper, we review related work with respect to comparative lab v. field studies. We then outline our research approach, describing the generic approach common to all three of our studies, before describing the particulars of each of our three environments. The penultimate section of this paper contrasts and compares the results we obtained from each environment, before we conclude with a discussion of the implications of our findings and further work.

2. RELATED WORK

Kjeldskov *et al*. [8] compared lab- and field-based approaches to the evaluation of a mobile Electronic Patient Record (EPR) system. They simulated a hospital ward in their lab and compared the usability problems identified by participants in that environment to the problems identified by participants in the field study – an actual hospital ward. Surprisingly, significantly more serious and cosmetic usability problems were discovered in the lab. Only one problem identified in the field was not similarly identified in the lab; this problem was not directly related to the usability of the system, but rather the integrity of data and its storage. Kjeldskov *et al*. noted that the field study posed challenges with respect to the collection of data: in contrast to lab participants, nurses operating in real life were (unsurprisingly) unable to accommodate the notebook-based method of recording data. The results of this study indicate that if the real-world context of use is taken into consideration in the design of a lab-based study environment, a lab-based approach *may* be at least as effective as a field evaluation. Some researchers have been quick to refute the findings of this study based on differences in task assignments and/or differences in quantitative and qualitative data collection techniques used in both evaluations [15]; others suggest that no clear definition of usability problems were given, and that the field study involved events that decreased control over the study [7].

In 2005, Kaikkonen *et al*. [6] similarly compared the results of a usability evaluation conducted in a typical usability lab with those returned from a field study (specifically, a shopping mall and train station). In both cases, the same pre-defined set of tasks was used to ensure that the context was the only changing variable. Contrary to their expectations, Kaikkonen *et al*. found that the same usability problems were found in both evaluation environments and task completion times were no different across study settings. Interestingly, however, Kaikkonen *et al*. noted that their lab set-up did not permit them to observe certain aspects of user behavior that were apparent in the field; that said, the lab environment did not adequately represent the real-world context of use – it did not include external interruptions, environmental distractions, varying lighting conditions, or other such factors that are likely to be present when performing tasks on a mobile device within a shopping centre or train station, for example. In a later reflection on their previous study, Kaikkonen *et al*. [7] concluded that their field study results were more related to user behavior and experience than usability and user interaction with the device *per se*. They suggest that field studies are only useful when the purpose is to gain knowledge about user behavior in a natural environment, and that they present no benefit in terms of understanding user interaction.

Duh *et al*. [3] also compared lab and field evaluations in terms of the usability problems identified with respect to a mobile phone-based application. The tasks studied related to the typical activities that users would engage in while using a mobile phone on public transportation; participants' interactions were recorded using a think-aloud protocol. In contrast to the study by Kjeldskov *et al*. [8], Duh *et al*. found significantly more critical problems in the field than in the lab-based setting. They suggest that these differences were due to external factors associated with the real-world environment of use, such as noise, the movement of the train, lack of privacy, and mental and physical demands that affected participant performance. Once again, the real-life (field)

environment was not sufficiently replicated in the lab – participants were seated in a quiet room and simply asked to "imagine" that they were on a train. Conversely, participants in the field study reported feeling increased stress and discomfort – likely the result of having to describe out loud everything they were doing in a public location which does not realistically reflect how most users interact with a mobile device in public.

Baillie and Schatz [1] evaluated a multimodal application using a combination of a lab study and a field study. The lab was free from interruption and noise, while the field study was outside within an area nearby shops and a train station. To their surprise, they observed that participants took less time to complete study tasks in the field than in the lab; although more problems were found in the lab, there were no differences in the critical problems identified in both environments. The overall usability of the application – in terms of effectiveness, efficiency, and satisfaction – was rated more highly in the field than the lab. Baillie and Schatz hoped that these results would go some way to refuting the claims of Kjeldskov *et al*. [8]; they hoped their results would demonstrate that undertaking usability evaluations in the field *is* "worth the hassle".

Nielsen *et al*. [15] compared a lab and field evaluation of the same context-aware mobile system in terms of their ability to identify usability problems. Contradicting the claims posited by Kjeldskov *et al*. [8], Nielsen *et al*. argue that, where both evaluations are conducted in exactly the same manner, field-based usability evaluations are *more* successful. Conducting usability evaluations in the same manner in both environments can, however, reduce the realism of the field itself, thus rendering the field-generated results less valuable. Nielsen *et al*. suggest that, in order to compare lab and field evaluations, the latter have to be less realistic than one might anticipate or want because the users' tasks must be designed beforehand. Others argue that reducing the realism of a field evaluation to this extent takes away the purpose and true definition of a field evaluation.

Holtz Betiol and de Abreu Cybis [5] conducted a study of mobile interfaces based on three different evaluation approaches: a lab test with a PC-based mobile phone emulator; a lab test with an actual mobile phone; and a field test using a mobile phone. A comparison of the results of these studies showed that there were no statistically significant differences between the lab and field tests when the mobile device itself was used; unfortunately, none of the studies introduced mobility so the findings are somewhat limited.

Whilst the work discussed above represents considerable progress with respect to our collective understanding of the benefits of lab v. field approaches to evaluating mobile technologies, there remain a number of unresolved issues as well as a need for more empirical data to further ground the ongoing debate. Aside from the fact that some of the previous studies adopted different data collection strategies in the different evaluation environments which calls into question the validity of the comparisons, most did not adequately replicate or represent the real world in the lab environment. At the very minimum, some of the lab-based studies omitted the inclusion of user mobility which has widely been recognized as having an impact on user performance [e.g., 2, 9, 14]; the lack of contextual or environmental relevance in many of the lab-based versions of the studies essentially means that the two environments were not compared on an equal footing. Furthermore, as previously mentioned, little focus was given to how best to incorporate environmental relevance within lab-based studies – and, in particular, no prior study tackled as environmentally challenging or novel a context as a lobster fishing vessel at sea!

What was also apparent in reviewing this body of literature is that the majority of the work thus far has concentrated on full-scale usability evaluations or studies of user behaviour and user performance relative to complete applications. There is little evidence of studies designed to focus, in a more abstract sense, on the feasibility of a given interaction component of a design (e.g., speech input). We feel that it is important to recognise that, as highlighted by the origins of our case study, sometimes we simply need to know the answer to a more abstract question such as "will speech prove feasible in this context?" as opposed to looking at the bigger picture of the application usability as a whole. As with our case, our collaborator wanted to be informed about the *potential* for speech input prior to committing the investment necessary to incorporate it within his application UI – only after which would a full usability evaluation be appropriate.

Thus, to restate our aims: (1) we wanted to investigate whether speech-based data input was a feasible data input technology for use on a lobster fishing vessel (and, in so doing, observe its impact on users' environmental cognizance and workload); and (2) we wanted to reflect on our experiences of bringing a complex environmental context into a lab to conduct such an investigation – as well as to observe the similarities and differences in the data obtained when we conducted our investigation across our different study set-ups.

3. RESEARCH APPROACH

Based on interviews with members of the lobster fishing industry, review of video (including audio) footage taken directly from a lobster fishing vessel during a typical fishing trip, and ethnographic observations of the work environment (including ambient noise levels), we analyzed the environmental conditions within the cabin of a diesel-engined lobster fishing vessel that were of relevance to the effective use of speech technology.

We identified three primary aspects as having the potential to impact the efficacy of speech-based data input, namely: ambient noise levels; vessel motion; and the need for users to multitask in terms of interacting with, or monitoring, other (typically electronic) equipment. It is important to note, at this juncture, that our context of use comprised recording catch data whilst a trap line was being hauled in; in such situations, a lobster vessel is typically idling or moving very slowly, and physical navigational activities (i.e., steering) are minimal.

On the basis of our observations and the assumptions outlined above, we developed our generic study protocol. In the field study, the three key aspects came with the environment – in our lab studies, we engineered ways to incorporate them in a meaningful and representative way.

As with Kaikkonen *et al*. [6], our intent was to keep the experimental tasks and mechanisms for data collection identical across all three of our study environments. Furthermore, since our intention was to evaluate the feasibility of speech in a general sense within our specific context of use prior to it being embedded within a given application, we designed a simple system to prompt for, and record, users' speech-based input (i.e.,

it was entirely focused on measuring the feasibility of speech as opposed to the usability of an application incorporating speech). We used this system as part of a generic experimental protocol which we administered in each of our three evaluation environments. This section describes the generic protocol; subsequent sections outline its administration relative to each environment.

3.1 Method

With the exception of some minor study-specific (typically, administrative) procedural issues, the generic protocol outlined here was followed identically in each of our studies such that we could equitably compare the results from each.

3.1.1 Equipment

Based on efficacy data returned by a previous study [11] we selected to use the Shure QuietSpot© Boom QSHB3 condenser microphone with noise cancelling properties. We used this with a Transit© M-Audio external USB audio interface to eliminate electric interference from surrounding electronic components, including those comprising the Toughbook mobile computers we used for our study.

Figure 1: Cabin of typical lobster fishing vessel, showing extent of electronic devices in situ.

Figure 2: Speech-based data input application (left) and distractions application (right) on Panasonic Toughbooks©; set-up shown is for a right-handed person.

Our ethnographic analysis highlighted the fact that lobster fishing crew members who might be required to use a software application in the cabin of a vessel would likely also be required to simultaneously monitor and react to other surrounding electronic displays (see Figure 1). To meaningfully evaluate the efficacy of speech within this context of use we felt that it was, therefore, imperative that we also assessed users' ability to simultaneously remain aware of, and react to, their physical environment – specifically, other electronic systems; the feasibility of speech-based interaction would be questionable if, in a safety critical environment such as this, it demanded so much of a user's cognitive resource that he/she could not effectively monitor his/her environment.

Figure 3: User interfaces to the speech-based data input (left) and distractions (right) applications.

We developed two simple applications: (a) a speech-based data input application; and (b) a 'distractions' application. Both were installed on Panasonic Toughbooks© running Windows XP; these Toughbooks are designed to be used in field environments, and as such have inbuilt resistance to shock, spills, vibration, and dust, making them a logical and safe choice. Figure 2 shows the two Toughbooks set up for a right-handed user.

Our speech-based data input application (see Figure 3) was designed to run on the multimodal Opera™ browser, which incorporates IBM's ViaVoice© speaker-independent speech recognition engine (which returns the best speaker-independent accuracy rates for mobile speech-based input [10]). We adopted a push-to-talk strategy for our application; allowing users to explicitly direct commands to a system is generally deemed more appropriate in noisy environments than a continuous monitoring strategy [18]. We used a local database (IBM DB2 Everyplace©) to capture experimental application usage data.

For each data item, participants were shown what to say (e.g., "Thirty Six" as shown in Figure 3, or "Set Trap Line" for example) and were given three attempts by which to achieve a successful data input entry; the results of their input were displayed in the field immediately below the data input instruction (e.g., "36" in Figure 3). If, after three attempts, a participant had not managed to enter an item correctly, the system automatically progressed to the next input. For the purpose of our study, we evaluated a 79 item data set; the items were selected on the basis of vocabulary appropriate (in a generic sense) to the lobster fishing industry as well as commands typical for vessel navigation. Each participant had to complete the same data entry items in the same order.

Our distractions application was designed to abstractly mimic the need to monitor ancillary technology while interacting with the speech-based software; it was designed to be run and used simultaneously to our data input application. It contains three rows of four buttons (or squares) as shown in Figure 3. In a preset (pseudo-random) pattern of location, interval time, and display duration, the application displays a sequence of red buttons; participants were required to acknowledge each red button by tapping the appropriate region of the touchscreen on the Toughbook. When successfully tapped, the red button would disappear; the same was true if the time duration for display of the button elapsed without the button being acknowledged. By observing how many distractions were acknowledged, we were able to assess the impact of speech-based input on participants' environmental cognizance. One might argue that we would have achieved a more natural or realistic set-up by simply introducing the speech-input application into the working environment onboard the vessels (for our field study); this would not, however, have been replicable in our other study environments and we would not have been able to retain control over the intensity and volume of distractions (it was also not advisable from an ethical/safety perspective onboard the vessels).

We set up the two Toughbooks such that the distractions application was situated on participants' dominant-hand side; this meant that their dominant hand was available for interaction with the distractions application and engaging the push-to-talk button on the speech-input application, whilst their non-dominant hand could be used to steady themselves (given the motion of the physical environment) if necessary.

3.1.2 Data Collection

During our studies, we electronically (within our experimental applications) recorded a range of measures to assess the efficacy of speech and its impact on users' environmental cognizance. These measures included details of participants' responses to the distractions and details of the data they entered as recognized by the speech recognition engine. Additionally, using questionnaires we manually recorded anonymous demographic information about our participants; we also required that participants completed subjective assessments of workload using the NASA Task Load Index (TLX) which measures workload according to six dimensions, namely: frustration levels; performance levels achieved; effort expended; mental demand; physical demand; and temporal demand [4]. We also took sound level readings so that we could measure and monitor (and replicate) the ambient noise levels in our studies.

3.1.3 Generic Procedural Issues

Appropriate to each study, participants were provided with information on the purpose of the study, including the study objectives and motivations (i.e., to assess the efficacy of speech in the context of a lobster fishing vessel) and what they would be required to do. Once participants had been given ample opportunity to read and ask questions about the provided information, they were required to review and sign a consent form to participate; once they had consented to participate, they were asked to complete a short demographic questionnaire.

Participants were then given targeted training relevant to the technologies they would be required to use during the studies; this training was delivered in conditions identical in all aspects to those in the study sessions themselves. Participants were first trained in how to use the microphone – specifically, how to use the push-to-talk strategy. They were then instructed in how to enter data into the data-input software application using speech; they were given an opportunity to try a series of 8 training data inputs. Following this, participants were trained in how to use the distractions application, and were given a chance to practice using it. Once participants had completed the training sessions for each application separately, they were given an opportunity to practice and familiarize themselves with using both applications in parallel, as they would be required to do during the course of the study sessions. This whole process took no more than 15 minutes in total.

Only once participants were comfortable with what they would be required to do, and had no further questions, did an actual study session commence. During the study sessions, participants were presented with the series of 79 data entry items displayed on the screen of the Toughbook running the speech-input application; the participants were required to enter, using speech, each item as it appeared. Whilst completing these data entry tasks, participants were required to simultaneously monitor and react to the distractions application running on the second Toughbook. After completing the study session tasks, participants were asked to rate, using the NASA TLX questionnaire, their subjective assessment of the workload involved.

4. FIELD TRIALS

We were invited to join the crews of lobster fishing vessels off the New Brunswick coast of the Bay of Fundy. The participants in our field study comprised the crews of the lobster fishing vessels; we collected usable data from 8 participants. All our participants were male, ranging in age from 18 to 50 years; all were native English speakers with a Canadian accent (this matched the recruitment criteria in our lab studies, albeit we had no control over the participants in our field study). We accompanied fishing crews on scheduled fishing trips. Fishermen participated in our study sessions at times when they were not otherwise engaged in mission critical tasks; that is, we took advantage of the physical environment but did not expect participants to complete study tasks in a manner that impeded on their primary activities. Participation took approximately 45 minutes per person.

The field study sessions were performed within the enclosed cabins of the lobster fishing vessels (as informed by our ethnographic studies). Our two Toughbooks were set up on the dash in each cabin – as shown in Figure 4 – and participants were required to stand throughout the duration of their participation (this being the *modus operandi* in the cabins of such vessels). As is typical for a field trial, we had limited control over the physical environment – specifically, the prevailing weather and sea conditions. That said, whilst not ideal given the winter conditions during which we conducted our study, the prevailing weather was relatively consistent and typical of the conditions in which the vessels normally operate.

Figure 4: Toughbooks set up in situ within the cabins of two different vessels; evaluator (left) demonstrates how participants were positioned while performing the tasks.

Figure 5: Dry-ground lab set-up.

5. LAB ('DRY-GROUND') STUDY

We adopted the set-up depicted in Figure 5 to allow us to abstractly incorporate each of the three key environmental aspects in our lab environment.

To introduce motion, we used a BOSU© platform. This is a standard piece of exercise equipment which we selected because (a) it is *designed* to provide an unstable platform on which to stand that causes a user to have to work to maintain balance, and (b) in consultation with several kinesiologists, it was

recommended to us that this equipment would be safe for people to use and would best replicate the average environmental motion inherent on a lobster fishing vessel when idling at sea in typical fishing-friendly weather conditions. Additionally, as a purely visual distraction, and to add further realism, we projected a looping clip of video footage taken from a lobster fishing vessel (as viewed from the cabin) onto the wall in front of participants. It was not our intent to sync the motion of the vessel in the footage to the physical motion experienced by participants, as this had no bearing on the efficacy of speech recognition and would have run the risk of inducing motion sickness. Previous studies have highlighted the impact of motion alone on the efficacy of speech recognition software [12, 17], so our key objective was to ensure that participants were on an unstable, moving platform on which they had to exert effort to remain balanced and stable – just as the fishermen have to do on the lobster fishing vessels. To introduce relevant ambient background noise, we used an audio recording of the ambient noise taken within the cabin of one of the lobster fishing vessels during our field study. This background noise was played within the lab space using the lab's 7.1 surround sound system. For the purpose of more detailed comparison in our lab study, we segmented our consideration of background noise into 3 noise levels (based on the observed range of actual ambient noise) with the result that we had three experimental conditions: A – our control, or baseline, level which was essentially a quiet environment; B – ambient noise introduced with an average of 76dB(A) and a maximum of 83dB(A); and C – ambient noise introduced with an average of 86dB(A) and a maximum of 93dB(A). We employed this segmentation to (a) allow us to compare the effect of the ambient noise to a quiet environment, and (b) allow us to determine if there is a threshold at which speech input may become impractical on a lobster fishing vessel; obviously, this level of environmental control was not possible in our field studies. The need to multitask was already incorporated in our generic protocol.

We used a between-groups study protocol, where 24 participants were assigned to one of the three study conditions – giving us a total of 8 participants per condition. Participants were recruited from the local community, including staff and students of the university. We restricted participants to persons with a Canadian accent and for whom English was their native language on the grounds that speech recognition engines are typically optimized to native English speakers, and this profile best matched our field participants. Our participants included 18 males and 6 females (distributed equally across the three conditions) and ranged in age from 18 – 50 years.

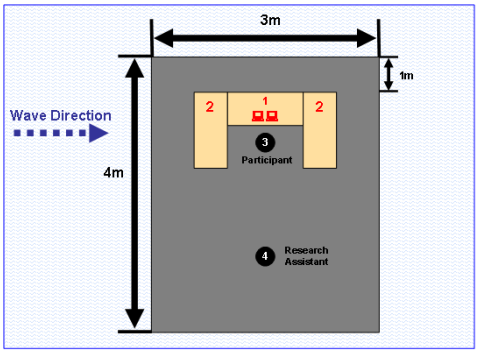

Figure 6: Schematic of floating platform.

6. OEB ('MIDDLE-GROUND') STUDY

The offshore engineering basin (OEB) is a 75m x 32m wave tank, with a water depth of approximately 3m. It is fitted with hydraulic wave generators which can create waves of up to 0.8m in height, and can be programmed to recreate different wave spectra. We fabricated a floating platform which was designed to reflect the 'dry-ground' set-up and field environment: it comprised dash-level surfaces on which the Toughbooks were located (a schematic is shown in Figure 6).

Figure 7: The floating platform in situ in the wave tank.

As can be seen from the right-hand photo in Figure 7, we situated the floating platform in the middle of the tank, and secured it in place with extendable ropes to allow it to move in response to waves but to prevent it travelling up the tank and ultimately colliding with the beach at the end opposite the wave generators.

The wave generators were programmed to send irregular wave patterns down the length of the tank; the wave spectrum used was based on real wave pattern data collected from a wave buoy off the coast of Newfoundland, Canada. The platform was located "port side", "beam on" to the waves such that the wave motion experienced by participants primarily comprised roll-induced-motion (i.e., side-to-side); this set-up was most representative of the kind of motion experienced on an idling vessel which will naturally orient itself (drift) "beam on" to waves. To introduce representative background noise, we played our background audio files (as used in the 'dry-ground' study) via speakers connected to one of the Toughbooks. A combination of mechanical and other environmental background noise, together with the audio files, resulted in participants being exposed to ambient noise in the range of 70dB(A) – 85dB(A). The need to multitask was already incorporated in our generic protocol. Participants were ferried to/from the floating platform for their sessions; a member of the research team was positioned on the platform at all times with the participant. Participants completed all their training whilst on the platform. For this study, the same recruitment strategy was adopted as was used in the 'dry-ground' study; our participants included 4 males and 4 females, ranging from 21 – 30 years old.

7. RESULTS & DISCUSSION

As can be seen, we began with rich ethnographic data about our complex environment of use. From this, we extracted the key environmental and contextual elements that had the potential to impact the feasibility of speech-based input. To retain control over aspects associated with multitasking across all our evaluation environments, we opted for an abstract distractions task (as opposed to relying on uncontrollable real-world distractions in the field study). Whilst motion came with the territory in the field study, we incorporated its effect in the lab studies using lab-specific (but, we believe, compatible) mechanisms. Ambient noise was also an environmental staple in the field study; we

relied on electronic replication of a recording of the field environment to bring this element in to the lab studies. In all cases, we used a dedicated test application to focus solely on the input of data using speech – hence our study was able to focus on the feasibility of an isolated data input technique (and its consequential impact on users' environmental cognizance and workload) as opposed to attempt a full usability study (which was not appropriate in our context).

For the purpose of observing differences and similarities in the data we obtained across our studies, we consider each of our 'dry-ground' lab-based study conditions independently – thus we refer to five studies or conditions/groups, namely: our field study; our OEB (or 'middle-ground') study; and our three 'dry-ground' lab-based study conditions – Lab A, B, and C. Where relevant, we use scatter plots of the actual data to demonstrate the pattern of distribution of results across our study groups; we feel that this, in the context of such a study, is of equal (if not more) value to reported statistical significance.

7.1 Data Entry Accuracy Rates

We adopted two measures of data input accuracy: (a) an *average* accuracy rate; and (b) a *first entry* accuracy rate. In terms of (a), we calculated an overall accuracy rate for each data input item by analyzing each word (or data item) and assigning a score of 1, ⅔, ⅓, or 0 if the data was entered correctly on the first, second, or third try, or not at all, respectively. For each participant, we totaled the weighted accuracy scores and divided the total by 79 to determine the *average* accuracy rate per participant, which we then represented as a percentage of the maximum possible score (i.e., 79). We calculated our second measure of accuracy – *first entry* accuracy – to reflect the fact that in safety critical systems it would be essential that correct data entry was achieved on first attempt. For each participant, we totaled the weighted accuracy scores for all items where the participant achieved a score of 1 (i.e., correct entry on first attempt) then divided this by 79 to give us participants' average *first entry* accuracy, which we also represented as a percentage of the maximum possible score (i.e., 79)

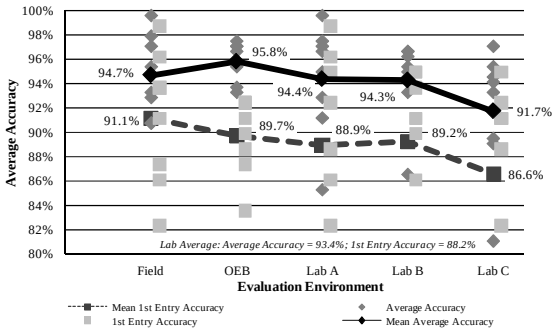

Figure 8: Mean average and first entry accuracy rates according to group (showing scatter plot of actual results).

Figure 8 shows the mean *average* and *first entry* accuracy rates across our five study groups. Across all 40 participants, irrespective of study group, we observed a mean *average* accuracy rate of 94.2%; this dropped to an average of 89.1% for *first entry* accuracy. There were no statistically significant differences between the results from our five study groups with regards *average* accuracy rates. We had anticipated that the loud ambient background noise prevalent on lobster fishing vessels would have severely impacted the *average* accuracy achievable; this does not, however, generally appear to have been the case. This suggests that the tonal make-up and/or volume of the specific background noise is such that it does not typically occlude the human voice as picked up by a microphone, or evoke Lombard Speech to an influential extent – further research would be required to determine which. The similarity of results returned by each of our study groups suggests that our attempts to bring the "high seas" into the labs have been reasonably successful: at least superficially, our five evaluation environments appear to have returned commensurate results. Of particular interest, is that the increased control over ambient noise levels afforded us in the 'dry-ground' lab-based study allowed us to observe a drop off in accuracy in Lab Group C (the noisiest condition), such that we can see the potential ambient noise threshold at which speech-based input may begin to become less effective/feasible.

An ANOVA test revealed that *first entry* accuracy rates were significantly impacted by evaluation environment ($F_{4,394}$=3.63, p=0.007). Tukey HSD tests indicated that participants in Lab Group C returned significantly lower *first entry* accuracy rates (on average, 86.6%) than participants in the field study (on average, 91.1%). We believe that this observation may be due to one or both of two factors: (a) although the ambient noise levels used in Lab Group C reflected the maximal ambient noise measured in the field, we artificially maintained the background audio at a constant volume (~86dB(A)) for the entire duration of the lab-based study session – in the field study, the ambient noise fluctuated over time – and so Lab Group C represents a worse case scenario than the field with respect to the intensity of ambient noise; and (b) our field study participants were all lobster fishermen and so were naturally accustomed to accommodating the ambient noise levels when conversing with each other onboard the lobster fishing vessels. Whilst, as we discuss in section 7.6, we had little control over the greater-than-ideal heterogeneity of users between our field and lab groups, these observations show the two sides to lab-afforded control: on the one hand, we *can* (unlike in the field) hold ambient noise at given levels in order to identify thresholds of speech feasibility but, in so doing, we increase the artificiality of the environment and so have to question the meaningfulness (albeit not the interest value) of the results relative to the real-world context.

7.2 Speech Entry Errors

We felt that it was interesting, and important, to consider the types of errors that resulted in incorrect data inputs. Hence, for every *incorrect* data input attempt, we analyzed the corresponding audio files and classified the types of errors that led to the failure. We identified 5 classes of error as shown in Table 1. Since each applicable audio file was subjectively assessed in order to classify the nature of the error it embodied, we attribute no statistical significance to the results and caution that our findings in this regard be taken as indicative and informative rather than definitive. That being said, it is important to identify wherein the likely source of error lies in order that systems developed using speech can best be optimized for maximal accuracy.

Table 2 shows the number of input errors made according to error type and study group. Across all 40 of our participants, irrespective of study group, participants returned 554 failed input

attempts, 141 (or 25.5%) of which were the direct result of human error (error types 1-3); the majority of human errors were the result of problems with the push-to-talk facility (error type 3), and this was consistent across all study groups. Participants in Lab Group C returned the most *clear* and *distorted* errors; as previously discussed, we believe that this may be an indication of a threshold at which we begin to observe the presence of signal-to-noise-ratio (SNR) issues and Lombard Speech. Error types across all study groups were commensurate in all other respects.

Error Type	Description
Clear	Utterance was clear and correct to the human ear but the SRE was unable to interpret it correctly. This is considered a problem with the SRE.
Distorted	Participant spoke either too loudly, softly, or breathed too heavily into the mic, distorting the audio and making it hard for the SRE to interpret.
Type 1	Occurred when a participant spoke a different word than asked for.
Type 2	Occurred when the correct word was spoken but was mispronounced.
Type 3	Occurred when parts of an input were cut off due to a participant releasing the mic too early before finishing a word or starting to speak too soon before the mic was fully engaged.

Table 1: Error classification.

Error Type	Number of Instances					Total
	Field	OEB	Lab A	Lab B	Lab C	
Clear	78	50	76	76	103	383
Distorted	1	1	2	0	26	30
Type 1	1	4	4	0	3	12
Type 2	5	1	0	0	0	6
Type 3	16	24	25	33	25	123
Total	**101**	**80**	**107**	**109**	**157**	**554**

Table 2: Input errors according to type and study group.

7.3 Task Completion Times

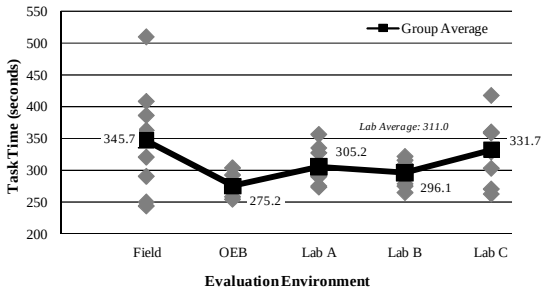

Figure 9: Average task completion time according to group (showing scatter plot of actual results).

Across all 40 of our participants, irrespective of study group, participants took an average of 310.8 seconds to complete all 79 data inputs (an average of 3.9 seconds per item). The task completion times ranged from a minimum of 243.0 seconds (3.1 seconds per item) to a maximum of 508.9 seconds (6.4 seconds per item) – see Figure 9. Both extremes were returned by participants in the field study; the maximum time (508.9 seconds) was considerably higher than the next longest task completion time (407.5 seconds) within this group, and so represents somewhat of an outlier. Although there are visible differences in the average task completion times returned for each of the study groups, the differences are not statistically significant. Although participants in the field study took longer, on average, to complete their tasks than the participants in the remaining study groups, the lack of statistical significance with respect to these differences leads us to conclude that, at least in terms of the measure of task completion time, the various evaluation environments were able to return commensurate results – i.e., that our abstract representations of the real world in the lab settings did not make the tasks any easier or faster, on average, to perform.

7.4 Distractions Identified

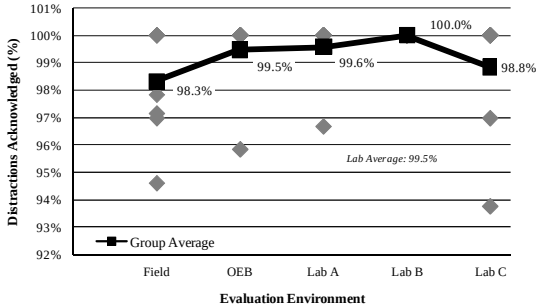

Figure 10: Average % of distractions acknowledged according to group (showing scatter plot of actual results).

Across all 40 of our participants, irrespective of study group, participants successfully reacted to an average of 99.2% of the distractions to which they were exposed during their study sessions. There were no statistically significant differences in terms of the percentage of distractions identified between any of our study groups. Figure 10 shows, according to study group, the average percentage of distractions identified by participants. Although we might have expected the complexity of the field study environment to have more substantially impeded participants' ability to monitor the distractions (i.e., there were more extraneous environmental distractions in the field than in our more controlled environments), it would appear that our lab replications were sufficient to return commensurate data.

7.5 Workload Experienced

Across all 40 of our participants, irrespective of study group, participants did not seem to consider workload excessive; on average, they rated the overall workload (which represents an average of all six workload dimensions, and uses the inverse of performance level achieved) as a mere 6.2 out of 20. Table 3 shows the average ratings according to dimension and group. As with the preceding measures, these results suggest that our abstract representations of the real world were sufficiently effective as to not make the tasks any easier or faster (on average) to perform compared to the field environment.

	Field	OEB	Lab A	Lab B	Lab C	*Average Lab*
Mental Demand	8.3	9.0	9.1	9.9	8.4	*9.1*
Physical Demand	4.4	3.4	4.9	6.8	4.0	*5.2*
Temporal Demand	5.5	6.8	5.8	8.3	8.0	*7.3*
Effort	7.0	8.9	6.3	7.5	7.8	*7.2*
Frustration	3.9	4.4	6.6	6.1	6.3	*6.3*
Performance	17.4	16.9	17.5	16.4	14.0	*16.0*

Table 3: Average workload ratings according to group.

7.6 Research Limitations

We acknowledge that our research is not without its limitations. We did not have complete control of the physical environment and participant recruitment during our field study; that said, in conducting our field study as we did, we worked with the in situ

population and prevailing weather conditions. Our field study therefore remained ecologically valid and so not only returned informative data that contributes to our understanding of the feasibility of speech in marine applications, but presents us with a ground truth (of sorts) against which to compare the effectiveness of our efforts to replicate the *relevant* aspects of the environment within our lab-based studies.

The use of fishermen was the ideal user group by which to investigate the feasibility of speech-based technology onboard lobster fishing vessels, and so contributed to *meaningful* field study data. That said, in terms of observing differences and similarities between the results returned by our various studies in order to reflect on the adequacy of our attempts to bring the novel, and complex, real-world environment 'into' the lab, we acknowledge that the ideal situation would have been to have used the same set of fishermen (representative users) in all our environments, counterbalancing their exposure to each of the environments to mitigate against learning effects. Unfortunately, however, this was not feasible given the geographical separation of our various evaluation venues together with the logistical impossibility of coordinating different orders of exposure for different users in the study given the weather conditions dictating field study sessions and the availability (and set up/tear down constraints) of a specialized resource such as the OEB. We are comfortable, however, that the restrictions we placed on our participant recruitment reduced, as far as was practically possible, the influence of individual participants.

One might, in light of the results obtained (i.e., the fact that, on average, participants correctly reacted to 99.2% of distractions), question the validity of the complexity of our distractions task. Further investigation would be necessary to determine whether the distraction requirement was sufficient. That being said, its impact appears to have been commensurate across all of our study sessions and, furthermore, previous studies have shown that speech permits users to remain more environmentally cognizant than other interaction mechanisms [12]; as such, our results may, indeed, be representative of participants' ability to use speech to effectively enter data while successfully monitoring other technology in their physical environ.

We note that, given the cavernous space in which the OEB is located, we found it hard with the technology available and environmental acoustics within our physical environmental constraints to bring the ambient noise levels up to the same maximum as the other study environments (the field and Lab Group C). We recognize this may have been an underlying reason for the slightly higher accuracy rates returned by participants in the OEB study, but also stress that this difference in accuracy was not statistically significant.

We recognize that the ethnographic observations which informed our environmental designs were based on a *typical* fishing day. As such, we appreciate that our results are limited to reflecting the contextual impact of a *typical* day on the ocean for a lobster fishing crew; they do not reflect the potential impact of more extreme weather conditions. That being said, we were informed that a lobster fishing vessel will not typically go out in more extreme conditions. We would, therefore, suggest that, albeit representative of a *typical* fishing day, we still based our study on what arguably represents (a) an extreme scenario for speech input, and (b) a complex evaluation environment in which to study speech input feasibility – that is, a small, diesel-engined fishing vessel in the winter in the Atlantic Ocean!

8. SUMMARY & CONCLUSIONS

In this paper, we have reflected on a comprehensive study of the feasibility of using speech within a complex real-world environment – namely, a lobster fishing vessel. In particular, we have discussed the process by which we attempted to recreate, within lab infrastructures that grew increasingly physically remote from the original environment of use, the environmental factors *relevant* to the use of speech within lobster fishing-based data-acquisition applications.

In reflecting on observed differences and similarities between the results returned in each of our studies, we believe that they suggest our attempts to bring the real world into the lab were successful. In the case of *first entry* accuracy and data input error types, we have additionally highlighted the paradox of control afforded by a 'dry-ground' lab-based study: it *can* support the identification of issues that would not otherwise be identifiable but, if over rigorously applied, can also lead to less meaningful results even when it is designed to maintain ecological validity or contextual relevance. We believe that our indication of a threshold ambient noise level at which speech may prove problematic is important, but that it is equally important to treat the finding with the caveat that, only rarely in reality, will such an extreme, sustained level of ambient noise be encountered!

In conducting this research, we were engaged with an application domain that represents an *extreme* context of use – and one that we believe is novel to the mobile HCI community; at the outset, we were not sure whether we *could* design a 'dry-ground' lab-based study that would hold its own against more ecologically true evaluation environments. We believe we have shown that it *is possible* to meaningfully abstract and bring relevant aspects of the real world into the lab in order to evaluate speech feasibility for use on lobster fishing vessels and, what's more, to seemingly do so such that the results obtained are commensurate with not only an environment that is closer to the real world, but also with the real-world context itself.

We acknowledge that the case study we have presented here is just *one* example of bringing a complex – and novel – context of use into play within a lab-based study. That said, to the best of our knowledge, our case study is unique and, thus, it contributes interesting, novel understanding to the ongoing evaluation debate. Furthermore, whilst our focus is not commensurate with prior studies that have contributed to the debate, we believe we bring another valid perspective to the discussion.

We would remind readers that we were focused on a *feasibility* study of a *given* technology relative to a *given* context of use: our primary intent here is to demonstrate that it is possible, even for environmentally complex scenarios, to develop appropriate lab-based studies based on diligent observation of the real-world context that return meaningful (in that they are commensurate with field study) results. Our hope is that, by reflecting on our case study, other researchers will be encouraged to explore means by which to bring other complex real-world environments into the lab with the confidence that, if appropriately executed, the lab-based studies can potentially return reliable and meaningful data. Thus, we can establish a body of knowledge to increasingly guide

the development of environmentally relevant lab-based mobile evaluation studies.

Furthermore, we hope that we have exposed a need to look at methods for mobile evaluations at a more fine grained or focused level: it is not always our intent to holistically evaluate usability, user performance, or user behaviour with respect to a complete system – sometimes, we just need to know if a component will work within a given context of use. In this sense, whilst with this paper we don't contribute to the ongoing evaluation debate in terms of the effectiveness of lab v. field studies for *usability* or *user performance measurement*, we hope that we have provided food for thought in terms of the adoption of lab and field approaches to the investigation of the feasibility of specific technological elements within mobile UI designs.

All evaluation methods have their advantages and disadvantages; the problem for the researcher is to pick the most appropriate method for a given evaluation purpose. By reflecting on our own experience of designing, and thereafter comparing, studies of speech input feasibility for lobster fishing-based application we hope that we have at least contributed to the community knowledge base such that we have, in some small way, plugged some of the holes we encountered when trying to select an evaluation method for our purpose!

9. REFERENCES

[1] Baillie, L. and Schatz, R., (2005), *Exploring Multimodality in the Laboratory and the Field*, in Proc. of ICMI'05, ACM, 100-107.

[2] Crossan, A., Murray-Smith, R., Brewster, S., Kelly, J., and Musizza, B., (2005), *Gait Phase Effects in Mobile Interaction*, in Extended Abstracts of CHI'05, ACM, 1312 - 1315.

[3] Duh, H.B.-L., Tan, G.C.B., and Chen, V.H.-h., (2006), *Usability Evaluation for Mobile Deivce: A Comparison of Laboratory and Field Tests*, in Proc. of MobileHCI'2006, ACM, 181 - 186.

[4] Hart, S.G. and Wickens, C., (1990), *Workload assessment and prediction*, in *MANPRINT: an approach to systems integration*, Booher, H.R., (Ed.), Van Nostrand Reinhold: New York, p. 257 - 296.

[5] Holtz Betiol, A. and de Abreu Cybis, W., (2005), *Usability Testing of Mobile Devices: A Comparison of Three Approaches*, in Proc. of INTERACT'2005, Springer, 470-481.

[6] Kaikkonen, A., Kallio, T., Kekalainen, A., Kankainen, A., and Cankar, M., (2005), Usability Testing of Mobile Applications: A Comparison Between Laboratory and Field Testing, *Journal of Usability Studies*, 1(1), 4-16.

[7] Kaikkonen, A., Kekalainen, A., Cankar, M., Kallio, T., and Kankainen, A., (2008), Will Laboratory Tests be Valid in Mobile Contexts?, in Handbook of Research on User Interface Design and Evaluation for Mobile Technology, Lumsden, J., (Ed.), Information Science Reference: Hershey, p. 897-909.

[8] Kjeldskov, J., Skov, M.B., Als, B.S., and Høegh, R.T., (2004), Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field, in Proc. of MobileHCI'04, 61 - 73.

[9] Kjeldskov, J. and Stage, J., (2004), New Techniques for Usability Evaluation of Mobile Systems, *International Journal of Human Computer Studies (IJHCS)*, 60(5-6), 599 - 620.

[10] Lumsden, J., Durling, S., and Kondratova, I., (2008), *A Comparison of Microphone and Speech Recognition Engine Efficacy for Mobile Data Entry*, in Proc. of OTM 2008 Workshops, Springer-Verlag, 519–527.

[11] Lumsden, J., Kondratova, I., and Durling, S., (2007), *Investigating Microphone Efficacy for Facilitation of Mobile Speech-Based Data Entry*, in Proc. of British HCI'2007, BCS, 89-98.

[12] Lumsden, J., Kondratova, I., and Langton, N., (2006), *Bringing A Construction Site Into The Lab: A Context-Relevant Lab-Based Evaluation Of A Multimodal Mobile Application*, in Proc. of MAPS'2006, IEEE, 62 - 68.

[13] Lumsden, J., Langton, N., and Kondratova, I., (2008), *Evaluating the Appropriateness of Speech Input in Marine Applications: A Field Evaluation*, in Proc. of MobileHCI'2008, ACM, 343 - 346.

[14] Mustonen, T., Olkkonen, M., and Hakkinen, J., (2004), *Examining Mobile Phone Text Legibility While Walking*, in Extended Abstracts of CHI'2004, ACM, 1243 - 1246.

[15] Nielsen, C.M., Overgaard, M., Pedersen, M.B., Stage, J., and Stenild, S., (2006), *It's Worth the Hassle!: The Added Value of Evaluating the Usability of Mobile Systems in the Field*, in Proc. of NordiCHI'06, ACM, 272-280.

[16] Oviatt, S., (2000), Taming Recognition Errors with a Multimodal Interface, *Communications of the ACM*, 43(9), 45 - 51.

[17] Price, K., Lin, M., Feng, J., Goldman, R., Sears, A., and Jacko, J., (2004), *Data Entry on the Move: An Examination of Nomadic Speech-Based Text Entry*, in Proc. of UI4All'04, Springer-Verlag LNCS, 460-471.

[18] Sawhney, N. and Schmandt, C., (2000), Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments, *ACM Transactions on Computer-Human Interaction*, 7(3), 353 – 383.