Gesture and Voice Prototyping for Early Evaluations of Social Acceptability in Multimodal Interfaces

Julie Rico and Stephen Brewster
Glasgow Interactive Systems Group
University of Glasgow, Glasgow G12 8QQ
{julie, stephen}@dcs.gla.ac.uk

ABSTRACT
Interaction techniques that require users to adopt new behaviors mean that designers must take into account social acceptability and user experience otherwise the techniques may be rejected by users as they are too embarrassing to do in public. This research uses a set of low cost prototypes to study social acceptability and user perceptions of multimodal mobile interaction techniques early on in the design process. We describe 4 prototypes that were used with 8 focus groups to evaluate user perceptions of novel multimodal interactions using gesture, speech and non-speech sounds, and gain feedback about the usefulness of the prototypes for studying social acceptability. The results of this research describe user perceptions of social acceptability and the realities of using multimodal interaction techniques in daily life. The results also describe key differences between young users (18-29) and older users (70-95) with respect to approach to understanding and preference of these interaction techniques.

Categories and Subject Descriptors
H.5.2 User Interfaces: Evaluation/Methodology, Prototyping.

General Terms
Design, Human Factors.

Keywords
Gesture, speech, social acceptability, prototyping, multimodal interfaces.

1. INTRODUCTION
Goffman describes every action that takes place in a public setting as a *performance* [7], and as mobile phones become increasingly integrated into our personal appearance, mobile phone usage becomes a performance. The variety of places where mobile phones are used means that performances are constantly changing and being reevaluated. With respect to multimodal mobile interfaces, the performative aspects of these interactions are accentuated given the often visible or audible nature of these interactions. Because of this, the users' perceptions of the social acceptability of multimodal interactions must be evaluated in order to create interfaces that provide a comfortable and enjoyable experience.

Conference'10, Month 1–2, 2010, City, State, Country.

The evaluation of social acceptability is made difficult by the fact that multimodal systems often require sophisticated sensing and recognition techniques, which are time consuming and expensive to produce. However, the use of experience prototypes [3] is a low cost way to gather feedback about the social acceptability of an interaction before any implementation has occurred, allowing designers to test divergent designs and choose to implement only those that are approved by users. This research investigates the use of a variety of experience prototypes in order to gain feedback about a divergent set of voice and gesture interactions that have been used in previous systems, could be supported using existing technology, and other techniques that might be used in the future.

2. PREVIOUS WORK
Since the early success of voice and gesture in an interface with the "Put that There" system [2], these interaction modalities have not seen the widespread success that many predicted [15]. A variety of technical challenges and human factors have been identified that affect the usability and experience of gesture and speech-based interfaces. Crangle describes the role of conversation in a speech-based interface [6], demonstrating the importance of speech exchange as an integral part of acceptance of such interactions. Crangle identifies additional aspects of conversation that are important parts of a speech-based interface, such as codified speech. Codified speech, which occurs naturally in everyday speech, refers to the style and use of words that emerges in different cultures and locations. This kind of codified speech could be used to create dialogues specifically suited to interaction with an interface. Attwater *et al.* describe the challenge of providing a natural speech-based interface given the complexity in implementation and unpredictability of users' behavior [1].

Väänänen and Böhm describe the challenges of gesture-based interfaces by comparing the use of gestures in conversation to gestures for interaction. Conversational gestures often have an implicit meaning, whereas gestures in an interface must have an explicit one. This leads interface designers to choose gestures which are often more similar to arbitrary hand positions than natural, conversational gestures. Cassell argues that users' have the same natural ability to use gestures in this way as they do to use DOS commands [5] or similar systems where arbitrary sets of commands are used and affordances are unclear. Gestures also present challenges in reliable recognition, for example the segmentation problem [13]. The issue of knowing when gestures intended for the interface begin and end makes the use of everyday conversational gestures as part of an interface difficult.

3. SOCIAL ACCEPTABILITY
While many challenges of using gesture and speech in an interface have been identified, they do not entirely address the issues of

multimodal mobile interface acceptance. The challenges that have been investigated in the literature thus far have not addressed the realities of using these interaction techniques as part of everyday mobile phone usage. As the mobile phone becomes integrated into everyday activities and individuals' appearances, the social acceptability of taking part in an interaction that uses these modalities plays a major role in interface acceptance [10]. The fact that multimodal mobile interfaces require users to change their normal behavior in public places highlights the need for designers to understand the social acceptability of taking part in such interactions. This means understanding the decision process of choosing what the appropriate actions are and how they should be executed based on one's personal opinions and ideas about that action and the perceived opinions and ideas of spectators. Thus, social acceptability is a combination of the personal and social forces that influence interface acceptance. Together, user experience, performer/spectator roles, and social acceptability, play an important role in the acceptance of mobile multimodal interfaces.

Previous work on social acceptability has mainly revolved around gestures and the possible scenarios in which they might be used. Ronkainen *et el.* completed a survey that asked respondents to assess the usefulness of different gestures given the locations where they might be used and the tasks the might be used for [11]. Rico and Brewster completed an on-the-street study that examined a set of gestures in both a public and a private setting over multiple trials [10]. This study discussed a variety of reasons why participants liked and disliked gestures. For example, the results showed that device-based gestures, that is gestures that directly manipulate a device, were significantly more acceptable than those that didn't because the visible presence of a device gave clear clues that explained one's actions to spectators. These studies, however, focused primarily on gestures even though these issues exist with a variety of modalities.

3.1 User Experience, Performance, and Social Acceptability

Although the exact scope and definition of user experience is still debated, it is clear that an understanding of an individual's thoughts, feelings and reactions to an interface are important factors that designers must consider [8]. With respect to multimodal interaction, an understanding of the user experience of these interactions is especially important because these interactions often require users to try new and possibly unfamiliar actions. The experience of using an interface develops and changes over time as the user is continually exposed to the interaction and experiences it in different settings with different people. User experience, however, is essentially an individual experience [8]. Although other people and spectators heavily influence the social context where an interaction takes place, the decision to interact and the experience of doing so is a personal and individual experience.

Because mobile phones are commonly used in public settings, the presence of spectators and the performative aspects of multimodal interactions play an important role in user acceptance. Following from Goffman's assertion that all actions done in a public setting are performances [7], the performance of an interaction with a mobile device can range from unconscious, automatic actions to explicit and deliberate performance on a stage. The presence of spectators and their affect on the performers has a major influence on the type of interaction the performer will experience [9]. Because of this, performer and spectator roles should play an important part in the design of multimodal mobile interfaces and the evaluation of social acceptability.

The combination of performer/spectator roles and user experiences form the two important aspects of social acceptability. This is manifested in the decision process when an individual chooses to take part in based on his/her own thoughts, and feelings and the perceived thoughts of spectators. This is clearly a process that changes over time as performers gain more experience with the interaction and gather more feedback from spectators. On a much longer time scale, social and cultural ideas also develop, spread, and change the definition of what is acceptable and what is not.

3.2 Experience Prototyping

Given the impact of social acceptability on interface acceptance, a wide variety of interaction techniques must be examined in order for designers to choose only those techniques that are both acceptable and usable. However, the technologies required to successfully implement complex multimodal interactions involving gestures, speech, haptics, etc. are often expensive, requiring extensive development time and sophisticated sensors and detection techniques. In order to make early evaluations of these interfaces possible, this research uses experience prototypes as a low cost tool for social acceptance evaluation. Experience prototypes refer to "any kind of representation, in any medium, that is designed to understand, explore, or communication what it might be like to engage with a product, space, or system" [3]. In this study, 4 such prototypes were used in a focus group setting in order to gain feedback about a set of multimodal interaction techniques using a minimal amount of development cost.

4. FOCUS GROUP STUDY

In order to better understand how individuals make decisions about social acceptability, this research used a focus group study that utilized a variety of experience prototypes. The goal of these groups was not only to explore the some of the possible factors of social acceptability but also to discover how the prototypes would be used and analyze the benefits and detriments of each one.

4.1 Study

The focus groups each examined a set of 16 gestures and 16 voice commands, shown in Figure 1. These modalities were chosen because of their often highly visible or audible nature. The performative aspect of these modalities makes them interesting interaction techniques from a social acceptability point of view. These modalities were investigated on an individual basis, rather than in combinations, to gather specific feedback about individual techniques before combining them.

With respect to the gestures, 4 categories where used; emblematic [7], device-based, arbitrary, and body-based. Emblematic gestures refer to those gestures that have a widely accepted meaning outside of the context of speech within a given culture. The device-based gestures are those that directly manipulate a device, such as a phone tap. Arbitrary gestures are hand positions that do not necessarily have an explicit meaning and may be open to interpretation. The arbitrary gestures used in this study were chosen based on their previous use in gesture-based systems [14]. Body-based gestures are movements of the body that do not directly involve manipulating a device, although external sensors might be manipulated for these gestures. Between these categories, there will be some amount of overlap where gestures may belong to multiple categories. With respect to the voice commands, 3 categories were used: command, speech, and non-speech. Command

inputs included one-word commands that related to phone tasks, such as “call” or “lock”. Speech inputs included short, commonly said phrases. These inputs were included to evaluate the acceptability of speaking to a mobile phone in a way that is not obviously related to a device or phone tasks. Non-speech inputs included a variety of sounds and noises, some of which occur normally in everyday life, such as whistling, and some of which do not like buzzing or popping.

For each focus group, the gesture and voice order was randomized, with half of the groups looking at gestures first and half the groups looking at voice first. Each focus group used one of the experience prototypes to familiarize the members with each interaction techniques. They then filled out a short worksheet with rankings and acceptance information, and participated in a semi-structured group interview. The interview topics included discussion of input preferences, locations where these inputs might be used, and the tasks for which these inputs might be used.

Gesture	Category	Voice	Category
OK Gesture	Emblematic	Say "Close"	Command
Money Gesture	Emblematic	Say "Open"	Command
Peace Sign	Embematic	Say "Call"	Command
Shrugging	Emblematic	Say "Lock"	Command
Device Stroke	Device-Based	Say "I'm Fine"	Speech
Device Shaking	Device- Based	Say "Bad Weather"	Speech
Device Flick	Device-Based	Say "That's Nice"	Speech
Device Rotation	Device-Based	Say "So Busy"	Speech
Upright Fist	Arbitrary	Humming	Non-Speech
Hook Finger	Arbitrary	Buzzing	Non-Speech
Sideways Fist	Arbitrary	Say "Chh"	Non-Speech
Open Palm	Arbitrary	Doo Doo Doo	Non-Speech
Shoulder Rotation	Body-Based	Say "Psst"	Non-Speech
Wrist Rotation	Body-Based	Whistling	Non-Speech
Foot Tapping	Body-Based	Clicking	Non-Speech
Head Nodding	Body-Based	Popping	Non-Speech

Figure 1. Table showing all gestures and voice commands organized by category.

4.2 Prototypes

Each of the experience prototypes was designed to allow participants to visualize and try each of the interaction techniques. The four prototypes used were videos, live demonstration, a Wizard-of-Oz phone prototype with vibrotactile feedback, and a phone-shaped prototype without feedback. Images of each prototype are shown in Figure 2.

The video prototypes portrayed a male actor sitting in a plain setting. Each gesture or voice command was displayed 3 times with a short pause in between. Demonstration prototypes included a live demonstration by the focus group leader. Again, this included 3 repetitions of each gesture or voice command, The Wizard-of-Oz phone prototype included a simple interface with an image of a stoplight. Participants would be asked to perform each gestures or voice command, When performed correctly, the on screen stoplight would turn green, the phone would vibrate briefly and play a tone. Two experimenters at a nearby laptop controlled this wirelessly while the participants took turns using two running prototypes. The shape prototype involved a set of phones, which were switched off, and small black clips that were used to represent a clip-on microphone. Participants were asked to clip this on the collar of their shirt and perform each gesture or voice command with the phone in their hand. The phone did not provide any feedback in this case.

Below: Demonstration prototype showing open palm gesture.

Above: Wizard-of-Oz Prototype.

Above: Video prototype showing hook finger gesture.

Below: Shape prototype using Nokia N95 and small black clip.

Figure 2. Experience prototypes used during focus groups.

4.3 Participants

The study was made up of 8 focus groups of 2-4 participants each. For the first 6 focus groups, which totaled 19 participants, the participants were selected from local university students, ranging in age from 18 to 29. These groups included two groups looking at video prototypes, two groups looking at demonstration prototypes, one group looking at shape prototypes, and one group looking at Wizard-of-Oz prototypes. The study continued with another 2 focus groups, which totaled 6 participants, where the participants were recruited from members of the local community ranging in age from 70 to 95. These groups included one group looking at video prototypes and one group looking at demonstration prototypes. Group participants in different age ranges were recruited to explore the differing perceptions of social acceptability and mobile phone usage between different generations.

Of the 25 participants, 56% of the participants were from the UK, 20% were from Asian countries, 8% were from Europe, and the remaining 16% declined to state. Focus group participants also answered questions about their mobile phone usage habits. Of the first set of focus group participants, aged 18 to 29, 100% of participants used their mobile to make phone calls and send text messages, 58% used the phone Web browser, 53% used their mobile to play games, and 37% used an email client on their mobile phone. Of the second set of focus group participants, aged 70 to 95, 83% used their mobile phone to make phone calls, 33% used

the phone to send text massages, and 17% used email, Internet, or the alarm clock on their phone.

4.4 Results

The results of these focus groups will be divided into two groups. First, the results from the set of focus groups including participants aged 18 to 29 will be discussed. Second, the results the focus groups including participants aged 70 to 95 will be discussed and compared to the first set of focus groups. For each of the focus group the results were gathered from the worksheets filled out by participants and the audio recordings of each session. Each participant filled out a worksheet with their own rankings for gesture and voice based commands separately as well as their acceptance or rejection of each of the interaction technique individually. After ranking each list independently, participants were asked to cross out any of the techniques that they would not be willing to perform in a public setting such as a busy pavement.

4.4.1 Focus Groups: Ages 18 to 29

The first six focus groups in this study involved 19 participants aged 18 to 29. Based on the worksheets filled out by these groups, Using the Friedman Test [4] to compare ranking data for the gesture and voice categories, these results show that both of these modalities have significant differences in rank of $p<=.001$ between their respective categories. Using the Wilcoxon signed-rank test for pair-wise significance tests, statistical significance was determined for comparisons between the categories. These are shown in Figure 3.

Comparison	p-value
Device - Body	< 0.02
Device - Emblematic	< 0.001
Device - Arbitrary	< 0.0005
Emblematic - Arbitrary	< 0.001
Body - Arbitrary	< 0.009
Body - Emblematic	**< 0.29**

Comparison	p-value
Command - Non-Speech	< 0.0007
Command - Speech	< 0.0001
Non-Speech - Speech	**< 0.45**

Figure 3. Pair-wise comparisons for gesture and voice commands, adjusting for Bonferonni's correction. Italicized items are not statistically significant.

This result shows that device-based gestures were significantly more acceptable than any other kind of gestures, with arbitrary gestures being the least acceptable, emblematic gestures the second least acceptable, and body-based gestures the second most acceptable.

As shown in previous research, device-based gestures have higher acceptance rates because they provide audience members with clear cues explaining the performer's actions [10]. Performers appear to feel more confident using these kinds of gestures because they are less likely to be misunderstood by audience members. Figure 4 shows the average rankings for each gesture, grouped by gesture category. The rankings ranged from 1 to 16, where 1 is the highest rank and 16 is the lowest.

These results also show how emblematic and arbitrary gestures rank as compared to the previously studied device and body-based gestures. The arbitrary gestures were ranked significantly lower than any other kind of gesture, with emblematic gestures being significantly more acceptable than arbitrary gestures and significantly less acceptable than device-based gestures. The arbitrary gestures were often described as unacceptable because they had unclear or confusing meanings for observers. This led participants to be unsure how they would be perceived by spectators and uncomfortable performing these commands.

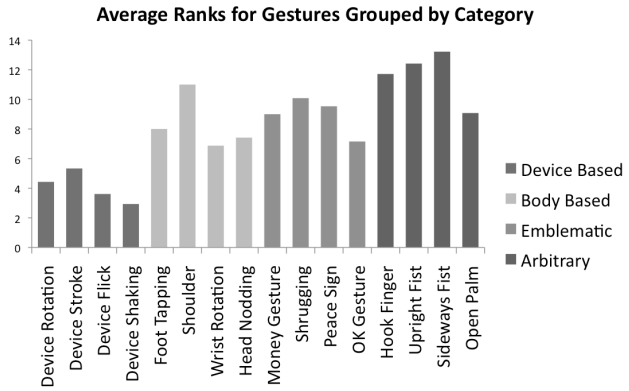

Figure 4. Gesture command average rankings for participants aged 18 to 29.

It is clear from these results that the imagined interpretations of other influences how a performer chooses to interact, and gestures that facilitate easy explanations are generally more acceptable than those that don't. This ability to demonstrate explanations of one's actions does not only apply to gestures, but also to speech. Command speech inputs were significantly higher ranked for acceptability than speech and non-speech sounds. One participant stated that "I don't mind all these easy to relate commands, lock, open, close, I don't even mind if I have to say them aloud." The clear indication that these commands were related to an interface made them more socially acceptable.

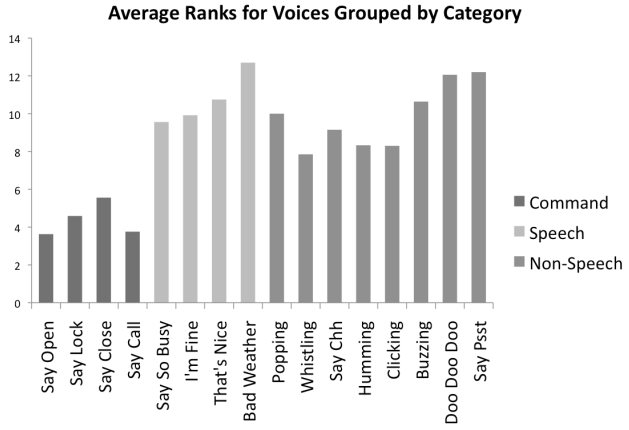

Figure 5. Voice commands average rankings for participants ages 18 to 29.

Although some categories or gesture or voice commands can lend themselves to a less ambiguous interpretation, the ability to demonstrate an interaction was not limited to command-based speech or device-based gestures. Participants described how unusual gestures could also indicate that actions were part of an interface. For example, one participant stated that "It's [hook finger] not to be mistaken, people won't connect that with a tick, it will be obvious you are steering something." One participant stated that "it is still a bit strange to start talking to yourself in the street. Whereas the things like the doo doo doos, they follow a pattern so someone would pick up that you were doing it for a certain reason. If you've got headphones on you have an excuse." Even a

different configuration of the technology could be enough to demonstrate an interaction. One participant described how they would be unwilling to perform voice commands with a hidden microphone, but that a visible microphone with a flashing light would be enjoyable, stating that "you want people to notice you are saying it to your phone rather than a hidden mic."

Given the wide variety of gestures and voice commands investigated in this study as compared to previous work, participants described new reasons for liking or disliking gestures or speech. One new reason identified for liking gestures or speech was the satisfaction involved in completing a task in this way. One participant stated that "For cancelling something, shaking is incredibly satisfying." Similarly, participants described other input techniques as fun, silly, and cool. Participants discussed their thoughts on how comfortable it would be to converse with a device. One participant stated that "I'm not sure if I want to be talking to my device as if it's a pet or a creature." While in some cases this may be due to the lack of responses and natural conversation [6], participants discussed their discomfort with the idea of a phone responding as well. One participant asked "if I say 'bad weather, what does it [the device] do? Console me? I don't see how that would work with the phone." Participants also described some situations where they disliked intentionally hidden gestures. One participant stated that "if I was saying 'I'm Fine' to my phone I wouldn't want people to think I was on my phone, it's sneaky, like having a fake phone call with my phone."

The focus group setting encouraged participants to consider each other's differing opinions, leading to discussions about personality differences and how social acceptability might change over time. One participant stated that "I think it's one of these things that maybe doesn't seem okay at the moment, but the more people that use it the more ok it would be. I think at the start of using your capacitive things when the iPhone first came out, flicking them, it seemed a bit strange at the start but it's okay." Talking about the future, participants discussed how they would feel if multimodal inputs became more widely accepted. One participant felt that "I don't want to look like an idiot doing it, but if everyone was doing it you would like an idiot if you weren't doing it." In contrast, another participant stated that "it's hard to say if I would be more comfortable doing it just because other people we doing it."

4.4.2 Focus Groups: Ages 70 to 95

Two focus groups were completed with a total of 6 participants ageing from 70 to 95. The comparison of ranking data for gesture categories between the this set of focus groups and those groups aged 18 to 29 is shown in Figure 6.

The biggest difference between these two age groups can be seen in the differing rankings for device-based gestures. Previous research has shown that device-based gestures are significantly more acceptable than body-based gestures [10], but these studies have not compared a broad range of ages. In focus groups with participants ranging in age from 18 to 29, the average ranking for device-based gestures was 4.2 as compared to 13.3 in focus groups whose participants ranged in age from 70 to 95. When describing dislikes, one participant stated that "I couldn't imagine stroking the device at all, that's alien to me." Because gestures directly involving the device were often unfamiliar to the older adults, these gestures did not have a clear meaning and were generally disliked.

Another clear distinction between these two groups was the ranking differences with emblematic gestures. In this case, focus groups with younger participants ranked emblematic gestures at 8.6 as compared to 5.3 in focus groups with older adults. This preference for emblematic gestures is further demonstrated in the rejection rates gathered from the worksheets, as shown in Figure 7. Emblematic gestures had a much lower rejection rate amongst the older adults, with a rejection rate of 4% as compared to 29% with the younger adults. Because the emblematic gestures all had easily recognized meanings, their familiarity made them more acceptable amongst the older adults. When describing preferences, one participant stated that she liked "the money gesture, I use that all of the time."

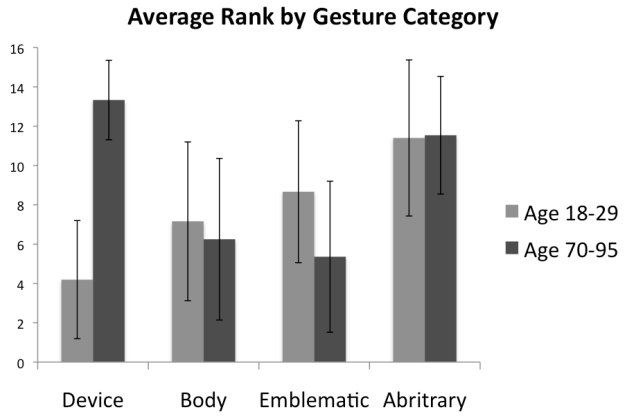

Figure 6. Comparison of focus groups with average rankings for gesture categories.

With respect to the voice input, the main difference demonstrated between the two age groups was with the speech-based input that was not directly related to phone commands. The average rankings for the voice command categories are shown in Figure 8. These inputs were generally highly ranked because they were often described as the familiar and common phrases in everyday life. One participant stated that "you would say 'That's nice,' that's an expression" The fact that there was a clear and unambiguous meaning associated with these sayings shows again how meaning was more important for the older adults than a clear association to the device or phone tasks.

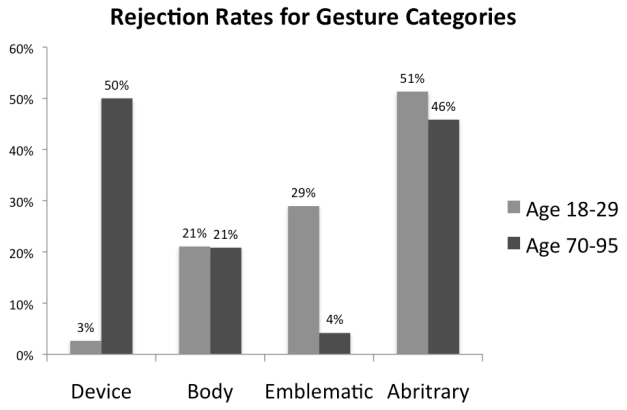

Figure 7. Comparison of gesture rejections rates for gesture categories and focus group age ranges.

The issue that repeatedly came up during these groups was what different gesture or voice commands meant to each individual.

While the association to the phone was important, the older adults tended to focus more on what these things meant personally before they would consider how they might be used on a phone. When describing voice commands that were liked, one participant stated that "they are familiar. A lot of these things convey nothing to me, Doo doo doo? I've never heard anyone say that." The lack of familiarity or meaning led to situations where gestures or voice commands couldn't be put into scenarios that made sense. One participant stated that "I don't know the implication of this hook finger. I don't see how I would use many of these." In the case where gestures or voice commands had multiple meanings or interpretations, this was an even bigger issue. When discussing why the open palm gesture was disliked, one participant stated that "I felt there could be two meanings. It could mean hi [open palm held at shoulder level], but if someone was being very aggressive to me, I would say stop [open palm held out with arm extended]." The fact that this gesture could mean different and possibly negative things given a slight change in performance made it less acceptable as a gesture to use in public. Multiple meanings were also discussed when these commands were performed in different contexts. One participant stated that "whistling, one might do to entertain the neighbors, but to whistle at someone, to catch their attention, might be rather rude."

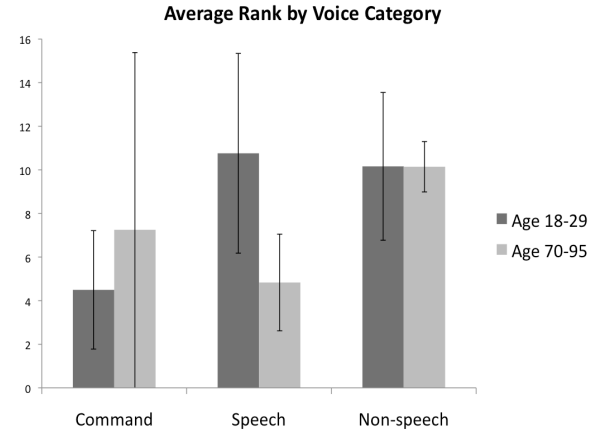

Figure 8. Comparison of focus groups with average ranking for voice categories.

Other issues discussed in these groups included gender differences with respect to the acceptability of certain gestures or voice commands. One participant described several gestures as simply unladylike. She said "what's the saying? Cackling hens and something… whistling women. No, it's not a good idea. At least that's what I was taught." In some cases, participants said they weren't embarrassed to do anything, or that if gestures or speech became popular with younger generations they would also be unembarrassed to adopt the behaviors. One participant stated that "if everyone else was doing it [popping, whistling, psst, doo doo doo], the young people were doing it… I mean, I was never going to get a mobile phone. I was never going to get and iPod, I've got both."

4.4.3 Concerns About Usability in the Wild

In all of the 8 focus groups, participants brought up specific concerns about how gesture or voice input might work in the wild and discussed some of their anxieties about how this might work.

False Positive Recognitions. The greatest concern participants discussed was the possibility of false positive recognitions given that many of the inputs could be accidentally performed during everyday activities. One participant stated that "the gestures were things you would do without thinking, like with your shoulders, you might accidentally activate the device." In this case, unusual gestures were sometimes favored as possible steering gestures. One participant stated that "it's [shoulder rotation] a very unnatural gesture, it's not something you would normally do, you can keep it as a steering gesture. This is a gesture that is sufficiently easy to do but not a normal part of your behavior that you could do it by mistake."

Audience False Positive. The second most commonly discussed concern was the possibility that spectators might mistake one of the inputs as an action directed towards them. One participant stated that "if you say something like 'Psst' somebody might think you are trying to talk to them." In particular, this was seen as a problem for inputs that did not involve a device. One participant stated that "it makes it even more ridiculous if you don't even have a phone in your hand, like playing with my iPod, I can shake it and it makes sense because I'm holding it. But if I'm walking down the street and suddenly [demonstrates a shaking motion] people might think I'm casting spells on them."

Distance of Device. For scenarios where the device was not directly manipulated, participants described a feeling of disconnect from the interaction and uncertainty about successful execution. One participant pointed out the importance of having a "clear connection with the action and the device." Another participant stated that "I think for the ones where you are actually holding the phone, you have some control over it." Another participant was worried that being distant from the device would make it more difficult to know if commands were interpreted correctly, which highlights the importance of feedback and the challenges of affective feedback when the device is distant.

Failure to Recognize Inputs. In a scenario where intended inputs were not correctly recognized, participants were concerned about the necessity to repeat movements or voice commands until a successful recognition was achieved, leading to frantic and embarrassing behavior. While demonstrating an erratic shrugging gesture, one participant stated "if it happens that when you try to do it and there's no execution, you keep on shrugging."

4.4.4 Experience Prototypes

One of the goals of this study was to evaluate different types of low-cost prototyping methods with respect to their usefulness and cost.

The cheapest prototype to create was the demonstration prototype. This prototype only required a demonstrator who had practiced and was familiar with each technique. The benefits of this prototype include the ability to either require participants to practice each command, request specifically that they only watch, or observe how participants respond to demonstration and which commands they choose to practice. This gives the focus group leader more control over the experience created by the prototype. The detriments of this prototype include the need for a trained demonstrator and the possible lack of consistency in performance, especially if different people act as the demonstrator. Also, the appearance and cultural background of the demonstrator may also have an effect of how the demonstrations are perceived. For example, an individual with a different accent from the focus group participants might be difficult for them to understand.

The second cheapest prototype to create was the shape prototype. This prototype only required a set of mobile phones and objects that approximated peripheral devices such as external micro-

phones. The benefits of this prototype include the ability to try each command with a consistent object for each participant. Especially for gestures that involve the device, this allows participants to experience how it feels to perform the command with a realistic object in their hand. For voice commands, the shape prototype allows participants to explore a variety of configurations easily and cheaply. For example, configurations including a hidden microphone in the collar and talking to the phone held to the ear could be easily experienced using the shape prototype. The detriments of this prototype include the lack of consistency between individual participants' performances. Because the shape doesn't provide feedback, it is difficult to ensure each participant performs the commands consistently. Also, this prototype may be uncomfortable for participants who are shy or nervous about performing commands in front of others.

The second most expensive prototype to create was the video prototype. This prototype required an actor to perform each command as well as video capture and editing equipment. This prototype also required a way of projecting the videos to be viewed during the focus group. The benefits of this prototype include a perfectly consistent performance since each participant will see the same videos. Like the demonstration prototype, the experimenter can also control the experience by requesting users to respond in a specific way to videos or by recording their automatic responses. The detriments of this prototype include the need for a projector or other way of showing videos to a group. Like the demonstration prototype, the choice of an actor may affect how individuals perceive performances as well.

The most expensive prototype to create was the Wizard-of-Oz prototype. This prototype requires a set of phones, laptop or other controlling equipment, and development time. Depending on the desired level of sophistication for the prototype, the development time can vary significantly. The benefits of this prototype include the ability to provide a hands-on experience that provides feedback to the participants and thus ensures of consistent performance for each participant. The detriments of this prototype include the fact that not all participants were able to use the prototype at once, which led to a slow and less smooth experience for participants. Also, participants who are shy or embarrassed performing these commands in front of others might be made uncomfortable when using this prototype.

Figure 9 shows the frequencies for three tags that were used to code the focus group transcripts, where each dot on the figure represents an occurrence of that tag within each focus group. These tags and frequencies demonstrate the topic of discussion and approach in understanding the commands in each group. The 'device' tag was used when any command was discussed in relation to a device, in both positive and negative ways. The 'meaning' tag was used whenever a command was describe by a certain meaning. This included cultural meanings, the meanings of familiar actions, or the lack of a meaning. The 'usage' tag was used whenever commands were discussed within a usage scenario or when usage scenarios were unknown or unimaginable. Examples of quotes that have been tagged show the types of things included in each category. For focus groups with younger adults, these frequencies show that discussions were evening distributed, with some groups focusing on each tag. For example, one group that used the demonstration prototype focused of discussions about the device while the other demonstration prototype groups hardly talked about the device at all. Stewart *et al.* describe a variety of factors that influence the dynamics of a focus group, including individual personalities, age, and appearance [12]. These factors are most likely the cause of variance between groups rather than the prototype used. This is also indicated in the difference between the older adult groups, which focused mainly on meanings, and the younger groups that focused each of the tags. This focus on meaning is also reflected in the previous result with the difference preferences for gestures and voice commands between these two age groups.

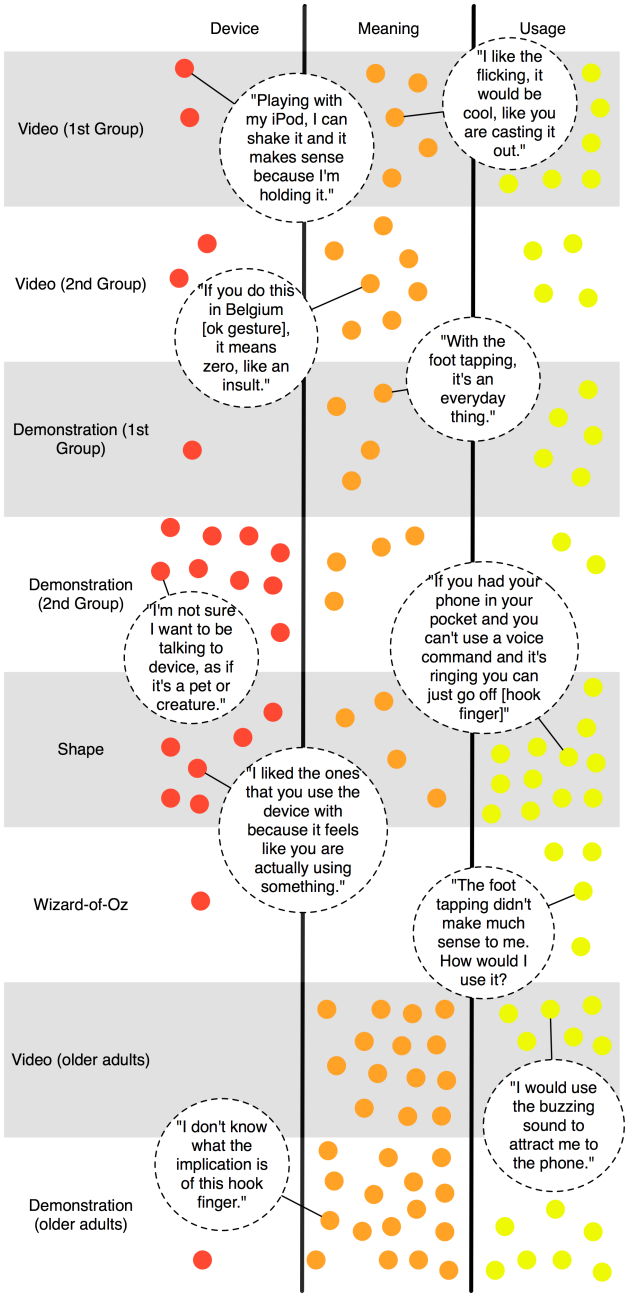

Figure 9. Selection of tag frequencies with example quotes for each focus group.

4.5 Discussion

There were some important differences that were observed between the younger adult and older adult groups. This includes a preference for emblematic and speech-based inputs rather than

device and command-based inputs. This was mainly due to the fact that the older adults focused more on personal meaning and understanding rather than a clear connection or application to the phone. This result was different than that of previous studies [10], however those studies did not incorporate a broad range of ages. In comparison to previous studies, this focus group study produced comparable results to the on-the-street study. Of the 8 reasons described for liking and disliking gestures from the on-the-street study [10], participants of these focus groups discussed each of these reasons. Although the focus group was able to elicit the same results due to the social atmosphere and collaborative discussion, it is still unclear whether this could be a complete replacement to on-the-street studies since the emergent behavior of spectators was not replicated in the focus group.

An important aspect of gesture and speech input design was considering the possible ways in which interaction could be demonstrated to spectators. Participants discussed the importance of avoiding confusion about why they were gesturing or speaking when such actions might be misunderstood by spectators. While this can often be achieved by simply holding a device, other methods, such as making peripheral devices more visible, were also identified as a successful was to demonstrate interaction. A possible area of interest in future studies of social acceptability includes a better understanding of the many ways this kind of demonstration can be achieved.

These focus groups also highlighted the anxieties of participants about using these gestures as part of a daily routine with respect to possible failures. This included failures not only on the part of the system, but perceptual failures on the part of both users and spectators as well. This issue creates an interesting opposition between acceptable and unacceptable gestures because those that were often ranked poorly were also cited for being unmistakable by spectators. For example, the hook finger gesture was ranked in the bottom five, but also described as usable because spectators would think it was obviously some kind of command.

Participants also discussed a variety of cultural differences and how they thought acceptance of gesture and voice inputs might change over time. Cultural differences with gesture showed how many of the movements could be impolite or even offensive in some areas. For example, the OK gesture was also described as a gesture for meaning zero, worthless or filthy in some cultures. Participants also discussed the differences in the appropriate level of noise making in public places in different cultures. One participant described how whistling in public can be considered improper in some cultures. These kinds differences were brought up by many of the focus group participants.

5. CONCLUSION

The results of this study showed that qualities of social acceptability previously investigated with respect to gestures are also applicable to voice-based input. The ability to demonstrate interaction to spectators, whether through a clear connection to a device or by observable aspects of the interaction, plays a major role in acceptability. This study also demonstrated the differing approach to understanding and accepting multimodal techniques between different generations of users. For the older adults, familiarity and personal meaning was more important than clear connection to a device when evaluating social acceptability. The variety of experience prototypes used in this study also provided insight into prototype design with respect to the cost of the prototype, the ability to control the experience, and the consistency provided.

6. ACKNOWLEDGMENTS

This work was supported by a National Science Foundation Graduate Research Fellowship, and the EPSRC funded GAIME Project (EP/F023405). Equipment and additional funding was provided by Nokia.

7. REFERENCES

[1] Attwater, D., McGrail, L., and Sargent, N. 2000. There's Nowt So Queer As Folk!. *BT Technology Journal* 18, 1 (Jan. 2000), 93-95.

[2] Bolt, R. A. 1980. “Put-that-there”: Voice and gesture at the graphics interface. In *Proc. SIGGRAPH 1980,* ACM Press (1980), 262-270

[3] Buchenau, M. and Suri, J. F. 2000. Experience prototyping. In *Proc. of DIS 2000*. ACM Press (2000), New York, NY, 424-433.

[4] Conover, W. J. Practical Non-parametric Statistics. John Wiley & Sons, Inc., New York pp 199-208 (1980)

[5] Cassell, J. (1998). A framework for gesture generation and interpretation. In R. Cipolla & A. Pentland (Eds.), *Computer vision in human-machine interaction* (pp. 191-215). Cambridge, UK: Cambridge University Press.

[6] Crangle, C. 1997. Conversational interfaces to robots. *Robotica* 15, 1 (Jan. 1997), 117-127.

[7] Goffman, Erving. The Presentation of Self in Everyday Life. Penguin Books, London (1990)

[8] Law, E. L., Roto, V., Hassenzahl, M., Vermeeren, A. P., and Kort, J. 2009. Understanding, scoping and defining user experience: a survey approach. In *Proc. CHI 2009,* ACM Press (2009), New York, NY, 719-728.

[9] Reeves, S., Benford, S., O'Malley, C., and Fraser, M. 2005. Designing the spectator experience. In *Proc. CHI 2005*. ACM Press (2005), New York, NY, 741-750.

[10] Rico, J., and Brewster, S. Usable Gestures for Mobile Interfaces: Evaluating for Social Acceptability. To appear in *Proc. CHI 2010,* ACM Press (2010), New York, NY.

[11] Ronkainen, S., Häkkilä, J., Kaleva, S., Colley, A., and Linjama, J. 2007. Tap input as an embedded interaction method for mobile devices. In *Proc. TEI 2007*, ACM Press (2007) New York, NY, 263-270.

[12] Stewart, D., W., Shamdasani, P., N., and Rook, D., W. Focus Groups: Theory and Practice. Sage Publications, California, USA. 2007.

[13] Strachan, S., Murray-Smith, R., and O'Modhrain, S. BodySpace: inferring body pose for natural control of a music player. *Ext. Abstracts CHI 2007,* ACM Press (2007), 2001-2006.

[14] Väänänen, K., and Böhm, K. Gesture Driven Interaction as a Human Factor in Virtual Environments – An Approach to Neural Networks. Virtual Reality Systems. R.A. Earnshaw, M.A. Gigante, and H. Jones. London, Academic Press Limited (1993).

[15] Wexelblat, A. Research Challenges in Gesture: Open Issues and Unsolved Problems. In *Proc. of the International Gesture Workshop* (1997). I. Wachsmuth and M. Fröhlich, Eds. LNCS, vol. 1371. Springer, London, 1-11.

Columns on Last Page Should Be Made As Close As Possible to Equal Length