Shoehorning in the Name of Science

Jens Edlund
edlund@speech.kth.se

ABSTRACT

This provocation paper calls for a deeper understanding of what spoken human-computer interaction is, and what it can be. Its given structure by a story of humanlikeness and fraudulent spoken dialogue systems - specifically systems that deliberately attempts to mislead their interlocutors into believing that they are speaking to a human. Against this backdrop, a plea that conversational user interfaces are viewed from the perspective of conversation and spoken interaction first, and from the perspective of GUIs and interface design second, lest we impose the limitations of one field onto the possibilities of another, rather than the other way around.

CCS CONCEPTS

• **Applied computing**; • **Computing methodologies** → *Artificial intelligence*; • **Human-centered computing**;

KEYWORDS

spoken interaction, speech science, speech technology

ACM Reference Format:
Jens Edlund. 2019. Shoehorning in the Name of Science. In *CUI '19: Conversational User Interfaces, August 22–23, 2019, Dublin, Ireland.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

This is a provocation paper, so it should be a paper that explores "controversial, risk taking or nascent ideas". It is submitted to be presented at the very first instalment of the International Conference of Conversational User Interfaces. This context seem to call for something refreshing, something new. Yet I will use a not very new, but quite true, story to provide the backdrop for my plea. As a result of this choice of mine, the text may sound surreptitiously similar to a history lesson. If it does, so be it - maybe one is called for. (Relax, it says âĂIJprovocation paperâĂİ on the box!)

2 THE STORY

2.1 Prelude: an insight

The story starts with an insight. Anyone who has built speech interface from the ground up will recognise the inevitable question "How do we get the text out of the speech?" The same holds for just about any area where people attempt to work with speech computationally. A majority of linguists taking an interest in speech takes the same route. So: we often attempt to encode speech as text, through some manner of transcription. Here is the insight: speech is not text. A transcription of the words spoken is a very poor representation of a spoken interaction. It misses nearly everything of importance: the roles of timing, prosody, other modalities, the surroundings, the environment, and the situation in which the interaction evolves. These are all lost, and more generally, the emergent, interactional and social nature of speech is gone as well.

2.2 Capitalising on expectation

Armed with this insight, we looked for tasks to test spoken interactions that did not rely on lexical content. In the early 2000s, we took inspiration from Nigel Ward [11] and built a Hummer, a system that simply inserted acknowledgements at reasonable points in a person's narrative [4]. The system responded to non-lexical cues, mainly prosody and pauses. It soon became clear that this system was able to stand-in for a person, for a while, in phone calls that were already started. If you doubt this, just think of the way we sound when we talk to someone who spends a little too long on a narration on the phone. Most of us are quite able to maintain such a conversation more or less without listening, but by simply emitting sounds as if we were listening, while we at the same time read the newspaper or engage in some other activity that is more interesting to us. For those curious, we later verified effects of timely feedback in a spoken human-computer interaction were verified in a large-scale experiment in [7].

If our Hummer provided encouraging acknowledgements such as "uh-huh" when the narration slowed down, our next mini-project, a Husher, inspired by [9], did the opposite. It would read a script, line by line, but leave an opportunity for an interlocutor to respond after each line. If no response came, the Husher would simply read the next line after a brief pause. If there was a response, however, the Husher would immediately and loudly hush the interlocutor before continuing. The system was extremely efficient. It turns out that it is very, very hard to continue speaking when one is being hushed, even when one is being hushed by a mindless automaton.

The general consensus about talking machines, then and now, is that conversing freely about anything is too complex a task. Consequently systems do better when the expectations of what they can handle are restricted to a specific domain. The successes of our Hummer and Husher pointed us to a somewhat dramatic interpretation of this idea: Given sufficient control of the situation (and consequently of the human interlocutors' expectations), we might be able to exceed Cassel's "machine that acts human enough that we respond to it as we respond to another humanâĂIJ [5] and build a machine behaving sufficiently similar to a human, in that situation, to be perceived as a human. We believed this to be true within the limits of speech technology state-of-the-art in the early 2000s. No deep learning involved.

CUI 2019, August 22–23, 2019, Dublin, Ireland

ACM ISBN 978-1-4503-7187-2/19/08.
https://doi.org/10.1145/3342775.3342794

Note that this is neither the imitation game described by Alan Turing [10], nor one of its rather different "Turing test" popularisations, although it bears some resemblance. But to be honest, our statement was considerably less bold than one might think. You may recall that game people used to play with their answer phones?

"Yeah? [...longish... pause] Got ya! You have reached the answer phone of Steven Smith [...]"

That, briefly, behaved "sufficiently similar to a human in that situation that it was perceived as a human". It did so by virtue of having complete control of its situation, and through that, the expectations. Be it for but a moment.

In 2005 (that is the earliest mention I can find in emails), I applied these insights and built a machine specifically to waste time for telemarketers. It was a hobby project intended in part to experiment with designing humanlike talking machines for specific situations, but also to scratch an itch: the prevalence of telemarketers. I used a fresh "Skype in" number that I planted in a few well-chosen web-based order forms to quickly get it into the telemarketing databases. This worked a charm.

The telemarketers' role is very well defined. They work from scripts, and are taught to always stay on with people who are actually willing to talk. This situation is very limiting and highly conventionalised. The telemarketer trap I built capitalised on this in the simplest of manners.

It would start out with the answer phone trick: "yeah?". As soon as the telemarketer spoke, it interrupted, Husher style, with something like "oh sorry wait wait wait just one minute i need to get this out of my hands!" in a thoroughly stressed out voice. It then went silent. It would wait, in silence, until the telemarketer spoke again. This could be a while - often up to 60 seconds. As soon as the telemarketer spoke again (e.g. "uh, hello?") the Husher-style interruption was triggered again, returning another canned response with similar meaning. This went on as long as the telemarketer stayed on, or maximally for ten turns, after which I played a recording of loud swearing as a large glass object broke against the ground. The system then finished by rapidly stating "i really have to go, i'm very sorry, you're going to have to call back later!" before hanging up. In the few days I tested the system, a fair proportion of telemarketers stayed on for all ten turns, and more then one did in fact call back later, although they hung up quickly the second time. Although I was uncomfortable talking about this with outsiders for some time, concerned about the ethics involved, I then started mentioning it in talks. The feedback was very positive, and I have been comfortable writing about it for years now. Times change.

The success of this machine made us look for other, similar situations, and shortly after, I and a few colleagues who shall remain nameless built a telemarketer, which could call up and pretend to represent a charity. We built around poor quality speech synthesis by using a Steven Hawking style voice and weave its use into the background story with which the fake telemarketer opened the conversations. It even became a selling point in the narrative. The system produced speech incrementally, and would allow the human to barge in to take the turn. It would then refrain from responding, and rather go back to its scripted routine. Much like any telemarketer. We all agreed that this system, however, was far too unethical to use on unsuspecting recipients, and only tried it out on colleagues a few times. It did the trick at least once.

2.3 Closing: times change

I have used examples from my personal experience in part because they were early attempts, but more importantly because with these, I have full insight. They were controversial to us when we worked on them, for various reasons. Today, they might still stir up some debate, perhaps, but not to the extent that their development is stifled in any way. This type of conversational machine that was pioneered by Nigel Ward et al., the Hummer and the Husher, is its own research area, more or less: active listening, with several large international projects focusing on little else. A much more elegant and well-designed software that traps telemarketers is available through the Phone Pirate Company [2], and has been showcased both on a TED talk [3] and on the television show the Dragon's Den [1]. And telemarketing bots, of course, are no more ethical now than they were then, yet they are legion. And they do work quite well.

3 PROVOCATION

I will now attempt to project this story into a short, succinct provocation, and to tie back to my introduction.

It is worth noting that we call this meeting "conversational user **interfaces**". Ever since the unfortunate anthropomorphism debate instigated by Ben Shneiderman and Pattie Maes way back when [8], HCI seems stuck in a fallacy: Steeped in a GUI tradition, it equates "interaction" with "manipulation of an interface". And the calls for theory and methodology surrounding spoken human-computer interaction are often calls for a reasonable, fair and methodologically sound comparison between, say, a physical button or a GUI on the one hand and a speech control on the other.

Occasionally, this is precisely how speech technology is used, and these calls make sense. In the vast majority of cases where speech (and other natural languages; note that writing really is not a natural language though) excels and provides exciting solutions, however, there simply is no keyboard or mouse based counterpart.

If the interpretation of "conversational user interfaces" on the CUI stage is to be "speech interfaces that substitute a keyboard, mouse or button with a speech command", then (a) let us be clear about that fact, and (b) let us truly limit ourselves to precisely that. We can then make minor adaptions to existing HCI paradigms to accommodate for some of the peculiarities speech. This is not a bad path to go down. The world is in dire need of better design of this type of speech interface, and it is a viable solution to many real-world problems, most notably in accessibility.

But we can do better. We can view conversational user interfaces more broadly, and include hummers, hushers, fraudulent bots, entertainment systems, companions, game characters, teaching assistants, life loggers, practice patients, patient tutors, and the rest of the endless range of possibilities afforded by spoken interaction. And if we do this, we should take pages from the books of other fields, where these applications have been studied extensively, from technical, pragmatic, and theoretical standpoints; refrain from passing judgement without understanding the issues at hand; and be very weary about shoehorning these technologies into methods that are clearly a poor fit.

4 ACKNOWLEDGEMENTS

To the extent that they annoy, the opinions expressed here are my own. For whatever crumbles of value you may find in this text, I owe thanks and gratitude to the long lines of pioneers whose efforts within speech science and speech technology have put us where we are today - from Thomas Edison, who immediately upon inventing the first machine that could accurately record and replay speech saw more clearly its potential than many do today[6]; through Gunnar Fant, who saw the value in marrying speech science and speech technology; to the many current researchers and scholars with whom I have had the privilege to work. Special thanks to Speech, Music and Hearing at KTH Royal Institute of Technology, for remaining innovative and inspirational for more than 6 decades and counting.

REFERENCES

[1] 2019. Shark Tank, season 10, eposide 15.

[2] Roger Anderson. [n. d.]. Jolly Roger Telephone | Revenge Has Never Been So Sweet. https://jollyrogertelephone.com/

[3] Roger Anderson. 2016. Telephone spam/scam problem? Bring in the robots. | Roger Anderson | TEDxNaperville. https://www.youtube.com/watch?v=UXVJ4JQ3SUw

[4] Jonas Beskow, Rolf Carlson, Jens Edlund, Björn Granström, Mattias Heldner, Anna Hjalmarsson, and Gabriel Skantze. 2009. Multimodal Interaction Control. In *Computers in the Human Interaction Loop*. Springer London, London, Chapter 14, 143–157. https://doi.org/10.1007/978-1-84882-054-8_14

[5] Justine Cassell. 2007. Body Language: Lessons from the Near-Human. In *Genesis Redux: Essays on the history and philosophy of artificial life*, Jessica Riskin (Ed.). Chicago, Chapter 17, 346–374. https://doi.org/10.7208/chicago/9780226720838.003.0017

[6] Thomas Alva Edison. 1878. The phonograph and its future. *The North American Review* 126, 262 (1878), 527–536. https://doi.org/10.1038/018116g0

[7] Joakim Gustafson, Mattias Heldner, and Jens Edlund. 2008. Potential benefits of human-like dialogue behaviour in the call routing domain. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 5078 LNCS. Springer Berlin Heidelberg, Berlin, Heidelberg, 240–251. https://doi.org/10.1007/978-3-540-69369-7_27

[8] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61. https://doi.org/10.1145/267505.267514

[9] Nikko Ström and Stephanie Seneff. 2000. Intelligent barge-in in conversational systems. In *Sixth International Conference on Spoken Language Processing (ICSLP 2000)2000*. Beijing, China, 652–655. https://www.isca-speech.org/archive/icslp{_}2000/i00{_}2652.html

[10] Alan M. Turing. 1950. Computing Machinery and Intelligence. *Mind* LIX, 236 (1950), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[11] Nigel Ward and W. Tsukahara. 1999. A Responsive Dialog System. In *.Machine Conversations. The Springer International Series in Engineering and Computer Science, vol 511.* Springer US, Boston, MA, Chapter 14, 169–174. https://doi.org/10.1007/978-1-4757-5687-6_14