Investigating microphone efficacy for facilitation of mobile speech-based data entry

Lumsden, Joanna; Kondratova, Irina; Durling, Scott

Publisher's version / Version de l'éditeur:

Proceedings of The 21rst British Computer Society (BCS) HCI Annual Conference, 2007

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC·CNRC

*Investigating Microphone Efficacy for Facilitation of Mobile Speech-Based Data Entry **

Lumsden, J., Kondratova, I., and Durling, S
2007

* Proceedings of the British HCI Conference. Lancaster, UK. 3-7 September 2007. NRC 49369.

Canada

Investigating Microphone Efficacy for Facilitation of Mobile Speech-Based Data Entry

Joanna Lumsden, Irina Kondratova
National Research Council of Canada
46 Dineen Drive, Fredericton, N.B.,
Canada, E3B 9W4
1.506.444.{0382,0489}
{jo.lumsden,irina.kondratova}@nrc-cnrc.gc.ca

Scott Durling
University of New Brunswick
Fredericton, N.B.,
Canada, E3B 5A3
1.506.444.0481
scott.durling@unb.ca

ABSTRACT
Despite being nominated as a key potential interaction technique for supporting today's mobile technology user, the widespread commercialisation of speech-based input is currently being impeded by unacceptable recognition error rates. Developing *effective* speech-based solutions for use in *mobile* contexts, given the varying extent of background noise, is challenging. The research presented in this paper is part of an ongoing investigation into how best to incorporate speech-based input within mobile data collection applications. Specifically, this paper reports on a comparison of three different commercially available microphones in terms of their efficacy to facilitate mobile, speech-based data entry. We describe, in detail, our novel evaluation design as well as the results we obtained.

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, microphone efficacy, mobile technology, evaluation.

1. INTRODUCTION
Speech-based input has been nominated as a key potential interaction technique for supporting today's *nomadic* or mobile users of technology [17, 20, 24, 28]. Offering a relatively hands-free means of interaction, it is argued that speech-based input has the capacity to heighten the functionality of mobile technologies across a broader spectrum of usage contexts [20]. Compared with other input techniques, speech has been shown to enhance mobile users' ability to be cognizant of their physical environment while interacting with mobile devices [9]. Given the typical multitasking nature of mobile users' behaviour – where an ability to monitor their surroundings is often essential to their safety – this increased capacity, in itself, suggests there is considerable merit to further investigating and improving speech as an input mechanism for use with mobile technologies.

To date, *"the Achilles' heel limiting widespread commercialisation of* [speech technology] *is the rate of errors and lack of graceful error handling"* [14, pg. 46]. It is estimated that a drop in recognition rates in the range of 20%-50% can occur when speech is used in a natural field setting as opposed to a controlled environment [9, 14, 24, 28]. Given that accuracy is a significant determinant of users' perception of speech recognition usability and acceptability [15, 20], 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 [14] – is challenging [16, 20]. Karat *et al* [7] suggest that errors encountered with speech recognition technology are fundamentally different to errors that arise with other interaction techniques, and that speech-based interaction will evolve as we identify, and strive to address, the problems encountered by users of current mobile application designs.

Acknowledging that recognition errors are problematic, researchers argue that much can be done to increase recognition accuracy and/or the ease with which users can correct errors in order to render speech-based interaction more reliable and functional [5, 9, 14, 20, 23]. Approaches taken include supporting the use of complementary input modes to achieve mutual disambiguation of input under mobile conditions [5, 6, 13, 14], investigating mechanisms to predict and counter speaker hyperarticulation [15, 18, 22], undertaking empirically-based context-specific selection of speech recognition engines to maximise accuracy [25, 27], investigating the effect of the enrolment (recogniser training) environment [20, 22] and variances in microphone between enrolment and actual use [1], and identifying the effect of background noise [9, 14, 20] and mobility [9, 20] on speech recognition accuracy and usability.

To develop effective and efficient speech-based user interfaces for mobile technologies, it is essential that mechanisms are found to decrease the recognition error rate; only then will the perceived usability of such systems be increased and user frustration decreased [20]. The research presented in this paper is part of an ongoing investigation into how best to incorporate speech-based input within mobile data collection applications. Specifically, we report on a comparison of different microphones in terms of their efficacy to facilitate mobile, speech-based data entry. Section 2 outlines relevant related work. Sections 3 and 4 then describe our experimental design

and discuss our results, respectively. We conclude, in Section 5, with a discussion of further work.

2. RELATED WORK

Two main problems are recognised as contributing to the degradation of speech recognition accuracy in mobile contexts [14]:

- people speak differently in noisy conditions; and
- background noise contaminates the speech signal.

In noisy environments, speakers exhibit a reflexive response known as the Lombard Effect which results in targeted speech modifications [14, 15, 18, 22]. These modifications are not limited to increase in volume, but instead include changes in pronunciation (hyperarticulation) which vary between speakers and based on the amount and type of noise [15, 18, 22, 28]. Studies suggest that the Lombard response is primarily automatic, and is not typically under volitional control [14, 18]; it has even been suggested that speech variation caused by ambient noise can be more degrading in terms of speech recognition accuracy than the noise itself [6]. Research has, consequently, demonstrated that it is not possible to eliminate or selectively suppress the effect of Lombard speech [14, 18]. As such, it is argued that traditional algorithmic approaches to speech recognition need to be adapted to accommodate dynamic stylistic changes in speech signals that are brought about by mobile speech input under noisy conditions [6, 14, 15]. The fundamental implication for speech recognition systems in the future is that they will have to be capable of handling variation in speech signals that come with mobile speech input [15].

As previously noted, under mobile conditions background noise can confuse, contaminate, or even drown out a speech signal; as a result, speech recognition accuracy has been shown to steeply decline in even moderate noise [14, 28]. Speech signals are picked up and delivered to speech recognition systems via microphones. Even under stationary conditions, variations in microphone type, placement, and quality have been shown to lead to different levels of user performance [1]; when the complexities of user mobility and non-static usage environments are introduced, the influence of the microphone itself becomes even more pronounced [4, 14, 23, 25, 27].

McCormick states that "*a quality microphone and quiet workplace are key to acceptable speech recognition*" [10, pg. 1]. Clearly, a quiet workplace is unlikely for mobile users. On the supposition that appropriate evolution of speech recognition algorithms is being addressed within the speech recognition research community, we focused our attention on how best to accommodate the combined effect of user mobility and background noise in order to make speech-based input on mobile devices a truly viable option. More specifically, the study presented in this paper focuses on the efficacy of different microphones – under different levels of background noise – to support *mobile* speech-based input. To situate our research study within the broader field, the remainder of this section outlines previous research that has been conducted to investigate (a) the impact of mobility and background noise on speech recognition, and (b) the influence of microphone type on speech recognition.

Price *et al* [20] investigated the effectiveness of speech-based input while walking. In particular, they assessed the effect of motion and enrolment condition (seated v. walking) on speech recognition accuracy for a speaker-dependent recognition engine coupled with an acoustic boom microphone; their primary concern was the effect on speech recognition accuracy when the enrolment condition differed to the actual use condition. They incorporated mobility by means of a treadmill, and negated possible effects of noise introduced by the treadmill by playing a recording of the treadmill noise during seated tasks. They found that recognition accuracy was significantly higher for seated tasks, but that there is potential to mitigate the effect of walking if enrolment is conducted under mobile (i.e., more demanding) rather than seated conditions. Their study stresses the importance of encompassing mobility in the design and testing of mobile speech-based input and, in particular, in enrolment (training) strategies; it was, however, limited to a single microphone and a speaker-dependent recognition engine. Furthermore, it utilised a fairly artificial environment in which little was done to incorporate environmental stress beyond the need to be mobile.

As part of a study considering the ability of multimodal interaction to support disambiguation of, and recovery from, speech recognition errors, Oviatt [14] compared a close-talking microphone incorporating noise-cancellation with the microphone built into the handheld PC which lacked noise-cancellation; the study also compared the effect of mobility (stationary v. walking) and background noise (42dB in a quiet room v. 40-60dB in a moderately noisy public cafeteria). They found mutual disambiguation led to a substantial improvement in recognition robustness, and that recognition rates were significantly less when users were mobile in a noisy environment than when they completed their tasks under stationary, quiet conditions. The advantage of mutual disambiguation was more pronounced under mobile conditions, but the extent of this advantage was dictated by the microphone being used. The comparison of the two microphones would not appear to have been Oviatt's primary concern and, as a result, limited detail is provided specific to the impact of the microphones themselves. Furthermore, at worst case the noise level introduced into the noisy condition reflects standard conversational speech (60dB) [11] and so this does not help determine the potential effect of, and microphone ability to handle, considerably higher background noise such as a noisy restaurant or highway traffic (70dB) or city traffic (90dB) [11].

In a previous study [9], we compared the use of speech input to stylus-based input for data entry in a mobile application designed for use on a construction site. We tested both input techniques for mobile data entry under three different levels of background noise – 70dB-80dB, 80dB-90dB, and 90dB-100dB – typical of a construction site. Our results showed that noise level significantly affected data entry precision when using speech and suggested that there may even be a threshold of approximately 80dB beyond which the accuracy achievable with speech input may prove unacceptable or indeed unusable [9]. Participants were significantly more satisfied with their own performance when using a stylus to enter data than using speech. Like other studies [14], we too found speech input to be the less stable input mode.

Sebastian [25] compared several commercial microphones in terms of their suitability for use with a mobile, voice activated ultrasound device. Six people participated in this study; their speech was recorded under static conditions in an acoustic chamber, and the resulting signal was fed (together with a separate feed of moderate background noise) to laptops running continuous speech recognition software. Sebastian determined that, for her specific context of use, whilst other microphones

showed some promise, the most appropriate microphone was a standard acoustic microphone. Unfortunately, because Sebastian restricted her voice capture to an acoustic chamber under *static* conditions, and only considered a very restricted set of commands specific to the ultrasound equipment, it is impossible to generalise her findings beyond the limited context of her study.

Vinciguerra [27] investigated which combination of speech recogniser and microphone would work best to support speech-based interaction with applications within a police car. Vinciguerra relied on a software application to test speech recognition and microphone combinations; the system played recorded speech files together with recorded (appropriate) background noise files to simulate a police officer issuing commands to the system within a police car. This study focused on three array microphones and one desktop microphone, all of which could be mounted on the dash of the police car. Although this study showed that each microphone performed differently depending on the recognition engine with which it was paired, none of the microphones tested are suitable (on the basis of physical form and function) for situations where the *user* is mobile and must physically carry the microphone.

Chang [1] sought to determine the effect of microphone variation. She noted that speaker, environment, and microphone can all contribute to variations in input signal to a speech recognition system, and that the position of the microphone relative to the speaker can cause distortion. To achieve the lowest error rates possible, most speech recognisers are trained and tested using high quality, head mounted, close-talking, noise cancelling microphones; Chang set out to determine what happens when the actual use microphone is substantially different to the training and testing microphone. She determined that most speech recognisers lack robustness to microphone variations and cannot, therefore, be used satisfactorily with microphones that do not match the microphone used in training.

Finally, Huerta [4] found that by making improvements in the speech codec used on GSM cellular networks, speech recognition error rates could be significantly reduced. This suggests that, by improving the quality of the signal reaching a speech recogniser, recognition accuracy may be improved.

Motivated by the results of our previous study [9], together with the research outlined here, we aimed to empirically compare the ability of three different microphones (appropriate in terms of form and function) to support accurate speech-based input under *realistic mobile, noisy conditions*. In doing this, we drew together into *one novel* evaluation, many of the constituent – and previously unconnected – elements of previous studies. The remainder of this paper discusses the design and findings of our evaluation.

3. EVALUATION DESIGN & PROCESS

Mobile computing generally relies on condenser microphones which operate on acoustic principles, picking up any sound waves with which they come in contact. The goal of our study was to compare the efficacy of two commercially available condenser microphones (each with different noise cancelling properties) with a commercially available bone-conduction microphone – see Figure 1 – in terms of their efficacy to facilitate mobile speech input.

Figure 1. Microphones evaluated (from left to right): NextLink Invisio Mobile (Bone Conduction) [12]; Shure QuietSpot QSHI3 [26]; and Plantronics DSP-500 [19].

The most novel of all three microphones, the bone-conduction microphone fits in the outer ear and detects audio signals conducted through bone vibration as a person speaks; this theoretically reduces the amount of background noise picked up since acoustic (environmental) vibrations are not detected. With all three microphones, we used an external USB audio interface to eliminate electrical interference from surrounding electronic components on the mobile computer.

To avoid testing the microphones relative to a specific application (or application domain), we developed a very simple data input application which allowed us to evaluate speech-based input of different data types (numbers without decimal points, decimal numbers, sequences of independent digits, and dates) and selection of drop-down list box items and radio buttons. The application was designed to run on a tablet PC running Windows XP. We used IBM's ViaVoice speaker-independent speech recognition engine on the basis that speaker-independent recognition systems have been shown to be more robust to noisy environments, and less susceptible to Lombard speech, than speaker-dependent systems [6, 15]. Furthermore, we adopted a *push-to-talk* strategy since, in noisy environments it is generally considered more appropriate as it enables users to explicitly direct speech to the system (or, conversely, deactivate recognition altogether) [24, 25].

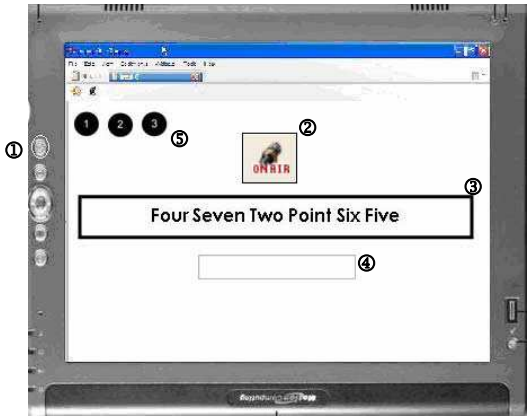

Figure 2. A screen dump of the evaluation application.

Figure 2 shows an annotated screen dump of the evaluation application. Whenever participants pressed and held the push-to-talk button (1) on the tablet PC, a large "on air" logo (2) was displayed in the centre of the screen to reinforce the fact that the system was ready to receive input. Participants were shown a data item (in terms of what they were to physically say) on screen (3) and were required to input that item using speech. The results of their speech input were reflected in an appropriate input field (4). Participants were required to achieve an accurate entry; upon an accurate entry, 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 in the top left corner of the screen (5)); 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/select each data type prior to commencing the study tasks. They were trained in conditions identical in all aspects (including microphone type) to those used in the study sessions themselves (see below).

Appropriately designed lab-based studies, which incorporate user mobility, have proven to be a viable means by which to meaningfully assess the usability of mobile applications under controlled experimental conditions [8, 9]. We therefore designed our study to reflect (albeit, abstractly) realistic environmental conditions in which mobile technologies are typically used, and required that our users were mobile within this environment. Our study adopted a counterbalanced between-groups design, with groups partitioned on the basis of microphone type. Using the application described above, participants were required to enter a series of simple data input tasks using speech while mobile. They were required to do this once in a quiet environment, to provide us with a baseline measurement of microphone efficacy, and once in an environment designed to represent a typical city street (a common use-case scenario of mobile technologies). The order in which these sessions were completed was counterbalanced across participants to mitigate against potential learning effect.

To establish the city street condition, we used the 7.1 surround sound system in our lab to deliver recorded city street sounds (at 70dB) around the participants. When a person is using mobile technology while walking down a city street, for instance, he must remain cognizant of his physical surroundings to avoid potential hazards. To ensure the results of our study would be meaningful, we therefore incorporated environmental awareness into participants' mobile data entry activities. To do this, we made use of an abstract (safe) "hazard avoidance" system which we developed for use in our lab. More detail on this novel experimental technique for abstractly recreating mobile real-world situations in the lab is provided in a companion paper in these conference proceedings [2].

In essence, participants were required to walk from one end of a grid (approx. 12m long) of coloured mats (typically used in children's play areas) to the other – see Figure 3.

Figure 3. Grid of mats and colour projections.

As they were doing this, we projected a sequence of blocks of colours (corresponding to the colours of the floor mats) on the walls facing them. Occasionally, we projected a colour block with the word "Avoid" written on it. This represented a "hazard". While this was being projected, participants were to avoid stepping on any floor mats which matched that specific colour (see Figure 4). Participants were given training on, and a chance to try out, the hazard avoidance system before they began the study tasks to make sure they were comfortable with the process. This technique required participants to be cognizant of their physical environment in a manner similar to the real world. We effectively incorporated a dynamically changing route based on hazard avoidance to ensure meaningful experimental results, without exposing participants to the risks associated with field-based trials (and allowing us to maintain control of the experimental environment).

Figure 4. Participant using the tablet PC to enter data via speech whilst avoiding "hazards".

Participants were assigned one of the three microphones. They were required to enter 10 data items per audio condition (silent v. noisy). We established two data sets, each with a unique set of 10 data entry items. We took care to balance the breakdown of data types and complexity across the two data sets as far as possible; the order in which data types were presented differed between the two sets to again mitigate against the learning effect, but the order in which participants were exposed to the data sets themselves remained constant to eliminate any potential bias that may have arisen due to some data elements being perceived as 'easier' than others. During the experiment, a range of measures was taken to assess the efficacy of the three microphones. We recorded the length of time participants took to complete their data entry tasks and the details of the data they entered. We also recorded their walking speed – both natural (when not completing data entry tasks) and when performing the study tasks – to allow us to determine the impact on walking speed of interaction with the technology. Additionally, after each session (silent v. noisy) we asked participants to indicate their subjective experience of workload (using the NASA TLX scales [3]).

Studies have shown that speech recognition is significantly affected by native speaker status [13]; speech recognition rates are typically much lower for accented speakers [14]. Additionally, the Lombard speech of female speakers has been shown to be more intelligible than the Lombard speech of male speakers and that, surprisingly, the opposite holds for normal speech [6]. In order to reduce the number of extraneous factors that *may* have impacted on our results, and in so doing focus on the effects of the *microphones* rather than the speakers, we therefore restricted our recruitment to participants who were native English speakers with a Canadian accent. We recruited an equal number of male and female participants but restricted our age range to 18 – 35 year olds (young adults) to limit speaker variation that comes with age. Finally, due to the colour-centric nature of our hazard avoidance system, we were not able to include colour blind participants. Twenty four people participated in our experiment, 8 per microphone/group.

We hypothesised that, in line with previous research, participants would take longer and achieve less satisfactory data entry (in terms of both subjective participant response and actual data input accuracy) under the noisy condition than the quiet condition. As a consequence of the theoretical reduction in the extent of background noise picked up by the bone conduction microphone, we hypothesised that it would return better accuracy and user satisfaction results than the two condenser microphones. As already noted, female Lombard speech has been shown to be more intelligible than male Lombard speech, so we hypothesised that female participants would return higher accuracy rates under the noisy condition than male participants; we anticipated seeing the reverse for the quiet condition.

4. RESULTS & DISCUSSION

The following sections reflect on analysis of the measures we took to assess the efficacy of the three microphones.

4.1 Accuracy

Our primary accuracy measure was calculated as a ratio of the total number of correct entries divided by the total number of attempts, expressed as a percentage. Data entries were awarded a correctness score of 1 if correct (irrespective of number of attempts) and 0 if incorrect (after three attempts). A multiple factor ANOVA showed that microphone ($F_{2,468}$=67.22, p<0.001), gender ($F_{1,468}$=3.91, p=0.04), and the combination of microphone and gender ($F_{2,468}$=19.29, p<0.001) had a significant affect on accuracy. Tukey HSD tests showed that the accuracy was significantly less for participants using the Invisio (bone conduction) microphone (avg.=36%) than for the QSHI3 (avg.= 76%, p<0.001) and DSP-500 (avg.=78%, p<0.001) microphones; the difference in accuracy for the QSHI3 and DSP-500 microphones was not statistically significant (see Figure 5).

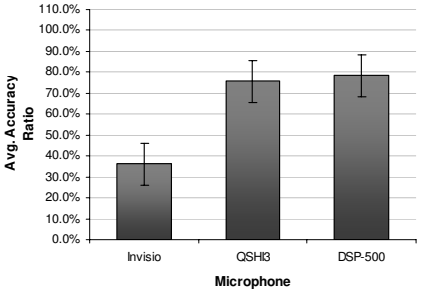

Figure 5. Accuracy according to microphone type.

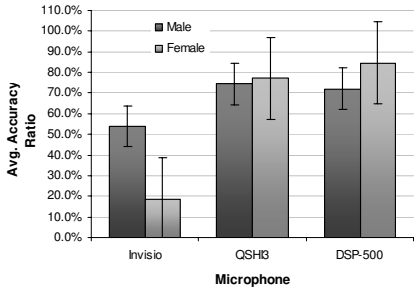

Figure 6. Accuracy according to gender and microphone.

Although not separated by a large margin, the average accuracy achieved by male participants (67%) was significantly higher (p=0.04) than for female participants (60%). Also, on average, females using the Invisio microphone returned significantly lower accuracy results (19%) than not only males using the same microphone (54%, p<0.001) but also all other gender-microphone pairings (see Figure 6). Likewise, on average, males using the Invisio microphone returned significantly lower accuracy results than either gender using either of the other microphones.

We included a second measure of accuracy based on the confidence value attributed to each entry – this measure (which reflects the recogniser's level of confidence that an utterance matches the item it selects from its grammar) was provided directly by the speech recognition engine, and is again represented as a percentage. A multiple factor ANOVA revealed similar results to those for accuracy itself, as one might expect. Specifically, microphones were shown to significantly differ ($F_{2,468}$=72.76, p<0.001) in terms of confidence measures returned by the speech recognition engine: the Invisio microphone resulted in significantly lower average confidence measures (66%) than both the QSHI3 (avg.=95%, p<0.001) and DSP-500 (avg.=92%, p<0.001) microphones; the difference in confidence measures for the latter two microphones was not statistically significant. Gender was shown to have a significant impact on confidence measures ($F_{1,468}$=15.73, p<0.001) – see Figure 7 – with males returning a significantly higher confidence rating (89%) than females (80%). The combination of gender and microphone was also shown to be significant ($F_{2,468}$=37.45, p<0.001) and, in essence, reflected the same pairing differences as the accuracy rates.

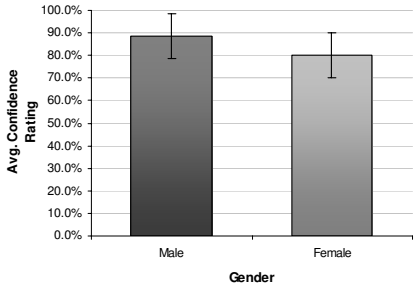

Figure 7. Confidence ratings according to gender.

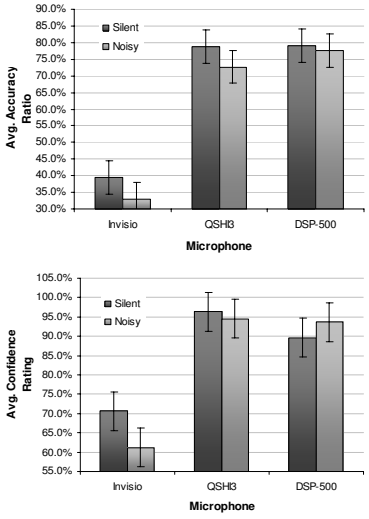

Figure 8. Accuracy and confidence ratings according to microphone and audio condition.

Our analysis highlighted a lack of demonstrable effect of audio condition. Both accuracy rates and confidence ratios appear to have been unaffected by the introduction of background noise (despite that noise being substantial, at 70dB) – see Figure 8 and overall values in Figure 9.

Given our initial hypothesis that accuracy rates would be less under noisy conditions, we were surprised that, and unsure as to why, we did not see a noticeable effect of background noise. As previously noted, female Lombard speech is considered more intelligible than male Lombard speech, but the opposite is true for normal speech [6] and it was on this basis that we had hypothesised that females would return higher accuracy ratings under noisy conditions and males would return higher ratings under quiet conditions.

Figure 9 shows that, across both audio conditions, accuracy and confidence measures reflect higher values for males than females. The higher values for males in the silent audio condition reinforces (albeit without statistical significance) the intelligibility observation for normal speech; the fact that the opposite is not demonstrated for noisy speech would suggest that Lombard speech might not have been a major factor in this study (although further tests/analysis would be required to confirm this).

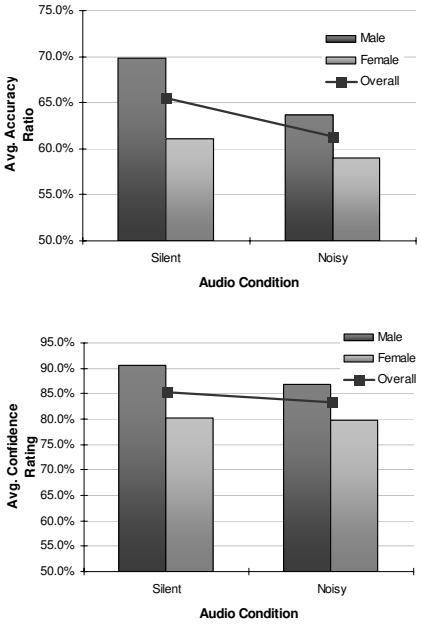

Figure 9. Accuracy and confidence ratings according to gender and audio condition.

An alternative conclusion might be that all three microphones were sufficiently able to mitigate against background noise, albeit some were clearly better than others. On the basis of participant accuracy and speech recogniser confidence, there is little to differentiate the QSHI3 and DSP-500 microphones, both of which are condenser microphones. Conversely, the bone conduction (Invisio) microphone performed consistently worse than the others. Furthermore, on the basis of this study, it would appear to be more susceptible to gender differences across speakers than both condenser microphones, which do not appear to be substantially affected by gender. These findings contradict our initial hypothesis that, as a result of the theoretical reduction in pick-up of background noise, the Invisio (bone conduction) microphone would return significantly better data entry accuracy than the other two microphones. This raises the question, therefore, as to whether or not the poor accuracy is a result of the mobility itself – that skeletal vibration brought about by physical motion was, in some way, distorting the signal for this microphone. Alternatively, it may be the case that the Invisio microphone has been optimised for the pitch of male speakers and is fundamentally less effective for female speakers. Further investigation will be required to determine the precise cause of the poor accuracy returned by this microphone.

4.1.1 Data Entry Type

Although it was not the primary focus of our study, we did consider the effect of data entry type in terms of accuracy and confidence ratings. Unfortunately, on the basis of our high level data, we were unable to draw any meaningful conclusions in this regard. To determine a clearer picture of the effect of data entry type, we propose to conduct further analysis of the data in order to look at the specific nature of errors but, for the time being, we assign this to future work.

4.2 Walking Speed

Prior to participants beginning their experimental sessions we timed them each walking, at a pace that was comfortable, 10 laps of the grid of mats while avoiding hazards (as described previously) and carrying, but not using, the mobile technology. We used this to calculate a baseline average preferred walking speed (PWS).

For each participant, for each experimental condition (silent v. noisy), we recorded the time it took them to complete their 10 data entry tasks as well as the number of laps they walked. We used this to calculate the percentage of their preferred walking speed (PPWS) at which participants walked under each audio condition. A multiple factor ANOVA analysis of these results showed that only the combination of microphone and gender had a significant effect on PPWS ($F_{2,36}$=5.37, p=0.009); microphone by itself did not seem to significantly affect the PPWS nor did audio condition.

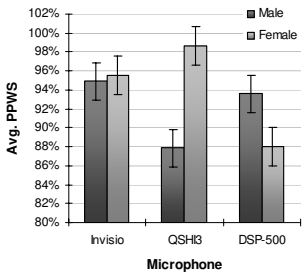

Figure 10. PPWS according to microphone and gender.

Figure 10 shows the average PPWS achieved by participants according to microphone and gender. Tukey HSD tests showed that females using the QSHI3 microphone (avg. PPWS=99%) were able to maintain a significantly higher walking speed than males using the same microphone (avg.=88%, p=0.04) and than females using the DSP-500 microphone (avg.=88%, p=0.04). No other comparisons were statistically significant. As seen from the accuracy rates shown in Figure 6, the ability of females using the QSHI3 to maintain a walking speed so close to their preferred walking speed would not appear to be at the expense of accuracy; using this microphone, both genders returned comparable accuracy rates but it would appear that males had to significantly slow down to achieve an accuracy

rate comparable to females. Females achieved the highest accuracy rates overall using the DSP-500 microphone, but this would appear to have been at the expense of walking speed as on this microphone, they returned the lowest PPWS. Use of the Invisio microphone would seem to equally affect males and females in terms of their walking speed.

When we ran a multiple factor ANOVA over the average number of laps taken by participants to complete their data entry tasks, we discovered that only microphone had a significant affect ($F_{2,36}$=8.68, p=0.001). Tukey HSD tests revealed that participants using the Invisio microphone walked significantly more laps (avg.=9.1) than participants using the QSHI3 microphone (avg.=7.1, p=0.018) and participants using the DSP-500 microphone (avg.=6.3, p<0.001). The difference in average number of laps walked by participants using the condenser microphones was not significant. We also analysed, using a multiple factor ANOVA, the total time taken by participants to complete their data entry tasks. Only microphone was shown to have a significant impact on total task duration ($F_{2,36}$=6.65, p=0.003). Participants using the Invisio microphone (avg.=122.6secs) took significantly longer to complete their tasks than participants using the QSHI3 (avg.=97.6secs, p=0.044) and participants using the DSP-500 microphone (avg.=87.1secs, p=0.003). Once again, there was no significant difference between the two condenser microphones. In accord with the results shown in Figure 10, when combined, these results suggest that participants using the Invisio microphone walked more laps and took longer but, on average, walked at a pace that was not significantly different to that recorded for participants using the other two microphones. This implies, therefore, that completing the tasks themselves was a more onerous undertaking for participants using the Invisio microphone – i.e., it simply took longer.

Once again, we see that audio condition did not seem to significantly affect our results. As noted previously, we had hypothesised that participants would take longer to complete their data entry tasks under noisy conditions. Since accuracy itself did not appear to have been directly influenced by audio condition, it is perhaps unsurprising that participants' task speed was also unaffected.

4.3 Task Load Ratings

The NASA Task Load Index (TLX) assesses subjective workload according to six independent dimensions, namely: mental demand; physical demand; temporal demand; effort; frustration; and performance. These can be considered individually, or in combination as a measure of overall workload. Each dimension is rated on a 20 point scale: in each case – barring performance – the lower the rating, the better; the inverse holds for performance, since a higher rating reflects increased participant satisfaction in their own performance.

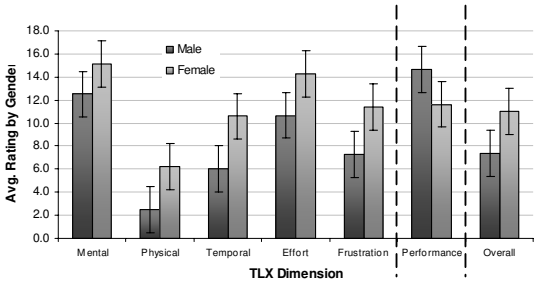

Figure 11. TLX ratings according to gender.

A multiple factor ANOVA revealed that ratings for overall workload were significantly affected by participant gender ($F_{1,36}$=27.83, p<0.001). On average, as can be seen from Figure 11, males rated overall workload significantly lower than females (7.4 v. 11.0 respectively). Neither audio condition nor microphone was shown to exert significant influence on the overall measure of workload.

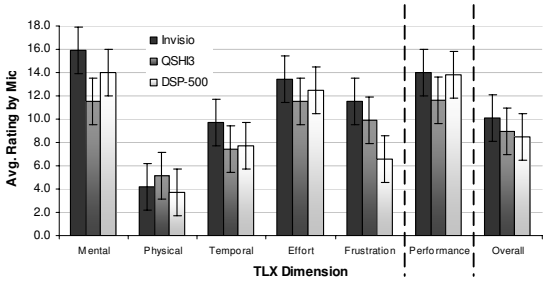

Figure 12. TLX ratings according to microphone.

When we looked at the individual factors contributing to overall workload, we saw some interesting results. For mental demand, a multiple factor ANOVA showed that both gender ($F_{1,36}$=8.46, p=0.006) and microphone ($F_{2,36}$=8.10, p=0.001) significantly effected participant ratings. On average, males found the mental demand (12.5) significantly less than females (15.1) – see Figure 11. Participants, irrespective of gender, generally found the Invisio microphone significantly more mentally demanding (15.9) than the QSHI3 microphone (11.5) – see Figure 12; there were no other statistically significant differences in terms of mental demand. These findings seem to reflect the accuracy differences identified between males and females as well as the lower accuracy ratings returned for the Invisio microphone.

A similar analysis of physical demand revealed that only gender had a significant impact on average physical demands ($F_{1,36}$=11.18, p=0.002), with females (6.3) finding the tasks significantly more physically demanding than males (2.5). Participants, as previously discussed, were required to carry and interact with a tablet PC whilst walking and, as noted in section 4.2, females – at least for one microphone – maintained a walking speed more consistent with their normal speed (i.e., faster); it is, therefore, perhaps unsurprising, that females rated the physical demands of the task higher than males.

Analysis of average temporal demand ratings is the one and only point at which we saw a significant impact of audio condition. A multiple factor ANOVA showed that both audio condition ($F_{1,36}$=4.62, p=0.038) and gender ($F_{1,36}$=14.53, p=0.001) had a significant affect on temporal demand.

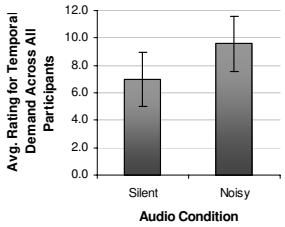

Figure 13. Temporal demand according to audio condition.

Figure 13 shows the average temporal ratings returned according to level of background noise. Under noisy conditions, participants clearly felt that they were under significantly more time pressure (9.6) to complete their tasks

than when performing the same data entry activities in the silent environment (7.0). Females also felt significantly more time pressure (10.6) than males (6.0) – see Figure 11. Research shows that humans exhibit a stress reaction to noise, and suggests that women may be more sensitive to noise stress than men [21]. At the level of conjecture, across all participants, it may be that the background noise induced a stress reaction that materialised as a feeling of time pressure. This reaction may have been heightened in female participants, and when added to the higher workload females reported experiencing across the board, it may account for the fact that the largest difference between the average ratings by gender was seen for temporal demand.

Gender was also found to significantly affect average ratings for effort ($F_{1,36}$=8.20, p=0.007) and frustration ($F_{1,36}$=8.67, p=0.006); once again (see Figure 11) female ratings were significantly higher than male ratings for both dimensions. Frustration levels were also found to be significantly affected by microphone ($F_{2,36}$=4.24, p=0.022) – see Figure 12. Participants using the Invisio microphone reported significantly higher levels of frustration (11.5) than participants using the DSP-500 microphone (avg.=6.6, p=0.019); frustration levels for the QSHI3 microphone (9.9) did not differ significantly from either of the other microphones.

Finally, consider participants' ratings of their own performance. In line with other ratings, participants' performance ratings were significantly affected by gender ($F_{1,36}$=6.39, p=0.016) with females rating their performance significantly lower (11.6) than males (14.7). Males' confidence in their own performance would seem justified given their significantly higher accuracy rates and lower workload experience – i.e., they generally found the tasks to be less demanding and were able to achieve more accurate results. In contrast, females achieved a lower accuracy rate – a fact of which they would seem to have been aware, given their lower self-assessment of performance – and they generally found the task to be harder than males.

In summary, therefore, we had hypothesised that participants would return significantly higher workload ratings when completing their data entry tasks under noisy conditions. In reality, it would appear that, with the exception of temporal demand, audio condition did not affect the tasks per se, and so did not, in turn, affect participants' perception of workload. Having hypothesised that the Invisio microphone would perform significantly better than the other two microphones, we had anticipated that participants would experience considerably less workload using this microphone. Unsurprisingly, given its poor performance in terms of accuracy, the Invisio microphone was actually found to increase workload (specifically, mental demand and frustration).

5. CONCLUSIONS & FUTURE WORK

Contrary to expectation, we found audio condition to have little impact on our results. Surprised by this, we measured the average audio level in our lab (corresponding to the level which we called "silent", respecting the fact that no natural environment is truly silent). Our lab is situated in the basement of our building and, to the casual observer, seems a very quiet environ. We were surprised, therefore, to find that the lab registered an average of 60dB, based largely on the background audio emitted from the (new) air conditioning. Referring back to Oviatt's study [11], the maximum audio level to which participants were exposed in her study was 60dB (although this seems questionable given that it was supposedly in a moderately busy public cafeteria). Oviatt compared this to a quiet condition at a reported 42dB and found increased background noise to have a significant impact on speech-based data entry. The question then arises as to whether – albeit our two audio conditions were *realistic* – a 10dB difference in audio level (or, more specifically, the precise difference between 60dB and 70dB) is perhaps not sufficient to identify statistically significant differences based on audio level. Alternatively, it may be the precise types of noise (i.e., air conditioning v. recorded city traffic) that have failed to support appropriate comparison. We would, therefore, propose to investigate this further.

As previously mentioned, we had expected females to return higher accuracy under noisy conditions than males, with the inverse expected for silent conditions. On the basis that audio condition was not found to impact our results at all, we did not observe this in our data. We previously concluded that perhaps Lombard speech was not a major factor in our study. If it does turn out that a 10dB difference specifically between 60dB and 70dB and/or for our particular audio types is insufficient to observe the impact of background noise, then it may be that either Lombard speech was in effect across both or neither of our audio conditions, irrespective of gender. Alternatively, akin to the Hawthorne Effect, it may have been that the introduction of the technology itself (i.e., the requirement to speak into a microphone per se) caused all participants to alter their natural speech, irrespective of background noise. Furthermore, it may be that, for experimental studies of speech recognition, it is impossible to avoid Lombard speech since users will naturally alter their speech until such time as they are comfortable/confident that the speech recognition technology will correctly interpret their input every time. Once again, we would propose to investigate this further.

Finally, we intend to examine the affect of the Invisio microphone in greater detail. Theoretically, it should have worked far better than the other two microphones and so we are intrigued to determine why we did not find this to be the case.

Although, as a result of our study, we have been forced to reject all of our hypotheses (all of which were based on theory and/or preceding work of others) we feel that we have contributed substantially to the corpus of knowledge in this field. In essence, as they currently stand, there is little to differentiate the two condenser microphones, and they have been empirically shown to out perform the newest (bone conduction) microphone technology. Interestingly, of the three microphones, the QSHI3 was by far the least expensive making it a realistic or viable yet potentially effective option for facilitating mobile speech-based data entry.

6. ACKNOWLEDGMENTS

We would like to thank our participants for their enthusiastic participation in this study.

7. REFERENCES

[1] Chang, J. *Speech Recognition System Robustness to Microphone Variations*. M.Sc. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1995.

[2] Crease, M. and Lumsden, J., A Technique for Incorporating Dynamic Paths in Lab-Based Mobile Evaluations. *In Proceedings of Human Computer Interaction'2007 (HCI'2007)*, (Lancaster, UK, 2007), 2007.

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

[4] Huerta, J. *Speech Recognition in Mobile Environments*. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2000.

[5] Hurtig, T., A Mobile Multimodal Dialogue System for Public Transportation Navigation Evaluated. In *Proceedings of 8th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'06)*, (Helsinki, Finland, 2006), ACM Press, 2006, 251 - 254.

[6] Junqua, J. The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers. *Journal of the Acoustical Society of America, 93* (1). 510 - 524.

[7] Karat, C.-M., Halverson, C., Karat, J. and Horn, D., Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems. In *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'99)*, (Pittsburgh, Pennsylvania, USA, 1999), ACM Press, 1999, 568 - 575.

[8] Kjeldskov, J., Skov, M.B., Als, B.S. and Høegh, R.T., Is It Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field. In *Proceedings of 6th International Symposium on Mobile Human-Computer Interaction (MobileHCI'04)*, (Glasgow, Scotland, 2004), 2004, 61 - 73.

[9] Lumsden, J., Kondratova, I. and Langton, N., Bringing A Construction Site Into The Lab: A Context-Relevant Lab-Based Evaluation Of A Multimodal Mobile Application. In *Proceedings of 1st International Workshop on Multimodal and Pervasive Services (MAPS'2006)*, (Lyon, France, 2006), IEEE, 2006, 62 - 68.

[10] McCormick, J. Speech Recognition. *Government Computer News, 22* (22). 24 - 28.

[11] Miyara, F. *Sound Levels*. Date Accessed: 18 March, 2006, http://www.eie.fceia.unr.edu.ar/~acustica/comite/soundlev.htm.

[12] NextLink. *Invisio Pro*. Date Accessed: 19 March, 2006, http://www.nextlink.se/.

[13] Oviatt, S., Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. In *Proceedings of Conference on Human Factors in Computing Systems (CHI'99)*, (Pittsburgh, USA, 1999), ACM Press, 1999, 576 - 583.

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

[15] Oviatt, S., MacEachern, M. and Levow, G. Predicting Hyperarticulate Speech During Human-Computer Error Resolution. *Speech Communication, 24* (2). 87 - 110.

[16] Pakucs, B., Butler: A Universal Speech Interface for Mobile Environments. In *Proceedings of 6th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'04)*, (Glasgow, Scotland, 2004), Springer-Verlag, 2004, 399 - 403.

[17] Picardi, A.C. IDC Viewpoint: Five Segments Will Lead Software Out of the Complexity Crisis, IDC, 2002.

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

[19] Plantronics. *DSP-500 Headset*. Date Accessed: 19 March, 2006, http://www.plantronics.com/north_america/en_US/products/cat640035/cat1430032/prod440044.

[20] Price, K., Lin, M., Feng, J., Goldman, R., Sears, A. and Jacko, J., Data Entry on the Move: An Examination of Nomadic Speech-Based Text Entry. In *Proceedings of 8th ERCIM Workshop "User Interfaces For All" (UI4All'04)*, (Vienna, Austria, 2004), Springer-Verlag LNCS, 2004, 460-471.

[21] Rhudy, J. and Meagher, M. Noise Stress and Human Pain Thresholds: Divergent Effects in Men and Women. *Journal of Pain, 2* (1). 57 - 64.

[22] Rollins, A., Speech Recognition and Manner of Speaking in Noise and in Quiet. In *Proceedings of Conference on Human Factors in Computing Systems (CHI'85)*, (San Francisco, USA, 1985), ACM Press, 1985, 197 - 199.

[23] Sammon, M., Brotman, L., Peebles, E. and Seligmann, D., MACCS: Enabling Communications for Mobile Workers within Healthcase Environments. In *Proceedings of 8th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'06)*, (Helsinki, Finland, 2006), ACM Press, 2006, 41 - 44.

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

[25] Sebastian, D. *Development of a Field-Deployable Voice-Controlled Ultrasound Scanner System*. M.Sc. Thesis, Worcester Polytechnic Institute, Worcester, MA, USA, 2004.

[26] Shure. *QuietSpot QSHI3*. Date Accessed: 19 March, 2006, http://www.sfm.ca/quietspot/qshi3.html.

[27] Vinciguerra, B. *A Comparison of Commercial Speech Recognition Components for Use with the Project54 System*. M.Sc. Thesis, University of New Hampshire, Durham, NH, USA, 2002.

[28] Ward, K. and Novick, D., Hands-Free Documentation. In *Proceedings of 21st Annual International Conference on Documentation (SIGDoc'03)*, (San Francisco, USA, 2003), ACM Press, 2003, 147 - 154.