Journal of Usability Studies
Vol. 12, Issue 4, August 2017 pp. 155–163

Usability Testing of Spoken Conversational Systems

Susan L. Hura
CEO
discourse.ai
susan@discourse.ai

The philosophy behind usability testing for speech-enabled systems is shared with general usability practices, but many usability practitioners have little or no experience testing speech interfaces, and the specific techniques required for collecting valid and reliable data are not widely understood. Spoken language and conversation have a number of properties that should influence the methods used to test speech user interfaces. Many excellent books exist on how to design effective speech user interfaces (Balentine & Morgan, 2001; Cohen, Giangola, & Balogh, 2004; Harris, 2005; Lewis, 2011), and these books offer some how-to information on testing voice user interfaces. However, the missing element is a detailed exploration of why speech interactions require alternate testing methods. The purpose of this editorial is to share practical experience, introduce concerns particular to speech, and point out potential issues.

Speech, Voice, and Conversation

The terms speech, voice, speech-enabled, and conversational are used interchangeably and inconsistently. There are nuances to the terminology that signals the kind of technology involved and the kinds of user interactions that are supported.

Speech technology itself is not new and commercial products with speech user interfaces have been widely available for many years (Pieraccini, 2012). The biggest challenge in the past has been accurate decoding from the acoustic signal (that is, the user's spoken input), and therefore the capabilities of speech systems have been quite limited. You have likely experienced examples of the earlier cohort of speech technologies in the form of dictation programs, interactive voice responses (IVR), and speech-enabled automotive systems. These technologies have been widespread since the late 1990s but tended to perform poorly for a variety of reasons (which are beyond the scope of this paper). Speech technology itself was often blamed for the poor experiences users had, but the failure was as much due to poor design practices and lack of user feedback mechanisms as to the recognition algorithms themselves. Also complicit in the failure were organizations that deployed speech-enabled IVRs as a blockade to keep users away from human customer service representatives, which led to interactions in which users rightfully saw themselves as victims of automated speech technology systems (Attwater, Edgington, Durston & Whittaker, 2000).

A new conversational ecosystem is emerging today, embodied by dedicated devices such as Amazon's Alexa and Google Home. The public was introduced to the new generation of speech-enabled systems by Apple's Siri in 2011, which was arguably the least reviled speech system in wide use (Pieraccini, 2012). A host of other intelligent conversational assistants are emerging at a rapid pace today including IBM's Watson (Markoff, 2011), Microsoft's Cortana (Warren, 2014), Apple's Homepod (Pierce, 2017), and Baidu's Duer (Gibbs, 2017). Within the past decade, a combination of factors including increased computing power, huge amounts of training data, and application of highly skilled resources by large corporations have improved speech recognition to near human levels of performance (Protalinski, 2017), which has in turn ushered in the beginning of a wave of new speech user interfaces.

These systems are moving into the previously unattainable realm of natural language processing and are marked by user-initiated interactions. Users are volunteering to use these systems rather than being forced into it (Attwater et al., 2000). Both Google and Amazon have given third party organizations the ability to create custom speech-enabled applications for evolving conversational ecosystems without many of the constraints that hampered previous generations of speech systems.

Spoken Language as a Modality

Speech user interfaces use spoken language for both user input and output[1]. The system presents content to the user via recorded speech prompts, text-to-speech, or a combination of the two. Users speak to interact with the system. Their spoken input is collected via microphone and undergoes signal processing algorithms that improve the quality of the signal for the purposes of speech recognition (end-point detection, echo and noise cancellation). Speech recognition algorithms decode user input to text, which can be used directly to determine system response in some systems or may be subject to further natural language processing to extract user intent and other linguistic details. Users process the system's responses via the same auditory processing channel used for other spoken interactions.

Speech interactions thus differ in obvious ways from graphical user interfaces in which the system presents information visually via text and images and user input is provided by typing, tapping, or mouse clicks. However, there are also different emergent properties to voice and visual systems that are worth noting.

[1] Many systems combine speech with other modalities. These multimodal systems are complex, and the interaction between modalities is not always intuitive from unimodal perspective (Dahl, 2017).

Speech interfaces are

- **sequential** in that speech must be produced one word at a time,
- inherently **dynamic** because sequentially-produce speech is constantly changing, and
- **transient** because speaking leaves no permanent and available record of what was said.

In contrast, graphical interfaces are

- **simultaneous** in that a great deal of information can be presented at once,
- more **static** because text and images do not need to change continuously, and
- **permanent** because graphical interfaces allow users to see current information and refer back to previously-presented information.

These emergent properties suggest that some tasks are inherently better suited to visual interfaces than to voice, and vice versa (Novick, Hansen, Sutton, & Marshall, 1999). More relevant for this paper are the ways in which the inherent characteristics of speech interfaces necessarily influence the choice of usability testing method. Certain common usability testing techniques must be modified, or sometimes abandoned, when testing speech interfaces.

Implications for Testing

An obvious case that fails for speech interactions are think-aloud techniques. Users simply cannot speak to describe their reactions and opinions to a spoken interaction while that interaction is in progress. It is tempting to consider solutions that would pause the human-computer conversation in order to allow the user to provide feedback. However, the sequential and transient nature of speech means that users may have difficulty picking up the thread of the conversation after a pause. Asking the user to pause mid-task in a graphical user interface will not necessarily change the nature of the interaction because visual output from the system is relatively permanent and static. If the user forgets exactly what he was doing while speaking to the researcher during a pause, he can simply look up at the screen again and quickly remember his place. Because spoken interactions don't offer the same ability to refer back to the previous interactions, any usability testing technique for a speech system must allow the user to complete some defined chunk of the interaction before interrupting.

One technique I have used successfully is retrospective think aloud, in which the researcher takes careful notes during the interaction and uses this to interview the user after the fact. This method requires the researcher to pay close attention to user responses (informed by knowledge about how the system works) and specifically aim to appear curious rather than judgmental. Some researchers record the user's spoken interactions with the system and play these recordings to the user to facilitate the retrospective think aloud. In my experience, this can be unsettling to some users because it seems to highlight the fact that they are being observed and may make them believe they are being judged for their participation in the interaction. I suspect the choice between these two methods of retrospective think aloud depend a great deal on the personality of the researcher and the test participant. Both seem equally valid and follow the general rule of allowing users to complete the interaction before soliciting their reactions and opinions.

A Long Aside About Early Prototyping

For graphical user interfaces, usability researchers rely on wireframes or paper prototyping to collect early user feedback before development and without the final visual design elements (colors, fonts, images). These methods offer a way to test the baseline interaction minus the critical, but separable, look and feel of the final application.

For speech interfaces, Wizard of Oz (WOZ) techniques occupy the same position in a timeline but are quite different because the presentation layer and interaction layers are more tightly bound for spoken interactions. The presentation layer elements in voice user interfaces (VUIs) include the voice of the system, word choice, speaking rate, and sentence structure, which

together represent the personality of the system[2]. It is not possible to isolate the interaction elements (i.e., the dialog of the system, the back and forth questions and responses) from the presentation layer elements that characterize the personality.

To state the obvious: Even early prototypes of speech interactions must include dialog spoken by some voice. Any voice is perceived by users as the voice of the system, and any spoken dialog forms the basis for developing theory of mind for their interlocutor. I have argued that there is no such thing as a system with no personality (Hura, 2008) and that attempts to create a system without a personality result in systems that are perceived as robotic or incoherent. In a conversation, users attribute characteristics and form impressions of the capabilities of the interlocutor (Nass & Brave, 2005). The research on the human tendency to anthropomorphize computerized interactions has come under question, but the broad point that we interact with speech systems as an interlocutor is not controversial.

Implications for Testing

I am not arguing against Wizard of Oz testing; it is a good tool for collecting early feedback on speech systems prior to development (Sadowski, 2001). However, there are several caveats that researchers must be cognizant of.

To conduct WOZ testing for a speech system, the interaction model and specific paths for task completion must be complete (just like GUI). For speech systems, this means that the voice prompts must be fully scripted to realistically represent the personality of the system. Hastily written voice prompts have a greater impact on baseline usability because they are more transient and sequential than a hastily sketched wireframe. If the placement or label of a visual element is not entirely clear, the user can study it at length and refer to the rest of the visual interface to try to understand how to interact with it, and then continue with his task. A hastily scripted voice prompt lacks this permanence and so is much more likely to throw the user into error recovery states which often lead to the inability to complete the task.

I strongly recommend testing with the voice that will be used in the production application. Vocal qualities like pitch, intonation, pace, and pause duration could arguably be characterized as the presentation layer of a voice. Every voice has these qualities, and they affect not only the sound and feel of a speech interaction but the user's ability to understand and therefore interact with the system. Listeners extract meaning from extra-linguistic elements like intonation and pausing, which leads to the conclusion that the voice of a speech system is inseparable from the interaction. If a system will use a computerized (text to speech or TTS) voice, use that voice in early testing. (Note that substituting one computerized voice for another likely has the same impact as substituting one human voice for another.)

Whether or not the production voice is available for prototype testing, favor recorded prompts over reading the prompts aloud during test sessions. As argued above, extra-linguistic factors like intonation, pace, and pause duration have a significant impact on the interaction itself. Sadowski (2001) cautioned that subtle changes in, "voice intonation and human-like reactions," can influence user responses during live WOZ testing. I agree with Sadowski's cautionary note but suggest that extra-linguistic factors will undoubtedly change from one instance of a spoken prompt to another, and that these changes have the potential to affect user perceptions and performance. In short, reading prompts live during WOZ testing introduces a source of variability that is similar to what would result from a paper prototype test in which the researcher sketches the interface anew for each participant. As a preliminary ideation technique among colleagues, a back-of-the-napkin sketch offers some value to the researcher, but this is clearly a very rough approximation of the production application. Reading prompts live in WOZ testing delivers a similarly rough approximation of a speech interface, so one's results should be interpreted with the same degree of caution.

The wide availability of free audio recording and editing software makes recording prompts a trivial investment of time and effort. With recorded prompts, the researcher knows that each user will hear the prompt produced with identical the tone, intonation, and timing—thus

---

[2] The personality or characteristics of a speech system are often referred to as persona. I avoid using the word persona this way (in this paper and in general) because persona has a different meaning in user research outside the speech community.

removing a source of variability from the results. Using recorded prompts makes WOZ more akin to the simple wireframes testing, in that the representation of the system remains consistent unless the researcher deliberately changes them.

Speech Is Natural

Beyond the difference attributed to modality of spoken[3] versus visual interfaces, there are inherent differences that relate to speech itself. Specifically, I would like to address the common claim that speech offers a more "natural" mode of interaction than alternative visual-manual interactions (like reading, typing, and tapping). The idea is that the "naturalness" of speech automatically enables UIs to be more intuitive and comfortable than visual-manual interactions and therefore easier to use. I find this use of the term "natural" more divisive than useful and disagree with the conclusion that speech interfaces are inherently more natural.

Some claims about the naturalness of speech revolve around how ubiquitous it is. Spoken language is clearly a highly overlearned behavior: Overlearned behaviors become automatic and occur without conscious volition on the part of the individual (MacKay-Brandt, 2011). Spoken language becomes automatic for most preschool-aged children so much so that it becomes impossible for individuals not to comprehend speech presented to them in their native language. We tend to be surrounded by spoken language and participate in it nearly continuously throughout our lives. The ubiquitous nature of speech has only increased in the digital age when many of us choose to consume content that is rich in spoken language via radio, television, and online content like videos and podcasts. However, there is little distinction to be drawn between spoken and written language on these grounds. Literate individuals are also surrounded by huge amounts of written language and have an increasing number of digital platforms through which to consume and produce written language (social media sites, blogs, texting).

One way in which spoken and written language clearly differ is in the mechanism by which each is acquired. Spoken (or signed) language is acquired innately, without any instruction other than simple interaction with caregivers. The human auditory system is fully developed in the third trimester of pregnancy, and there is a growing body of evidence that infants are acquiring specific linguistic knowledge before birth. For example, infants are born recognizing their mother's voice (DeCaspar & Fifer, 1980) and specific words presented to them in utero (Partanen et al., 2013). The implication is that infants have significant receptive knowledge of spoken language well before they are physically capable of the motor control required to produce speech, a process that develops over the first two or three years of life. In contrast, individuals must be explicitly taught to read, and full competence in reading and writing tends to require a longer period of development.

The claim that speech is natural also ignores the common use of communicative gestures that serve the same pragmatic functions as spoken language (Clark, 2004). For instance, a speaker may offer an item to the listener by saying, "Would you like one?" while extending a plate of cookies. A natural response from the listener who wants a cookie would be to extend his hand, to say "yes," or both. In fact, the gestural behavior may be even more natural than speech in this case because an interchange like this could easily occur between individuals who do not share a language, hinting that this is a social behavior that predates language itself.

Furthermore, even if we concede that the use of spoken language in human-human interactions may be conceived of as natural because it is innate, spoken human-computer interactions should not be deemed natural by association. There is evidence that people do not speak to computers in the same way they speak to people (Hauptmann & Rudnicky, 1988), and computers simply do not behave like real humans in conversation at this point in history. The limitations present in every existing speech system (including the new generation of conversational applications) require users to modify their beliefs and expectations about how a

[3] In this section, many of the points I raise about spoken language are also true for signed languages. Like spoken language, signed languages tend to be overlearned, omnipresent, innately required, and governed by unconscious rules.

conversation works. These modifications effectively remove spoken interactions with computers from the category of conversation in general.

Implications for Testing

The most significant implication of the purported naturalness of speech is the tendency for naturalness to be used as an argument against strong user-centered design practices. If speech is natural, and pretty much all of us know how to speak, then why do we need to spend time on design and user testing? This specious line of argumentation is commonly heard at organizations considering new speech applications and is as demonstrably false as a claim that knowing business processes or understanding marketing initiatives qualifies a person to design a company's website.

Another version of this argument combines the speech-is-natural idea with expertise of visual interaction design. If I'm already a designer, why can't I just design this speech application? Expertise in interaction design is hugely beneficial to designing good speech interfaces because of the shared user-centered design philosophy. However, expertise in design is almost exclusively visual today, meaning it is not sufficient, irrespective of how natural speech may be.

Conversational Interactions Are Different

Spoken language interactions between humans do not typically occur as single-turn interactions (one person asks, the other answers) but are embedded in larger conversations. Conversations are a particular instance of spoken interaction "involving multiple participants, shared knowledge, and a protocol for taking turns and providing mutual feedback" (Schmandt, 1994, p. 6). Conversation is a social manifestation of language; we don't converse simply to exchange information but to perform social actions as well (van Dijk, 1997). In conversation, we make judgments about and form a theory of mind of our interlocutors (Wilde Astington & Baird, 2005).

In fact, conversation may be among the methods young children use to bootstrap their understanding of others and themselves (de Rosnay & Hughes, 2006). Extra-linguistic elements of conversation such as turn-taking are present in newborns (Rutter & Durkin, 1987) and are a foundational part of mother-child bonding (DeCaspar & Fifer, 1980). Anyone who has conversed with an infant too young to produce actual speech can attest to the fact that conversation can happen without it.

Finally, I will turn to Schmandt's point about conversations requiring a protocol for taking turns and providing mutual feedback. The term protocol does not convey the scope of the set of rules that govern conversations. These rules are not intuitively obvious to users and are not explicitly taught, yet they shape every spoken interaction. Philosopher Paul Grice was the first to attempt to codify the rules that govern conversation (1975). He described an overarching *cooperative principle* that describes how speakers and listeners typically behave with mutual cooperation to achieve effective communication. Individuals who obey the cooperative principle speak in a way that furthers the purpose of the conversation and allows the listener to make inferences about what was said.

For example, if a speaker asks, "Do you have the time?" a perfectly logical response would be "yes," but this is a distinctly uncooperative reply. The speaker's purpose in asking "Do you have the time?" is not to find out whether the listener in fact knows what time it is. Instead, the speaker is requesting that if the listener does indeed know the time, that he please share that information with the speaker. This example illustrates two important points about the cooperative principle: first, that you have the ability to judge how cooperative a response is, and second, that you were unaware of this ability until you came upon this uncooperative response. This is typical of individuals in conversation; we implicitly know how to behave cooperatively in conversation, but the rules of conversation are so ingrained that we are largely unaware of that knowledge until we come upon an example that violates a rule.

Implications for Testing

As researchers, we cannot ask participants whether a speech system is behaving cooperatively because they lack conscious awareness of the rules of conversation. Test participants are typically good at spotting elements that fail to follow the rules, but they rarely describe them as such.

Instead, participants tend to describe broken conversational experiences in terms of politeness or personality attributed to the system. If a user asked, "Do you have the time?" and the system responded "yes," test participants would know that this was not a good response. However, the reasons they give are likely to describe the system as "rude" or say that they didn't like "her." Some participants may be able to explain that the system seems rude because she failed to answer the question, but even these individuals are not consciously aware of the underlying conversational principles.

Moving Forward

If conversational interfaces have half the impact that futurists and science fiction writers predict, many more people will need to be versed in testing speech systems. Voice interaction design has been a niche profession, and there are simply not enough qualified individuals with experience in speech. My plea to usability researchers and interaction designers who have only worked on graphical user interfaces is to educate yourselves on spoken language and speech systems before jumping in with both feet. The best chance to create great speech experiences is to cultivate as many informed professionals as possible. I encourage you to learn what you can about design and testing for speech technologies and conversational systems through resources like design guidelines published by the Association for Voice Interaction Design (www.avixd.org) and numerous online groups and in-person meetups related to speech technologies.

References

Attwater D., Edgington M., Durston P., & Whittaker S. (2000). Practical issues in the application of speech technology to network and customer services applications. *Speech Communication, 31*(4), 279–291.

Balentine, B., & Morgan, D. P. (2001). *How to build a speech recognition application: A style guide for telephony dialogues* (2nd ed.). San Ramon, CA: EIG Press.

Clark, H. H. (2004). Pragmatics of language performance. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 365–382). Oxford, UK: Blackwell.

Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Boston, MA: Addison-Wesley.

Dahl, D., Ed. (2017) *Multimodal interaction with W3c standard*. Switzerland: Springer International.

DeCaspar, A., & Fifer, W. P. (1980) Of human bonding: Newborns prefer their mothers' voices. *Science*. 1174–1176.

de Rosnay, M. & Hughes, C. (2006). Conversation and theory of mind: Do children talk their way to socio-cognitive understanding? *Developmental Psychology, 24*(1). 7–37.

Gibbs, S. (2017, June) *Baidu launches Duer digital assistant to take on Siri, Cortana and Google Now*. The Guardian. Retrieved from https://www.theguardian.com/technology/2015/sep/09/baidu-duer-digital-assistant-siri-cortana-google-now

Grice, Paul (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Harris, R. A. (2005). *Voice interaction design: Crafting the new conversational speech systems*. San Francisco, CA: Morgan Kaufmann.

Hauptmann, A., & Rudnicky, A. (1988). Talking to computers: An empirical investigation. *International Journal of Man-Machine Studies, 28*, 583–604.

Hura, S. (2008). Voice user interfaces. In P. Kortum (Ed.), *HCI beyond the GUI: Design for haptic, speech, olfactory, and other nontraditional interfaces* (pp. 197–227). Burlington, MA: Morgan Kaufmann.

Lewis, J. R. (2011). *Practical speech user interface design*. Boca Raton, FL: CRC Press, Taylor & Francis Group.

MacKay-Brandt, A. (2011) Automaticity. In J. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (pp.328–328). New York, NY: Springer-Verlag.

Markoff, J. (2011) *Computer wins on 'Jeopardy!': Trivial, it's not*. The New York Times. Retrieved from http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html?pagewanted=all

Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. Cambridge, MA: MIT Press.

Novick, D. G., Hansen, B., Sutton, S., & Marshall, C. R. (1999). Limiting factors of automated telephone dialogues. In D. Gardner-Bonneau (Ed.), *Human factors and voice interactive systems* (pp. 163–186). Boston, MA: Kluwer Academic Publishers.

Partanen, E., Kujala, T., Näätänen, R., Liitola, A., Sambeth, A., & Huotilainen, M. (2013). Learning-induced neural plasticity of speech processing before birth. Proceedings of the National Academy of Sciences, 110(37), 15145-15150.

Pieraccini, R. (2012). *The voice in the machine*. Cambridge, MA: MIT Press.

Pierce, D. (2017, May) *Apple's Homepod puts Siri in a speaker*. Wired. Retrieved from https://www.wired.com/2017/06/apple-homepod/

Protalinski, E. (2017, May). *Google's speech recognition technology now has a 4.9% word error rate*. VentureBeat. Retrieved from https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/

Rutter, D. R., & Durkin, K. (1987). Turn-taking in mother–infant interaction: An examination of vocalizations and gaze. *Developmental Psychology, 23*(1), 54–61.

Sadowski, W. (2001). Capabilities and limitations of Wizard of Oz evaluations of speech user interfaces. In *Proceedings of HCI International 2001: Usability evaluation and interface design* (pp. 139–142). Mahwah, NJ: Lawrence Erlbaum.

Schmandt, C. (1994) *Voice communication with computers*. New York, NY: Van Nostrand Reinhold.

van Dijk, T. (1997) Discourse as interaction in society. In T. van Dijk (Ed.) *Discourse as social interaction* (Vol. 2, pp. 1–37). London, UK: Sage.

Warren, T. (2014, April) *The story of Cortana, Microsoft's Siri killer*. The Verge. Retrieved from https://www.theverge.com/2014/4/2/5570866/cortana-windows-phone-8-1-digital-assistant

Wilde Astington, J., & Baird, J. (2005). Introduction: Why language matters. In J. Wilde Astington, & J. Baird (Eds.) *Why language matters for theory of mind* (3-25). New York, NY: Oxford University Press.

About the Author

Susan L. Hura, PhD
Dr. Hura is CEO of discourse.ai, which provides a platform to create intelligent customer conversations. Her consultancy, SpeechUsability, focuses on collecting user feedback to deliver intuitive and appealing voice interactions. Susan served as co-chair of the SpeechTEK conference 2008-2016, and she is Past President of the Association of Voice Interaction Design.