Evaluating the Appropriateness of Speech Input in Marine Applications: A Field Evaluation

Joanna Lumsden
National Research Council of Canada
46 Dineen Drive, Fredericton, N.B., Canada, E3B 9W4
+1.506.444.0382
jo.lumsden@nrc.gc.ca

Nathan Langton
National Research Council of Canada
46 Dineen Drive, Fredericton, N.B., Canada, E3B 9W4
+1.506.444.0533
nathan.langton@nrc.gc.ca

Irina Kondratova
National Research Council of Canada
46 Dineen Drive, Fredericton, N.B., Canada, E3B 9W4
+1.506.444.0489
irina.kondratova@nrc.gc.ca

ABSTRACT
This paper discusses the first of three studies which collectively represent a convergence of two ongoing research agendas: (1) the empirically-based comparison of the effects of evaluation environment on mobile usability evaluation results; and (2) the effect of environment – in this case lobster fishing boats – on achievable speech-recognition accuracy. We describe, in detail, our study and outline our results to date based on preliminary analysis. Broadly speaking, the potential for effective use of speech for data collection and vessel control looks very promising – surprisingly so! We outline our ongoing analysis and further work.

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

General Terms
Human Factors, Performance, Experimentation, Measurement.

Keywords
Speech Input, Evaluation, Field Study.

1. INTRODUCTION
As mobile HCI evolves as a discipline, a number of novel evaluation approaches are being conceived [e.g., 1, 3]. Simultaneously, the benefits of lab evaluations over field evaluations are subject to much ongoing debate [e.g., 6, 7].

It is argued that usability evaluations of mobile applications should always be conducted in the field to increase the likelihood of a realistic evaluation context (although this is not always the case [8, 11]) and thereby secure more meaningful results. Field evaluations have, however, some obvious disadvantages such as difficulties regarding data collection and limitations of control over the experiments [7, 8, 11].

MobileHCI 2008, September 2–5, 2008, Amsterdam, the Netherlands.
ACM 978-1-59593-952-4/08/09.

Conversely, it is argued that lab-based evaluations afford greater experimental control and easier data capture (and therefore greater data integrity) than field evaluations because it is possible to incorporate high quality data collection methods whilst ensuring the safety of study participants [8, 11]. Additionally, researchers are increasingly generating novel means by which to enhance the contextual relevance of lab studies [2, 7, 10].

Some initial studies have suggested that both field and lab evaluations report the *same* usability problems [5], some claim that the two environments test *different facets of usability* [6], while others have demonstrated that lab evaluations can identify *more* usability problems, including context-specific problems, than field evaluations [7]. That said, the findings of the latter have been subjectively refuted on the grounds of differences in the task assignments and data collection techniques used in both environments [11].

Researchers are only beginning to explore the pros and cons of lab v. field usability evaluations. The relative infancy of this discipline, and the corresponding debate, means that literature on experimental comparisons of field v. lab evaluation methods is scarce, with the debate often being a matter of opinion [11].

Our overall research agenda has two goals: (1) to compare the effect of evaluation environment (based on three different levels of abstraction of the same context-of-use), when applied to identical task assignments and when utilizing identical data collection techniques, on the experimental results obtained during a mobile usability study; and (2) to ascertain whether speech-based input is a viable interaction mechanism for use in a marine environment – specifically, a lobster fishing boat. In this paper we discuss the first of a trio of studies designed to meet these goals. At time of writing, we are actively engaged in the second study (a lab study), and are preparing for the third (a pseudo-lab study in a wave tank); in this paper, however, we focus on the performance and results of the first – a field study. Section 2 briefly outlines the usability issues associated with speech-based technologies. Sections 3 and 4 then describe our experimental design and discuss our *preliminary* results, respectively. We conclude, in Section 5, with a discussion of further work.

2. SPEECH INPUT
Nominated as a key interaction technique for supporting mobile users of technology, speech-based input offers a relatively hands-free means of interaction and, it is argued, has the capacity to extend the functionality of mobile technologies across a broader

range of usage contexts [14, 15]. Compared with other input techniques, speech has been shown to enhance mobile users' cognizance of their physical environment while interacting with mobile devices [10]. It is, however, estimated that a 20%-50% drop in recognition rates can occur when speech is used in a natural field setting as opposed to a controlled environment [12, 15]. Given that accuracy is a significant determinant of users' perception of speech recognition usability and acceptability [14], developing *effective* speech-based solutions for use in *mobile* contexts – where users are typically subjected to a variety of additional stresses, such as variable noise levels, a need to multitask, and increased cognitive load [12] – is challenging [14].

Two main problems contribute to the degradation of speech recognition accuracy in mobile contexts: (1) people speak differently in noisy conditions; and (2) background noise can contaminate the speech signal, with the result that recognition accuracy has been shown to steeply decline in even moderate noise [12]. In noisy environments, speakers exhibit a reflexive response known as the Lombard Effect which results in targeted speech modifications [12, 13], such as changes in volume and pronunciation (hyperarticulation). Space does not permit us to elaborate on ongoing research to improve speech recognition accuracy for mobile contexts; we would, however, refer readers to [9] for a review of the field. In the research presented in this paper, our focus is to determine the effect of context – specifically, that of a lobster boat – on the accuracy obtained when using speech-based input.

3. STUDY DESIGN & PROCESS

Our field study was designed on the basis of (a) previous ethnographic studies of the environment – a lobster fishing boat – and (b) the requirement that our experimental tasks and data collection mechanisms remain constant across our three experimental set-ups. We are focusing on lobster fishing boats because: (a) we are working with a client who is developing software for use on lobster fishing boats and is seeking empirical evidence that speech would be a viable input option; and (b) in many senses, a relatively small, diesel-engine fishing vessel in the middle of winter in the Atlantic Ocean arguably represents a worst case marine scenario for use of speech input!

To avoid testing speech relative to a specific software application, we developed a very simple data input application which allowed us to evaluate speech-based input of different data types (see Figure 1 (left)). The application was designed to run on a Panasonic Toughbook running Windows XP. Informed by earlier studies [9, 15], we used IBM's ViaVoice speaker-independent speech recognition engine, adopted a *push-to-talk* strategy, and employed a Shure QSHB3 condenser microphone.

Figure 1 (left) shows a screen dump of the evaluation application. Whenever participants pressed and held the spacebar on the toughbook, the mic logo changed to an "on air" logo to reinforce the fact that the system was ready to receive input. Participants were shown a data item (in terms of what they are to physically say) on screen ("Five" in Figure 1) and were required to input that item using speech; the results of their input were reflected immediately underneath in the input field. Participants were required to achieve an accurate entry, upon which the system automatically moved them on to the next data entry item. In the interests of time (and to mitigate against potentially fuelling high user frustration), we restricted participants to three attempts per item (the number of available attempts was always shown by a counter in the top left corner of the screen); if, on their third attempt, participants still failed to achieve a correct entry, the system automatically moved onto the next item and the attempts counter was reset. Participants were given training on how to use speech to enter data prior to commencing the study tasks. They were trained in conditions identical in all aspects to those used in the study sessions themselves.

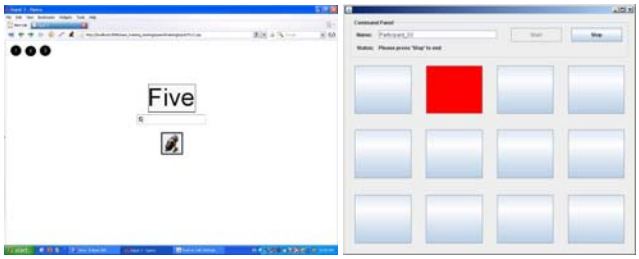

Figure 1. Speech input application (left) and distraction application (right).

As can be seen from Figure 2, lobster boat crews are required to simultaneously monitor and interact with a plethora of electronic devices. To reflect this in our experimental set-up we developed a very simple 'distractions application' (see Figure 1 (right)): it runs simultaneously to the data input application and mimics the need to monitor ancillary technology by displaying a sequence of red squares, in a preset (pseudo-random) pattern of location, interval, and duration, which the study participants are required to acknowledge by tapping (via the touchscreen). Although one could argue that we could have more realistically achieved the same effect by simply introducing the speech-input application to the working environment onboard the boats, this would not have been replicable in our other study environments; it was also not possible from an ethical/safety perspective. Figure 2 (right) shows the two toughbooks (one for each system) set up in the cabin of one of the boats on which we conducted our field studies.

Figure 2. Cabin of a lobster boat (left) and Toughbooks set up in situ (right).

Participants were first trained (in situ) on the speech-based application, then the distractions application, and then were given an opportunity to familiarize themselves with using both in parallel. During the experiment, we took a range of measures to assess the efficacy of speech: we recorded the length of time participants took to complete each of their (79) data entry tasks and the details of the data they entered; we recorded details of participants' responses to the 'distractions'; finally, we asked participants to subjectively rate the workload involved with the study tasks (using the NASA TLX scales [4]).

Acknowledging lack of complete control over the environment of field studies, we recruited (on a voluntary basis) the assistance of

lobster fishermen to conduct our field experiments: we had no control over the demographics of the crews on the boats, nor were we able to control the prevailing weather conditions – we merely accompanied crews on (typically scheduled) fishing trips. We were able to complete 3 separate trips on lobster fishing boats during which we were able to run 8 fishermen in total through the experimental session; prevailing weather, whilst not ideal given the winter conditions, was relatively consistent and typical of the conditions in which the boats normally operate. Fishermen participated in our study sessions at times when they were not otherwise engaged in mission-critical activities (e.g., en route to/from trap lines as opposed to when hauling in/laying the traps). Participation took approximately 45 minutes in total per person; participants were not compensated for their time. Our participants were all males, aged 18-45 years.

4. RESULTS & DISCUSSION

In the following discussion, we highlight our preliminary findings; additionally, we outline the ongoing analysis in which we are actively engaged.

Engine noises on the fishing boats were in the range 80dB(C) – 100dB(C) (approx. 70db(A) – 95dB(A) adjusted for human hearing). On the basis of previous research which has shown the negative impact of ambient noise on speech-recognition accuracy, we anticipated realizing very low accuracy rates in our field study. We adopted a simple measure of accuracy per data input: namely ultimate success (a Boolean value of 1 for correct and 0 for incorrect after exhausting all available tries) divided by the number of tries used. We observed an average accuracy rate of 94.7% which was much higher than we anticipated. Given that in safety critical systems it would be essential that correct entry was achieved on *first* attempt, we re-assessed our results to look at the extent to which participants achieved successful entry on their initial attempt: on average, 91.1% of data inputs were correctly interpreted first time. This is by no means ideal, but it *is* extremely encouraging.

As previously mentioned, our data set comprised 79 data items. The items were selected on the basis of vocabulary appropriate to the lobster fishing industry as well as commands typical for vessel navigation. Our data types, therefore, included: dates; digits (numbers entered by saying, for example, "One Five" for 15); decimals; numbers (numbers entered by saying, for example, "Fifteen" for 15); and a series of control commands which we subdivided according to function – fishing v. navigation.

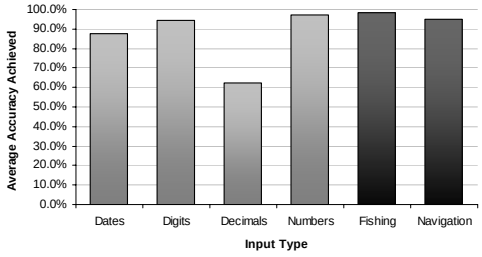

Figure 3. Average accuracy achieved according to type.

Figure 3 shows the average accuracy rates that were achieved according to our type classification. An ANOVA test showed that accuracy was significantly affected by data type ($F_{5,78}$=8.74, p<0.001). Tukey HSD tests showed that decimals were significantly less accurately entered than all other data types; there were no other statistically significant differences between data types. Interestingly, the command inputs (darker grey bars in Figure 3) achieved high levels of recognition accuracy; in the case of navigation commands, one particular command ("Aft") caused problems for the majority of participants; with it removed from the calculations, the average accuracy rate rises from 95.1% (as shown in Figure 3) to 99%. These results would suggest that there is real potential to use speech to issue commands to marine systems; in fact, it would appear that commands have the accuracy edge on data input. On average, it took participants 342.9sec to complete all 79 data inputs (an average of 4.3sec per item). Task completion times ranged from a minimum of 241.4sec (3.1sec per item) to a maximum of 502.5sec (6.4sec per item). Whilst the aforementioned data accuracy rates are obviously encouraging, it is not so immediately obvious whether the time associated with achieving these rates is equally encouraging. Further study would be necessary to draw conclusions in this respect.

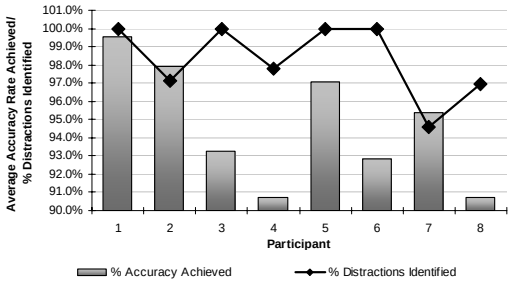

Figure 4. Average accuracy achieved and distractions identified across participants.

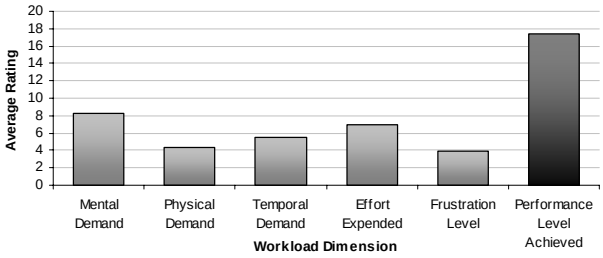

Figure 5. Average workload ratings according to dimension.

On average, participants successfully reacted to 98.3% of the distractions to which they were exposed during their study session. As can be seen from Figure 4 – in which average accuracy rates are shown alongside the percentage of distractions identified by participants – while some participants seemed to achieve similar performance on both systems, there is considerable disparity between others' performance across the data input and distraction tasks. We are actively engaged in detailed analysis in order to model the actual patterns of activity and voice nuances underpinning these observations in order to determine, if possible, cause and effect – for example: did participants make mistakes on speech entry *because* they were physically reacting to a distraction or vice versa?; did participants demonstrate a tendency to concentrate more fully on one application or the other?; were there specific nuances in participants' speech that impacted their input?; and what is the breakdown of error type (e.g., human v. mechanical)?

Finally, participant responses with respect to subjective assessment of workload are encouraging (see Figure 5). For all bar the performance level (see darker grey bar), a lower value is preferable (representing lesser perceived workload). It would seem that participants were generally very happy with their performance (as well they should be given the accuracy rates achieved); they also did not consider the workload excessive (average workload across the first 5 dimensions being a mere 5.8).

The above workload ratings are exemplified in the following comment made by one of the participants: "*This was quite a bit easier than I thought it would be. It didn't seem to be hard at all. Worked Well.*" – and encouraging reaction.

5. CONCLUSIONS & FUTURE WORK

Contrary to expectation, we found that the ambient noise prevalent in our field studies had little noticeable impact (relative to the anticipated 20%-50% drop) on recognition accuracy. Surprised by this, we tightened up our measure of accuracy to consider only those items entered correctly on first attempt; this, too, returned a higher than expected, and encouraging, accuracy rate. Similarly, the accuracy achieved according to data type suggests that there is potential scope for vessel control as well as data collection using speech. Speech did not appear to impose excessive workload on our participants, even with the combined need to monitor and react to the distractions application – an encouraging indicator of potential workload in real use. In summary, therefore, our *preliminary* analysis indicates there is considerable promise for the effective use of speech in marine environments.

In the previous section, we highlighted a series of questions which we are actively investigating. Pursuing answers to these questions will allow us to form a more complete picture of the appropriateness of speech in marine environments. Furthermore, we hope to complete our additional evaluation studies in the coming months; thereafter, we will be able to engage in analysis of the results to determine the effect of evaluation environment.

6. ACKNOWLEDGMENTS

Our thanks to our collaborative partner (Trapster Inc.) and to the crews of the lobster fishing boats, without whose assistance this field study would not have been possible.

7. REFERENCES

[1] Alexander, T., Schlick, C., Sievert, A. and Leyk, D., 2008. Assessing Human Mobile Computing Performance by Fitt's Law. in Lumsden, J. ed. *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, Information Science Reference, Hershey, 830-846.

[2] Crease, M., Lumsden, J. and Longworth, B., 2007. A Technique for Incorporating Dynamic Paths in Lab-Based Mobile Evaluations. In *Proceedings of the British HCI'2007 Conference*, (Lancaster, UK, 2007), BCS, 99-108.

[3] Crossan, A., Murray-Smith, R., Brewster, S. and Musizza, B., 2008. Instrumented Usability Analysis for Mobile Devices. in Lumsden, J. ed. *Handbook of Research on User Interface Design and Evaluation for Mobile Technology*, Information Science Reference, Hershey, 927-944.

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

[5] 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.

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

[7] 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 *Proceedings of the 6th International Symposium on Mobile Human-Computer Interaction (MobileHCI'04)*, (Glasgow, Scotland, 2004), 61 - 73.

[8] 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.

[9] Lumsden, J., Kondratova, I. and Durling, S., 2007. Investigating Microphone Efficacy for Facilitation of Mobile Speech-Based Data Entry. In *Proceedings of the British HCI'2007 Conference*, (Lancaster, UK, 2007), BCS, 89-98.

[10] 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 *Proceedings of the 1st International Workshop on Multimodal and Pervasive Services (MAPS'2006)*, (Lyon, France, 2006), IEEE.

[11] 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 *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles (NordiCHI'06)*, (Oslo, Norway, 2006), ACM Press, 272-280.

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

[13] Pick, H., Siegel, G., Fox, P., Garber, S. and Kearney, J., 1989. Inhibiting the Lombard Effect. *Journal of the Acoustical Society of America*, *85*(2). 894 - 900.

[14] 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 *Proceedings of the 8th ERCIM Workshop "User Interfaces For All" (UI4All'04)*, (Vienna, Austria, 2004), Springer-Verlag LNCS, 460-471.

[15] 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.