Designing Speech and Language Interactions

Cosmin Munteanu
National Research Council Canada and
University of Toronto
cosmin.munteanu@nrc-cnrc.gc.ca

Matt Jones
Swansea University
matt.jones@swansea.ac.uk

Steve Whittaker
University of California at Santa Cruz
swhittak@ucsc.edu

Sharon Oviatt
Incaa Designs
oviatt@incaadesigns.org

Mathhew Aylett
CereProc
matthewa@cereproc.com

Gerald Penn
University of Toronto
gpenn@cs.toronto.edu

Stephen Brewster
University of Glasgow
stephen.brewster@glasgow.ac.uk

Nicolas d'Alessandro
University of Mons
nda@numediart.org

ACM CHI 2014 April 26, 2014, Toronto, Canada.

Abstract

Speech and natural language remain our most natural forms of interaction; yet the HCI community have been very timid about focusing their attention on designing and developing spoken language interaction techniques. While significant efforts are spent and progress made in speech recognition, synthesis, and natural language processing, there is now sufficient evidence that many real-life applications using speech technologies do not require 100% accuracy to be useful. This is particularly true if such systems are designed with complementary modalities that better support their users or enhance the systems' usability. Engaging the CHI community now is timely – many recent commercial applications, especially in the mobile space, are already tapping the increased interest in and need for natural user interfaces (NUIs) by enabling speech interaction in their products. This multidisciplinary, one-day workshop will bring together interaction designers, usability researchers, and general HCI practitioners to analyze the opportunities and directions to take in designing more natural interactions based on spoken language, and to look at how we can leverage recent advances in speech processing in order to gain widespread acceptance of speech and natural language interaction.

Keywords
Speech and Language Interaction; Automatic Speech Recognition; Natural Language Processing; Natural User Interfaces.

ACM Classification Keywords
H.5.2 [User interfaces]: Voice I/O, Natural language, User-centered design, and Evaluation/methodology.

Introduction and Motivation

Our senses, such as touch, sight, hearing, or speech, allow us to interact with objects, information, or other humans. While such interactions are only slightly altered by technological progress, digital technologies are now reshaping the way we interact with our environment. We are no longer in direct control over such interactions; instead, we project them through a virtual layer. During the past decade we have witnessed dramatic changes in the way people access information and store knowledge, mainly due to the ubiquity of mobile and pervasive computing and affordable broadband Internet. Such recent developments have presented us the opportunities to reclaim naturalness as a central theme for interaction. We have seen this happen with touch for mobile computing; it is now time to see this for speech as well.

Unfortunately, humans' most natural forms of communication, speech and language, are also among the most difficult modalities for machines – despite, and perhaps, because these are the highest-bandwidth communication channels we have. While significant efforts, from engineering, linguistic, and cognitive sciences, have been spent on improving machines' ability to understand speech and natural language, these have often been neglected as interaction modalities, mainly due to the usability challenges arising from their inherently high error rates and computational complexity.

The challenges in enabling such natural interactions have often led to these modalities being considered, at best, error-prone alternatives to "traditional" input or output mechanisms. However, this should not be a reason to abandon speech interaction[1] – in fact, people are now exposed to many more situations in which they need to interact hands- and eyes-free with a computing device. Furthermore, achieving perfectly accurate speech processing is a lofty goal that is often nothing short of a fairy tale – a system that scores 100% in accuracy against an arbitrary standard such as a manual transcript is not guaranteed to be useful or usable for its users. There is significant research evidence pointing to the fact that proper interaction design can complement speech processing in ways that compensate for its less-than-perfect accuracy (Oviatt, 2003, and Munteanu, 2006), or that in many tasks where users interact with spoken information, verbatim transcription of speech is not relevant at all (Penn and Zhu, 2008).

Recent commercial applications (e.g. personal digital assistants) have brought renewed attention to speech-based interaction. However, as illustrated by the reviews and opinion pieces in popular media (Gizmodo, 2011), such technologies are receiving mixed reviews, often due to unexpected low or inconsistent accuracy for some tasks – "voice recognition isn't that good", or

[1] Throughout the rest of this document we will use the term speech and speech interaction to denote both verbal and text-based interaction, where the textual representation has been obtained from its original spoken source.

due to perceived lack of usefulness – "more of a gimmick than a useful tool" (Business Insider, 2012). This can be in part attributed to speech being marketed as an input/output modality, while in fact speech can assist with a wider range of tasks that are not limited to direct interactions.

Figure 1: Popular media view of speech-enabled mobile personal assistants (Gizmodo, 2011)

If we are to pick just one example where such research is desperately needed, it would be the area of access to multimedia repositories. 72 hours of video are uploaded to Youtube each minute (Youtube, 2012). At this rate, it is humanly not possible to consume the amount of data being generated, and it is becoming increasingly difficult to search for information or navigate through such large and often multilingual collections. Technologies that assist with such tasks include summarization of text or audio/video documents, browsing/searching through and indexing of large multimedia repositories, secure user authentication, natural language generation, speech synthesis, or speech-to-speech machine translation. Unfortunately, there is very little HCI research on how to leverage the engineering progress being made in these areas into developing more natural, effective, or accessible user interfaces.

Goals

In light of such barriers and opportunities, this workshop aims to foster an interdisciplinary dialogue and create momentum for increased research and collaboration in:

- Formally framing the challenges to the widespread adoption of speech and natural language interaction,
- Taking concrete steps toward developing a framework of user-centric design guidelines for speech- and language-based interactive systems, grounded in good usability practices, and
- Establishing directions to take and identifying further research opportunities in designing more natural interactions that make use of speech and natural language.

Topics

We are proposing to build upon the discussions started during our lively-debated and highly-engaging panel on speech interaction that was held at CHI 2013 [5]. As a natural follow-up to the significant interest that was generated by the panel, we propose several topics for discussions and activity among workshop participants:

- What are the important challenges in using speech as a "mainstream" modality?
- What opportunities are presented by the rapidly evolving mobile and pervasive computing areas?

- Given the penetration of mobile computing in emerging markets, are there any specific usability or technology adoption issues surrounding speech interaction?
- What opportunities exist to improve users experiences by using speech-based technologies that are not limited to input or output?
- Can we broadly characterize which interfaces/applications speech is suitable for?
- What can the CHI community learn from the Automatic Speech Recognition (ASR) and the Natural Language Processing (NLP) research, and in turn, how can it help the ASR and NLP communities improve the user-acceptance of such technologies? For example, the speech research community is mainly driven by engineering puzzle-solving – what else should we be asking them to extract from speech beside words/segments?
- How can we bridge the divide between the evaluation methods used in HCI and those in speech processing (which are mostly based on Artificial Intelligence practice)? ASR practitioners prefer to improve systems with respect to concrete metrics – can the CHI community help propose meaningful alternative measures of ASR quality that are still machine-implementable?
- How can speech be combined with other modalities to increase usability and robustness of interfaces?
- What are the contexts (commercial, literacy support, assistive technology, etc.) in which we can expect to see spoken language processing expand the most in the future?
- Shouldn't speech be more expressive as well as natural? How/in which contexts can we avail ourselves of that expressiveness (e.g. can you sing your search query?) Speech and pointing/deixis are a natural combination – what else can be combined with speech to make it more expressive/natural?
- What are the usability challenges of synthetic speech? How can expressiveness and naturalness be incorporated into interface design guidelines, particularly in mobile contexts where text-to-speech could potentially play a significant role in users' experiences?

References

[1] Business Insider (2012). Frankly, It's Concerning that Apple is Still Advertising A Product as Flawed as Siri. http://www.businessinsider.com, 2012.

[2] Gizmodo (2011). Siri is Apple's Broken Promise. http://www.gizmodo.com, 2011.

[3] Munteanu, C. et al. (2006). Automatic speech recognition for webcasts: how good is good enough and what to do when it isn't. Proc. of ICMI.

[4] Munteanu, C. and Penn, G. (2013). Speech-based interaction. Course, ACM SIGCHI 2011, 2012, 2013.

[5] Munteanu, C. et al. (2013). We need to talk: HCI and the delicate topic of speech-based interaction. Panel, ACM SIGCHI 2013.

[6] Oviatt, S. (2003). Advances in Robust Multimodal Interface Design. IEEE Comput. Graph. Appl. 23-5.

[7] Penn, G. and Zhu, X. (2008). A critical reassessment of evaluation baselines for speech summarization. In Proc. of ACL-HLT.

[8] Youtube (2012). Upload statistics. http://www.youtube.com/t/press_statistics