Empirical Evidence for a Diminished Sense of Agency in Speech Interfaces

Hannah Limerick
Dept. of Computer Science
University of Bristol, UK
hannah.limerick@bristol.ac.uk

James W Moore
Dept. of Psychology
Goldsmiths, University of London
j.moore@gold.ac.uk

David Coyle
Dept. of Computer Science
University of Bristol, UK
david.coyle@acm.org

ABSTRACT
While the technology underlying speech interfaces has improved in recent years, our understanding of the human side of speech interactions remains limited. This paper provides new insight on one important human aspect of speech interactions: the sense of agency - defined as the experience of controlling one's own actions and their outcomes. Two experiments are described. In each case a voice command is compared with keyboard input. Agency is measured using an implicit metric: intentional binding. In both experiments we find that participants' sense of agency is significantly reduced for voice commands as compared to keyboard input. This finding presents a fundamental challenge for the design of effective speech interfaces. We reflect on this finding and, based on current theory in HCI and cognitive neuroscience, offer possible explanations for the reduced sense of agency observed in speech interfaces.

Author Keywords
Speech interfaces; voice commands; the sense of agency.

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

INTRODUCTION
At ACM CHI 2014 Aylett et al. discussed the ups and downs of the relationship between HCI research and speech technology [1]. They argue that disillusionment within the HCI community with speech interfaces is partly due to a mismatch of expectations. Speech technologists have often presented speech interfaces as providing a *"natural means of communication"*, whereas in reality technical limitations such as high error rates, recognition latency and issues with ambient noise can reduce their effectiveness. Aylett et al. make a strong case for significant progress in tackling these limitations. However, they also recognise that substantial non-technical challenges remain. For example, Shneiderman has argued that *"speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks"* [9]. He further argues that our understanding of the human side of speech interactions is insufficient and that there is a need to address design challenges in speech interfaces by increasing this understanding.

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-3145-6/15/04…$15.00
http://dx.doi.org/10.1145/2702123.2702379

This paper provides new insight on one important aspect of the human side of speech interactions: the sense of agency. We focus on the sense of agency when interacting with voice command interfaces. The sense of agency can be defined as the experience of controlling one's own actions and, through this control, affecting the external world [2]. It is a crosscutting experience that links to concepts such as free will, causality and responsibility. In the context of HCI the importance of agency is illustrated by Shneiderman's 7th rule for interface design, which recommends that designers strive to create interfaces that *"support an internal locus of control"* [10]. This is based on the observation that users *"strongly desire the sense that they are in charge of the system and that the system responds to their actions"*.

Agency has been extensively studied in the field of cognitive neuroscience [4]. More recently Coyle et al. have applied methods developed in cognitive neuroscience to investigate peoples' sense of agency when interacting with computers [2]. They have shown, for example, that on-body interfaces can engender a greater sense of agency than keyboard interactions. A more detailed review of early HCI research on the sense of agency is also available in [6]. In this paper we describe two experiments comparing peoples' sense of agency in voice command and keyboard interfaces. Our aim is to determine if the sense of agency when interacting with speech interfaces is different to that experienced in more traditional input methods. Our results lead us to conclude that people experience less control over their environment when interacting via speech interfaces.

INTENTIONAL BINDING
Both of our experiments use intentional binding as an implicit metric for the sense of agency. Intentional binding is the name given to a temporal phenomenon that occurs when a person takes a voluntary action that causes an outcome [3]. In this case actions are perceived to happen later than they actually did, while outcomes are perceived as happening earlier. The overall effect is a binding, whereby the interval between an action and its effect is perceived as shorter than is actually the case (Figure 1).

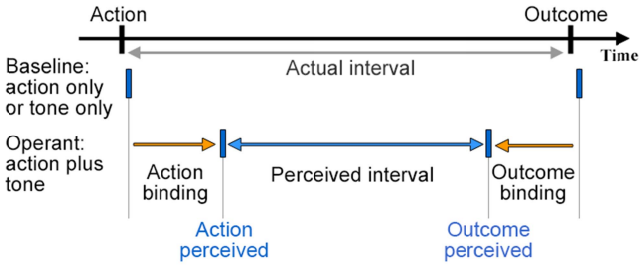

Figure 1: The intentional binding effect.
Intentional binding = action binding + outcome binding.

Research has shown that for binding to occur, actions must be intentional and must lead to an outcome. Involuntary or unintended actions have the opposite effect, with the interval between actions and outcomes perceived as longer than the actual interval. Overall, there is a strong scientific consensus that intentional binding is a robust implicit metric for the sense of agency [3, 8]. Larger binding values correlate to a greater sense of agency. Other methods to assess agency do exist, e.g. fMRI studies or self-report questionnaires. Binding studies cost significantly less than fMRI studies, are more practical in most contexts, and provide an implicit means to investigate agency. A key advantage of the binding measure is its implicit nature, sidestepping issues with explicit measures (e.g. questionnaires) such as demand effects and introspection.

EXPERIMENT 1

This experiment compared participants' sense of agency for two action conditions: a key press and a voice command. Previous research has shown that the predictability of action-outcome relationships can have an impact on the sense of agency [8]. In order to remove voice recognition errors as a potential confound we required the recognizer to have a very low error rate. We implemented our voice interface using Sphinx 4 [11], an open source, hidden Markov model-based recognition system, with continuous recognition capabilities. Our task required recognition of just one word, "Go", so our grammar file contained only this word. This approach proved effective. In the study there was no block of trials with more than 4 errors (less than 10% error rate). Previous work shows that an error rate below 10% is unlikely to cause a confound [8].

Procedure

Coyle et al. [2] describe two methods for measuring participants' temporal perception in binding experiments: the Libet Clock and interval estimation methods. The Libet clock method offers a more detailed insight into the agentic experience and is the one applied here. Participants report their perception of time by recording the position of the hand on a clock that rotates at a rate of one rotation every 2560ms. In order to measure intentional binding, baseline and operant measures are taken for both actions and outcomes. In the operant blocks actions cause an outcome. For baseline blocks only an action or an outcome occur. Table 1 shows the full set of measures taken and the calculations used for intentional binding.

A within-subjects design was used, with one independent variable: action modality - voice command or key press. Figure 1 illustrates the procedure for individual trials. Participants watched a screen showing a Libet clock and used a footswitch to initiate trials. Once initiated, the clock hand started to rotate. For key press trials, participants pressed the 'enter' key on the keyboard as an action. For the voice command condition, participants said the word 'Go'. We used the end of the utterance as the point of action in the voice condition. In both conditions participants were asked to make an action whenever they were ready.

As is common in intentional binding experiments the outcome used in our experiment was a tone. A fixed action-outcome interval of 500ms was used to ensure sufficient time for word recognition by the voice recognizer. This is a slightly longer interval than is typically used in intentional binding experiments, but it is well within the interval range for which binding is expected [8].

Blocks of trials were completed for each input condition as outlined in Table 1. Each block included 40 trials, with mean values used to determine intentional binding using the calculations shown in Table 1. This resulted in a total of 320 trials per participant. The blocks for each condition were grouped together and the order of the input conditions was alternated for odd and even numbered participants. The

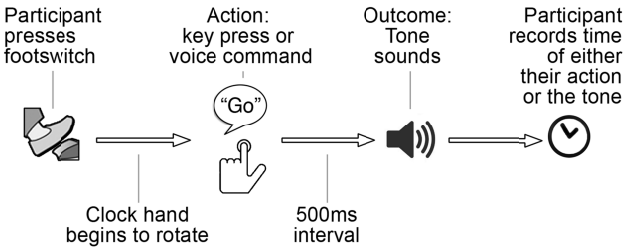

Figure 2: The procedure for trials in Experiment 1.

Measurement Blocks			
Block Type	**Action**	**Outcome**	**Participant report**
Action Baseline	Key-Press or Voice command	None	Perceived time of action
Action Operant	Key-Press or Voice command	Tone	Perceived time of action
Tone Baseline	None	Tone	Perceived time of outcome
Tone Operant	Key-Press or Voice command	Tone	Perceived time of outcome
Intentional Binding Calculations			
Action Baseline Error = actual time – perceived time			
Action Operant Error = actual time – perceived time			
Outcome Baseline Error = actual time – perceived time			
Outcome Operant Error = actual time – perceived time			
Action Binding = Action Operant Error - Action Baseline Error			
Outcome Binding = Outcome Baseline Error - Outcome Operant Error			
Total Binding = Action Binding + Outcome Binding			

Table 1. The temporal measurements, in milliseconds, and calculations used to estimate intentional binding.

	Action binding	Outcome binding	Total binding
Button	23.00ms (16.25ms)	22.57ms (34.28ms)	45.56ms (36.47ms)
Voice command	40.12ms (64.10ms)	5.78ms (39.16ms)	45.90ms (93.80ms)

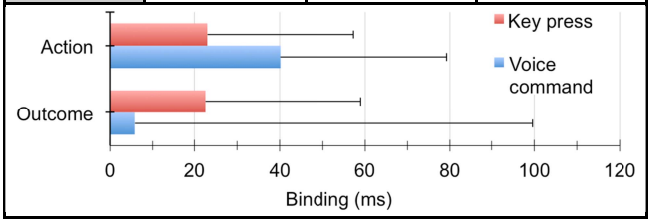

Table 2: the mean action, outcome and total binding times for Experiment 1. Standard deviations in brackets.

order of the blocks was randomised within the input conditions, but was balanced for odd and even participants.

Participants

14 participants, all right-handed and aged 20-40, took part. All had normal or corrected to normal vision and hearing. The study was approved by the University of Bristol ethics committee. All participants gave written, informed consent and received a £15 retail voucher for participating.

Results

Table 2 shows the mean action, outcome and total binding values for the key press and voice command conditions across 14 participants. The full operant and baseline measures and calculations for these binding values are included in Appendix 1 (see online auxiliary materials). To test if intentional binding occurred in either the key press or voice condition, we conducted separate 2x2 repeated measures analysis of variance on participants' perceived times for each input modality, with factors event (action vs. outcome) and context (operant vs. baseline). For the key press condition there was significant intentional binding: $F(1,13)=20.293$, $p<.001$. For the verbal command we did not find significant binding: $F(1,13)=3.112$, $p<.101$.

Analysis

Intentional binding is the shift of *both* the perceived action and outcome towards one another. This binding was present for the key-press condition, with a binding value consistent with prior literature. Binding was not found for the voice command. The mean action and outcome binding for the speech interface indicate that insufficient outcome binding occurred to elicit a significant overall binding effect.

Further analysis of the temporal measures in the voice condition revealed an important issue. Participants' action baseline error (see online Appendix) indicates that they perceived their speech action occurring 316ms before the point recorded by the computer. This is an unusually large baseline error for action estimation. Subsequent recordings suggest that saying the word "Go" takes ~300ms. Taken together this suggests that our participants perceived their speech actions as occurring at the beginning of their utterance, rather than at the end – the point we chose as the action reference point for the computer. In effect this means the action-outcome interval for the voice command was ~800ms, rather than the 500ms we intended. While intentional binding has been observed at intervals of 800ms, it is less likely than at 500ms [8]. It is thus possible that a longer action-outcome interval may explain the absence of binding in the voice condition.

	Action binding	Outcome binding	Total binding
Button	47.37ms (122.34ms)	47.75ms (117.11ms)	94.12ms (176.17ms)
Voice command	-5.70ms (94.90ms)	-4.36ms (34.51ms)	-10.06ms (95.44ms)

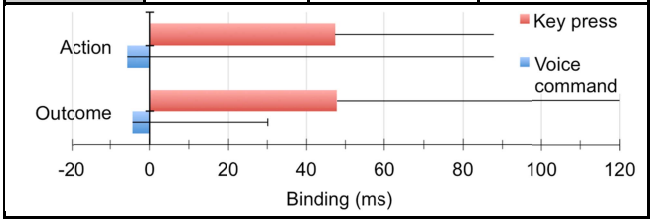

Table 3: the mean action, outcome and total binding times for Experiment 2. Standard deviations in brackets.

EXPERIMENT 2

To address our concerns with Experiment 1, a second experiment was conducted with two alterations. First, the 'StartSpeech' signal (a built-in component in Sphinx4 which indicates that speech has started) was used as the point of speech actions. This change addressed our observation in Experiment 1 that participants perceived their actions as occurring at the start of the utterance. Second, to address the possibility that a longer action/outcome interval resulted in no intentional binding for the speech condition, the action/outcome interval was shortened to 300ms. Aside from these changes the procedure was the same as Experiment 1.

Participants

Again we had 14 participants; all right-handed, aged 20-40, with normal or corrected to normal vision and hearing. The university ethics committee approved the study and participants received a £15 retail voucher. One participant was excluded from our analysis due to an equipment fault.

Results

Table 3 shows the mean action, outcome and total binding effects for the key press and voice command conditions. To test whether intentional binding was occurring, we again conducted separate 2x2 repeated measures analysis of variance tests on participants' perceived times for each input modality, with factors of event (action vs. tone) and context (operant vs. baseline). For the key press condition there was a trend toward significant intentional binding: $F(1,12)=3.496$, $p< .086$. Although the mean binding scores where higher than in Experiment 1, a large variance between participants rendered the score insignificant. For the verbal command we did not find significant binding: $F(1,12)=.144$, $p< .711$.

Analysis

Experiment 2 addressed a potential reason for the absence of intentional binding for voice commands in Experiment 1. We found that these alterations did not elicit intentional binding for voice commands. The key press condition showed a trend towards significance and given a larger sample size would likely be consistent with Experiment 1 and a wealth of prior intentional binding literature.

DISCUSSION AND CONCLUSION

Our results suggest that intentional binding does not occur for voice command interfaces. This in turn suggests that the sense of agency is lower for voice commands than for input techniques such as a keyboard. We believe this finding reveals an important underlying limitation of speech interfaces and presents a challenge for designers of speech interfaces. Users will experience a reduced sense of control over their environment when interacting via voice interfaces. Voice interfaces will feel less responsive and as a result users may experience a reduced sense of ownership or responsibility for the outcomes of their actions. Overall users will have a reduced sense that they are in charge of the system. In this context it is worth noting that the simplified speech interface used in our experiments allowed us to minimise recognition errors and latency. Therefore, even with continuing improvements in the technology underlying speech interfaces, the issue of a reduced sense of agency is likely to remain.

An obvious question that arises from our results is: Why do speech interactions provide a diminished sense of agency as compared to keyboard interactions? We are not yet in a position to offer definitive answers to this question. However we can offer two possible explanations, both of which have implications for designers.

One explanation from prior HCI research relates to the allocation of cognitive resources during tasks. It has been suggested that usability issues for speech interfaces are due to the fact that working memory is a cognitive resource that is shared between the processes of problem solving, recall and speech, and further that limb movements do not compete for the same cognitive resources [9]. This explains why humans find it difficult to speak and think at the same time, but can easily walk and talk simultaneously. This is interesting as recent research in cognitive neuroscience finds that increasing a person's cognitive working memory load reduces their sense of agency [5]. An implication of this finding is that voice command interfaces should only be deployed with care in situations that have high cognitive working memory loads, but also require users to maintain a strong sense of control.

An alternative explanation for a reduced sense of agency in speech interactions is based on a theory in cognitive neuroscience – cue integration. This theory holds that various cues surrounding actions and outcomes are integrated optimally and are weighted by their reliability to give rise to sense of agency [7]. This includes internal sensorimotor cues, e.g. proprioception, and contextual cues such as an intention to make a certain action. In [2] participants experienced significantly greater intentional binding for skin-based input than for keyboard input. The present study and [2] currently represent the only two investigations into agency and non-conventional input techniques. However from these we see a potential continuum arising. Skin-based input provides a greater sense of agency than keyboard input, which in turn provides a greater sense of agency than voice commands. It is possible that the graded degree of agency across these interfaces may be ascribed to the varying number of cues and/or the reliability of these cues. For speech interfaces the main agency cues available are auditory and proprioceptive. By comparison a keyboard offers a wider array of cues, including auditory, proprioceptive, visual and haptic. Over and above this, the skin-input modality provides further cues, through the body itself acting as the input device.

This explanation of the reduced sense of agency in our speech interface is intriguing, as it also offers a possible solution for designers. It suggests that the sense of agency in speech interfaces – or indeed any input modality - could be improved by offering users increased contextual cues (e.g. haptic feedback) regarding their interactions.

ACKNOWLEDGEMENTS

This research was funded by the Wellcome Trust Neural Dynamics PhD programme at the University of Bristol.

REFERENCES

1. Aylett, M.P., Kristensson, P.O., Whittaker, S. & Vazquez-Alvarez, Y. (2014) *None of a CHInd: relationship counselling for HCI and speech technology*. ACM CHI EA 2014. 749-60.
2. Coyle, D., Moore, J., Kristensson, P.O., Fletcher, P. & Blackwell, A. (2012) *I did that! Measuring users' experience of agency in their own actions*. ACM CHI 2014. 2025-34.
3. Haggard, P., Clark, S. & Kalogeras, J. (2002) *Voluntary action and conscious awareness.* Nature Neuroscience **5**(4) 382-85.
4. Haggard, P. & Tsakiris, M. (2009) *The Experience of Agency: Feelings, Judgments, and Responsibility.* Curr Dir Psychol Sci. **18**(4) 242-46.
5. Hon, N., Poh, J.H. & Soon, C.S. (2013) *Preoccupied minds feel less control: Sense of agency is modulated by cognitive load.* Consciousness and Cognition. **22**(2) 556-61.
6. Limerick, H., Coyle, D. & Moore, J.W. (2014) *The Experience of Agency in Human-Computer Interactions: A Review.* Frontiers in Human Neuroscience. **8**:643.
7. Moore, J.W. & Fletcher, P.C. (2012) *Sense of agency in health and disease: A review of cue integration approaches.* Consciousness and Cognition. **21**(1) 59-68.
8. Moore, J.W. & Obhi, S.S. (2012) *Intentional binding and the sense of agency: a review.* Conscious Cogn. **21**(1) 546-61.
9. Shneiderman, B. (2000) *The limits of speech recognition.* Commun. ACM. **43**(9) 63-65.
10. Shneiderman, B. & Plaisant, C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 2004.
11. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P. & Woelfel, J. (2004) *Sphinx-4: a flexible open source framework for speech recognition*.