User Research Considerations in the Design of Speech-enabled Mobile Devices

by

Mark Smolensky

David Attwater

Susan Hura

Catherine Zhu

Table of Contents

Introduction

The last five years has witnessed automated speech recognition (ASR) technology and speech output evolving beyond their most frequently utilizing application, the interactive voice response (IVR) system, to other interaction domains such as automotive, television remote controls, and mobile devices. Indeed, the latest incarnation of speech technologies embedded in the Apple iPhone 4s (together with artificially intelligent logic), incorporates a degree of robustness that has peaked mainstream interest. It appears the promise of speech technologies to become an accepted, and perhaps an expected, user interaction modality might have finally arrived.

When the general population begins to show an enthusiasm for adopting a particular technology, there is an unfortunate propensity for companies to exploit it anywhere and everywhere that can be imagined. There is little doubt that companies will also attempt to do this with speech technologies. Our recommendation can be summed in two words – DO NOT – lest you want to risk creating a poor user experience for your customers.

As is true with other user interaction methods, companies should carefully consider whether and where to make use of speech technologies in a product by taking into account the user population, the target task(s), and other factors. This paper will outline briefly what factors companies should consider with regard to incorporating ASR and Text-to-Speech/pre-recorded, digitized audio prompts (together here on in referred to as speech technologies in this paper) into mobile devices. These factors can only be considered through the lens of user research.

Why User Research?

User research helps companies avoid the pitfalls that come from designing based on a product-centric view of the world and instead design based on a customer-centric view of the world. In order to design a superior customer experience, a company needs to understand the following points: (a) who its users for a target product will be, (b) what their goals are, and (c) how they currently interact (or are likely to interact) with the product to achieve their goals.

User Research Methods

A variety of user research methods have been developed (or borrowed from other disciplines) to address the points outlined in the previous section. All of these methods seek to answer key questions around the users of the product being considered for design or redesign.

Rohrer (2008) categorizes various user research methods along **3 dimensions**:

- Attitudinal vs. Behavioral
- Qualitative vs. Quantitative
- Context of Website or Product Use

The attitudinal vs. behavioral dimension contrasts "what people say" with "what people do". The purpose of attitudinal research is usually to understand, measure, or inform change of people's stated beliefs. While is it critical to observe users' behavior, it is often useful to gather self-reported user insights about their expectations, perceptions, or understandings of a domain. As such, surveys, card sort techniques, and interviews can often prove useful in revealing such insights. Conversely, user research methods that focus on human behavior seek to reveal how people will actually act when performing a task. Ethnographic observation, prototype testing, the Wizard of Oz technique, live testing, and data mining are but some of the strategies for getting at these user behaviors.

The qualitative vs. quantitative dimension contrasts data that is gathered directly (qualitative) versus indirectly (quantitative). In ethnographic studies, for example, the researcher directly observes how people use technology (or not) to meet their needs. Researchers can then ask questions of users to understand better why they behaved a certain way or how to fix a problem. Conversely, quantitative insights are gained typically through statistical analyses of data sets gathered from surveys and performance logs. Quantitative insights are particularly good at revealing the scope of a certain set of behaviors or a range of behavioral patterns.

The context of product use deals with the technology level of fidelity that the user study is incorporating. For example, how are study participants going to be using the target product? Will they be afforded the opportunity to interact with the product in a naturalistic way or will their interaction be scripted somehow? Alternatively, perhaps the study is even more basic and the product will not even be involved.

The usage context of the study determines the types of questions that the researcher can address so it is best for the researcher to give some forethought to this. For example, if the researcher's goal is to minimize interference from the study in order to understand behavior or attitudes as close to reality as possible, then the context of use should focus on naturalistic interactions such as ethnographic observations and data mining from live testing. Contrastingly, a study based on a scripted interaction is usually done to focus on very specific insights, such as on a redesigned logic. Studies that do not make use of the product at all are conducted to examine issues that are broader than usage and usability, such as a study of larger cultural behaviors.

Most user research methods can move along one or more of the aforementioned dimensions, usually to satisfy multiple goals. For example, field studies can focus on what people say (ethnographic interviews) or what they do (extended observation); desirability studies and card sorting have both qualitative and quantitative versions; and eye tracking can be scripted or unscripted. The overall point, however, is that the user researcher should know what information they are seeking about the user population, their beliefs, their natural propensities; done well, this information will prove valuable to successful product design and help the product organization mitigate the risk of rolling out a product that will, at best, be underutilized, or at worst, abandoned altogether for a competing product.

Usefulness Factors for Considering ASR

Several key questions that companies should ask when considering speech technologies for their products can be answered by user research:

1. Who is using speech technologies, how often, and for what kinds of tasks?
2. Determine what non-speech input methods (if any) ASR users employ, how often they employ them, and for what purpose.
3. Determine the performance and satisfaction of current users of ASR and systems.
4. Assess the performance costs related to ASR errors.
5. Compare usage, performance, and satisfaction for speech technologies and non-speech input/output methods.

We present a brief discussion of some of the factors that can inform, through user research, *whether* speech technologies will be useful in a product. If, after user research results support the use of speech technologies in a product, the following factors will also inform the reader as to *how* speech technologies could best be incorporated into the product:

1. Context of Use
2. Device Capabilities/Constraints
3. Human Capabilities/Constraints
4. Activity
5. Propensity

These factors and some of the issues that may need to be considered for each are presented pictorially in Figure 1. Note that these factors are somewhat inter-related and the reader should not be surprised if some issues arise in discussion of two or more factors.

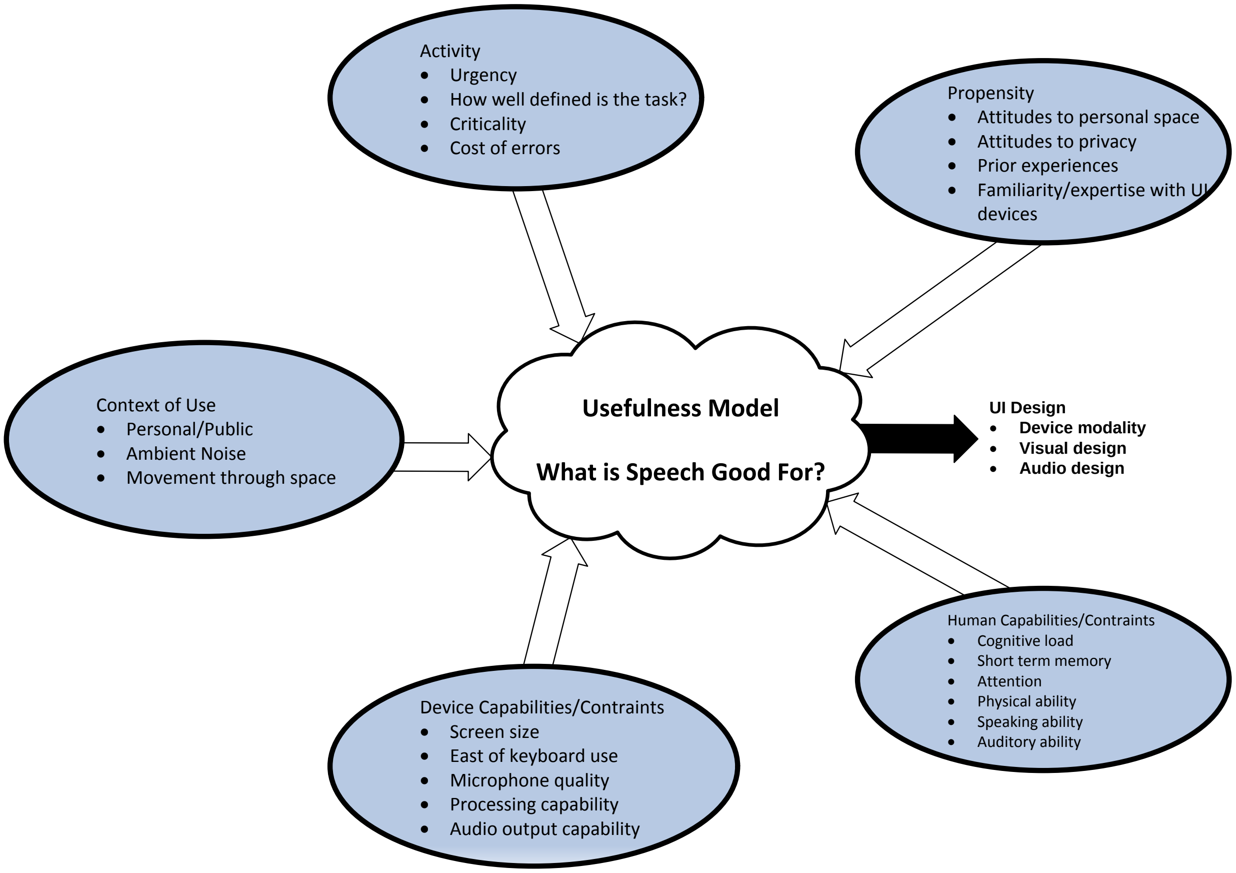

Figure 1: Speech Technologies Usefulness Model

Consideration Factor 1: Context of Use

Whether to provide for ASR and output modality will also depend on the context in which the user is performing the task. One contextual use point to consider is the ambient noise within which the user will likely perform the task. ASR accuracy degrades in noisy environments. Additionally, noisy environments can interfere with the intelligibility of text-to-speech or pre-recorded digitized audio prompts. Users may become frustrated having to fix mistakes continually due to misrecognitions. Koester (2004) found that 75% of users surveyed reported that fixing ASR recognition mistakes was the largest cause for dislike of any product that makes use of ASR as its primary user input modality.

The need for privacy in performing a task should be a crucial contextual consideration. Whereas users who before were accustomed to only having to be concerned with others peering over their shoulders to see what was on their device displays, speech technologies now presents even more concerns for privacy. Users may not want to speak a personal identification number (PIN) or have their account balance provided to them using audible output when in public, as this would compromise personal identification security. Similarly, users may not want to speak their credit card numbers or even speak their text messages (as in speech-to-text entry) because of privacy concerns. Indeed, Koester (2004) also found that the second most disliked feature of ASR involve privacy issues.

Users may be more apt to rely on speech technologies in a mobile device if they are performing a task where they are on the move, such as walking or driving. This is because their hands and eyes are often not free for long periods to be able to interact with gestures on a mobile device display. Additionally, tasks performed on the move are often more time critical and pertinent to the immediate context (e.g., users who are attempting to navigate from one point to another; users who need to notify someone they are about to meet with that they will be late). We will discuss this point a bit further under the Activity factor.

To sum, in considering whether to incorporate the speech technologies in mobile devices, attention should be paid to the contextual usage of the task(s). Where privacy issues exist, or ambient noise is too overwhelming, it may be best to allow for a gestural input alternative and a visual output alternative. Where mobile device applications are expected to be used while users are “on the go”, it would seem that speech technologies may be an appropriate solution to allow users to focus on their primary task of walking or driving. These factor elements, however, should be weighed against each other as well as against the other factors outlined in this paper.

Consideration Factor 2: Device Capabilities/Constraints

Users will be more inclined to opt for the speech modality channel if they are using a communication device where the screen size is small and the keyboard cumbersome AND the time and effort to complete the target task is perceived to be faster with speech than with gesture/visual modalities (Koester, 2004).

In considering whether to incorporate speech technologies in a mobile device, it is important to assess the quality of the device’s microphone, as poor quality will affect ASR accuracy. Most cell phones today, for example, have eletret condenser microphones that are omnidirectional. They capture sounds from all around without discretion. These microphones, however, do utilize some cancellation technology that helps eliminate some ambient noise. A better solution would be to include two microphones – one for speakerphone use that is a standard omnidirectional electret microphone, and a hyper-cardioid microphone for up close use. There are many organizations currently working on this very technology (e.g., Jawbone, Capcom, Motorola, etc.).

Similarly, if a mobile device carries a rudimentary sound generator or poor speakers then aurally presented information will be apt to be misunderstood. This issue is also compounded if the mobile device does not possess or interact with a robust pronunciation dictionary. Under these conditions, users would be more likely to view the mobile device's display unless their eyes were occupied elsewhere.

An addition consideration is the processing speed of the device or network for speech technologies. If significant latencies exist between the callers’ utterance and the device response, users will likely not opt for the speech modality, except if their need is urgent for hands free/eyes free.

Consideration Factor 3: Human Capabilities/Constraints

As with any other user interface (UI) design element under consideration, so too must human capabilities and constraints be considered when deciding on what aspects of a UI will use ASR or TTS/DS. For example, people have a natural tendency to try to minimize cognitive loading of information. This is why people write things down instead of trying to maintain them in memory. It is also why people prefer to see output that is lengthy or information-rich displayed rather than spoken back to them. Care should be taken to consider the amount of information that will be provided back to the user so as to determine the best method for presenting it.

Few tasks are purely cognitive. Physical ability is critical to the success of performing a task. Speech is often more desirable for user populations with physical constraints that make other modes of interaction difficult or impossible. Text-to-speech output has long been used by the visually-impaired to give them access to printed information that would otherwise be unavailable to them. Similarly, ASR may afford greater access to information to users with motor impairments. Speech technologies may also be attractive to users with age-related declines in vision and motor abilities. Older users may have trouble reading text displayed on small mobile device screens or accurately hitting small targets on touch screens, and may therefore be more willing to use speech.

Speaking and listening also have critical physical components. The degree of diverse abilities of the target population must be considered. Models of ASR are built on fluently produced speech, so potential users with motor impairments or illnesses that reduce their ability to produce fluent speech may not be successful using speech applications. Similarly, non-native speakers of a language often produce non-fluent speech that is quite different from the models used to train ASR engines. In some cases, non-native speakers use a speech application in their non-native language because the native language version is not available, but it has been noted anecdotally that some users choose the non-native speech application even when one is available in their language. The reasons for such choices are not well-documented, but may be evidence of cultural factors influencing users' language choice. Several ASR vendors have created an ASR model for "Spanglish," (i.e., English as produced by native speakers of Spanish). Such custom speech models tuned to the speech production of a non-fluent population are possible, but this is a significant undertaking and not easily accomplished with "off the shelf" speech components.

To state an obvious point, the auditory capability of the target population should also be taken into consideration. Users must be able to hear and comprehend the spoken output of a speech application. Pure auditory ability is a consideration for older users, who often exhibit hearing loss with increasing age, as well as age-related declines in auditory processing related to listening comprehension. Non-native speakers may also have more difficulty with listening comprehension than native speakers. For both users with hearing loss and those who may have trouble with auditory comprehension, the remedy is the same and relatively easy to accomplish: text-to-speech or pre-recorded, digitized audio prompts can be presented more loudly and more slowly, the ability to repeat any audible output should be simple and obvious, and if possible within system constraints, non-audio version of the information should be made available to the user.

Consideration Factor 4: Activity

The nature of the activity to be accomplished via an application has a significant impact on how well-suited it is for speech technology in several ways. Well-defined activities that have predictable inputs and outputs tend to be amenable to speech technologies. When the range and variety of user input is well-defined, ASR systems are more likely to correctly recognize it. Similarly, if the output of a system is well-defined, it is possible to use pre-recorded, digitized audio prompts (which tend to be more aesthetically pleasing and easier to comprehend) or to fine-tune text-to-speech output for proper pronunciation and inflection.

The urgency and criticality associated with a task can be factors in favor of speech technologies. We are defining urgent activities as those that must be completed within a short timeframe that is not under the user's control. Critical activities are those for which the user may face unwanted consequences if he fails to complete them. If an urgent or critical activity arises when the user's eyes and/or hands are busy with other important tasks, speech may enable the user to attend to the activity in a timely way without disengaging from other tasks. A good example of this sort of critical activity would be the act of stocking inventory in a warehouse setting. Having to continually rely on gestural input or visual output on a mobile job inventory aid, will impede the warehouse stocker's progress. Conversely, the stocker asking for the next item to be stocked and being instructed what the item is and where it should be stored, allows the stocker to carry on their task with their hands and eyes.

As briefly mentioned earlier under context of use, urgent and critical tasks may arise in situations where other modes of interaction are not available to the user. A user who might normally favor another modality may be willing to use speech technologies if it enables him to accomplish a critical task quickly and simply in his current environment. For example, workers might be open to using a speech application that allowed them to submit their timecards via speech in time to get paid, while still performing their job tasks. Note, however, that possible ASR errors must be taken into consideration for critical and urgent tasks. ASR may be more error-prone than tolerable for such tasks (the difference between fifty and fifteen hours would be significant in the hypothetical timecard application for both the employee and the organization.) Applications that rely on speech for critical and urgent activities can ameliorate the effects of misrecognitions by requiring the user to ecplicitly confirm recognition results ("You worked fifteen hours during the last pay period, is that correct?"), but such confirmation strategies make the interaction significantly longer and may make them less attractive to users.

Another factor that makes ASR less viable is a high cost of errors for the activity. In some cases, it is possible to recover from ASR errors in a way that is acceptably efficient and effective with no costs other than additional time in the interaction. For other activities, the cost of recognition errors may be too high for speech to be a good fit. Some activities may occur at a pace at which explicit confirmation takes too much time, so the initial recognition result must be used. In such cases, incorrect recognition results may lead to financial, safety or health repercussions for the user. The difference between fifteen and fifty hours on a timecard is important, but the difference between fifteen and fifty milligrams of a medication may dramatically impact a patient. When the costs of recognition errors are high, and when explicit confirmation of recognition results is impractical or impossible, speech may not be the preferred mode of interaction.

Consideration Factor 5: Propensity

The suitability of speech interaction in an application is also affected by the users' propensity to use speech in the typical context of use of the application. Compared to the other factors discussed in this paper, propensity is more difficult to determine objectively and with certainty, because it deals exclusively with users' attitudes, expectations, and behaviors. A user who has had numerous negative experiences with ASR in the past will be less likely to willingly use speech again as compared to a user who appears similar on measurable, objective criteria.

Propensity to use speech technologies is also bound up in cultural norms and values in a way that other factors are not. As mentioned earlier in this paper, speech is a more public and obvious modality because we can't shut our ears the way we shut or avert our eyes. Interacting via speech is "out loud" and available for others to overhear, and thus propensity to use speech is affected by notions of personal space and privacy. Users will tend to be less comfortable and therefore have fewer propensities to use speech technology when others are within the boundaries of their personal auditory space, and when the interaction concerns personal or private information. Speech is not likely to be the preferred modality for financial transactions or healthcare applications when others are within earshot of the user.

Propensity to use speech technology is also affected by users' expertise with speech and other possible modes of interaction. Although speech offers many benefits to users, some will choose to continue to use the touchtone IVR system that they have memorized through long use. When the target user population has an alternate modality available that is over-learned, they may have fewer propensities to choose speech. On the other hand, users who have successfully interacted with other speech applications may have greater propensity to try speech in new contexts because they have firmer expectations for how and when to speak.

Motivations for ASR in Mobile Devices

Whether the user will use speech technologies, whether the user will find it beneficial, and whether the user will find it desirable will depend on many factors. In the end, it is a matter of weighing enough of these factors that should determine the use these technologies in human-system interactions.

Scheiderman (1992) outlines several key categories of systems. While not specifically discussed with regard to speech technologies, it is nevertheless helpful to talk about speech technologies within the additional context of these system categories:

- Life-critical systems
- Industrial and commercial systems
- Office, home, and entertainment applications
- Exploratory, creative, and cooperative systems

Life-critical systems require high degrees of reliability and effectiveness. Types of mobile systems include power utilities technician job aids, aircraft inspection devices, municipal emergency services, and medical assistive devices/patient charting, among others. For life-critical systems, user entry of information into the system and system display of results is paramount as the cost of errors is high (i.e., people can die when errors are made). An aesthetically pleasing user experience is less important within these types of systems because the users are well trained and are highly motivated to maintain the safety of the system. For this reason, users must be able to have trust in the speech technologies. For this to happen, ASR had better be accurate or allow for fast correction when ASR mis-recognitions occur, lest users will opt out of using it.

Industrial and commercial systems include banking, insurance, inventory management, reservations, billing, and point of sales devices. For these systems, costs control is paramount even if controlling costs means settling for a lower level of reliability than would otherwise be achievable. Often, controlling costs within these systems means increasing users' speed of performance. The tradeoffs for speed of performance and error rates are decided by the total cost over the system's lifetime. Speech technologies for these systems would be beneficial to the extent that ASR and audible output can aid in the acceleration of task performance.

An office, home, and entertainment application is another type of system that includes email, text messaging, smartphone applications, GPS navigation systems, automotive controls, and television remote controls. For these systems, usefulness, ease of use, low error rates, and an aesthetically pleasing experience is crucial because use is usually optional and if the user cannot succeed quickly, they are likely to abandon the use of the product and try a competing product. It is with these types of systems where the largest exploitation of speech technologies is likely to occur.

Exploratory, creative, and cooperative systems include mobile search, financial decision-making, writer's workbenches, and medical expert systems. In these systems, the users may be well versed about the task domain but may also be novices in how to perform their desired task with the system. At best, designers can pursue the goal of having the system become transparent as users focus on the task at hand. This goal seems to be met most effectively when these systems provide a natural affordance for users to interact directly with them, such as a natural language ASR dialog interaction, followed by immediate audible feedback and a next set of steps presented by the system.

The purpose of having presented these different types of systems is to get the reader to think about a product according to one of these broad categories because they inform about motivational tradeoffs as well as the user population base.

Conclusion

The specific methods for answering the key user research questions we posed at the beginning of this paper are beyond the scope of this paper. Solid user research methods, however, do exist for addressing them and careful consideration should be given to determining *whether* and *how* speech technologies should be incorporated in a target product.

The key takeaways of this paper should be:

- Product designers should first consider the general system category for which they are designing; each category carries with it likely characteristics of the user population, design priorities, and consequences of human error. Defining the type of system category you're designing for can help orientate the designer to answering the aforementioned questions about whether to make use of speech technologies in a product.
- Having categorized the type of general system category they are designing for, user researchers should consider, in greater detail, the speech technologies usefulness factors necessary to understand the user, their goals, and their naturalistic ways of going about accomplishing target tasks (Figure 1). Consideration of these factors will aid in answering the key user research questions at the beginning of this paper and will shed light on whether speech technologies provide a natural affordance for completing the target task.

Done well, this information derived from a systematic user research effort will prove valuable to successful product design and help the product organization mitigate the risk of rolling out a product that will, at best, be underutilized, or at worst, abandoned altogether for a competing product.

References

Rohrer, Christian (2008). When to use Which User Experience Research Methods. Jakob Nielsen's Alertbox, October 6. 2008.

Koester, Heidi Horstmann (2004) Usage, performance, and satisfaction outcomes for experienced users of automatic speech recognition. Journal of Rehabilitation Research & Development, Volume 41 Number 5, September/October 2004, Pages 739 — 754.

Scheiderman, Ben (1992), Designing the User Interface. Addison-Wesley Publishing Company, Inc.