Designing Speech, Acoustic and Multimodal Interactions

Cosmin Munteanu
University of Toronto
Mississauga
cosmin.munteanu@utoronto.ca

Pourang Irani
University of Manitoba
Pourang.Irani@cs.umanitoba.ca

Sharon Oviatt
Incaa Designs
oviatt@incaadesigns.org

Matthew Aylett
CereProc
matthewa@cereproc.com

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

Shimei Pan
University of Maryland,
Baltimore County
shimei@umbc.edu

Nikhil Sharma
Google, Inc.
nikhilsh@google.com

Frank Rudzicz
Toronto Rehabilitation Institute
University of Toronto
frank@spoclab.com

Randy Gomez
Honda Research Institute
r.gomez@jp.honda-ri.com

Benjamin Cowan
University College Dublin
benjamin.cowan@ucd.ie

Keisuke Nakamura
Honda Research Institute
nakamura@hri.jp

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
Copyright is held by the owner/author(s).
CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3027086

Abstract
Traditional interfaces are continuously being replaced by mobile, wearable, or pervasive interfaces. Yet when it comes to the input and output modalities enabling our interactions, we have yet to fully embrace some of the most natural forms of communication and information processing that humans possess: speech, language, gestures, thoughts. Very little HCI attention has been dedicated to designing and developing spoken language, acoustic-based, or multimodal interaction techniques, especially for mobile and wearable devices. In addition to the enormous, recent engineering progress in processing such modalities, there is now sufficient evidence that many real-life applications do not require 100% accuracy of processing multimodal input to be useful, particularly if such modalities complement each other. 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 especially with mobile and wearable devices, and to look at how we can leverage recent advances in speech, acoustic, and multimodal processing.

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

Introduction and Motivation

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 with 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 other interaction modalities as well.

At the same time, as wearable devices gained prominence among the ecosystem of interconnected devices, novel approaches for interacting with digital content on such devices will be necessary. Such device form factors present new challenges (and opportunities) for multimodal input capabilities, through speech and audio processing, brain computer interfaces, gestural input and electromyography (EMG) interaction, among other modalities.

Unfortunately, communication through speech and language, while undoubtedly natural, are also among the most difficult modalities for machines, as these are the highest-bandwidth two-way communication channels we have. While significant effort, in 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 [1].

The challenges in enabling such natural interactions have often led to these modalities being treated as error-prone alternatives to "traditional" input or output mechanisms. This should not be a reason to abandon speech interaction[1] – people now face many more situations requiring hands-and eyes-free interaction. In fact, achieving 100% accurate speech processing may not be necessary: proper interaction design can complement speech processing in ways that compensate for its lack of accuracy [4,8]; and, in many tasks where users interact with spoken information, verbatim transcription is often not relevant [9].

Recent commercial applications (e.g. personal digital assistants such as Siri), made possible particularly by advances in speech processing and machine learning, have brought renewed attention to speech- and multimodal-based interaction. Yet, these are still perceived as "more of a gimmick" [2] or simply "not that good" [3], which can be partly attributed to a mismatch between the affordance of speech as input modality and the reality of speech as content creation modality. Additionally, new consumer devices such as the Amazon Alexa have limited functionality and do not always handle multimodal interactions, e.g. speech in challenging acoustic environments such as loud music. Unfortunately, there is very little HCI research, with rare exceptions such as our CHI 2014 and 2016 workshops on which we are building upon [8], on how to leverage the recent engineering progress into developing more natural, effective, or accessible UIs.

Concurrently, we are seeing how significant developments in miniaturization are transforming wearable devices, such as smartwatches, smart glasses, smart textiles and body worn sensors, into a viable ubiquitous computing platform. The varied usage contexts and applications in which such emerging

[1] We use the term speech and speech interaction to denote both verbal and text-based interaction.

Figure 1: Popular media view of speech-enabled mobile personal assistants [3]

devices are being deployed in create new opportunities and challenges for digital content interaction.

While the wearable computing platform is at its early developmental stages, what is currently unclear or may be ambiguous is the range of interaction possibilities that are necessary and possible to engage device wearers with digital content. The diverse usage contexts, form factors, and applications, call for a multimodal perspective on content interaction through wearables. Because of the explosion of new modalities and sensors on smart phones and wearables, the multimodal-multisensor fusion of information sources is now possible. It is thus important to consider the broad range of input modalities, including speech and audio input, touch and mid-air gestures, EMG input, and BCI, among other modalities. The 'one size fits all' paradigm may be less applicable; instead we need to explore crossing multiple input modalities and identify which modalities may co-exist for varied usage contexts.

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, audio, and language interaction,
- Taking concrete steps toward developing a framework of user-centric design guidelines for speech-, audio-, and language-based interactive systems, grounded in good usability practices,
- Establishing directions to take and identifying further research opportunities in designing more natural interactions that make use of speech and natural language, and
- Identifying key challenges and opportunities for enabling and designing multi-input modalities for a wide range of wearable device classes.

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 [6], which was followed by two successful workshops (20+ participants each) on speech and language interaction, held at CHI 2014 and 2016 [7]. Additionally, a course on speech interaction that has been offered at CHI for the past six years [5] by two of the co-authors of the present proposal has always been well-attended. As such, we are proposing here to broaden the domain of this community-building activities to all manners of human forms of acoustic-based communications (audio and speech/language), and to all types of interfaces for which such communications may be more suitable: desktops, mobiles, wearables, personal assistant robots, smart home devices. We propose several topics for discussions and activity:

- What are the important challenges in using **speech as a "mainstream" modality**? Speech is increasingly present in commercial applications – can we characterize which other applications speech is suitable for or has the highest potential to help with?
- What **interaction opportunities** are presented by the rapidly evolving **mobile, wearable, and pervasive computing** areas? How and how much does **multimodal processing increase robustness over speech** alone, and in what contexts?
- Can speech and multimodal increase usability and robustness of interfaces and **improve user experience beyond input/output**?

- What can the CHI community learn from Automatic Speech Recognition (ASR), Text-to-Speech Synthesis (TTS), and Natural Language Processing (NLP) research, and in turn, how can it help these communities **improve the user-acceptance of such technologies**? For example, what should we be asking them to extract from speech beside words/segments? How can work in context and discourse understanding or dialogue management shape research in speech and multimodal UI? And can we bridge **the divide between the evaluation methods used in HCI and the AI-like batch evaluations** used in speech processing?
- How can UI designers make better use of the acoustic-prosodic **information in speech** beyond simply word recognition, such as **emotion recognition** or identifying users' cognitive states? How can this be translated into the design of **empathic voice interfaces**?
- What are the **usability challenges of synthetic speech**? How can expressiveness and naturalness be incorporated into interface design guidelines, particularly in mobile or wearable contexts where text-to-speech could potentially play a significant role in users' experiences? And how can this be generalized to **designing usable UIs for mobile and pervasive** (in-car, in-home) **applications that rely on multimedia response generation**?
- What are the opportunities and challenges for speech and multimodal interaction with regards to **spontaneous access to information afforded by wearable and mobile devices**? And can such modalities facilitate access in a secure and personal manner, especially since mobile and wearable interfaces raise significant privacy concerns?
- Are there particular challenges when interacting with emerging devices such as **smart home / ambient personal assistants** (e.g. Amazon Echo) or when interacting with **social robots**?
- What are the **implications for the design of speech and multimodal interaction presented by new contexts for wearable use**, including hands-busy, cognitively demanding situations and perhaps even unconscious and unintentional use (in the case of body-worn sensors)? Wearables may have form factors that verge on being 'invisible' or inaccessible to direct touch. Such reliance on sensors requires clearer **conceptual analyses of how to combine active input modes with passive sensors** to deliver optimal functionality and ease of use. And what role can **understanding users' context** (hands, eyes busy) play **in selecting best modality** for such interactions or in **predicting user needs**?

CHI Contributions and Benefits

The proposed topics address speech and multimodal interaction, as well as its application to areas such as wearable and mobile devices. This is currently receiving significant attention in the commercial space and in areas outside HCI, yet it is a marginal topic at CHI. Only a limited number of research papers, panels, workshop and tutorial are presented by other researchers than the authors of this proposal (to our knowledge, the last CHI workshop on this topic not organized by the present authors has been organized in 1997). As such, we hope that the proposed workshop will increase the collaboration between researchers and practitioners belonging to the CHI community and those mostly dedicated to speech, language, and multimodal technologies. As in previous years, we anticipate continuing working with industry partners and sponsors

that will facilitate the presence of industry guest speakers and further opportunities for collaboration between academia and industry.

Our workshop aims to develop speech, audio, and multimodal interaction as a well-established area of study within HCI, aiming to leverage current engineering advances in ASR, NLP, TTS, multimodal, gesture recognition, or brain-computer interfaces (BCI). In return, advances in HCI can contribute to creating processing algorithms that are informed by and better address the usability challenges of such interfaces.

We also aim to increase the cohesion between research currently dispersed across many areas including HCI, wearable design, ASR, NLP, BCI complementing speech, EMG interaction and eye-gaze input. Our hope is to energize the CHI and engineering communities to push the boundaries of what is possible with wearable, mobile, social robots, and pervasive computing, but also make advances in each of the respective communities. As an example, the recent significant breakthroughs in deep neural networks is largely confined to audio-only features, while there is a significant opportunity to incorporate into this framework other features and context (such as multimodal input for wearables). We anticipate this can only be accomplished by closer collaboration between the speech and the HCI communities.

Our ultimate goal is to cross pollinate ideas from the activities and priorities of different disciplines. With its unique format and reach, a CHI workshop offers the opportunity to strengthen future approaches and unify practices moving forward. The CHI community can be a host to researchers from other disciplines with the goal of advancing multimodal interaction design for wearable, mobile, and pervasive computing. The organizing committee for this workshop (the authors list) is living proof that CHI is the most appropriate venue to initiate such inter-disciplinary collaborations. We believe that the rich diversity of this committee's professional networks, spanning several research communities, will ensure that this workshop will attract many new participants to CHI 2017.

Workshop plan

Before the workshop

[November to December] Publicize the workshop through e-mail, distribution lists, social media, and through posters and in-person at conferences. Set up Google drive and forms for paper submission and reviews. Launch the 2017 version of the workshop website at: www.dgp.toronto.edu/dsli2017

[By Jan 25th] Send acceptance notifications.

[February 17th]: Camera-ready deadline

[March] Upload papers on the workshop website and create preliminary break-out groups. Invite participants to read the position papers of those in the same group.

[April] Send the workshop agenda to participants.

During the workshop

(Participants: 20; Duration: 7 hours + breaks)
Activities:

- Introductions from each participant, including a 2 minute overview of their research [1 hour]
- Plenary presentations of select papers [1 hour]
- "Birds of a feather" group sessions, each addressing one of the workshop's goals, and discussing position papers for members of each group [1.5 hours]
- Reporting from break-out sessions [1 hour]
- "Matchmaking" – Facilitate future collaborations between HCI and speech / acoustic processing

researchers by looking at challenges encountered in one area that can be solved with help from the other area. This will be conducted following the template of Robert Dilts' "Disney Brainstorming Method" in which groups progress through stages of idea analysis, generation, evaluation, and planning [1.5 hours]
- Drafting of workshop conclusions, proposed framework, findings, and determine an after-workshop action plan (e.g. a special-issue journal proposal) [1 hours]

After the workshop
[1 month] Finalize the action plan and send a workshop follow-up report to all participants.
[2-4 months] Follow-up: development of a proposal for a special edition journal or a proposal for a grand challenge-type workshop.
[5-8 months] Carry out the proposed action plan.

Participants

The workshop aims to attract position papers and participation from a diverse audience within the CHI community. We aim to bring togeether a group of participants, varied in their background, with an interest or record of activity in research, design, or practice in areas related to the core themes and challenges of the workshop. We look for papers relevant to the areas of speech, acoustics, natural language interaction, multimodal interaction, brain-computer interfaces (as used to complement speech), natural user interfaces, especially as applied to mobile, wearable devices, smart home personal assistants, or social robotic contexts. Participants with a speech recognition, synthesis, acoustic processing, or general language processing background, but with an interest in the HCI aspects of speech-based interaction, will be particularly welcome (as emphasized in the call for papers). Based on our past experiences (panel, previous workshops, and related CHI courses), we expect that the workshop will attract contributions from a truly multidisciplinary audience. The organizers have themselves backgrounds that complement each other, ranging from significant work in core speech processing issues, to natural and multimodal interfaces, and to designing intelligent wearable and mobile interfaces. We are confident that this diversity will ensure that the proposed workshop will be engaging and fruitful.

Sample Call for Papers

This workshop aims to bring together interaction designers, usability researchers, and general HCI and speech processing practitioners. Our goal is to create, through an interdisciplinary dialogue, momentum for increased research and collaboration in:
- Formally framing the challenges to the widespread adoption of speech, acoustic, and natural language interaction,
- Taking concrete steps toward developing a framework of user-centric design guidelines for speech-, acoustic-, and language-based interactive systems, grounded in good usability practices,
- Establishing directions to take and identifying further research opportunities in designing more natural interactions that make use of speech and natural language, and
- Identifying key challenges and opportunities for enabling and designing multi-input modalities for a wide range of emerging devices such as wearables, smart home personal assistants, or social robots.

We invite the submission of position papers demonstrating research, design, practice, or interest in areas related to speech, acoustic. language, and multimodal interaction that address one or more of the

workshop goals, with an emphasis, but not limited to, applications such as mobile, wearable, smart home, social robots, or pervasive computing.

Position papers should be 4-6 pages long, in the ACM SIGCHI extended abstract format and include a brief statement justifying the fit with the workshop's topic. Summaries of previous research are welcome if they contribute to the workshop's multidisciplinary goals (e.g. a speech processing research in clear need of HCI expertise). Submissions will be reviewed according to:

- Fit with the workshop topic
- Potential to contribute to the workshop goals
- A demonstrated track of research in the workshop area (HCI or speech/multimodal processing, with an interest in both areas).

Important Dates:

- January 25th, 2016: Submission of position papers
- February 3rd, 2016: Notification of acceptance
- February 17th, 2016: Camera-ready submissions

Workshop URL: http://www.dgp.toronto.edu/dsli2017

Authors' Biographies

Prof. Cosmin Munteanu is an Assistant Professor at the Institute for Communication, Culture, Information, and Technology at University of Toronto Mississauga. His research includes speech and natural language interaction for mobile devices, mixed reality systems, learning technologies for marginalized users, usable privacy and cyber-safety, assistive technologies for older adults, and ethics in human-computer interaction research. http://cosmin.taglab.ca

Prof. Pourang Irani is a Professor in Human-Computer Interaction at the University of Manitoba and Canada Research Chair in Ubiquitous Analytics. His research sits at the crossroads of mobile/wearable computing and data visualization. http://www.cs.umanitoba.ca/~irani

Dr. Sharon Oviatt is internationally known for her multidisciplinary research on human-centered interfaces, multimodal and mobile interfaces, spoken language and digital pen interfaces, educational interfaces, technology design and evaluation, and the impact of interface design on human cognition. Her latest textbook, The Paradigm Shift to Multimodality in Contemporary Computer Interfaces (co-authored with Phil Cohen) was published by Morgan Claypool in 2015.

CereProc Chief Science Officer Dr. Matthew Aylett has over 15 years' experience in commercial speech synthesis and speech synthesis research. He is a founder of CereProc, which offers unique emotional and characterful synthesis solutions and has been awarded a Royal Society Industrial Fellowship to explore the role of speech synthesis in the perception of character in artificial agents. http://www.cereproc.com

Prof. Gerald Penn is a Professor of Computer Science at the University of Toronto. He is one of the leading scholars in Computational Linguistics, with significant contributions to both the mathematical and the computational study of natural languages. Gerald's publications cover many areas, from Theoretical Linguistics, to Mathematics, and to ASR, as well as HCI.

Dr. Shimei Pan is an assistant professor at University of Maryland, Baltimore County (UMBC). Before joining UMBC, Dr. Pan was a research scientist at IBM T. J Watson research center in New York. Her primary research interests include Natural Language Processing, Multimedia Multimodal Conversation Systems, User Modeling and Human-Computer Interaction.

Dr. Nikhil Sharma is a Senior User Experience Researcher at Google. He leads the user experience research for Voice Search. He's worked on both multimodal and voice only experiences on a range of devices including smartphones, smartwatches and cars.

Dr. Frank Rudzicz is a Scientist at the Toronto Rehabilitation Institute and an Assistant Professor in Computer Science at the University of Toronto. His research focus is on machine learning and signal processing in atypical speech. He is the President of the ACL/ISCA SIG on Speech and Language Processing for Assistive Technologies and Young Investigator of the Alzheimer's Society. http://www.cs.toronto.edu/~frank

Dr. Randy Gomez is a senior scientist at the Honda Research Institute Japan (HRI-JP). His interests include speech recognition, speech enhancement, multi-modal interaction and intelligent systems. Currently he oversees the research work in robust multimodal interaction with context awareness at HRI-JP.

Dr Benjamin R. Cowan is an Assistant Professor at University College Dublin's School of Information and Communication Studies. His primary research focuses on the psychological aspects of speech interface interaction. In particular he studies how the design of speech interfaces affect user's perceptions, their mental models as well as their language production. http://www.benjamincowan.com

Dr. Keisuke Nakamura is a senior scientist at the Honda Research Institute Japan. His research interests are in the field of robotics, control engineering, signal processing, computational auditory scene analysis, multi-modal integration, and robot audition.

References

1. Aylett, M. et al. (2014). None of a CHInd: relationship counselling for HCI and speech technology. In Alt.CHI '14
2. Business Insider (2012). Frankly, It's Concerning that Apple is Still Advertising A Product as Flawed as Siri. http://www.businessinsider.com, 2012.
3. Gizmodo (2011). Siri is Apple's Broken Promise. http://www.gizmodo.com, 2011.
4. 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.
5. Munteanu, C. and Penn, G. (2016). Speech-based interaction. Course, CHI 2011, 2012, 2013, 2014, 2015, 2016
6. Munteanu, C. et al. (2013). We need to talk: HCI and the delicate topic of speech-based interaction. Panel, ACM SIGCHI 2013.
7. Munteanu, C. et al (2016). Designing Spoken Language Interaction. CHI 2014 & 2016 Workshops
8. Oviatt, S. (2003). Advances in Robust Multimodal Interface Design. IEEE Comput. Graph. Appl. 23-5.
9. Penn, G. and Zhu, X. (2008). A critical reassessment of evaluation baselines for speech summarization. In Proc. of ACL-HLT.