Conversational Agents in Experiential and Ludic Design

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

ABSTRACT

An alternative to traditional user centred design is to focus on the role of experience and ludic elements in human computer interaction. We argue that conversational agents (CAs) are a powerful design tool for adding ludic elements to a system such as irony and personification. Furthermore, the ability to personify devices offers designers a means to 'superbrand' as personification reaches deep into the human psyche creating a powerful felt experience of technology. We consider the importance of design in a CA interface and as an example we consider how an audio book can be used to create a conversational agent, and explore the application of such an agent as a cultural probe, a personal assistant, and as a means of subverting a sat-nav application.

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

Ambiguity is a core part the complexity and subjectivity of our lives. For thousands of years, art, music, drama and story telling has helped us understand, come to terms with, and express the complexities of our existential experience. Technology has long played a pivotal role in this artistic process, for example the role of optics in the development of perspective in painting and drawing, or the effect of film on story telling.

As information technology becomes ever more present, and ever more powerful in mediating our relationship with the world, it becomes part of this ambiguous sense of experience. In response, human computer interaction (HCI) research has begun to look at alternatives to the dominant approach of user-centred design in order to refocus on the emotional and aesthetic elements of technology. Two alternatives to the traditional HCI approach are ludic design, which focuses on the importance of encouraging playfulness in a design[15], and experience-centred design, which focuses on the sense of experience that a system would like to engender in a user[14]. In these design approaches, ambiguity, normally avoided in interface design, can be harnessed to encourage intrigue, mystery and delight[5].

However, current computer interfaces that help us understand, come to terms with, and mediate the explosion of electronic data, and electronic communication that now exists are generally limited to the mundane. Whereas the ability to get the height in metres of Everest is a trivial search request (8,848m by the way from a Google search), googling the question 'What is love?' returns (in the top four), two popular newspaper articles, a YouTube video of Haddaway and a dating site. It is, of course, an unfair comparison. Google is not designed to offer responses to ambiguous questions with no definite answers. In contrast, traditional forms of art and artistic narrative have done so for centuries.

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 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.

PERSONIFICATION

Personification: the attribution of personal qualities; especially : representation of a thing or abstraction as a person or by the human form. Merriam-Webster.

Personification is the act of giving a non-human object human qualities or abilities. *Personification Technologies* [1] can be regarded as a combination of speech technology, multimodal interfaces, embodied conversational agents, knowledge representation and inference and human language technology, that are required to produce an agent which can be personified by a user during a task.

Users readily adopt a social view of computers [16] and previous research has shown how this can be harnessed in applications such as: health advice [2], tutoring [13], or helping children overcome bullying [6]. These applications attempt to simulate the characteristics of a person (allowing *personification*), so that users can interact using the day-to-day strategies of human communication.

Speech technology is a key enabling technology for CAs. Without speech synthesis CAs cannot produce dynamic output, and are limited to pre-recorded prompts. Without dynamic output they cannot react appropriately to users. Without speech recognition the system cannot converse and interpret input speech.

IMPORTANCE OF INTERFACE DESIGN

Putting together the different components of a CA is complex and requires many design choices. Without a high level view of what the CA should achieve and what sort of user interactions are envisaged a CA will not perform successful. Interface design must play a key role in choosing the role of different output and input technologies, tuning the underlying technologies to the CA system design, and dictating the evaluation strategy.

Choosing Technologies

Language technologies can play a varied role in a CA system. However current systems have tended to be all or nothing. In fact there is enormous scope for mixing and matching technologies in order to fulfil a design objective. For example, perhaps speech recognition is not required, typed input may be satisfactory, or gestures or changes in facial expressions. Alternatively perhaps speech synthesis can be replaced with sound effects or movement from some sort of embodied device. Pre-recorded prompts might be seamlessly integrated with dynamic content rendered with speech synthesis, multimodal input might support recognition systems.

Inevitably if such systems are designed by speech and language engineers these technologies will dominate the resulting system. However HCI, with its broader interest in user experience and other interface technologies can offer a much richer design environment where speech and language technology becomes part of a more compelling whole.

Tuning Technology

If we can predict what a user will say, speech recognition is easy. Often speech recognition is added as a black box module to a CA. No work is carried out to constrain the language model, or train models from data similar to the input data the system will be exposed to. Doing these things will make the recognition better, also tuning the recognition to give very best result for the application rather than being tuned to a very best word error rate will have a more direct impact on the final user experience.

If we can predict what a system will need to say speech synthesis is easy. In fact if there is no dynamic information to render then we can use pre-recorded prompts. Often the requirement will be somewhere between the two. Designing and building the synthetic voice with the application in mind will produce much more natural and accurate output. More fundamentally, if we know in advance what sort of personality we want the system to have then we can customise the voice to realise this design criteria.

Speech technology engineers have no incentive for producing customised systems that will perform exceptionally in a specific design without designers demanding such systems. This requires resources, collaboration and the willingness to remove off-the-shelf black box modules and replace them with something more specific and more appropriate. In turn this requires compelling design ideas that can be effectively realised with the help of this technology.

Evaluation

Traditionally, speech technology has tended to evaluate recognition and synthesis outwith an application context. However the tuning of technology should be linked to the evaluation of an overall system. This can vary from the trivial (e.g. a system can't pronounce words used by the language system so they are added), to the more profound (e.g. allowing users to re-score recognition results to allow adaptive training). HCI methodology offers a framework for relating system wide evaluation to user experience and thus a means of tuning the system appropriately.

CASE STUDY: A TALKING BOOK

In order to explore the idea of ambiguity in a language based interface we will look at a concrete example: a CA based on a novel.

A novel has a style and often a strong narrative voice. For example Pride and Prejudice has a strong narrative voice that we associate with all of Austin's novels. Some of this style is connected with the use of language and some with the themes and cultural context of the novel. By leveraging the user's knowledge of a well known novel we could create a strong personified character that can converse in the manner of the book, and can be identified as being a personification of the book.

Why is this an interesting thing to do? By basing our CA on a well known work of fiction we give it a cultural context as well as offering users a strong set expectations on the experience of interacting with such a CA. This it turn can offer designers a very concrete sense of what role personification may or may not play in a system.

Why is this a useful thing to do? In a conventional sense, perhaps it is not useful. It will not tell you the height of Everest any more accurately than Google search, nor help you find a pizza restaurant any better than Google maps. However, it is a useful thing to think about. Considering such a system can give designers an insight into how personification using speech technology may fit into ambiguous and experiential interfaces, and what technical challenges there may be in speech and language technology to make such a system a reality. Such a system would require, on top of a standard speech recognition system, and language understanding system, a characterful natural language generator, a customised speech synthesis system. In the next sections we consider the

background of these technologies and how they might fit together in such a engineering project.

Characterfull Natural Language Generation

Recent work in language generation [12, 8] has begun to look at ways of explicitly controlling language generation which automatically conveys a sense of personality and character. PERSONAGE[12], originally built as a restaurant recommender system, has been parametrised to allow the control of a set of language generation features that reflect differences in personality as described by *the big five*[4]. In Lin & Walker[11] this approach was extended by using movie scripts from the IMSDb website to automatically generate differences in character for the SpyFeet role playing game. By using the *the big five* as a basis for parametrisation PERSONAGE could control some the perceived personality traits of the conversational agent, for example how extrovert or introvert it appeared. However it did not harness other elements of character, for example the use of dialect, idiosyncratic fillers and catch phrases.

A simple language model, such as a ngram model, can model much of the idiosyncratic content in text. Within literary criticism such models have been used to characterise a writers style, or a characters way of speaking[17]. One means of applying such statistical models in language generation is to use an overgenerate and scoring method[8]. In this concept system the language generation generates many alternative possible outputs and each is compared to a statistical model to select the most appropriate. The advantage of this approach is that it can be used to control stylistic difference less strongly connected with a big five analysis such as individual preferences in rhetorical structure. However as Walker points out[12] this is computationally expensive.

In this previous work the objective is to take a character trait and use it to generate language in a new domain, such as a restaurant or film review. An alternative approach might be to take a work of art, such as novel, and use this to produce both language style **and** dialogue content. In this approach the content would tend to support the language style and the perceived characterisation. For example *Treasure Island* might be used for a pirate like speech style as well as initiating dialogue on the subject of doubloons and buried treasure.

New domains could then be grafted more gradually onto such a conversational agent allowing our pirate conversational agent to also discuss the height of Everest and where to buy pizza within a dialogue context of buried treasure. In doing so we merge the ill defined ambiguous content of a novel with our factual information. Potentially offering a more satisfying answer to the question 'What is Love?'. In this case perhaps the wide ocean, the wind in your sails and the ability to bear down on a helpless Spanish Galleon.

Characterfull Speech Synthesis

Modern speech synthesis uses a corpus approach to produce a voice where a corpus of audio from a target voice is used to either produce a statistical model of the speakers speech (parametric speech synthesis) [19], or used directly to be recombined into target speech (unit selection speech synthesis)[7].

Audio books offer a rich source of expressive speech and have been a focus of recent speech synthesis research looking at retaining vocal style and character. Blizzard is a speech synthesis challenge were different systems are built based on the same input audio and evaluated within an identical framework[10]. In 2012/13 one of the tasks was to build a voice from a very large corpus of audio book recordings.

Audio book data is termed *found data* in speech synthesis research, to distinguish is from data that is recorded specifically for creating a synthetic voice. Producing high quality synthesis from found data is challenging: data may not exactly match accompanying text, speakers may change there voice quality and vocal style to give the impression of different characters or to convey drama and excitement and recording environments may vary.

However this approach has been successfully applied to mimic well know speakers, such as Barack Obama[1] and George Bush[2]. It was also a means of recreating the film critic Roger Ebert's voice for his own use[9].

When a voice talent reads an audio book they typically act in order to read the text in an appropriate vocal style, and with appropriate expressive speech to reflect the underlying linguistic content. By using an audio book as a source for a voice it is possible to model the perceived character of the speaker's voice and to control it with reference to linguistic content.

There is an interesting parallel between using an audio book to create a speech synthesis voice, and using the text of the book to effect the character of the linguistic output from a dialogue system. Both systems can create a sense of character and, potentially, if combined could reinforce this sense of character.

DEPLOYING A PERSONIFIED CA

Personified CA as a Cultural Probe

Conversation Piece [18], a speech based interactive art exhibition used the idea of of a CA to explore visitors relationship with art. In this work a podium with a small sculpture on top of it would ask the visitor what they thought about the art. The different podium's had different dialogue strategies that presented different personalities. Visitors were very willing to engage in conversation in this context and found the system playful and thought provoking. A personified CA will produce a strong user reaction (though not always a positive one!). Art installations are an ideal way of challenging a user with surprising or unusual personalities within a cultural context, for example: we could contrast a CA based on Trainspotting by Irvine Welsh with one based on Eat, Pray, Love by Elizabeth Gilbert. These types of cultural probes can give designers a clearer idea of the possibilities of this type of interactive technology.

Personified CA as a Personal Assistant

The conventional approach to an automated personal assistant is to produce a CA with a helpful, attitude free, educated

[1] www.nutbots.net/talking_head

[2] www.idyacy.com/cgi-bin/bushomatic.cgi

female voice. But what if the personality of the assistant is contrary and opinionated? You can already replace sat-nav speech output with celebrity or character based voices. Why not replace Siri with a Pirate? Such a strongly personified CA might not be appropriate but it may enrich the experience of using the technology.

Subverting Sat-Nav

In Android it is possible to replace the default synthetic voice. The synthesiser will accept text and then produce the audio output. This gives us the opportunity to subvert any application on Android which uses speech output. Rather than speak the text the application gives to the system, we can re-phrase it using characterful language generation and then speak it with a characterful synthesis voice. So "Turn left" might become "Port side my hearties!". We would urge readers to experiment with this idea by taking, for example, the Cereproc Glaswegian voice (available on Google Play for free) to replace the sat-nav in their current phone. It will then give you directions in a threatening manner.

CONCLUSION

In this paper we have argued that speech technology can play a vital part in ludic and experiential designs. Personification is a powerful tool in producing systems which encourage strong reactions from users. Currently there is very little work on how we can leverage this technology to produce compelling and exciting systems. Bland personalities in CAs are safe, complex and ambiguous personalities are not. But in our view HCI research should embrace dangerous and thought provoking design ideas. We have given an overview of the technology required to implement a personified CA. The task is challenging but tractable. An open question remains in how personification might fit into designs and systems, and how multi-modal approaches may support and complement personification language technologies.

INTEREST IN THE WORKSHOP

Dr Matthew Aylett is currently a Royal Society Research Fellow at the University of Edinburgh focusing on the relationship between personification and speech synthesis. He sees this workshop as an ideal forum to discuss his ideas from a speech technology background, with international HCI experts to see how the disciplines might work together to produce the next generation of interactive systems.

REFERENCES FORMAT

REFERENCES

1. Benyan, D., and Milval, O. Landscaping personification technologies: From interactions to relationships. In *Proceedings of Human Factors in Computing Systems* (2008), 3657–3662.
2. Bickmore, T., Pfeifer, L., and Jack, B. Taking the time to care: Empowering low health literacy hospital patients with virtual nurse agents. In *SIGCHI Conference on Human Factors in Computing Systems* (2009).
3. Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. *Embodied Conversational Agents*. Harper Collins, MIT Press, 2000.
4. Funder, D. C. *The Personality Puzzle*, second ed. W. Norton and Company, New York, 1997.
5. Gaver, W. W., Beaver, J., and Benford, S. Ambiguity as a resource for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03 (2003), 233–240.
6. Hall, L., Jones, S., Paiva, A., and Aylett, R. Fearnot!: providing children with strategies to cope with bullying. In *8th International Conference on Interaction Design and Children* (2009).
7. Hunt, A., and Black, A. Unit selection in concatanative speech synthesis using a large speech database. In *ICASSP*, vol. 1 (1996), 192–252.
8. Isard, A., Brockmann, C., and Oberlander, J. Individuality and alignment in generated dialogues. In *4th International Natural Language Generation Conference* (Sydney, Australia, 2006), 22–9.
9. Jones, C. What roger ebert cannot say. *Esquire Magazine* (2010).
10. King, S., and Karaiskos, V. The Blizzard Challenge 2012. In *Proc. Blizzard Challenge Workshop* (Barcelona, Spain, September 2012).
11. Lin, G. I., and A.Walker, M. All the worlds a stage: Learning character models from film. In *Conference on Artificial Intelligence and Digital Entertainment*, AAAI Press (2011).
12. Mairesse, F., and Walker, M. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics* (2011).
13. Massaro, D., Liu, Y., Chen, T., and Perfetti, C. A multilingual embodied conversational agent for tutoring speech and language learning. In *Interspeech* (2006).
14. McCarthy, J., and Wright, P. *Technology as Experience*. MIT Press, 2004.
15. Morrison, A. J., Mitchell, P., and Brereton, M. The lens of ludic engagement: evaluating participation in interactive art installations. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07 (2007), 509–512.
16. Nass, C., Steuer, J., and Tauber, E. Computers are social actors. In *SIGCHI conference on Human factors in computing systems* (1994).
17. Siemens, R., Schreibman, S., and Unsworth, J., Eds. *A Companion to Digital Humanities*. Blackwell, Oxford, 2004.
18. Wright, A., Linney, A., Evans, A., and Lincoln, M. Conversation piece: a speech-based interactive art installation. In *ACM Multimedia* (2007), 377–378.
19. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. 6th ISCA Workshop on Speech Synthesis (SSW-6)* (Aug. 2007).