Hit The Road, Jack

P4: Evaluation

Amit Garg
Rachel LeRoy
Sahib Singh
John Crisp
Bryan Bennett

Table of Contents

Evaluation Design and Rationale

Our user testing sessions involved testing different sensory modes of receiving information within our communication paradigm. We tested the efficiency differences between 3 different sensory modes of receiving messages: visual, auditory, and tactile. Each testing participant participated in 3 separate trials, randomized in order to reduce learning effect:

1. visual, receiving hand and arm signals via line of sight (LOS).
2. auditory, receiving messages via bone/headphones.
3. tactile, receiving messages via a tactile headband installed in a baseball hat.

Our defined independent variable is sensory mode for receiving information, while our dependent variables are total task time and perceived usability.

Summarizing our evaluation process:

- We established an obstacle course with a set of 15 soccer cones, in grid format. The evaluation was conducted outdoors in low light conditions, on Tech Green, on the evenings of Thursday, November 19 and Monday, November 23.
- Each participant participated in 3 separate trials, visual, auditory, and tactile. These were randomized in order to reduce learning effect.
- Before the visual trial, the participant was trained in the hand signals used, followed by a brief test to confirm that the participant had learned the signals correctly.
- Before the auditory trial, the participant was fitted for the bonephones, and the sounds were played to ensure the participant could hear them. This was followed by a brief test to confirm that the participant had learned the auditory signals correctly.
- Before the tactile trial, the participant was fitted for the tactile hat, the vibration motors were tested on the participant to ensure the participant could feel them (and comfortably), and the participant was trained in the meaning of each vibration motor. This was followed by a brief test to confirm that the participant had learned the tactile signals correctly.
- For each trial, the participant wore a pair of ski goggles that had a spray tint applied. This simulated a much darker environment, without use of a blindfold, yet allowed the evaluation team to observe the trials effectively, and ensure the safety of the participants.
- A member from our team acted as leader. The leader walked through the grid systematically through the shape of a digital 8, transmitting directional signals to the participant. The shape of the course, and the path through the course, was designed to reduce the possibility that the participants could “guess” or learn the pattern, which could make the trial times not a fair predictor of actual performance.
  For visual trials, the leader used the hand signals as trained.
  For auditory trials, the leader used a smartphone to operate a web app that transmitted auditory signals via bluetooth to the bonephones worn by the participant.
  For tactile trials, the leader used a smartphone to operate a native app that transmitted signals via bluetooth to the tactile hat, which engaged the vibration motors in the headband of the hat worn by the participant.

- We collected the time to complete the entire obstacle course for each trial.
- We administered post-testing surveys to assess the usability, using the System Usability Scale (SUS)

Participants (n=6, 4 male, 2 female) had ages ranging from 18-26. Three were Georgia Tech undergraduates, one was a recent graduate from the University of Michigan School of Information MSI, and two were current MS HCI graduate students.

We chose total task time because it is a general indicator of efficiency among each of our prototypes. We used the SUS because it is established as a "quick and dirty" method of evaluating a system's usability. Although the SUS is generally used for digital interfaces, the 10 questions of the SUS were applicable to the wearable prototypes we created as well as the Hand and Arm signals. Due to complexities in conducting the trials, we were not able to record response time (RT) as we had hoped and planned for in our P3 report. We did not have the technology in place to do so, and it was not feasible to manually capture that information without verbal communications (between leader and observers) that might have distracted the participants and impacted our ability to gather accurate total task time. In a future user study, we would include data loggers that automatically register timestamps between when information is sent as well as when an action was completed by the participant. Having response time would give a deeper insight into the perceptual and cognitive load for each testing condition.

Results Discussion (Time to complete course)

In order to quantitatively evaluate our prototypes, we collected the time it took for each participant to complete the obstacle course. Our hypotheses were as follows:

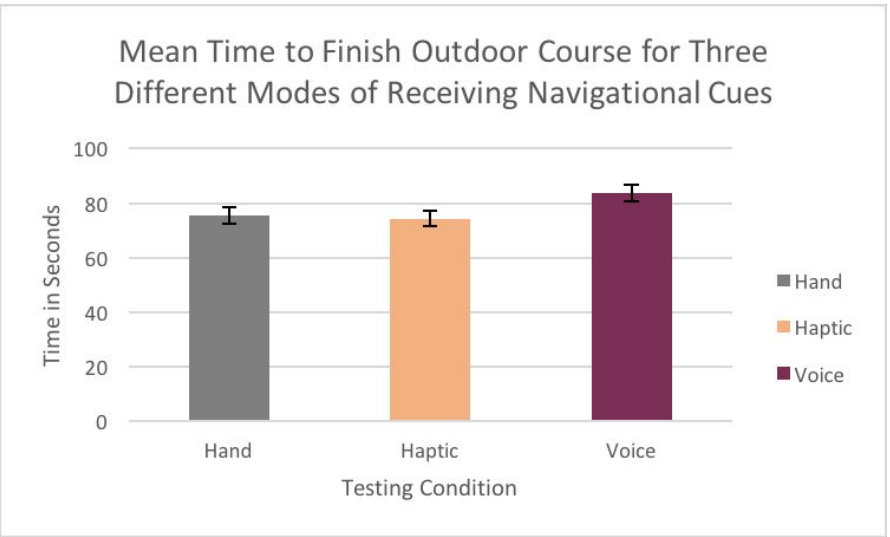
Null Hypothesis: All means are equal ($\mu1 = \mu2 = \mu3$)
Alternative Hypothesis: Not all means are equal

The following table displays the raw time data collected per user and per testing condition. It is important to note that subjects are separated by columns, and testing conditions are separated by rows. This will be important later when discussing the results.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
Line of Sight	75.62	74.94	65.8	78.56	82.75	75.96
Haptic Hat	82.2	77.54	80.86	63.49	80.38	61.48
Bonephones	79.85	97.61	83.7	79.05	86.45	75.96
			All Data in Time (sec)			

Because our study was a within subjects design with one independent variable (sensory mode of receiving information), we performed a repeated measures, one-way ANOVA

at the 5% level of significance to determine if there were any significant differences between our three testing conditions. Below are our results.



Means: 75.6 s (Hand), 74.3 s (Haptic), 83.8 s (Voice)

Repeated Measures, One-Way ANOVA Table:

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	315.0273	2	157.51365	3.05781665	0.09199477	4.10282102
Columns	372.857667	5	74.5715333	1.44765915	0.28873708	3.32583453
Error	515.118033	10	51.5118033			
Total	1203.003	17				

As mentioned earlier, testing conditions in the raw data table were separated by rows, while subjects were separated by columns. Therefore, in the ANOVA table, we are only interested in the variation between rows, not between columns. This is why we have highlighted the p-value corresponding to the source of variation between rows.

Based on our analysis, there was no statistical significance between any of the three testing conditions for time to complete the course ($F = 3.058$, $p>0.05$). Thus, we fail to reject the null hypothesis that all means are equal. There are many factors about our research study that could have caused the observed lack of statistical significance. First, because our sample size was low, the statistical power of our study was less than optimal. Consequently, our chance of Type II error is high. Second, we ran our user testing study on two separate days (3 participants on one day, 3 on another). There may have been unintentional variations in our study (e.g. environmental conditions were different, some different research personnel present, etc.) that caused variations in the data. We realized that this factor may indeed be part of the explanation when we looked at the raw data. Looking closely at the times, one can see that in the first three

columns, subjects consistently completed the course **fastest** in the **Hand and Arm signals** testing condition. However, in the second three subjects, subjects consistently completed the course fastest with the **Haptic Feedback prototype**. Concurrently, the first three subjects were tested on a night that happened to be a full moon and a clear night, and the second three on a different, cloudy night. The full moon and clear night may have allowed users to see the Hand and Arm signals more clearly than on the session during the cloudy night, which in turn could have seriously affected the accuracy of our data.

All confounding variables aside, our data show that the Haptic Feedback prototype, the Bonephones prototype, and the traditional Hand and Arm signals produced technically equal communication efficiency. However, further data analysis using the System Usability Scale gives greater insight.

Results Description (SUS)

System Usability Scale (SUS) Testing

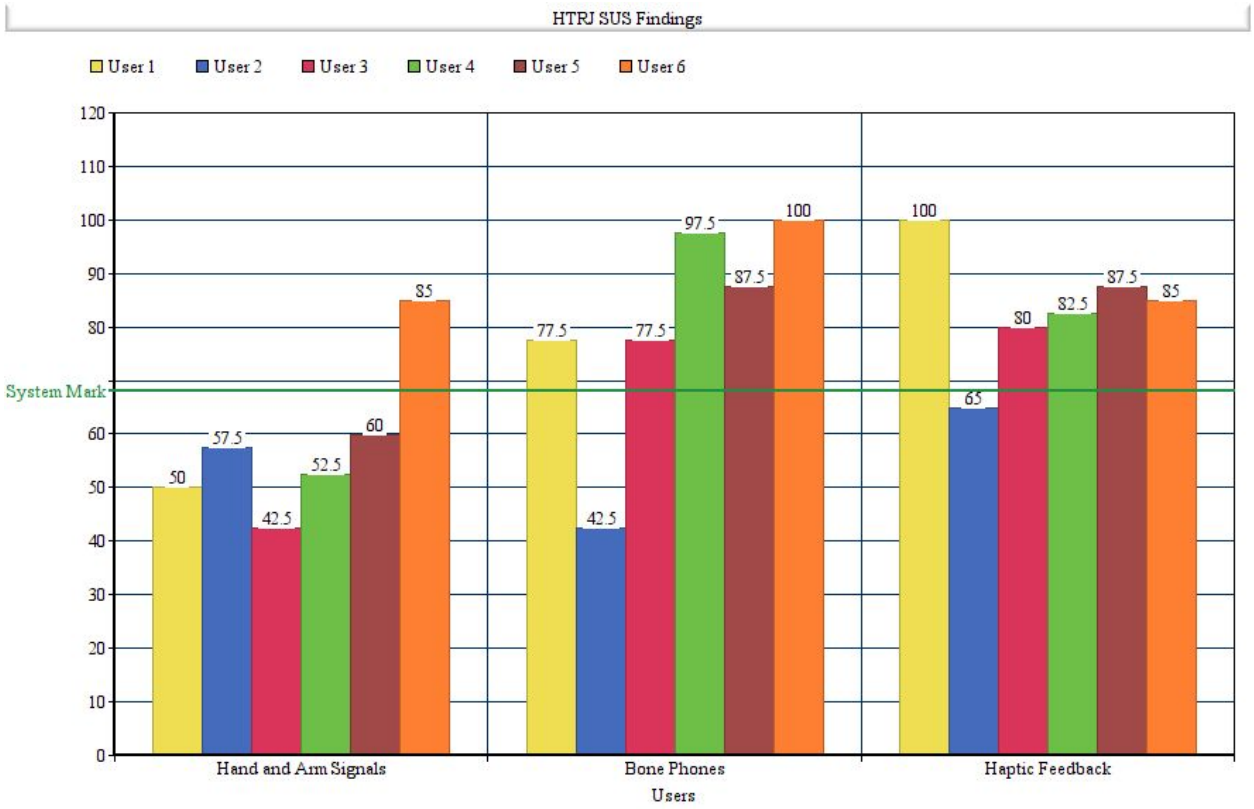
To understand users' preference of operating the three systems (Hand & Arm signals, Haptic Feedback, and Bonephones) under low visibility, we asked them to complete the SUS survey for each of the three systems at the conclusion of each participant's user testing session.

We had a total of 18 responses, which we categorized three different ways:

- grouped by each system
- grouped by each user
- grouped by average SUS scores across testing condition

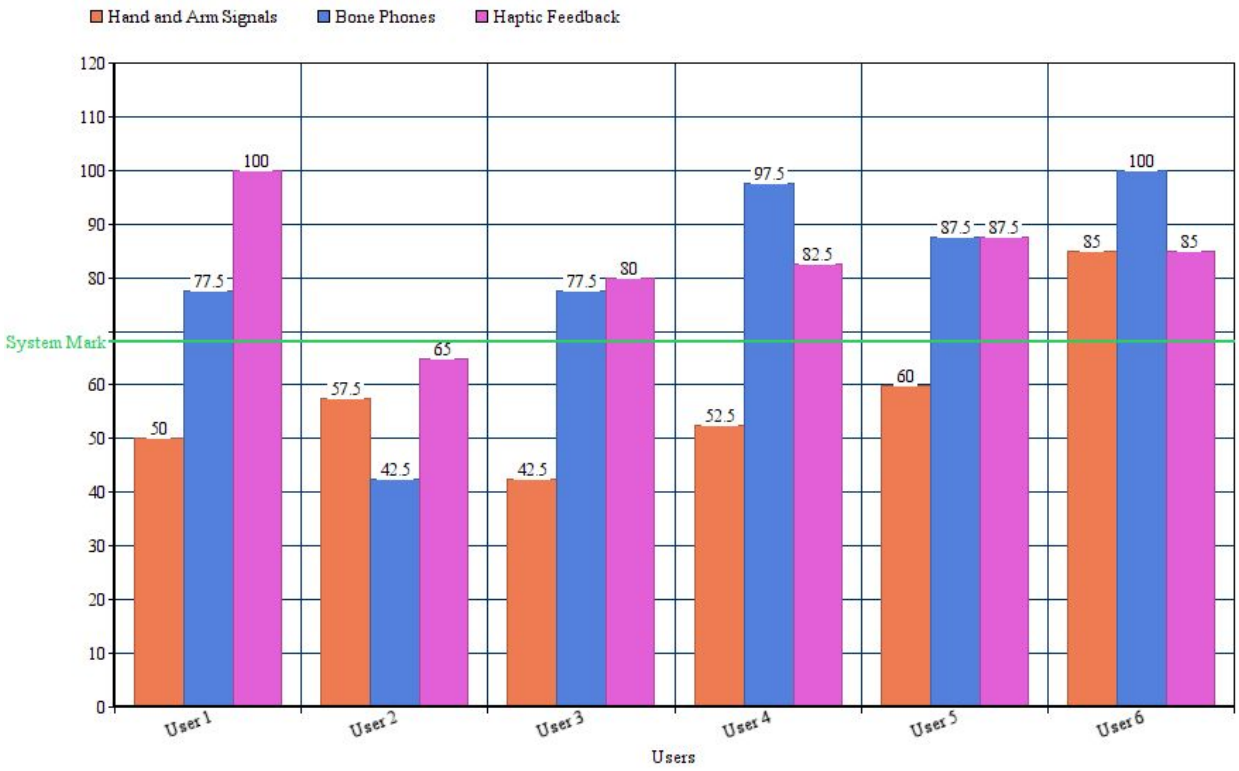
Each graph shows a green line horizontally across the graph known as the Good System Mark. This is represented by a score of 68 on the SUS, and is a threshold past which a system may be considered "usable".

Grouped by System



The most important finding we got from grouping the data by system was a clear indication of the fact that no system was an exclusive winner. In all 3 systems, there was exactly one user whose response was opposite that of the general trend about the group. For example, in the Haptic Feedback group, there is one user (in blue), whose response indicates that they found the system to be below the Good System Mark, whereas all other responses or scores were above that of the Good System Mark.

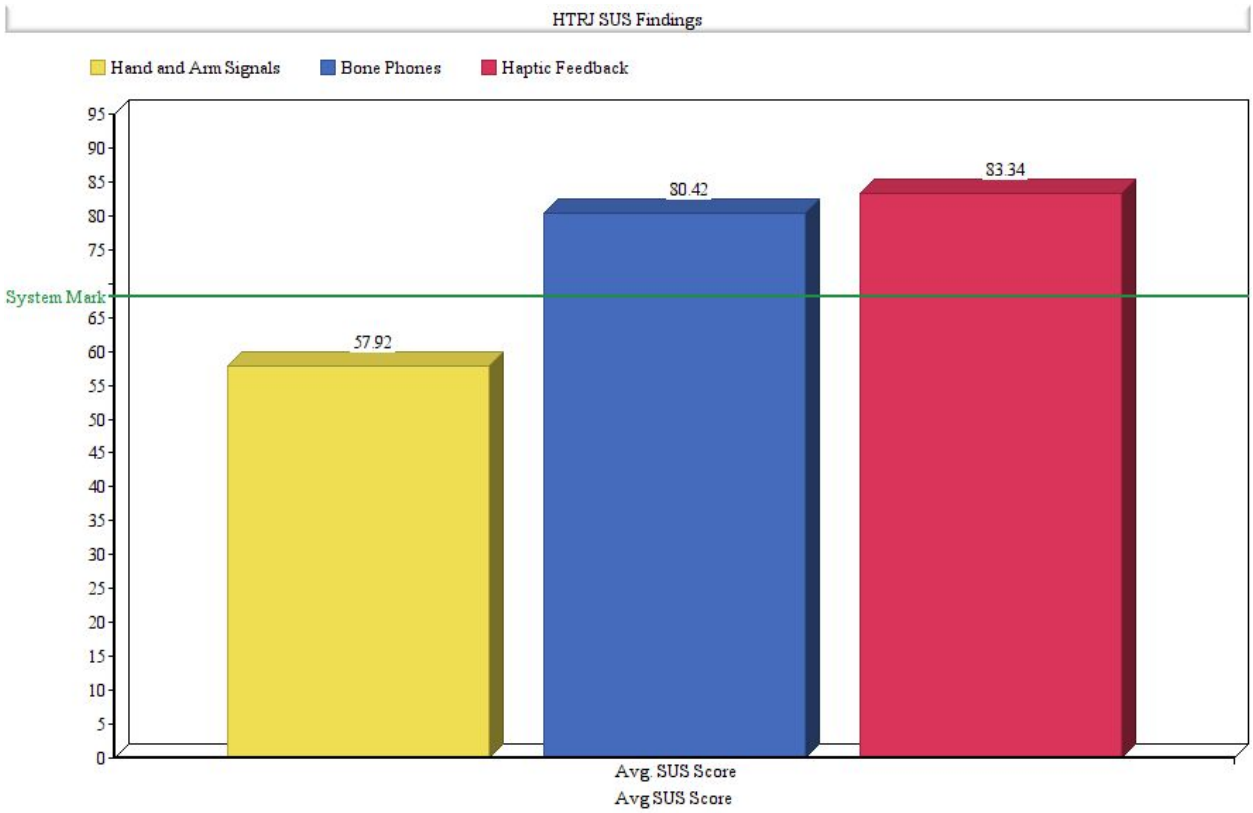
Grouped by User



On grouping the data based on each individual user, the most interesting finding was that there was a case where all of the user's responses for each of the system was below the Good System Mark and there was another case where all the of the user's responses for each of system were above the Good System Mark. That meant that there was one user who did not find all three of the systems to be comfortable and usable, while there was one user who found all three systems to be comfortable and usable. This gave us an indication on the variety of input we had and reiterated the fact that there was no exclusive winner. However, there was a clear loser. The Hand and Arm signals were consistently rated the lowest across all users, indicating that all users felt the Hand and Arm signals were the least usable out of the three testing conditions.

Average SUS Score of each individual system:

The average SUS Score for each of the systems gave an overall view of the perceived usability for each condition. The average score gave the final view after considering all the different factors and variations which occurred in the data patterns.



The overall average score of each of the three systems showed us which system qualified as a usable system with a high comfort factor and which did not. From the given data, we found that the **Hand and Arm Signal** system scored a meager **57.92** on the SUS, where a good system is denoted by a score of 68 or above. Hence, the overall consensus of our users was that the Hand & Arm Signal system is not the best system to use for communicating navigational information in low visibility. The other two systems, namely the **Bonephones** and **Haptic Feedback**, had a mean SUS score of **80.42** and **83.34** respectively, which both lie in the range of a usable system, outscoring the Hand and Arm signal by a **margin of approximately 22-25 points**. This meant that our users prefered both of our prototypes over Hand and Arm signals, with a slightly higher preference given to the Haptic Feedback prototype.

Finally, because the SUS survey is a collection of Likert items that outputs a combined score, the mean can be used as a measure of central tendency (as opposed to the median) and thus analysis using inferential statistics is possible. Our hypotheses were the same as before:

Null Hypothesis: All means are equal ($\mu 1 = \mu 2 = \mu 3$)
Alternative Hypothesis: Not all means are equal

To begin, we again performed a repeated measures, one-way ANOVA for the computed SUS score across all three testing conditions. Below we show the raw data as well as the resulting ANOVA table:

	Bonephones	Hand and Arm	Haptic
Subject 1	77.5	50	100
Subject 2	42.5	57.5	65
Subject 3	77.5	42.5	80
Subject 4	97.5	52.5	82.5
Subject 5	87.5	60	87.5
Subject 6	100	85	85
Mean	80.41666667	57.91666667	83.33333333
		Computed SUS Scores	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	2115.277778	5	423.0555556	2.378758298	0.1141202715	3.325834529
Columns	**2321.527778**	**2**	**1160.763889**	**6.526747364**	**0.01535734527**	**4.102821015**
Error	1778.472222	10	177.8472222			
Total	6215.277778	17				

This time, we were interested in the variation between columns, due to the fact that the raw data is organized where testing condition is separated by column. The ANOVA tells us that there is a significant difference ($F = 6.53$, $p<0.05$), therefore we reject our null hypothesis and accept the alternative hypothesis. Because there were only three testing conditions, post hoc analysis consisted simply of paired sample t-tests. The first t-test was between the Haptic Feedback and the Bonephone group. Below is the result:

Null Hypothesis: μ haptic = μ bonephone
Alternative Hypothesis: μ haptic > μ bonephone

t-Test: Paired Two Sample for Means (haptic v bonephone)

	Variable 1	*Variable 2*
Mean	80.41666667	83.33333333
Variance	436.0416667	129.1666667
Observations	6	6
Pearson Correla	0.5934255446	
Hypothesized Me	0	
df	5	
t Stat	-0.4242813867	
P(T<=t) one-tail	**0.3444983653**	
t Critical one-tail	2.015048342	
P(T<=t) two-tail	0.6889967306	
t Critical two-tail	2.570581835	

There is no significant difference between the SUS scores for Haptic Feedback and Bonephone testing conditions (*t=0.42, p>0.05,* fail to reject hypothesis). Participants did not show significant preference between the two prototypes.

The second t-test was between the Bonephone group and the Hand and Arm signals group. The reasoning behind testing these two groups is as follows: if the t-test is non-significant, then the significant difference (observed from the ANOVA table) must lie between the Haptic Feedback group and the Hand and Arm signals group. However, if the t-test is significant, then we can conclude that the Bonephone group and the Haptic Feedback group, although they do not differ themselves, both have significantly higher SUS scores than the Hand and Arm signals group. This is because the mean SUS score for the Bonephone group is lower than that of the Haptic Feedback group; i.e. if the Bonephone group has significantly higher SUS scores than Hand and Arm signals group, then the Haptic group must also have significantly higher SUS scores than the Hand and Arm signals group as well.

Below is the result of the t-test. The observed significance was below the 5% significance level ($p<0.05$).

Null Hypothesis: μ hand = μ bonephone
Alternative Hypothesis: μ hand < μ bonephone

t-Test: Paired Two Sample for Means (bonephones v hand)

	Variable 1	*Variable 2*
Mean	57.91666667	80.41666667
Variance	213.5416667	436.0416667
Observations	6	6
Pearson Correla	0.3516095661	
Hypothesized Me	0	
df	5	
t Stat	-2.642490988	
P(T<=t) one-tail	**0.02291928364**	
t Critical one-tail	2.015048342	
P(T<=t) two-tail	0.04583856728	
t Critical two-tail	2.570581835	

We reject the null hypothesis and accept the alternative hypothesis (*t=2.64, p<0.02)*, and thus conclude that the Bonephone prototype and the Haptic Feedback prototype were both perceived to be significantly more usable for receiving navigational information than the Hand and Arm signals.

Future Improvements

Implications of Results With Respect to Design

The purpose of the first prototype was to evaluate among three different methods of receiving communications (visual, auditory, and tactile). From constructing and evaluating our first prototype, we learned the following, which will impact our fully envisioned system:

- The average SUS scores were higher, and the average course completion times were lower, for verbal and tactile (as compared to visual). The implication of this finding is that our initial assumption was correct: the fully envisioned system should include either verbal or tactile receiving methods, rather than depending on visual line-of-sight receiving.

- Correctly fitting the tactile hat to the evaluator was a critical part of initiating an evaluation. Too loose a fit will lower the wearer's ability to detect the tactile vibrations, which could increase course completion time (and possibly reduce user satisfaction as measured by SUS). Too tight a fit may increase the "jarring" effect of the tactile vibrations, which possibly could increase course completion time and reduce user satisfaction. The time necessary to fit the hat correctly, and test if the evaluator could notice the vibrations, was acceptable when performing an evaluation of a prototype. The implication of this finding is that, if tactile is selected for reception in the fully envisioned system, fitting correctly should be considered a key usability factor. While initial fit time may not be critical in the fully envisioned system, the ability to put it on quickly with a perfect fit every time (after the initial fit) would also be a key usability factor. Both of these key usability factors should be considered when developing the evaluation of the fully envisioned system.

  While the initial prototype evaluation did not identify a similar concern with auditory, fitting correctly should also be a key usability factor for auditory (even though it might be easier to achieve).

- The first tactile prototype seemed a bit fragile, but was mostly stable during evaluation, with a few minor issues that could be resolved quickly. However, going back to the initial requirements defined with the ROTC program, "ruggedness" was identified as a critical requirement for the fully envisioned system. The implication of this finding is that our early information on the importance of ruggedness, together with the experience during evaluation of the first prototype, confirms that ruggedness should be a key usability factor in the fully envisioned system, especially if tactile is included.

Description of How Prototype Design Could Be Improved

While there are several improvements that would be appropriate to the first prototype, it would be more appropriate to define the objectives of the next prototype, and include improvements to the portions of the first prototype that would be carried over as part of the next prototype.

The primary objective for the next prototype would be to evaluate technologies to send messages. Two technologies/methods to prototype as sending methods were identified in our initial studies, a motion-sensing glove (interpreting hand and arm signals) and a pressure sleeve (tapping to send a signal).

- Related to sending, improvements should be made to the first prototype by expanding the possible messages sent beyond the current four messages (basic directions of forward, right, left, and stop). We should conduct additional wants/needs analysis to select what additional messages should be sent. Further, the send prototype should be constructed in a manner to flexibly support the addition of new messages without substantial rework. (Adding more messages to the first prototype would require substantial effort.)

While the primary objective of the next prototype is to evaluate send technologies, the following improvements should be made to the receive aspects of the next prototype, which are carried over from the first prototype:

- Transmission of sound should not have the delay or lag that is inherent in the first auditory prototype. The first prototype uses a web-based app; pressing a command on the touchscreen requires a round-trip to a web server, which returns a sound file that is then played through bluetooth to the bonephones. The lag in the auditory prototype is variable based on many uncontrollable factors (wifi load, dynamic routing selection, uncertain caching), which affects the ability to effectively and consistently evaluate it head-to-head with the tactile and visual approaches. In the next prototype, a native app should be written which would store the sound files locally, and deliver them directly through bluetooth. This will improve the ability to directly compare the performance of auditory and tactile receive approaches, on a level playing field.

- The haptic technology must be more rugged, and must be easily adjusted to the user (both in terms of fit and sensitivity). Feedback on this issue is important to evaluate in the next prototype, in order to inform the design of the fully envisioned system.

- Construction of the next prototypes should include capturing information to streamline evaluation, such as accurate timing. This function can be accomplished by integrating a real time clock and data logger into the hardware. The real time clock will keep track of time, while the data logger will acknowledge when a command has been sent with a time stamp and organize this data for export and analysis. Capturing when a command has

been received and acted upon can be gauged by human observation. Since having a human record the receive time is not ideal, we should also consider the ability to capture motion through an accelerometer or other device, and carefully consider the cost of including that feature in the next prototype versus its benefits in data capture. Regardless, including a more accurate timing capability can reduce possible evaluator error during the evaluation, improving the ability of the evaluation to correctly assess competing technologies. The evaluation process for the next prototype should include a more complex course and set of instructions; those requirements should be considered when developing the prototype.

- Further modifications of the haptic prototype include the ability to define finer turning directions. The first prototype assumed all participant turns would be 90 degrees. However, when a participant turned more or less than 90 degrees, there was no easy way to correct their course because a vibration in the opposite direction encouraged them to correct by another 90 degrees (more or less) which is an additional overcompensation, and would contribute to total task time. It is possible that more vibration motors could be placed in the hat to delineate 30, 45, and even 60 degree turns. However, another alternative could be including a magnetometer which can register the direction that the participant is facing and continue to buzz until they have pivoted to the correct direction. This idea, though, assumes that a predetermined track has been developed. Ultimately, finer control of direction is needed in later prototypes.

- For the purpose of the semester, we decided to communicate only navigational cues in order to have a proof of concept. However, in future prototypes, we would increase the complexity of the messages being sent with special effort to communicate more commonly used Hand and Arm signals in the Army. This would then result in more complex tactile and/or auditory patterns in the prototypes.

Conclusion

During this project, team “Hit The Road, Jack” showed the ability to:

- Define a target audience, its needs, and determine a feature set that might improve on pain points
- Design a research study and submit an IRB for that study
- Redefine a target audience while still keeping the same parameters, in order to continue the study
- Design and construct a physical prototype, including the use of pre-made components, which held down the cost and time of prototype development
- Quickly develop physical prototyping skills, including soldering and Arduino programming
- Define usability for wearables (heuristics for wearable usability is an area where the industry needs to make serious steps forward; we felt like pioneers in this area)
- Define and conduct a prototype evaluation, including recruiting participants, and gathering data for later analysis

- Conduct appropriate statistical analysis on the data gathered during the prototype analysis, in support of conclusions
- Determine appropriate next steps, including features to add to the next prototype, improvements to the first prototype, and improvements to evaluation procedures for the next prototype.