Designing Intelligent Multimodal User Interfaces for Mobile Wireless Devices

Osamuyimen (Uyi) Stewart
IBM Research

Abstract
Recent study by Opinion Research has shown that although many people find mobile devices (like Smartphone) initially attractive, their longevity and widespread adoption lies in their usability, intuitiveness, and an enjoyable initial user experience. Against this background, the focus on mobile device development has been on speech and text input/output. Use of speech alone suffers from temporality and ambiguity; similarly, use of text alone faces the challenges of hands-free and eyes-free interaction. This work examines two important design issues regarding (a) modality synergy, i.e., how to combine speech and text input/output in an intelligent manner whereby they complement each other's weaknesses and (b) recovery from errors, i.e., how to shift the burden away from the user in a way that will result in a graceful task completion.

1.Introduction

The ubiquity of mobile devices has served as a catalyst in its tremendous global adoption, adaptation, and penetration. This global growth fuels the demand for intelligent and innovative services beyond current uses in: (a) personal communication (e.g., email or voicemail, etc.); (b) Personal Information Management (PIM) applications (e.g., calendar, to-do or task lists, address book, etc.); and (c) information queries for timely data (e.g., stock quotes, weather, directory assistance, etc.). As these applications gain in popularity, there has been unfortunate growing user dissatisfaction with their usability. For example, it has been observed that although many people find mobile devices (like Smartphone) initially attractive, their longevity and widespread adoption lies in their usability, intuitiveness, and an enjoyable initial user experience [1]. Therefore, it is crucial that we re-think how mobile devices are designed and optimized for usability in order to offer a usable, intuitive, and best of breed user experience. The remainder of the paper is organized as follows: In section 2, we examine each modality separately and show that any ad hoc decision devoid of empirical justification with respect to choice of modality or, in general, overall design and usability will result in sub-optimal user experience. Finally, in section 3 we examine how to combine the relative strengths of each modality to design synergistic next-generation best of breed user interfaces (intelligent multimodality user interface).

2.Designing User Interfaces for Mobile Devices

Designing user interfaces for mobile devices present interesting challenges and opportunities that must be carefully examined in order to realize an optimal user experience. At the core of the user experience is the basic issue of choice of modality or modalities for interacting with the mobile devices. Modality refers to the use of a medium, or channel of communication, as a means to express and convey information.

There are several choices in the universe of verbal and non-verbal human communication modalities: speaking, writing, visual, logographic, gestural, emotional, touching, eye gazing, etc. Thus far, only two of these have been used in commercial versions of mobile device development: Direct manipulation (typing/text with stylus or mini-keyboard) or Speech (speaking/voice). Typically, only one of these modalities is offered—text or speech, and very rarely a combination of the two modalities. There are some fundamental linguistic and cognitive differences between these two modalities that need to be carefully examined and understood before combining them in mobile devices. Here is a summary of five such differences (amongst others) that are relevant to the focus of this paper:

1. ***Primacy of speech***. This implies that most humans learn to speak ever before they learn to write. Indeed, it has been postulated that we are born with an innate (unconscious) predisposition for language (speaking) [2, 3]. By implication, this view is in contrast with direct manipulation (typing/writing) which may be considered as a conscious and learned habit.
2. ***Functionality***. Although speaking and writing are symbolic systems, yet they differ significantly in the ways in which users may select which modality to use. It has been argued that choice of modality (speech or text) is functionally derived from the prevailing communicative setting or tasks [4].
3. ***Cognition***. Speech is transient or ephemeral and requires focused cognitive attention. Unlike writing, the sounds we hear in a speaking event must be initially processed in our short-term memory, thus requiring us to focus cognitively on the sounds in order to be able to "hear" all that has been said and even to remember (unless they are repeated) [5].
4. ***Structure***. Written language differs from spoken language. While meaning and form (sounds) remain the same in both modalities, there are many structural aspects in which writing and speaking differ. For example, this difference has been documented with the structure of prompts in conversational interactive user interfaces [6].
5. ***Learnability***. Humans differ with respect to their basic approach to learning and usability of communicative channels such as touch, gaze, speech, pictures, etc. It has been proposed that how we learn new concepts or to do things is dependent on how we are "wired", whether we are visual, tactile, acoustic, etc. [5]. This, in turn, may influence our preference for a given mode of interaction with mobile devices.

Theses differences underscore an important and fundamental question of how to design intelligent user interfaces. The obvious answer lies in the designer's ability to combine the comparative strengths of both modalities for use in a mobile device. We will now examine the relative pros and cons for each modality.

2.1Speech-based Mobile devices

Speech is a natural form of communication that is pervasive, efficient, and supports hands-free and eyes-free interaction which fits very well with the context of mobile device use. As an illustration, consider the following scenario:

Bill Woods is driving to an important meeting but he is running late. He needs to let his colleague, Jim Boyd, know that he will be about 30 minutes late. He is not sure if Jim is still at home, or on the road, or if he has already arrived at the office for the meeting. The problem can be summarized as follows: Bill needs to dial multiple numbers within a short time, while using his hands to drive, keeping his eyes on the road, keeping his focus on driving, and maintaining the right speed, which are some of the attendant features common to the mobile environment.

Through the use of voice recognition, Bill may be able to make his calls by simply speaking voice commands such as "call Jim at home" or "call Jim at cell phone" or "call Jim at work". For example:

Bill:	Call Jim at home
System:	Ok. Calling Jim at home, if you need to stop the call simply say cancel. [Dialing]
Ann:	Hello, Ann speaking.
Bill:	Hi Ann, Bill here. I'm trying to reach Jim. Is he still at home?
Ann:	Oh, hello Bill. Good morning. Actually, Jim left really early this morning. You can try his cell.
Bill:	Thanks. I'll do that. Have a nice day. [Disconnects call]
Bill:	Call Jim at Cell phone
System:	Ok. Calling Jim at cell phone, if you need to stop the call simply say cancel. [Dialing]
Jim:	Hi, this is Jim. I'm sorry I can't take your call right now... [Disconnects call]
Bill:	Call Jim at work
System:	Ok. Calling Jim at work, if you need to stop the call simply say cancel. [Dialing]
Jim:	Hello, this is Jim.
Bill:	Hey buddy, I am stuck in traffic. I'm running about 30 minutes behind. Please go ahead and start the meeting promptly. I'll join you guys shortly.
Jim:	No worries. I understand. See you soon.

From this scenario, we observe that through the use of speech modality, Bill is able to meet the hands free and eyes free requirement without having to scramble through the maze and challenge of using his hands to dial these numbers on the telephone keypad, while needing to use same hands to drive, and also keep his focus and attention on safe driving. In light of this, the expectation in the last three decades of research is that speech-based interface will become the future of computing, although its application and adoption in current mobile devices is still in its infancy. For small, portable devices, speech-based interaction has several advantages including low-cost and small hardware, it can be used on the move or whilst the eyes and hands are busy, and it is natural and quick. However, speech is also fraught with some challenges that must be accommodated when using this modality. There are at least three of such challenges that must be addressed: temporality, ambiguity and propensity for recognition errors. Let us illustrate with some modifications of the scenario just described above:

Bill: Call Jim at home”
System: Ok. Calling Jim at home, if you need to stop the call simply say “cancel” [Dialing]
Bill: Stop, hold on. No, not that number
System: I’m sorry I didn’t understand

The problem illustrated by this modified scenario is all too familiar to most of the people who currently use speech recognition systems. Essentially, the problem centers on the users’ control of the interaction. In this particular instance, the system makes the point to inform or educate the user how to stop a call (e.g., by saying “cancel”). However, since speech is temporal and it cannot be immediately revisited except it is re-spoken, Bill does not remember what the exact command and tries other likely alternatives which fail to trigger the right response. Asking users to remember exact words or phrases for specific actions imposes high cognitive burden on a user, while tasking their short-term memory. The task of remembering *exactly* what to say is in competition with other things in the short-term memory such as the task of staying focused on the driving. In this typical scenario, the user (Bill) needs to focus a little more on the interaction; which, however, is difficult to achieve for mobile users who are mostly involved in multiple tasks at the same time.

Another issue is that speech user interfaces are fraught with ambiguities at several levels including sentential, lexical, acoustical, etc. [7]. Consider the following scenario:

Bill: Call Jim at home”
System: Ok. Calling Tim at work, if you need to stop the call simply say cancel. [Dialing]
Bill: Stop, cancel. CALL J-I-M at HOME
System: I’m sorry I didn’t understand

Once again, this is another familiar scenario with users of speech recognition systems. The user requests to speak with “Jim” but due to the large amount of homophones in English, which makes the already-far-from-perfect speech recognition accuracy even poorer, the system confuses “Jim” for “Tim” (because both names are in Bill’s address book). What happens next is really the point being made here because although acoustic confusability is a natural occurrence (even in human-human interaction), notice that due to the initial error of confusing “Jim” with “Tim” the user now attempts to help the system by slowly emphasizing the pronunciation of “Jim”(J-I-M). This results in an unnatural acoustic form “J-E-E-M” which does not match the intended form “JIM”. This is an instance of hypertalk which refers to speech that is over-enunciated and spoken more slowly and loudly, in an attempt to overcome communication problems in human dialog [8]. Unfortunately, while hypertalk is effective in recovering communication problems during human-human conversation, it has been shown to further degrade speech recognition performance, rather than helping the recognizer [9].

Finally, the performance of the speech recognizer deteriorates in mobile environments due to background noise, mobility of users, and other unpredictable factors. In conclusion, speech is a natural form of communication that is pervasive, efficient, and supports hands-free and eyes-free interaction which fits very well with the context of mobile devices use. However, due diligence is required in the design of mobile applications that use speech to take advantage of the strengths of this modality while avoiding its pitfalls such as temporality, ambiguity and the propensity for errors, in order to achieve optimal user experience.

2.2Direct-manipulation-based Mobile Devices

Direct manipulation (with stylus and/or mini-keyboard) is based on the visual display of objects of interest, the selection by pointing, rapid and reversible actions, and continuous feedback [10]. Going back to the problems described in the speech-based scenario above, i.e., Bill needs to dial multiple numbers within a short time, while using his hands to drive, keeping his eyes on the road, keeping his focus on driving, and maintaining the right speed, which are some of the attendant features common to the mobile environment. With direct manipulation, Bill would have to select from a list of icons on his mobile device to initiate the desired action (e.g., call Jim at home, call Jim at work, etc.). In this regard, the issues of temporality observed with speech-based interface are minimal with direct manipulation because the user is in control of the interaction; they decide when to press the icon, which one to select, etc. Moreover, unlike speech, there is no problem with ambiguity because of the near-isomorphism between the object (icon) selected and the action triggered. Therefore, it comes as little surprise that direct manipulation is currently the predominant interaction modality for mobile devices because it is assumed to be fail-safe, transparent, and intuitive.

However, like speech-based interfaces, the direct manipulation modality also has several constraints that affect the usability of mobile devices. For one thing, direct manipulation has only limited means of object identification that make it difficult to communicate complex actions freely. Users can select one icon to trigger fixed actions but are never able to express complex events or actions by combining multiple actions. For example, expressions like “Call Jim at work in Chicago” [in Chicago as opposed to New York], which involve the idea of combining two variables where a second entity is a modifier of the first, is nearly impossible with direct manipulation-based interface. This sort of constraint imposes considerable cognitive burden on users of mobile devices with only a direct manipulation modality and this greatly affects user adoption of these devices and usability.

Furthermore, direct-manipulation-based interaction requires the use of our hands and eyes free for input/output, which clearly does not fit the mobile context. The concept of focus of attention (FOA) has been proposed as a way for evaluating the demands on the user’s attention when using different mobile devices [11]. Each instance that requires an additional attention on the part of the user increases the FOA by one. An example of a single FOA scenario is an expert touch typist using a QWERTY keyboard to copy text from a nearby sheet of paper. Since the typist is an expert, he/she only attends to the source of the text without the need to look at the keyboard or display. In general, the goal should be to minimize the FOA in mobile devices since people often attend to their surroundings when using devices in the mobile context. In the illustrative scenario with

Bill, the use of direct manipulation modality scores very low on the usability scale on account of the increased FOA associated with the use of hands and visual engagement for dialing the multiple numbers and still being able to drive and focus on the road.
In conclusion, speech and direct manipulation modalities have comparative strengths and weaknesses that can be systematically and intelligently exploited for the design of optimal user interfaces for mobile devices. In section 3, we will focus on how to combine the relative strengths of each modality to design synergistic next-generation best of breed user interfaces (intelligent multimodality user interface).

3.Designing Intelligent Multimodal User Interfaces

Multimodality is the simultaneous or alternate use of several modalities such as speech and text input/output for communication. When properly designed, a multimodal interface can support natural, flexible, efficient, and powerfully expressive means of human-computer interaction that are easy to learn and use. And it can be used for applications, user groups, and usage contexts that either have not been available or have been accommodated poorly in the past [12], particularly in the mobile context.

Grasso et al. [13] have identified two essential principles relevant to the research in multimodal speech and direct-manipulation interfaces, summarized respectively below:

A. The complementary framework between speech and direct manipulation.

Cohen [14] has identified some of the complementary strengths of direct manipulation and speech interface:

Direct Manipulation	Speech Recognition
Direct engagement	Hands/eyes free operation
Simple, intuitive actions	Complex actions possible
Consistent look and feel	Reference does not depend on location
No reference ambiguity	Multiple ways to refer to entities

Table 1. The Complementary Strengths of Direct Manipulation and Speech

By combining multiple modalities, the strengths of one modality compensates for the weaknesses of the other. Speech fundamentally enables hands free and eyes free communication and can also complement direct manipulation in being able to specify simple as well as complex objects and actions by using verbal description; while direct manipulation enables users to learn which objects and actions are available in the system, and offers the means to overcome hard speech problems involving temporality, ambiguity, and puts the user in control of the interaction.

B. The contrastive functionality

Speech and text modalities can be used in different ways to designate a shift in context or functionality. For example, direct manipulation is used for entering original input and real data, while speech is reserved for correction and issuing commands.

Direct Manipulation	Speech Recognition
Visible References	Non-visible References
Limited References	Multiple References
Simple Actions	Complex Actions

Table 2. Proposed applications for Direct Manipulation and Speech

By providing complementary modalities in both input and output communication channels, multimodal interaction is theoretically superior to speech-based interaction which solely relies on speech for communication . Other empirical studies have also confirmed the superiority of multimodality, in regard to flexibility (i.e., expressive power), usability, and efficiency. For example, the combination of speech and direct-manipulation interactions has been shown to achieve more reliable performance in map-based tasks than unimodal interaction, especially in mobile environment [16, 17], because of the mutual disambiguation of the two modalities. Also, multimodal error recovery has been shown to be faster than unimodal correction by re-speaking . More recently, multimodality has been examined in access to email messages on a cell phone and found to be preferable to users than unimodal interaction . Multimodal error correction has also been evaluated in a prototype multimodal dictation system [18]. The results showed that multimodal error correction is more accurate and faster than unimodal correction by re-speaking.

4.Conclusion

The proliferation of mobile technologies brings the “anytime, anywhere” computing fantasy to a reality. But real ubiquitous computing won’t be realized till the human computer interaction epitomized by the user experience becomes natural, quick, intuitive, and reliable. In this paper, we systematically analyzed two candidates for multimodal mobile interaction: speech-based and direct manipulation-based. It was shown that each modality, used separately, fails to offer tangible solution to user interface issues. We proposed true multimodality by offering the synergistic and simultaneous use of the strengths of speech and direct manipulation modalities. This approach avoids the pitfalls (weaknesses) associated with each modality by decreasing the FOA to provide an optimal experience and enhanced product usability.

5.References

[1] Klie, L., Holiday smartphone buyers want their money back. *Speech Technology eWeekly*, January 30, 2008

[2] Chomsky, N., *Language and Mind*. New York: Harcourt Brace Jovanovich. 1972

[3] Baker, M, C., *The Atoms of Language: The Mind’s Hidden Rules of Grammar*. New York: Basic Books. 2001

[4] Stewart, O.T, Dai, L., Lubensky, D. Understanding generic user preference: A comparative study of multimodal and speech-based interface for mobile devices.

Workshop on Speech in Mobile and Pervasive Environments (in conjunction with ACM Mobile HCI) September 12, 2006, Espoo, Finland

[5] Pinker, S., *The Language Instinct: How the Mind Creates Language*. New York: HarperPerennial. 1995

[6] Cohen, M., Giangola, J., Balogh, J. *Voice User Interface Design*. Boston: Addison-Wesley. 2003

[7] Stewart, O.T., Blanchard, H. E. Linguistics and psycholinguistics in IVR design. In Gardner-Bonneau, D and Blanchard H.E. (eds) *Human Factors and Voice Interactive Systems.* Second edition. New York: Springer. Pp 81-115, 2008

[8] Suhm, B., IVR usability engineering using guidelines and analyses of end-to-end calls. In Gardner-Bonneau, D and Blanchard H.E. (eds) *Human Factors and Voice Interactive Systems.* Second edition. New York: Springer. Pp1-41, 2008

[9] Soltau, H, Waibel, A. Acoustic models for hyperarticulated speech. International Conference on Speech and Language Processing (ICASSP), Beijing, China, 2000

[10] Shneiderman, B. Direct manipulation: a step beyond programming languages. *IEEE Computer,* 16, 8 57-69.

[11] Dai, L. Revisiting the speed accuracy tradeoff: A study of informal note taking using mobile information technologies. University of Maryland Baltimore County PhD dissertation, 2008

[12] Oviatt, S., Wahlster, W. Introduction to the special issue on multimodal interface. *Human-Computer Interaction,* 12, 1-5.

[13] Grasso, M. A., Ebert, D. S., and Finin, T. W. The integrality of speech in multimodal interfaces. *ACM Transactions on Computer-Human Interaction,* 5, 4 303-325.

[14] Cohen, P. R., *The role of natural language in a multimodal interface*, in *Annual ACM Symposium on User Interface Software and Technology (UIST)*. 1992.

[15] Cohen, P. R., Dalrymple, M., Moran, D. B., Pereira, F. C. N., Sullivan, J. W., Jr, R. A. G., Schlossberg, J. L., and Tyler, S. W., *Synergistic use of direct manipulation and natural language*, in *ACM Conference on Human Factors in Computing Systems (CHI)*. 1989.

[16] Oviatt, S., *Multimodal System Processing in Mobile Environments*, in *Annual ACM Symposium on User Interface Software and Technology (UIST)*. 2000: San Diego, CA.

[17] Oviatt, S. Multimodal Interfaces for Dynamic Interactive Maps. In *Proceedings of the ACM International Conference on Human Factors in Computing Systems (CHI)* 1996, 95-102.

[18] Suhm, B., Myers, B., and Waibel, A. Multimodal Error Correction for Speech User Interfaces. *ACM Transactions on Computer-Human Interaction,* 8, 1 60-98.

[19] Suhm, B., Freeman, B., and Getty, D., *Curing the menu blues in touch-tone voice interface*, in *ACM Conference on Human Factors in Computing Systems (CHI)*. 2001.

[20] Lai, J., *Facilitating mobile communication with multimodal access to email messages on a cell phone*, in *ACM Conference on Human Factors in Computing Systems (CHI)*. 2004: Vienna, Austria.