Designing Speech and Multimodal Interactions for Mobile, Wearable, and Pervasive Applications

Citation for published version:
Munteanu, C, Irani, P, Oviatt, S, Aylett, M, Penn, G, Pan, S, Sharma, N, Rudzicz, F, Gomez, R, Nakamura, K & Nakadai, K 2016, Designing Speech and Multimodal Interactions for Mobile, Wearable, and Pervasive Applications. in Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. CHI EA '16, ACM, New York, NY, USA, pp. 3612-3619, 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, United States, 7/05/16. DOI: 10.1145/2851581.2856506

Digital Object Identifier (DOI):
10.1145/2851581.2856506

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems

Designing Speech and Multimodal Interactions for Mobile, Wearable, and Pervasive Applications

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
University of Toronto
frank@spoclab.com

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

Abstract

Traditional interfaces are continuously being replaced by mobile, wearable, or pervasive interfaces. Yet when it comes to the input and output modalities through which we interact with such interfaces, we are 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 and multimodal interaction techniques, especially for mobile and wearable devices. Independent of 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, two-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 with mobile and wearable devices, and to look at how we can leverage recent advances in speech and multimodal processing.

Keywords

Speech and Language Interaction; Automatic Speech Recognition; Speech Synthesis; Natural Language Processing; Natural User Interfaces; Multimodal Interaction; Mobile Interaction; Wearable Devices.

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 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.

At the same time, as wearable devices gain prominence among the ecosystem of interconnected devices we rely on for our daily activities, novel approaches for interacting with digital content on such devices will be necessary~~ to accommodate the growing range of applications and contexts in which they get used in~~. 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, 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).

[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.

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

Recent commercial applications (e.g. personal digital assistants), made possible particularly by advances in speech processing and new machine learning algorithms, have brought renewed attention to speech- and multimodal-based interaction. Yet, these continue to be perceived as "more of a gimmick than a useful tool" (Business Insider, 2012) or simply as not "that good" (Gizmodo, 2011). Unfortunately, there is very little HCI research on how to leverage the recent engineering progress into developing more natural, effective, or accessible user interfaces.

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. While the sales of mobile devices have remained constant, that of wearable devices is growing rapidly [**REF**]. The varied usage contexts and applications in which such emerging 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 on such devices. One response how seems clear: the diverse usage contexts, form factors and applications, with wearables, calls for a multimodal perspective on content interaction. ~~For example, during a meeting a user may interact with a wearable through subtle EMG gestures, but while driving, the same device may be primarily controlled via speech, as it may afford a wider input bandwidth than EMG signals.~~ We believe, it is 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. Furthermore, as such a platform offers a much broader usage contexts than desktops or current mobile device, we believe that the before long 'one size fits all' paradigm, may be less applicable and instead our communities need to pave the way for crossing multiple input modalities to identify which modalities may co-exist for varied broad 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 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.
- 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 [5] which was followed by a successful workshop (20+ participants) on speech and language interaction, held at CHI 2014 ([**REF**]). The interest attracted by the 2013 panel and the 2014 workshop has resulted in organizers receiving many requests for further development of the workshop. Additionally, a course on speech interaction that has been offered at CHI for the past five years ([**REF**]) by two of the co-authors of the present proposal has always been well-attended. As such, we are proposing here to expand the proceedings of such a workshop to two days and enlarge its scope to speech and multimodal interaction for mobile, wearable, and pervasive computing. We therefore propose several topics for discussions and activity among workshop participants:

- What are the important challenges in using **speech as a "mainstream" modality**? Can we broadly characterize which interfaces/applications speech is suitable for? (e.g. commercial, literacy support, assistive technology, in-car, in-home)

- What **interaction opportunities** are presented by the rapidly evolving **mobile, wearable, and pervasive computing** areas? Can speech and multimodal increase usability and robustness of interfaces and **improve user experience beyond input/output** within these areas?

- ~~Given the penetration of mobile computing in emerging markets, are there any specific **usability or technology adoption** issues surrounding speech interaction?~~

- What can the CHI community learn from the Automatic Speech Recognition (ASR), Speech Synthesis, and the Natural Language Processing (NLP) research, and in turn, how can it help the speech technology ~~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 work in areas such as context/discourse understanding or dialogue management can 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?

- ~~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 – can we **combine other such modalities (e.g. emotion) 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 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?

- 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)? ~~While overlap in usage contexts may exist with other platforms, differences are clear – w~~Wearables may have form factors that verge on being ‘invisible’ or inaccessible to direct touch (if worn under clothes or if implanted). 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 the context** of the user (hands-busy, 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. The workshop organizers are also in advanced stages of negotiating industry sponsorship for this that, to a minimum, allow us to invite keynote speakers which will attract further interest and opportunities for collaboration between academia and industry. Given the location of CHI 2016, with such close proximity to many of the “tech giants” who are vastly invested in natural user interfaces, we believe the timing and conditions are favourable for the organization of this workshop. ~~For this, we aim to reach three development objectives:~~

~~*Scientific Development*~~

~~The primary objective is the development of speech and multimodal interaction as a well-established area of study within HCI. Such an area will be concerned with the investigation, design, development, and evaluation of natural and effective interactive technologies that based on novel multimodal interfaces, leveraging current engineering advances in ASR, NLP, text-to-speech synthesis (TTS), multimodal/gesture recognition, or brain-computer interfaces. In return, advances in HCI can inform developments in its engineering counterparts, leading to adapting and designing NLP and ASR algorithms and methods that are informed by and better address the usability challenges of speech and multimodal interfaces that are employed in real-life, critical interaction tasks~~

~~*Agenda Development*~~

~~A second objective of this workshop is to gather researchers from diverse backgrounds, including but not limited to those advancing the fields of speech and audio input, EMG, BCI and gestural interaction, to define the future roadmap for multi-input modalities for~~

~~wearable devices. There has been less cohesion in working towards the theory and application of such advancements toward the aim of maturing a common technology platform. Our hope is that by focusing and challenging the research community on multi-input modalities for wearables, we will energize the CHI and engineering communities to push the boundaries of what is possible with wearable devices, but also make advances in each of the respective communities. As an example, the speech recognition community has had significant and recent breakthroughs via the application of deep machine learning techniques to domain specific concerns. Yet the results are mostly confined to tightly controlled conditions, and perhaps with very little 'perturbations' that are natural artefacts in the case of wearable use cases. Since most current research on deep neural networks is largely confined to audio-only features, there is a significant opportunity to incorporate into this framework other features and context (such as multimodal input for wearables), which we anticipate can only be accomplished by closer collaboration between the speech and the HCI communities.~~

~~*Community Development*~~

Developments in input and interaction modalities for wearable devices is dispersed across a broad range of areas including, human-computer interaction, wearable design and development, speech and audio processing, natural language processing, brain computing that acts as a complementary modality to speech, EMG interaction and eye-gaze input. Furthermore, design and development in wearables is happening in both academic as well as in industrial settings. Our goal is to cross pollinate ideas from the varied activities and priorities of these different disciplines. Researchers working in areas that experience a burgeoning scientific discourse may not be informed of the latest research in overlapping areas, or of their applicability to their own research. A key objective of this workshop is to initiate brainstorming across disciplines, a task not possible through the traditional format of research papers. By closely sharing past accomplishments in each of these fields, we can create the opportunity to strengthen future approaches and unify practices moving forward.

We want the CHI community to be a host to researchers from these other disciplines and communities with the goal of advancing interaction design and development on wearables.

Workshop plan

Before the workshop

[November] Establish the workshop review committee from amongst researchers and practitioners established in the proposed area of the workshop. The Easy Chair conference system will be used for paper submission and reviews, and a university-hosted workshop website will be created.

[November-December] Publicize the workshop through e-mail, distribution lists, social media, and through posters and in-person at conferences. Tentative date for distributing the CFP: Nov 20th. In addition, the workshop organizing committee may invite specific submissions from a small group of researchers and practitioners that the committee sees as valuable for the interactions and discussions during the workshop.

[Early December]: Review of early round of position papers and notification of early acceptance to authors (criteria: outstanding positions papers that the review and workshop organization committee deems as a priority, along with invited submissions).

[December] Review position papers

[Late January] Send acceptance notifications. (no later than January 25)

[February]: Camera-ready deadline (February 8)

[March] Upload papers on the workshop website and create preliminary assignments of participants to 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: 20-25

Activities for day one, with a focus on the theoretical and methodological challenges of designing speech and language interactions:

- Introductions from each participant, including a 2 minute overview of their research projects or interests as relevant to the workshop [1 hours]
- Brief presentations of position papers related to speech and language interaction design aspects [2 hours]
- Group discussions [0.5 hour]
- “Birds of a feather” sessions, each addressing one of the workshop's goals (specific goals may be adjusted based on participants' input) [1.5 hours]
- Reporting from break-out sessions [1 hour]

Activities for day two, with focus on specific application areas:

- Brief presentations of position papers related to the application of multimodal interactions (speech, language, EEG, gestures, etc.) to wearable and mobile interfaces [2 hours]
- “Matchmaking” – Facilitate future collaborations between HCI and speech 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 [2 hours]
- Drafting of workshop conclusions, proposed framework, findings, and determine an after-workshop action plan (e.g. a special-issue journal proposal) [2 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 of varied backgrounds, with an interest in, or record of, research, design, or practice in areas related to speech, natural language interaction, multimodal interaction, brain-computer interfaces as used to complement speech, natural user interfaces, especially as applied to mobile or wearable devices. Participants with a speech recognition, synthesis, 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 workshop, and related CHI courses), we expect that the workshop will attract contributions from a truly multidisciplinary audience. The members of the organizing committee have themselves backgrounds that complement each other, ranging from significant work in core speech processing issues, to natural and multimodal interfaces, to designing wearable and mobile interfaces, and to testing the social acceptance of speech applications. 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 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.
- Identifying key challenges and opportunities for enabling and designing multi-input modalities for a wide range of wearable device classes

We invite the submission of position papers demonstrating research, design, practice, or interest in areas related to speech, 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, or pervasive computing.

Position papers should be no more than 4 pages long, in the ACM SIGCHI Archival format, and include a brief statement from the author(s) justifying the interest in the workshop's topic. Summaries of research already presented are welcome if they contribute to the multidisciplinary goals of the workshop (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).

Authors' Biographies

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

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 recently been awarded a Royal Society Industrial Fellowship to explore the role of speech synthesis in the perception of character in artificial agents. 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.

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, 2014, 2015

[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.