The CogSIS Project: Examining the Cognitive Effects of Speech Interface Synthesis

Leigh Clark
University College Dublin
Dublin, Ireland
leigh.clark@ucd.ie

João Cabral
Trinity College Dublin
Dublin, Ireland
cabralj@scss.tcd.ie

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

Speech interfaces are becoming a more common dialogue partner. With the growth of intelligent personal assistants, pervasive and wearable computing and robot-based technologies, the level of spoken interactions with technology is unprecedented. However, while the technological challenges around the production of natural synthetic voices have been widely researched, comparatively little is understood about how speech synthesis affects user experience and behaviour. The CogSIS Project examines the psychological and behavioural consequences of synthesis design decisions in human interactions with speech technology. In particular, we explore how design decisions around politeness, accent, naturalness and expressivity impact the assumptions we make about speech interfaces as communicative actors (i.e. our partner models). The project fuses knowledge, concepts and methods from psycholinguistics, experimental psychology, human-computer interaction and speech technology to 1) understand how synthesis design choices impact users' partner models, 2) how these choices interact with partner models and impact user experience and evaluations and 3) how these choices impact users' own language production. The project will lead to a set of theory-driven practical and actionable guidelines for speech synthesis and speech interface design.

Speech interface, voice user interface, alignment, partner model, politeness, speech synthesis

1. INTRODUCTION

Speech has become a more prominent way of interacting with automatic systems. In addition to established interactive voice response (IVR) interfaces and satellite navigation systems, voice enabled intelligent personal assistants (IPAs) like Amazon Alexa, Apple Siri, Google Assistant, and Microsoft Cortana are widely available on a number of mobile and home-based devices. The technical infrastructures underpinning speech interfaces have advanced rapidly in recent years and have been widely researched (e.g. Chan, Jaitly, Le, & Vinyals, 2016). However, research on the user experience (UX) side of speech interfaces is said to be limited by comparison (Aylett, Kristensson, Whittaker, & Vazquez-Alvarez, 2014). While we are beginning to know about how people use devices like IPAs (Cowan et al., 2017; Luger & Sellen, 2016), there remain gaps in our understanding of the psychological and behavioural effects, particularly with speech synthesis design decisions. The CogSIS Project explores these concepts through a multidisciplinary approach and addresses several core areas of investigation. This paper discusses the experiments planned during the projected, focusing on partner modelling, language production of users, and the evaluation of speech interfaces.

2. PARTNER MODELS & USER LANGUAGE PRODUCTION

Previous literature in speech-based human-computer interaction (HCI) has discussed the concept of *partner models*. These are people's expectations of their partner's abilities, including how they are able to communicate. Partner models can drive users' own language production in HCI (e.g. Amalberti, Carbonell, & Falzon, 1993), similar to human-human communication (Branigan, Pickering, Pearson, McLean, & Brown, 2011; Brennan & Clark, 1996). The ongoing research in this project seeks to expand the understanding of how partner models affect user language production, and its link to several aspects of synthesis design choices such as language use and the type and quality of an interface's voice.

2.1. Using politeness strategies

One of these design choices is the use of politeness strategies – a of category linguistic strategies used to manage interpersonal relationships, often by avoiding imposition on other speakers (Brown & Levinson, 1987). Previous research has shown that people will use politeness strategies towards speech interfaces (e.g. Large, Clark, Quandt, Burnett, & Skrypchuk, 2017) that used a human playing the part of an interface in a

© 2018 Leigh Clark, João Cabral and Benjamin Cowan. Published by BCS Learning and Development Ltd.
Proceedings of British HCI 2018. Belfast, UK

Wizard of Oz interaction (see Dahlbäck, Jönsson, & Ahrenberg, 1993). These politeness strategies include apologising to the 'system' and asking permission before making requests. However, politeness strategies by users may be linked to the human voice being used in the interface design. This project will compare the extent to which politeness strategies may be mimicked or *aligned* with by users when interacting with either synthesised speech or a human voice.

2.2. Accent & Expressivity

This project also explores design choices of accent and expressivity on users' partner models and language production. Previous research has suggested there may be preferences towards the evaluation of interfaces that have similar accents to their users. This is often referred to as the *similarity-attraction effect* (Dahlbäck, Wang, Nass, & Alwin, 2007; Nass & Lee, 2000). However, less is understood regarding how the accent of speech interfaces affects users' own language production. Experiments currently being undertaken in this project compare users' language production towards Irish and American accented speech interfaces when describing objects that vary in Irish and American English dialects (e.g. 'nappy' and 'diaper'). These studies aim to understand the extent to which users' language production of these objects is affected by (a) the interface's accent and (b) users' prior knowledge of the dialectal terms for each accent. It may be the case that accent affects people's model of the computer partner's lexical knowledge and that this consideration is consequently influencing lexical choice with a speech interface.

Research in this project has also been conducted on the expressivity of synthesised speech – i.e. the personality and characteristics speech can project based on its prosodic features (e.g. surprise, joy, sadness). The project has provided some preliminary results with expressive synthesised speech. Results with talking virtual characters showed that expressive humanlike voices were perceived as more understandable, expressive, and likeable than expressive synthesised voices (Cabral, Cowan, Zibrek, & McDonnell, 2017). However, people's judgements of virtual characters (e.g. the character's appeal and credibility) did not vary depending on the voice that the character used.

3. EVALUATION OF SPEECH INTERFACES

Alongside users' language production, the CogSIS project explores users' evaluation of speech interfaces based on different design choices. These evaluations explore some of the concepts outlined in the previous section, as well as other aspects like context. For example, users will evaluate speech interfaces on concepts of trust, rapport, likeability, and assumed knowledge, and experiments will examine how synthesis design choices discussed in this paper affect these user evaluations.

3.1. Effect of Context

As well as the aspects of speech outlined in Section 2, this project will also assess the impact of context on speech evaluations. Evaluation of synthesised speech is often conducted using traditional listening tests implementing Mean Opinion Scores (MOS) of concepts such as naturalness (e.g. Sawada, Tokuda, King, & Black, 2017). While these are somewhat of an established standard, it is not fully understood how the context in which speech is presented affects users' evaluations. Recent research comparing traditional listening tests to interactions with a talking virtual avatar has suggested that actual interactions may be viable methods of evaluating speech (Mendelson & Aylett, 2017). However, the gap between listening to audio clips of synthesised speech, often in an online survey, and interacting with a virtual avatar is significant. This project examines users' evaluations of synthesised speech when *presented* with or without a context. These presented contexts refer to explaining to users what the synthesised speech they are listening to is being designed before and how it will be implemented (e.g. an IPA or navigation system). This serves as an intermediate gap between traditional non-contextual speech evaluations and actual interactions with various types of speech interfaces such as virtual avatars.

4. SUMMARY

This paper outlines the focal research areas of the CogSIS Project. This project explores the psychological, behavioural and user experience effects of design decisions around speech synthesis. In particular, this project addresses people's partner models towards speech interfaces and how this impacts their own language production. Furthermore, this project assess the differences in speech synthesis evaluations depending on the language used and the different qualities of voice presented in speech interfaces. In doing so, we aim to further contribute to theories underpinning speech-based interactions with computers, and provide practical and actionable theory-driven guidelines for speech synthesis and speech interface design.

ACKNOWLEDGEMENTS

This research was funded by a New Horizons grant from the Irish Research Council entitled "The COG-SIS Project: Cognitive effects of Speech Interface Synthesis" (Grant R17339).

7. REFERENCES

Amalberti, R., Carbonell, N., & Falzon, P. (1993). User Representations of Computer Systems in Human-computer Speech Interaction. *Int. J. Man-Mach. Stud.*, *38*(4), 547–566. https://doi.org/10.1006/imms.1993.1026.

Aylett, M. P., Kristensson, P. O., Whittaker, S., & Vazquez-Alvarez, Y. (2014). None of a CHInd: relationship counselling for HCI and speech technology (pp. 749–760). ACM Press. https://doi.org/10.1145/2559206.2578868.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(6), 1482–1493.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Cabral, J. P., Cowan, B. R., Zibrek, K., & McDonnell, R. (2017). The Influence of Synthetic Voice on the Evaluation of a Virtual Character (pp. 229–233). ISCA. https://doi.org/10.21437/Interspeech.2017-325.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4960–4964). https://doi.org/10.1109/ICASSP.2016.7472621.

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., … Bandeira, N. (2017). 'What Can I Help You With?': Infrequent Users' Experiences of Intelligent Personal Assistants (pp. 1–12). ACM Press. https://doi.org/10.1145/3098279.3098539.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz Studies — Why and How. *Intelligent User Interfaces*, 8.

Dahlbäck, N., Wang, Q., Nass, C., & Alwin, J. (2007). Similarity is More Important Than Expertise: Accent Effects in Speech Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1553–1556). New York, NY, USA: ACM. https://doi.org/10.1145/1240624.1240859.

Large, D. R., Clark, L., Quandt, A., Burnett, G., & Skrypchuk, L. (2017). Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics*, *63*, 53–61. https://doi.org/10.1016/j.apergo.2017.04.003.

Luger, E., & Sellen, A. (2016). 'Like Having a Really Bad PA': The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286–5297). New York, NY, USA: ACM. https://doi.org/10.1145/2858036.2858288.

Mendelson, J., & Aylett, M. P. (2017). Beyond the Listening Test: An Interactive Approach to TTS Evaluation (pp. 249–253). ISCA. https://doi.org/10.21437/Interspeech.2017-1438.

Nass, C., & Lee, K. M. (2000). Does computer-generated speech manifest personality? an experimental test of similarity-attraction (pp. 329–336). ACM Press. https://doi.org/10.1145/332040.332452.

Sawada, K., Tokuda, K., King, S., & Black, A. W. (2017). The blizzard machine learning challenge 2017 (pp. 331–337). IEEE. https://doi.org/10.1109/ASRU.2017.8268954.