Speech-based Interaction: Myths, Challenges, and Opportunities

Cosmin Munteanu
Institute of Communication, Culture, Information,
and Technology
University of Toronto Mississauga
cosmin.munteanu@utoronto.ca

Gerald Penn
Department of Computer Science
University of Toronto
gpenn@cs.toronto.edu

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3027117

Abstract

HCI research has for long been dedicated to better and more naturally facilitating information transfer between humans and machines. Unfortunately, humans' most natural form of communication, speech, is also one of the most difficult modalities to be understood by machines – despite, and perhaps, because it is the highest-bandwidth communication channel we possess. While significant research efforts, from engineering, to linguistic, and to cognitive sciences, have been spent on improving machines' ability to understand speech, the CHI community (and the HCI field at large) has been relatively timid in embracing this modality as a central focus of research. This can be attributed in part to the unexpected variations in error rates when processing speech, in contrast with often-unfounded claims of success from industry, but also to the intrinsic difficulty of designing and especially evaluating speech and natural language interfaces. As such, the development of interactive speech-based systems is mostly driven by engineering efforts to improve such systems with respect to largely arbitrary performance metrics. Such developments have often been void of any user-centered design principles or consideration for usability or usefulness.

The goal of this course is to inform the CHI community of the current state of speech and natural language research, to dispel some of the myths surrounding

speech-based interaction, as well as to provide an opportunity for researchers and practitioners to learn more about how speech recognition and speech synthesis work, what are their limitations, and how they could be used to enhance current interaction paradigms. Through this, we hope that HCI researchers and practitioners will learn how to combine recent advances in speech processing with user-centred principles in designing more usable and useful speech-based interactive systems.

ACM Classification Keywords

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

Authors’ Keywords

Designing speech-based interactions; Automatic speech recognition; Text-to-speech.

Instructors' Bio

Cosmin Munteanu

Institute for Communication, Culture, Information, and Technology, University of Toronto Mississauga, and Technologies for Ageing Gracefully Lab (TAGlab), University of Toronto
http://cosmin.taglab.ca

Cosmin Munteanu is an Assistant Professor at the Institute for Communication, Culture, Information, and Technology (University of Toronto at Mississauga), and Associate Director of the Technologies for Ageing Gracefully lab. Until 2014 he was a Research Officer with the National Research Council of Canada, where he designed mobile and immersive natural user interfaces for a wide range of applications. His area of expertise is at the intersection of Human-Computer Interaction, Automatic Speech Recognition, Natural Language Processing, Mobile Computing, and Assistive Technologies. He has extensively studied the human factors of using imperfect speech recognition systems, and has designed and evaluated systems that improve humans' access to and interaction with information-rich media and technologies through natural language. Cosmin's multidisciplinary interests include speech and natural language interaction for mobile devices, mixed reality systems, learning technologies for marginalized users, assistive technologies for older adults, and ethics in human-computer interaction research.

Gerald Penn

Department of Computer Science, University of Toronto
http://www.cs.toronto.edu/~gpenn/

Gerald Penn is a Professor of Computer Science at the University of Toronto. His area of expertise is in the study of human languages, both from a mathematical and computational perspective. Gerald is one of the leading scholars in Computational Linguistics, with significant contributions to the formal study of natural languages. His publications cover many areas, from Theoretical Linguistics, to Mathematics, and to Automatic Speech Recognition, as well as Human-Computer Interaction.

Outline and learning objectives

- How Automatic Speech Recognition (ASR) and Speech Synthesis (or Text-To-Speech – TTS) work and why these are such computationally-difficult problems
- Where are ASR and TTS used in current commercial interactive applications

- What are the usability issues surrounding speech-based interaction systems, particularly in mobile and pervasive computing
- What are the challenges in enabling speech as a modality for mobile interaction
- What is the current state-of-the-art in ASR and TTS research
- What are the differences between the commercial ASR systems' accuracy claims and the needs of mobile interactive applications
- What are the difficulties in evaluating the quality of TTS systems, particularly from a usability and user perspective
- What opportunities exist for HCI researchers in terms of enhancing systems' interactivity by enabling speech

Course history

This course proposal is based on conference tutorials and courses accepted for presentation at:

- Cosmin Munteanu and Gerald Penn: "Speech-based Interaction: Myths, Challenges, and Opportunities", course **presented at the CHI Conference**, 2011, 2012, 2013, 2014, 2015, and 2016.
- Cosmin Munteanu and Gerald Penn: "Hands-free Interfaces: The Myths, Challenges, and Opportunities of Speech-based Interaction", tutorial presented at the MobileHCI Conference, 2010, 2011, 2012, 2013, and 2014.
- Cosmin Munteanu and Gerald Penn: "Natural Interaction for Serious Games: Enhancing Training Simulators through Automatic Speech Recognition", tutorial presented at the I/ITSEC Conference, 2010, 2011, 2012, 2013, 2014, 2015, and 2016.

Course material and content

This CHI presentation is continuously updated to include more examples and analysis of recent commercial adoption of speech recognition, notably developments on mobile platforms (such as Apple's Siri, Google's Voice Search, or Microsoft's Cortana) or to smart home personal assistants (e.g Amazon's Echo). Similar to previous courses, and based on the feedback received from participants, the proposed CHI course material is currently being updated with new interactive examples and short group exercises.

The aim of the presentation is two-fold: present new concepts to the audience, and foster discussions and exchange of ideas. Slides will be used to introduce the main points, while videos and audio clips will be played to illustrate various examples. After (and during) each of the main concepts of the tutorial outline is presented, time will be allocated for interaction between presenters and audience.

NEW FOR 2017: A new sub-topic that was developed for the presentation at CHI 2015 is interactive speech-based applications centred around language translation, language learning support, and interacting across multiple languages. This will be updated and expanded for the proposed CHI 2017 tutorial. Recent advances in Deep Neural Networks have dramatically improved the processing accuracy of speech recognition systems; however, this requires powerful computational resources not available to all developers – we are updating our course materials for 2017 to include an analysis of its implications for the design of interactive systems. Additionally, even the most capable computation servers continue to struggle when acoustic, language, or interaction environments are adverse, resulting in large variations in the accuracy of processing speech – this is particularly relevant for home-based smart personal assistants such as Amazon Echo where unexpected interaction contexts (e.g. loud music) can negatively impact performance and thus user experience. We will update the 2017 materials with further discussion and analysis of such examples.

Benefits to CHI attendees

The course will be beneficial to all HCI researchers or practitioners without a strong expertise in ASR or TTS, who still believe in fulfilling HCI's goal of developing methods and systems that allow humans to naturally interact with the ever increasingly ubiquitous mobile technology, but are disappointed with the lack of success in using speech and natural language to achieve this goal. During the past CHI instances of this course many of the attendees were from the speech processing community, especially from relevant industries attending the course (and CHI) due to their interests in learning about how to incorporate their own engineering advances into better user-centred designs.

Duration

Course duration: 2 ½ hours.

Requirements and hands-on activities

Typical presentation support (e.g. projector), with an emphasis on requiring a working audio connection for the presenters' laptop as the tutorial relies extensively on audio and video examples.

The three-hour version of this course includes opportunities for classroom participation, discussion of recent uses of ASR and TTS in commercial hands-free products, as well as discussions of case studies.

The course also includes three interactive, hands-on activities. The first activity will engage participants in proposing design alternatives for the error-handling interaction of a smartphone's voice-based search assistant, based on an empirical assessment of the type of ASR errors exhibited (e.g. acoustic, language, semantic). For the second activity, participants will conduct an evaluation of the quality of the synthetic speech output typically employed in mobile-based speech interfaces, and propose alternate evaluation methods that better reflect the mobile user experience. **NEW FOR 2017**: The third activity will center around uncovering speech processing errors of a home-based personal assistant and designing interactions that maintain a positive user experience in the face of unexpected variations in speech processing accuracy.

Participants

Ideal number of participants: the course interactive activities are based on forming groups of 4-6 participants. There is no minimum or maximum recommended number of participants; in the past, attendance typically included 20 to 50 participants. No prior technical experience is required for the participants. The classroom activities will be conducted using the participants' smartphones (Android or iPhone) and the instructor-provided devices – no software download will be required (the built-in phone functions will be used). Participants will work in small groups, ensuring that even participants without smartphone are able to fully contribute.