Improving User Acceptance of Voice Recognition Technology and Voice Interface (in a Mobile Device context) by Daphne J. Lee

Introduction

Since the early 90s, speech technology has advanced vastly in accuracy and speed, generating many successful voice recognition software tools such as Dragon's Natural Speaking or IBM's ViaVoice [2]. Yet despite continuous technological improvements and apparent opportunity gain, organizations and consumers are not widely adopting VRT [3]. Voice technology, for all its promoted benefits, seems to have little impact on the average user interaction and pales in comparison to other interface systems. Thus, this article seeks to ask why voice recognition technology has yet to be fully integrated into mainstream consumer products and what can be done to improve user acceptance of voice technology.

Background on Voice Technology

Voice recognition software has existed for almost two decades, with articles proclaiming its benefits since the late 80s. Many companies have sought to take advantage of VRT and current day forerunners are manufacturing speech software that claims to achieve 95% accuracy rates [2]. The most popular software is Dragon System's Naturally Speaking, a $699 priced continuous-diction product that converts user voice input into text while automatically adjusting to a person's unique speech pattern and idiolect. Other software companies include IBM with ViaVoice and Lernout & Hauspie, a supplier of voice diction software to companies such as Microsoft and Unisys [2].

In general, voice recognition technology is divided into two categories: command recognition and voice diction [2]. Both categories have certain benefits and shortcomings, which are listed in Figure 1. These shortcomings are discussed in fuller detailed in the Challenges section.

Command recognition involves simple single word instructions such as "open," "exit," or "close" and achieves higher accuracy rates then voice diction software. However, when used with alternate modes of input (e.g. mouse and keyboard), command instructions can be slower than simply clicking a button. Task that require spatial location are difficult to explain in words ("close top-left window as oppose to just 'close'" or "open word file TC401 to 'open'"), as clicking the appropriate location with

a mouse is more efficient. Thus, command recognition is seldom found on devices designed for the everyday user and is mostly installed onto systems where the user does not use a mouse or keyboard.

The second VRT system is voice diction. This system differs from command recognition because textual inputs are longer, multi-word sentences. More popular than command recognition, voice diction systems can handle complex commands and theoretically support conversational dialog. Commonly used for essay-writing tasks, voice diction is the opposite of silent writing. Speech has been shown to surpass writing in various stages of the composition task. Admittedly, voice diction is not superior in all aspects but certain stages, such as drafting and free writing, showed marked increase in efficiency and quality compared to silent writing [1]. When speaking, a person is unhindered by trivial grammatical or spelling issues, allowing for the creative flow in idea development. In addition, writers can express inner emotions and character dialog more fluently in speech rather than silent writing [1]. However, voice diction still suffers the pit-falls of phonetic ambiguity. Distinctions between homonyms (i.e. knight vs. night) or word sequences (i.e. the sky vs. this guy) are difficult and would counter any efficiency gain from voice diction with laborious copyediting.

Types of voice-recognition systems

	Command recognition	Voice dictation	
		Discrete	Continuous
Input	Users speak one word to input a command.	Users must input words slowly and carefully.	Users input words at a normal conversational pace.
Advantages	Very accurate	Suitable for dictation-style input; large vocabulary; modest processing requirements	Suitable for dictation-style input; large vocabulary; easy and natural to use
Disadvantages	Limited applications; not suitable for dictation-style input; small vocabulary	Limited accuracy; slow, cumbersome, and unnatural to use	Limited accuracy; formidable processing requirements

Figure 1. Summary of VRT systems from 'Voice-Based Interface Make PCs Better Listeners'

Opportunities in Mobile Technology

Voice Recognition Technology can be effective in many fields; this article focuses on opportunities in mobile device applications. To begin, speech technology offers new means of communicating through Voice Interface instead of keyboard or touch-screens. Moreover, linguists acknowledge speech as the most 'natural' form of communication because humans primarily communicate through voice. All people in the world have some form of spoken language, while only half contain a written version. We do not communicate through smell, scent, or touch but through auditory signals accompanied by supplementary visual gestures and expressions. Furthermore, many linguist support the **Inateness Hypothesis**, which claims that all "humans are genetically predisposed to learn and use language." [9] With speech playing such a key part to our fundamental means of communication, human interaction should prosper under a voice-driven interface.

Besides our preference to communicate with speech, voice interfaces will offer a viable alternative to preexisting technologies and their limitations. Monitor and keyboard input/output devices require users' eye and hand attention. Users cannot be looking elsewhere or manipulating something when using these interfaces. A most obvious situation would be driving wherein users ideally must be looking at the road and not checking email on their Smartphone. In general, voice interfaces excel in hand-busy and eye-busy situations [8], allowing users to interact with their mobile devices while performing some other task (i.e. driving, cooking, carrying children).

Other than multitasking, voice interfaces allow interaction with users who have little familiarity or are uncomfortable with using traditional input/output modals. New markets could be tapped because new users do not need to know how to read to use the mobile device. Illiterate users may not seem like the ideal persona for a high-tech voice recognition device but research done by Jan Chipchase in Zimbabwe have shown a surprising phenomenon: shared phone usage. True, a lone student in Zimbabwe could not sustain a cell phone but by "pooling" resources with other students, they could all purchase a single airtime and distribute it among themselves [10]. In addition, observation in Zimbabwe showed that the shared phones purchased were "relatively high-end mobile devices"—thus, it goes to show that technology can travel to unintended user groups [10].

Other user groups that benefits from voice technology are the disabled and elderly. Most mobile devices create some hassle with their tiny buttons and small screens [4]. Voice interfaces circumvent these problems for users who have poor or failing eyesight and motor-dexterity. These users would probably benefit most from voice interfaces with the greatest need and motivation to learn the technology. VRT requires strong

motivation because there are still limitations and problem associated with VRT that could cause errors and test a user's patience. These challenges are discussed in the next section.

Challenges of VRT

A general consensus exists within most research articles concerning voice technology and its applications: VRT is highly desirable among users and shows great potential in various mobile tasks; however, its benefits seem limited to paper concepts. In practical application, computer technology fails to achieve our expectations in satisfactory communication. As pointed out earlier, leading VRT software claims a 95% accuracy rating in its continuous-diction abilities. 95% may seem reasonable but consider that this means five words out of one hundred will be incorrect. This document thus far contains around 2000 words; using current VRT software, 100 words would be incorrect from my actual intent. Those 100 words would have a serious impact on my writing experience, requiring intensive copy-editing to find those 'incorrect' words (VRT never misspells a word, so manual reading is need to pick out the incorrect phrase "an ice man" and substitute the actual "a nice man").

Another challenge in voice technology is user expectation. People either have grown up or are familiar with voice commanded computers such as the famous Star Trek computer or HAL from Space Odyssey 2001. These were interfaces designed in the realm of Science Fiction, with an emphasis on fiction, but those were the concepts people were exposed to and understood. People already have a pre-conceived notion of how a voice interface experience should be: they talk in a normal, conversational speech pattern to the computer and it response in turn. However, current software cannot handle such varied idiolects. Most VRT software requires no background noise, clearly enunciated words [1], and full sentences to provide context. This is not how people talk normally as we slur our words, introduce pauses (err and urms), and we rarely formulate entire sentences before talking.

I do not claim that VRT must handle conversation speech because that could be quite difficult considering the numerous varieties of individual speech patterns; however, I point out people's pre-conceive notions may hamper VRT adoption. Learning a new interface is difficult, even using a mouse for the first time would need practice and patience. However, there were no precedents to a mouse interface, resulting in lower expectation to its performance and more forgiveness to mistakes. In a mouse interface, the user might take some blame ("Oh no! I clicked the wrong icon") as oppose to a voice interface ("No! No! Don't type 'see Mabel'! Type 'seem able'! Useless software!").

Why is it so difficult to capture the verbal language? Some linguist would claim that the reason is paralinguistic [6]. Also known as psycholinguistics, analog acoustic expression, prosodic properties, and emotion words; it is our emotion in voice. Describe best as "not what you say but how you say it" [7], paralinguistic information is often omitted by VRT software. Apparently, developers were more focused on the literal translation rather than the paralinguistic being said. Examples of prosodic properties of speech are intonation and rhythm. VRT diction software removes a user's pauses, their errs and ums; their intonation and their emotion.

Thus far, speech researchers have commonly attributed emotion in speech as a way of conveying internal states and not the semantic meaning of a word. However, research by Campell (2005) challenges that view: claiming that our analog acoustic expression significantly changes the meaning of a word [6].

Perhaps this could be a clue towards making VRT more acceptable. Current diction technology focuses more on the literal meaning of a sentence rather than the acoustic presentation of the sentence. Emotion in speech may not be just a vanity addition to VRT but an important component towards achieving conversational speech.

Future Research Areas

Voice technology has been around for some time, since the late 80s. So why has it not caught on like all other technologies or simply died out? VRT seems stuck in limbo, with few interactions with the average-user but still enough for continued VRT existence. People seem drawn to the notion of talking to your computer, most famous of all being the Star Trek computer or HAL from Space Odyssey 2001, but language complications seem to hamper that dream. However, there are still many untapped areas in VRT that could help improve user adoption.

1) Task requirements [5]: Continued research should be done, not on where voice technology would fit, but rather when do users use voice rather than text to complete a task. Instead of trying to implement voice technology into daily life, designers should aim to supplement and enrich pre-existing user tasks.

2) Conversational behavior: Further research concerning how comfortable users are with speaking to a mobile device/computer. Should the interaction follow conversational speech (as in talking to another human being) or should there be a distinction because it is a machine.

References

[1] Honeycutt, L. (2003). Researching the Use of Voice Recognition Writing Software. *Computers and Composition,* 20, 77–95

[2] Edwards, J. (1997). Voice-based interfaces make PCs better listeners. *Computer, 30*(8), 14-17

[3] Kamm, C. (1995). User Interface for Voice Applications. (Human-Machine Communication by Voice). In *Proceedings of the National Academy of Sciences of the United States, 92*, p10031(7). Retrieved March 04, 2007, from Expanded Academic ASAP via Thomson Gale: http://find.galegroup.com/itx/infomark.do?&contentSet=IAC-Documents&type=retrieve&tabID=T002&prodId=EAIM&docId=A18143034&source=gale&userGroupName=wash_main&version=1.0

[4] Mynatt, E.D., & Melenhorst, A.S., & Fisk, A.D., & Rogers. W.A. (2004). Aware Technologies for Aging in Place: Understanding User Needs and Attitudes. *IEEE Pervasive Computing, 3*, 36-41

[5] Burrell J., Brooke T., Beckwith R. (2004) Vineyard Computing: Sensor Networks in Agricultural Production. *IEEE Pervasive Computing*, 3(1), 38-45.

[6] Campbell, Nick. (2005). Getting to the heart of the matter: speech as the expression of affect; rather than just text or language. *Language Resources and Evaluation,* 39(1), 109-119.

[7] Shintel, H., Howard C. N., and Okrent, A. (2006) Analog acoustic expression in speech communication. *Journal of Memory and Language,* 55(2), (August 2006): 167-177.

[8] Nielsen, J. (2003). *Voice Interfaces: Assessing the Potential*. Retrieved January 30, 2007, From the World Wide Web: http://www.useit.com/alertbox/20030127.html

[9] Tserdanelis, G., & Wong, W.Y.P. (2004). *Language Files*. Ohio State University.

[10] Chipchase, J. (2006) *Shared Phone Use*. Retrieved March 14, 2007, From the World Wide Web: http://www.janchipchase.com/sharedphoneuse