None of a CHInd: Relationship Counselling for HCI and Speech Technology

Matthew P. Aylett
CereProc, Ltd., University of Edinburgh
Crichton St.
Edinburgh, UK.
matthewa@inf.ed.ac.uk

Yolanda Vazquez-Alvarez
University of Glasgow
Lilybank Gardens
Glasgow, UK.
Yolanda.Vazquez-Alvarez@glasgow.ac.uk

Per Ola Kristensson
University of St. Andrews
North Haugh
St Andrews, UK
pok@st-andrews.ac.uk

Steve Whittaker
University of California at Santa Cruz
Santa Cruz, CA 95064
swhittak@ucsc.edu

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

ACM 978-1-4503-2474-8/14/04 ...$15.00.
http://dx.doi.org/10.1145/2559206.2578868

Abstract
It's an old story. A relationship built on promises turns to bitterness and recriminations. But speech technology has changed: Yes, we know we hurt you, we know things didn't turn out the way we hoped, but can't we put the past behind us? We need you, we need design. And you? You need us. How can you fulfill a dream of pervasive technology without us? So let's look at what went wrong. Let's see how we can fix this thing. For the sake of little Siri, she needs a family. She needs to grow into more than a piece of PR, and maybe, if we could only work out our differences, just maybe, think of the magic we might make together.

Author Keywords
Speech technology; pervasive systems; ambiguity; ludic design; human computer interaction

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction: Where did it all go wrong?
The root cause of many problems in a relationship is to do with mismatched expectations. The phrase "speech is a natural means of communication" is so ubiquitous (29700 hits on Google—not bad for a 7-word sentence

search term) that speech technologists rarely question the importance of their work. In science fiction, speech interfaces and the role of speech as a means of human-computer interaction stretches back many years. In movies, evil computers intent on taking over the world, or killing hapless astronauts, have long been given both artificial intelligence and the ability to both voice their evil plans and listen in on mere humans as they attempt to thwart them. Alas (or thankfully) neither evil intelligent computers, nor much in the way of conversational artificial systems are much in evidence in the modern world.

In the heady days of the eighties and nineties, HCI professionals could remain quietly sceptical. It was nice to see the enthusiasm of speech collaborators. It was charming to see how activities like booking a flight, following a recipe, or dictating a letter could fit into a speech environment. But, as in any relationship, after the first thrill of love and excitement, there comes a time to get stuff done. There comes a time to get the washing done and put the shelves up.

The sentence "speech is a natural means of communication" rings pretty hollow when you discover a few decibels of noise caused your recognition rate to drop below 50%. And it was hard to see the promise in a speech interface with unnatural and monotonous speech output.

Finally, as if this betrayal was not enough, they start flirting with other people, people who you don't respect and regard as *enemies of usability*. Yes, ladies and gentlemen, speech technology embraced the dark god of the CALL centre. The filthy lucre of commercial success tempted good upstanding engineers away from the light and into the dark usability hell of the automated telephone service. You take a step back, you see them for what they are. And you suddenly ask yourself, was there really any spark? Was there really anything you had in common? You start playing with GUIs, it's just a friendly thing to start with, but then you see how pretty they are, how un-complicated, how willing to make you happy.

The speech technologists don't even see it happening. They are too focused on getting that word error rate down and modelling that prosody, to see that the ardour has cooled, that there really is someone else. A few years later they meet in bars, and say how they never understood you, and how they did everything they could for you, and how they can't understand why you fell out of love with them.

Speech Technology is Marginalised

HCI is a large diverse community. There will always be many modalities and technologies competing for the attention of engineers and designers interested in producing new interfaces and interactive technologies. So is there any evidence for this tragic portrayal of the relationship between HCI and speech technology.

For example, if we look at speech recognition or speech synthesis as a search term in CHI publications, has there been a sharp drop in published work supporting this falling out of love scenario? In a simple search based on Google Scholar the answer appears to be no. Searching with the publication field set to CHI over the range 1990-2000, we find 6770 results of which only a small percentage have the terms *speech synthesis* or *speech recognition* in them (77, $<$0.02%). In 2000-2010, we find 11900 results and a a small increase in percentage terms but still a very small proportion of total search hits (188, $<$0.02%). This small percentage does not reflect a disregard for speech tech-

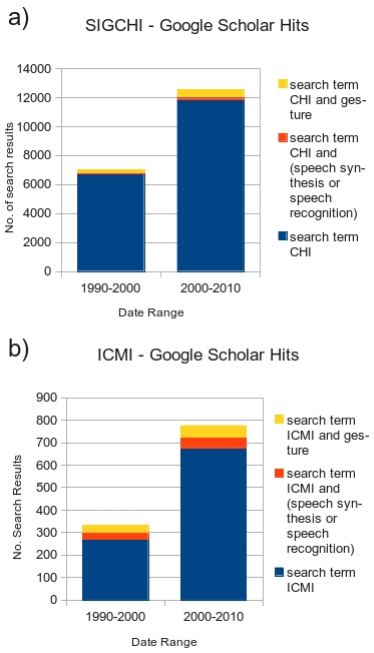

Figure 1: Comparison of search results for speech technology and gesture compared to overall hits for the conference.

ology as searching for the term *gesture* also produces mall percentage results. Rather this reflects wide remit f CHI as an *umbrella* conference for HCI work (See Figre 1a).

If we look at the International Conference on Multi-modal Interaction (ICMI), we see a much higher percentage of search results for the terms *speech synthesis* or *speech recognition*. Between 1990 and 2000 we have 270 results for publication ICMI, of which 32 are speech related hits (>10%) compared to 30 for the search term *gesture* (>10%). If we look at the period 2000-2010 we have a total of 676 search results for publication ICMI of which 48 are speech related (>7%) compared to 51 for "gesture" (>7%). So there is evidence of HCI and speech technology collaboration in the area of multi-modal interaction (See Figure 1b).

However there is a perceived problem. Both at CHI 2002 and CHI 2013 there was a panel discussion on the topic of speech technology and HCI[4][8]. At CHI 2002, the panel was asked why, *"the use of speech is so controversial in the HCI community."*. One summary of the take home message from this panel was *"...speech recognition is still restricted to special domains as I have learned in a panel about speech interfaces."* Gerd Waloszek, SAP AG[1]. A whole decade later a possible reason for the problem was expressed as follows: *"This may be due to a widespread perception that perfect domain-independent speech recognition is an unattainable goal."* [8].[2]

Speaking directly with researchers who have spent considerable time in the HCI and speech related field there is also informal feedback that a problem exists. Reasons for this problem include: repeated suggestion of a new dawn for speech technology right back to the 80s which never materialised, the difficulty in building error-free speech systems, speech technologists' disinterest in real systems, and the difficulty faced by non-speech experts in creating non-traditional speech interfaces. Worse, speech technology is perceived as dull and hard. *"I think some in HCI may see it as tricky to put together a playful wow speech interface/interaction that captures the imagination at CHI, where in my view wow design is becoming quite a driving force."* Dr. B. Cowan.

This perception of speech technology is supported by some serious reviews of the limitations of the technology. Perhaps the most important being Shneiderman [12], where he very explicitly pointed out some of the problems with speech interfaces in comparison with GUIs.

> "Human-human relationships are rarely a good model for designing effective user interfaces. Spoken language is effective for human-human interaction but often has severe limitations when applied to human-computer interaction. Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks."

Ouch! He does however accept that speech might be useful for voice mail and disabled users.

> "However, speech has proved useful for store-and-forward messages, alerts in busy environments, and input-output for blind or motor-impaired users."

[1] www.sapdesignguild.org/community/readers/print_reader_chi2002_gw.asp

[2] This panel will be followed up at CHI 2014 with a workshop exploring the issues. In contrast to the snarky comments and poorly substantiated opinion presented here, we expect the workshop to have a more positive, if less honest, approach to the subject.

In other words, its place is very much the periphery of HCI. In some senses, we can see this as a reaction to the assumption that speech interfaces are the natural solution to human-computer interaction, or as stated earlier the common sentence "speech is a natural means of communication". After all, so is punching someone in the face, it naturally communicates dislike and anger, but that doesn't make it a good model for an interface.

Has a decade of research work in speech technology as well as a decade of change in our use of computational devices changed this view? Not really. This critical view of speech technology extends to much more recent work. For example, a review of potential input and output technologies for *Always-Available* mobile interactions [7] in 2011 concluded:

> "Furthermore, significant technical limitations call into question the ultimate performance of speech interfaces in real-world environments, and the strong association between social interactions and speech has raised further criticism of the role of speech in UIs. Starner[13] breaks down some of these social and technical limitations in more detail."

This is quite a severe criticism. Yet the paper does not apply the same view to a gesture input system such as sixth sense[6], where a camera is tied around a users neck and used as a gesture input system. Arguably, this approach may also have technical limitations and complicate social interactions[3]. This raises the question of why speech technology is treated so negatively in comparison to other technologies. The Starner paper [13] referenced here is an excellent review of the challenges in the use of automatic speech recognition (ASR). Furthermore, the overall impression from Starner [13] is more positive than this reference suggests. His paper discusses the technical limitations and challenges but also presents various ways these technical limitations can be minimised with good engineering and design.

Another interesting example is Ni and Baudisch in 2009[9]. This paper sets out to review the technology that is relevant when extrapolating miniaturisation to the extent when devices allow invisible integration into arbitrary surfaces or human skin. They look at candidate technologies for supporting meaningful interaction with such devices, in particular gesture-based communication. The use of audio, and speech technology is mentioned and dismissed as follows:

> "Audio is an interesting alternative as well for input, e.g., in the form of speech recognition[12]. The inherent volume of speech input can limit its applicability in situations where others are around[12]."

This paper is referencing back to Shneiderman[12]. This issue of noise pollution is a serious one. However, historically, not one that appeared to prevent the widespread uptake of mobile phones. More interesting, is that the paper feels it is able to dismiss a technology which can arguably play an important role in eyes-free and hands-free interaction, which are both critical issues in very small mobile devices.

Yet there has been speech technology and HCI collaboration since the Starner [13] and Shneiderman [12] reviews, for example[16, 15]. Why is none of this more recent work referenced? Neither Morris et al. [7], nor Ni and Baudisch [9], are an attack on speech technology, they

[3] This isn't a criticism of this technology, we think its cool

just both conclude, based on reviews published nearly a decade earlier, that speech technology has problems and is not so relevant for the area they are looking at (wearables and miniaturised interfaces). This ability to dismiss speech technology so readily is a symptom of an underlying problem. A senior speech technologist summarised this impression of something being wrong in the relationship between speech technology and HCI as: *"what I would term 'academic disrespect' going both ways."* Prof. S. Renals.

To summarise we argue that the cause for this is rooted in:

- The historical difficulty in integrating and controlling ASR and speech synthesis technology.
- The very real challenges in using speech technology interactively in terms of cognitive load and unwanted intrusion.
- An impression of a series of false dawns, where suggested functionality and ubiquitous use failed to materialise.

Combined, these issues make it hard to produce compelling HCI engineering using speech technology that has an impact, or a 'wow' factor. Furthermore is has undermined any excitement HCI has in the field, leaving it to enthusiasts rather than the mainstream.

Obstacles in Combined HCI Speech Technology Research

Technical Challenges in Speech Technology Integration
The clue is in the name, "human-computer **interaction**". It's not called "human submits batch job to large parallel computer network". Building interactive interfaces is complex because it has to allow for interruption and multi-tasking. Engineers have been working on this with GUIs for over 40 years and we are sure readers will remember how fragile GUI systems used to be in the past. Speech interfaces are a much more immature technology and also introduces two further challenges for an interactive interface. Firstly the difficulty in producing acceptable latencies, and secondly, the requirement of allowing user interruption.

The computation required for speech recognition and (until recently), for speech synthesis, is large. Often a whole phrase (or even a sentence) must be processed at a time. Thus, to manage latencies of less than 200ms, an engine of at least a 10 times real-time is required. The more interactive you want the interface to be, the more important response time becomes. Approaches like incremental processing and clever dialogue management can make a big difference to speed and response, but you require an actual system to develop these approaches. Traditionally, speech technology has left system building to commercial organisations and HCI researchers.

Even with low latency systems, there is a real challenge in building a system which is sensitive to user feedback. For example, if the speech synthesis system is producing a sentence and the user speaks, the system must process the speech input, possibly halt speech output, and respond to the feedback within about 200ms. The sheer engineering complexity of a computationally intensive parallel architecture has been a serious barrier in interactive speech projects. For example take SAL, (the sensitive artificial listener from the Semaine project[5]). Here the main research objective was to deal with the problem of interacting with a user and did so with varying degrees of success. If you bear in mind this was a multi-million

Euro, multiple site, EU project and the true challenge in engineering terms becomes apparent.

The resources required are also so high because each component in a interactive speech system is complex and difficult to produce. In the past, sourcing decent pre-built systems and integrating them has been fraught with problems, especially for a fast latency, interactive system.

Compare this with writing a GUI. There are many toolkits and graphical development environments that support and help engineers create these interfaces. Although speech synthesis has been supported by Apple and Microsoft for some time, until recently the systems they provided were significantly below the industry standard. Meanwhile, ASR was just not available at all, except by purchasing extra commercial system which were expensive to deploy. Even today with speech recognition and synthesis available on iOS and Android, there are many constraints and limitations for 3rd party engineers. Historically, big companies have not been keen to share this technology.

These technical difficulties have a big impact on the ability of an independent HCI researcher to develop speech technology interfaces, not just because of the technical challenge but the social effect this has. Unlike many other interface systems, it comes with a lack of control of the core technology as well as a dependence on external researchers with different priorities and research methodologies.

Cognitive Load and Social Intrusion
As Shneiderman [12] rightly pointed out way back in 2000, there is a big issue in the intrusive nature of audio, especially speech and its effect on dividing attention and increasing cognitive load. Yet, as almost every speech-related HCI paper points out, speech is an ideal means of communication in an eyes-free or hands-free setting.

Because of the performance and architectural difficulties described above, speech interface systems are still an immature technology, thus compelling techniques for dealing with these problems are still in their infancy. (See for example Vazquez-Alvarez and Brewster [14], which explored issues in multi-tasking within audio interfaces). Furthermore, users have little experience of such interfaces, except in terms of using automated telephone systems to access call centres, and phone-based speech recognition recently came top in *Wired*'s 12 most annoying technologies[2]. All these factors add to the design difficulty.

Confidence
So you get the technology together, manage to design a system which avoid the pitfalls of cognitive overload and social intrusion. Oh, sorry, you find it just doesn't actually work. The ASR recognition rates promised by your ASR supplier go through the floor in a real environment, the language understanding cannot cope with the open-ended nature of user input, and the speech synthesis is so boring to listen to that you can only get undergraduates to use the system because you are paying them to. Way back since the 80s there have been false dawns where speech technology was presented as *the next big thing* and each time this was a common bitter experience for the naive HCI researchers who joined in.

Serial failure isn't sexy. If you combine all these issues, the fact HCI and speech technology have had a difficult relationship over the years seems less surprising. In fact, it would have been a miracle if it hadn't been this way.

But the time has come to fix this thing. We are entering a new dawn in mobile and pervasive interactive technologies. Speech technology is going to be central to interactive technologies over the next decade. (Yes really, this time it really will be, honest).

Why Now?

Siri, Smart Phones and Social Media

Controlling the direct information channel to the user, and dominating it, is one way of making a lot of money, Google have their search engine, Apple have iTunes, Amazon have their store. This direct relationship with users is seen as so powerful that companies like Twitter and Facebook, with moderate advertising incomes, have been valued in the billions. Maintaining this direct connection with the user means that these companies must respond to changes in the way we access Internet services.

In 2000 when Shneiderman [12] published his review of the limitations of speech technology, the dominant means of accessing Internet services was using a desktop or laptop computer. In 2008, laptop sales exceeded desktop sales (38.6m vs 38.5m - iSuppli, 2008). The mass market for tablets had yet to emerge, and netbooks were pipped to be the next big thing. But, in just 4 years, smartphone share of the handset market quadrupled from 12% (140m) to 58% (1bn)(Gartner), and in 2013 tablet sales are expected to exceed laptops sales (227m vs 134m desktops and 180m laptops, IDC, 2013). In 2013, nearly 80% of the devices sold that are used to access the Internet are smartphones and tablets.

As engineers, we often focus on the technology rather than the commercial drivers behind it. Speech technologists saw Siri as 90s speech technology done well, how useful Siri is and how many users are using it is a matter of contention. But companies like Google and Amazon (amongst many others) are taking this technology very seriously indeed. This is not because of a renewed evangelism for speech technology, it is because it is a means of controlling the direct channel to the user. Imagine how pleased Google must have been for Siri to use their search engine, and present the result to the user without the user seeing any of their ads.

Large US corporates have been on a buying spree of speech technology and speech technology-related companies. While academics within speech technology are welcoming this new interest in their field of expertise, they are seeing an aggressive recruitment drive of ASR researchers into industry. Apple set up its first ever R&D lab outside Cupertino in Boston, and it's a speech lab, Amazon has bought ASR and speech synthesis companies, while Google purchased a speech synthesis company as far back as late 2010.

Mainstream HCI researchers may still regard speech technology as dull but this view is not shared in the commercial world. For this reason alone, it is time for HCI and speech technology researchers to look again at constructive collaboration.

Things are Easier

The ease of deploying ASR and speech synthesis has improved. For example, on Android, many state-of-the-art speech synthesis systems can be purchased for a few dollars. The naturalness of these systems is significantly better than several years ago. The ASR system on Android can also be accessed with some constraints. Open source toolkits such as HTK[17], have been joined by Kaldi[10].

Things are Better
Google were pioneers at making use of big data to improve ASR. Microsoft recently published ASR results based on deep neural networks which were significantly higher than the state-of-the-art. Siri has cleverly made use of ASR within an application domain which mostly just works. These are incremental improvements, but improvements none the less.

Speech is Not Just Communication
Human communication is rich, varied and often ambiguous. This reflects the complexity and subjectivity of our lives. We might expect speech and language technology, dealing as it does with such a central form of human communication, to be at the forefront of applying technology to the interpretation of our ambiguous and multi-layered experience. In fact, much of the work in this area has avoided ambiguity and is often used as a tool to disambiguate information rather than as a means to interpret ambiguity. Take, for example, conversational agents (CAs)[3]: These are computer programs which allow you to speak to a device and will respond to you using computer generated speech. These systems can potentially harness the nuances of language and the ambiguity of emotional expression. However, in reality, we use them to ask them how high Mount Everest is or where you can find a nearby pizza restaurant. This raises the question of how we might extend such systems to help us interpret more complex aspects of the world around us. It is important for this technology to strive to do so for two fundamental reasons: firstly, technology has become part of our social fabric and as such this technology needs to be able to engender playfulness, and enrich our sense of experience, and secondly, applications which could perform a key role in mediating technology for social good require a means of interacting with users in much more complex social and cultural situations.

Speech technology offers a means to extend technology from the mundane to reflect the ambiguity, beauty and complexity of life. HCI has always taken up the challenge of not just looking at what is now, but trying to envisage what is next. Speech technology will be key in ambiguous and ludic systems because of the capacity of natural language to be both playful and ambiguous.

How do we Fix this Thing?

Pick Up the Phone
Earlier on in the paper we explored some of the challenges of working with speech technology. Yes it's hard, but its getting easier. Back in the 80s designers didn't complain about green screen terminals and waited for VGA displays—they worked with the technical limitations that existed. As Starner [13] discussed back in 2002, there are many clever ways of integrating speech technology which offsets the limitations. 100% recognition rates and perfectly natural speech synthesis are not required if you know what the user is doing, and can produce an elegant application which makes doing it easy, fun and compelling. Starner now works with Google Glass and will no doubt be following his own advice.

Customising speech technology for the application and context makes the technology work better. Speech technology is crying out for good design input. In addition there has been a sea of change in the use of mobile devices, which raises the challenge of eye-free and hands-free interfaces.

So pick up the phone, don't be proud, don't be scared, we understand that there is a history but we can get back

together if we only have the energy, tolerance and good will to do so.

Broaden the Scope of Speech-Related Work
AMI and AMIDA were two large EU projects that focused on using speech technology (in addition with other methods) for augmenting and supporting meetings [11]. Here ASR is in the background, pervasively capturing information and restructuring it to help users record, interpret and track the meetings they are part of. Recently CereProc released an iOS application (MyMyRadio aka Noozfeed), which aimed at aggregating social media and news and using characterful synthesis to present it as a personalised radio station[1].

These are examples where speech technology have not been used in a conventional dialogue setting, or for simple command, control and notification. There is a hunger within speech technology to investigate novel ways of using the technology and a creativity within HCI to explore new avenues of interaction with this technology. By broadening the scope of speech related work we can reignite the passion of collaboration.

Toolkits and Frameworks
Current ASR and speech synthesis APIs provide only limited functionality. For example, it is not possible to retrieve a lattice or word-confusion network unless you compile and train your own ASR engine (such as PocketSphinx). This requires considerable expertise in order to achieve acceptable recognition rates. The current inflexible speech APIs limit the amount of innovation that can be realistically achieved in an HCI research project without direct involvement of speech experts.

To lower the barrier of entry and to encourage more adventurous and rich user interfaces leveraging speech technologies, the HCI and speech communities could jointly design flexible and rich speech toolkits and frameworks, promoting innovative speech applications by enabling HCI researchers to tweak the inner workings of the algorithms and access the entire hypothesis space of the ASR engine without the need to build and train an engine from scratch. However, for such a project to succeed, HCI and speech researchers will need to start talking again.

Dip your Toe in the Water
Find a small program grant and develop a little project with some local speech engineers, something easy and fun. Maybe give a talk, maybe ask them to present some new work to you. Once both parties start to understand and become up to date with each other's research the possibility of making beautiful music together might emerge.

Conclusion
It started with a quick drink, just to talk about old times. Soon, we started talking about what we'd been doing since we split up. How the beautiful GUI became so needy and invasive, how we flirted with touch and gesture. And once you get talking you remember how it used to be, all the fun we had, how we used to talk about getting into robotics together.

So we met up again for dinner. We walked back together, you invited me in for coffee. Well, you know... you should never say never again. This time it will be different. This time there is an opportunity for real respect, understanding, and yes, maybe love.

Acknowledgements
This research was funded by the Royal Society through a Royal Society Industrial Fellowship.

References

[1] Aylett, M., Vazquez-Alvarez, Y., and Baillie, L. Evaluating speech synthesis in a mobile context: Audio presentation of facebook, twitter and rss. In *Proceedings of ITI2013* (2013).

[2] Baldwin, R. 12 most annoying technologies as chosen by wired commenters. *Wired Magazine* (10 2012).

[3] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. *Embodied Conversational Agents*. Harper Collins, MIT Press, 2000.

[4] James, F., Lai, J., Suhm, B., Balentine, B., Makhoul, J., Nass, C., and Shneiderman, B. Getting real about speech: overdue or overhyped? In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, ACM (2002), 708–709.

[5] McKeown, G., Valstar, M. F., Cowie, R., and Pantic, M. The semaine corpus of emotionally coloured character interactions. In *ICME 2010*, IEEE (2010), 1079–1084.

[6] Mistry, P., and Maes, P. Sixthsense: a wearable gestural interface. In *SIGGRAPH ASIA*, ACM (2009), 11.

[7] Morris, D., Saponas, T. S., and Tan, D. Emerging input technologies for always-available mobile interaction. *Foundations and Trends in Human–Computer Interaction 4*, 4 (2010), 245–316.

[8] Munteanu, C., Jones, M., Oviatt, S., Brewster, S., Penn, G., Whittaker, S., Rajput, N., and Nanavati, A. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, ACM (2013), 2459–2464.

[9] Ni, T., and Baudisch, P. Disappearing mobile devices. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, ACM (2009), 101–110.

[10] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* (2011).

[11] Renals, S., Hain, T., and Bourlard, H. Recognition and understanding of meetings the AMI and AMIDA projects. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, IEEE (2007), 238–247.

[12] Shneiderman, B. The limits of speech recognition. *Communications of the ACM 43*, 9 (2000), 63–65.

[13] Starner, T. E. The role of speech input in wearable computing. *Pervasive Computing, IEEE 1*, 3 (2002), 89–93.

[14] Vazquez-Alvarez, Y., and Brewster, S. A. Eyes-free multitasking: the effect of cognitive load on mobile spatial audio interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2011), 2173–2176.

[15] Vertanen, K., and Kristensson, P. O. On the benefits of confidence visualization in speech recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2008), 1497–1500.

[16] Vertanen, K., and Kristensson, P. O. Parakeet: A continuous speech recognition system for mobile touch-screen devices. In *Proceedings of the 14th international conference on Intelligent user interfaces*, ACM (2009), 237–246.

[17] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. The HTK book. *Cambridge University Engineering Department 3* (2002), 175.

Commentary

For alt.chi paper
None of a CHInd: Relationship Counselling for HCI and Speech Technology

Benjamin R. Cowan
HCI Centre
School of Computer Science
University of Birmingham
Edgbaston Campus
Birmingham, B15 2TT
b.r.cowan@cs.bham.ac.uk

I must commend the authors for writing a very thought provoking and timely paper. I think the paper highlights very well that work from the speech synthesis and recognition communities is not well integrated into the CHI literature. Yet there is also work in the area of human-computer dialogue that is also notable by its relative absence. Brennan (e.g. Brennan, S., The grounding problem in conversation with and through computers. In S. R. Fussell & R. J. Kreuz (Eds.), *Social and cognitive psychological approaches to interpersonal communication*. Hillsdale, NJ: Lawrence Erlbaum, (1998), 201-225.), Bell (e.g. Bell, L. and Gustafson, J. Interaction with an animated agent in a spoken dialogue system. Proceedings of the Sixth European Conference on Speech Communication and Technology, ISCA (1999), 1143–1146.) and more recently Branigan (e.g. Branigan, H.P., Pickering, M.J., Pearson, J.M., McLean, J.F., and Brown, A. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. Cognition 121, 1 (2011), 41–57.) have published highly relevant research discussing and exploring the nature of phenomena we see in our linguistic interactions with computers. Yet although CHI would seem to be a natural home for this work in terms of impact, the themes of the work have been more fully embraced by psychology in understanding human-human dialogue interactions, than in HCI.

It seems to me that researching speech interfaces as dialogue actors, looking at speech interaction through the prism of dialogue research, could be a useful neutral ground for the relationship counselling needed between the HCI and speech technology fields, and an area where HCI could give a strong contribution. What strikes me in the speech technology research I have been exposed to, as supported by some of the comments and reviews for the submission, is the constant striving for error rate reduction to try and improve speech interactions, to the detriment of seeing the user as a psychological entity in the interaction. The CHI community can add to this a rich theoretical and practical understanding of why we interact the way we do with speech interfaces and how system design impacts that. This sort of work could not only lead us to build theories of human-computer dialogue as well as practical design knowledge, but could also be incorporated with the recognition side of speech technology systems to increase efficiency, naturalness and indeed reduce error.

To increase the emphasis of the HCI side of speech technology, I feel it comes down to the need for 1) a heightened awareness of human-centered disciplines that can help us understand and develop theoretical views of this interaction and 2) a methodological awareness of how to go about studying this interaction, developing novel methods as well as "borrowed" methods from other disciplines to build on their previous work. As highlighted in the paper, 100% recognition and natural speech synthesis are not needed if you know what the user is doing. I completely agree that we need to understand what the user is doing in these interactions more fully. In my view, dialogue research is an important foundation step in this understanding.