“You, Move There!”: Investigating the Impact of Feedback on Voice Control in Virtual Environments

Mitchell Baxter
2193433B@student.gla.ac.uk
School of Computing Science, University of Glasgow
Glasgow, Scotland

Anna Bleakley
anna.bleakley@ucdconnect.ie
School of Information & Communication Studies, University College Dublin
Dublin, Ireland

Justin Edwards
justin.edwards@ucdconnect.ie
ADAPT Centre, School of Information & Communication Studies, University College Dublin
Dublin, Ireland

Leigh M. H. Clark
l.m.h.clark@swansea.ac.uk
Computational Foundry, Swansea University
Swansea, Wales

Benjamin R. Cowan
benjamin.cowan@ucd.ie
ADAPT Centre, School of Information & Communication Studies, University College Dublin
Dublin, Ireland

Julie R. Williamson
julie.williamson@glasgow.ac.uk
School of Computing Science, University of Glasgow
Glasgow, Scotland

Figure 1: We developed a room scale virtual reality game where players controlled toy army men using speech and gaze. The environment, shown above, included a range of game objects and everyday objects distributed in the environment that could be referenced by deictic expressions or name. The toy soldier agents could be commanded to complete tasks such as occupy a checkpoint, capture a flag, or interact with objects in the environment.

ABSTRACT

Current virtual environment (VEs) input techniques often overlook speech as a useful control modality. Speech could improve interaction in multimodal VEs by enabling users to address objects, locations, and agents, yet research on how to design effective speech for VEs is limited. Our paper investigates the effect of agent feedback on speech VE experiences. Through a lab study, users commanded agents to navigate a VE, receiving either auditory, visual or behavioural feedback. Based on a post interaction semi-structured interview, we find that the type of feedback given by agents is critical to user experience. Specifically auditory mechanisms are preferred, allowing users to engage with other modalities seamlessly during interaction. Although command-like utterances were frequently used, it was perceived as contextually appropriate, ensuring users were understood. Many also found it difficult to discover speech-based functionality. Drawing on these, we discuss key challenges for designing speech input for VEs.

CUI '21, July 27–29, 2021, Bilbao (online), Spain

ACM ISBN 978-1-4503-8998-3/21/07.
https://doi.org/10.1145/3469595.3469609

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; User studies.

KEYWORDS

Virtual Environments, Speech Input, Gesture Input

ACM Reference Format:

Mitchell Baxter, Anna Bleakley, Justin Edwards, Leigh M. H. Clark, Benjamin R. Cowan, and Julie R. Williamson. 2021. "You, Move There!": Investigating the Impact of Feedback on Voice Control in Virtual Environments. In *3rd Conference on Conversational User Interfaces (CUI '21),July 27–29, 2021, Bilbao (online),Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3469595.3469609

1 INTRODUCTION

Speech is often overlooked as an input technique for virtual environments (VEs). Current research and commercial products have typically focused on gesture, gaze, and locomotion, but input techniques that require *physical action* have limitations that make them unsatisfying in otherwise high fidelity VE interactions [11]. The mapping of continuous human motion to discrete controls presents serious challenges, and physical inputs have limitations including low information capacity[45], fatigue [46], and encumbrance [26]. Incorporating speech as a modality for VE interaction could overcome these challenges by adding a familiar and information rich input technique to existing physical inputs. Speech as part of a multimodal experience was demonstrated in the foundational "Put that there" [3] system, but exploring how speech input works in a modern multimodal VE presents a new series of challenges and opportunities.

State-of-the-art VEs, such as room-scale virtual reality applications using HTC Vive, go substantially beyond the affordances imagined in the original "Put that there" [3] system. VEs enable new kinds of multimodal feedback that make full use of visual and auditory modalities combined with spatial understanding in a fully three dimensional space. Speech is a potentially powerful addition to this kind of multimodal experience, which is particularly important in embodied VE experiences. A VE creates a space for spatially distributed references, including objects, virtual agents, and locations. Interaction in a VE can take place between multiple users, users and agents, and users and objects. By incorporating speech with traditional input techniques in VEs, we have an opportunity to develop experiences offering deeply instinctive control for a variety of situations and tasks.

Although speech shows promise as a control modality, there has been limited research about how to the design speech experiences for an effective and usable voice experience in a VE. This paper focuses on the challenge of feedback, making use of the visual, audio, and spatial capabilities of immersive virtual reality to provide feedback during speech interaction. We developed an immersive virtual reality game where users commanded virtual agents using speech to complete tasks like occupying checkpoints and capturing flags, as shown in Figure 1. We completed a qualitative lab evaluation to explore how three different styles of feedback provided by the virtual agents affected user experience when commanding these agents. Our results demonstrate that when using speech to command agents, participants prefer speech based responses from the agents as confirmation. Participants used the different feedback styles to characterise the agents and assess their competence. Although 3D VEs afford spatial contexts for natural language use, users still used command-like utterances to control the agents, because of the task context along with the desire to ensure they were understood. Our paper contributes by identifying key design preferences for speech based feedback when commanding agents in virtual environments as well as identifying user language choices in speech-based VE interaction.

2 RELATED WORK

2.1 Multimodality in Virtual Environments

Inspired by "Put That There" [3], the concept of multimodality has become a major research topic in VE research. Multimodality tends to be introduced with the goal of enriching the VE experience by increasing the immersion, realism and naturalness of interaction [6, 22, 44, 49, 52]. Especially for screen based (rather than immersive) VEs, the fusion of voice and direct manipulation based interfaces to bring effective multimodal interactions has been a key theme [12, 13, 53, 57]. Much of the research on multimodality in VEs centres around the usability impact and/or preference for specific input modality combinations. The fusion of speech with other modalities like gesture can lead to new meaning of actions, when compared to these modalities being used on their own. Multimodal research in AR highlights that, when speech and gesture are used in tandem, the meaning of gestures becomes intertwined with corresponding speech commands [31, 32].

Research on multimodal feedback in VEs is more limited. Using a number of feedback modalities can lead users to have a more realistic experience. Work using VR to examine behaviours in dangerous real world situations found that using multisensory feedback led people to find the experience more realistic than if using just using audio visual feedback channels [48]. Moreover, utilizing sound and haptic feedback in combination enhances positive user perceptions in VE experiences [5]. Although designing feedback to combine senses like smell [6, 41] have been researched, more commonly studies look specifically at one feedback modality, with haptics and vibrotactile feedback being the focus of much work [25, 28, 29, 43].

2.2 Virtual Environments, Speech and Design

Using speech in VE interactions is thought to hold a number of benefits [35]. When compared to different forms of control mechanisms, such as using icons as well as realistic 3D visual representation of objects to control functionality, speech is seen as easy to learn, uncomplicated, fast and a simple way of handling text input in VR space [23]. As with its wider use through intelligent personal assistants in mobile phones and smart speaker devices [9, 21, 34], speech in VEs is also perceived as effective in facilitating multitasking, freeing up other modalities of control [35]. It has been proposed that speech can lead to a more natural and efficient form of interaction in VEs, as users are already familiar with using speech as an interaction modality through conversation [35]. It would allow users to refer to objects and functions that are not directly visible, thus not limiting the interaction space to that which is visible [35]. The benefits of speech may also be especially apparent when using means of graphical input is less efficient [47]. Especially when combined with other modalities afforded by VR such as gesture, speech significantly improves user task performance and the VE experience [24]. Yet, design of the VE experience is also critical to

the success of using speech in VEs. For instance, designing agents to be embodied rather than non-embodied leads to higher ratings of social presence and relatability in speech based VR interactions [50, 54].

2.3 Speech in agent interaction and the importance of feedback

Although design may lead to more positive, natural VE experiences, the language used for VE control of agents is still likely to be highly constrained compared to natural language with human partners [35]. Work on language use with intelligent personal assistants [17, 40] and on human-machine dialogue research more widely [1] [30] emphasises that constrained language is common when talking to agents. Users tend to adapt their speech, using fewer fillers and tended to request confirmation more explicitly in dialogue with machine partners [1]. Agents are regularly seen as *at risk listeners* [39] and less able communicative partners compared to humans [4], leading users to adapt their speech to make sure that they are being understood [4]. Based on this, agent feedback may play a pivotal role in speech based VE user experience by allowing the agents to signal attention and comprehension, assuaging concerns over whether commands have been registered by the agent. In co-present speech interactions, partner feedback is a critical component of successful conversation [8]. It serves to facilitate the development of mutual belief between dialogue partners that they have understood each other, which helps to build common ground [7, 8, 18]. This feedback can be given through back channelling (e.g. "uh-huh", "yeah"), responding with specific requests or utterances that confirm understanding, or through providing visual evidence that they have understood the utterance (e.g. conducting an action based on a request) [8].

3 INVESTIGATING THE IMPACT OF FEEDBACK ON VOICE CONTROL IN VIRTUAL ENVIRONMENTS

We completed a qualitative lab evaluation of a VR application that used speech combined with gaze as a way of controlling agents in a VE. Speech has potential as a control modality for agent interaction in virtual environments [50, 54]. Yet there is limited research about how to design effective speech based VE interactions. Our work focuses on a critical component of design in these environments; the user feedback from agents that commands have been understood and completed. Our analysis focuses on user preferences of these feedback mechanisms, identifying their impact on user experience and how users perceived their own interactions with the system.

3.1 Application

We developed a room-scale VR game where players could command toy soldiers using speech to complete tasks such a capturing a flag and occupying checkpoints, as shown in Figure 2. The virtual environment was a casual home office space that included game elements like flags and shooting targets and everyday objects such as a lamp, books, and a radio. Within the game, the user was in embodied VR, and thus could move throughout the environment by walking or head position, thus choosing their own perspective in real time.

The application included three toy soldier figures coloured green, blue, and orange. Each toy soldier could be selected using colour or pronouns combined with gaze (e.g. "Blue soldier" or "You"). Toy soldiers could be commanded to move to a named object or by using a deictic expressions combined with gaze (e.g. "Blue soldier, move to the cup" or "Blue soldier, move there"). Players commanded the toy soldiers to complete actions such as capturing a flag, turning on a lamp, or shooting targets using speech and gaze.

The application included three different feedback techniques, speech, visual, and behavioural, to investigate how feedback would influence user experience and speech input in a virtual environment. Each feedback type implemented output for soldier selection and command comprehension.

Figure 2: Toy soldiers could respond to speech commands with speech, visual, or behavioural feedback. For speech feedback, soldiers would respond verbally to selection and command execution. For visual feedback, user interface elements would appear to indicate selection and command execution. For behavioural feedback, toy soldiers would animate behaviours for selection and command execution.

3.1.1 Speech Feedback. During game play, toy soldiers could respond with speech feedback after selection and while completing commands, as shown in Figure 2, Top. After selection, toy soldiers would respond with a randomly selected audio clip (e.g. "Ready!" or "Attention!"). After commands were parsed, toy soldiers would also respond verbally before completing the task (e.g. "Yes Sir!"). Speech feedback was implemented using directional audio, so players could hear the location of soldier without requiring visual attention.

Speech feedback in a VE creates the opportunity for more dialogue based interactions with the virtual agents and provides spatial cues to agents' locations. Speech feedback creates a clear connection between input and output, creating a sense of more continuous interactions between players and agents.

3.1.2 Visual Feedback. Visual feedback was displayed as user interface elements overlayed on top of the VE, as shown in Figure 2, Middle. A matching coloured bar was displayed above a toy soldier to indicate selection and a matching caret was displayed in the environment to indicate where a command would be executed. These elements were rendered so they were scaled correctly to the perspective projection but always drawn over other visual elements to guarantee visibility regardless of location.

Visual feedback is possible in a VE with more flexibility than traditional voice only agents. A VE makes it possible to embed visual feedback situated in 3D space, providing visual and spatial information. There is also significant flexibility in where and how feedback is displayed, especially when players use a headmounted display such as HTC Vive.

3.1.3 Behavioural Feedback. Toy soldiers could respond with animated behaviours as a form of feedback. Soldiers rotated towards the speaker to indicate selection, as shown in Figure 2 Bottom, and would rotate away to indicate de-selection. Soldiers also displayed animations such as shooting and capturing flags while commands were executed.

This kind of behavioural feedback aimed to mimic the kinds of social signals that would be "given off" during human-to-human interaction. Although such feedback is more subtle than speech or visual feedback described above, these behaviours can provide a rich source of feedback to players that supports individual interpretation and understanding.

3.1.4 Implementation. The application was implemented in Unity and designed for the HTC Vive headset. Visual assets for the home office scene[1] and toy soldier models[2] were purchased for the application and used in accordance with the Unity Asset Store terms and conditions.

So as to recognise user speech commands, we used a commercial automatic speech recognition tool (IBM Watson Speech to Text) which was integrated into the Unity implementation. As a commercial product, accuracy in real time was deemed of high quality, and reflective of the types of speech recognition tools likely to be used when using speech in commercial VR applications. Speech input was then processed using the Windows Dictation service, refined using a keyword resolution service, and finalised using a command resolution service. The application processed streams of speech as a single phrase using a 0.3 second threshold to segment strings.

Once speech was processed into string segments, the application completed a keyword resolution step to match phonetic pronunciations with valid vocabulary. When a string could not be matched to the application vocabulary, it was converted to its phonetic representation and its *Levenshtein* string distance [33] was computed across the valid vocabulary to find a nearest match. If there was a match within the valid threshold, the keyword resolution service would return a valid keyword. If no match could be found, the service returned "NOMATCH" to allow the probabilistic command resolver to take these unmatched keywords into account.

Processing the keyword resolved strings into executable commands was the final stage in processing speech input. We implemented a probabilistic model that used the keyword resolved string, current application and task state, and gaze position to construct a command that could be issued to the application.

3.2 Procedure

The evaluation was composed of three tasks; a training task, a structured task, and an open ended task. The evaluation then concluded with a semi-structured interview.

3.2.1 Training Task. Participants completed a training task using the speech input with a single toy soldier. Text positioned within the VE prompted participants to select the soldier, move them, and interact with different objects. This training introduced the system's ability to refer to places using deictic expressions (e.g. "move here;") or by naming locations (e.g. "move to the magazine"). Participants were also introduced to the capabilities of the toy soldiers within the environment, for example the ability to turn on the desk lamp.

3.2.2 Structured Task: Capture the Flag. Participants completed a structured "Capture the Flag" task using three toy soldiers. Each soldier was required to occupy a checkpoint corresponding to their colour before capturing a flag and returning it to the starting point. The checkpoints and flags were distributed throughout the environment. There was no time limit or required order to reach the checkpoints and capture the flags. The distance each agent had to travel to achieve all their goals was equally distant.

3.2.3 Open-Ended Task: Free Play. Participants completed an open-ended task using three toy soldiers with the purposefully vague goal of "Achieve whatever you can, commander." There was no time limit since the goal of the open ended task was to encourage participants to explore the affordances of the scene and test the capability and competence of the toy soldiers. The free play scene included a number of interactive objects such as a lamp, a blue police box [3], and enemy soldiers.

3.3 Experimental Design

During the training task, participants interacted with a single toy soldier that displayed all three feedback techniques. The colour of the toy soldier (blue, green, or orange) was counter-balanced across all participants.

[1]Gabro Media, HQ Suburban House, https://assetstore.unity.com/packages/3d/environments
/urban/hq-suburban-house-81890

[2]Mixaill, Toy Soldiers, https://assetstore.unity.com/packages/3d/characters/toy-soldiers-61368

[3]The TARDIS is a time machine in the form of a small blue "Police Box" from the popular science fiction series Doctor Who https://www.bbc.co.uk/programmes/b006q2x0

For the structured and open-ended tasks, participants interacted with three (blue, green, and orange) toy soldiers. For each participant, the three soldiers were assigned to a feedback technique (speech, visual, and behavioural) that remained fixed for the remainder of the experiment. The assignment of colours to feedback techniques was counter-balanced across all participants.

The experiment was held in a controlled room-scale VR lab setting using an HTC Vive and a Blue Snowball Ice microphone. 21 participants were recruited from a university campus population with an average age of 22.3yrs. Participants were recruited via email and through convenience sampling. Participants were not paid for participation. The study lasted approximately thirty minutes.

4 RESULTS

4.1 General language use in interaction

The voice interactions were transcribed using IBM Watson and were then verified and corrected manually. The transcripts were used to give insight into the types of commands that were being used in interaction and contextualise the qualitative analysis.

Across the experiment there were 2501 statements, with an average of 41.68 per experimental session. Across the tasks there were 30.15 statements in the training condition, 62.5 in the capture the flag task and 32.4 in the free play task. Much of the language used was command-based, with each statement on average being 3.54 words. Most of the common statements were specific references to the agent they wished to command using the colour (e.g. "blue"; N=130, "green"; N=135, "orange"; N=101) or the colour paired with a verb (e.g. "blue move"; N=55, "blue go"; N=9). Controlling the agent with a single command was also common (e.g. "move"; N=339, "fire"; N=52). Although not as frequent as the statements mentioned, participants also used spatial deixis (e.g. "green move here"; N=22, "move here"; N=20) slightly more than using specific nouns referencing objects in the environment (e.g. "move to the mug"; N=14, "turn on the light; N=16).

4.2 Relationships and Characterisation

At the end of the evaluation, participants often had strong characterisations of the toy soldier agents and preferences influenced by the feedback modality. 42% of participants participants preferred speech feedback (N=9), 29% preferred visual feedback (N=6), and 29% preferred behavioural feedback (N=6) conditions (see Fig. 3). However, there was a strong bias towards preferences for the toy soldier agent that participants completed their training task with, with just over half of participants favouring the agent colour from the training task (which was counter-balanced).

4.2.1 Familiar Agents. Even though the training task was relatively short, participants often characterised the training agent as being more familiar, as being important because they had trained them, developing a bond. For example, when one participant was asked which agent was their favourite they responded *"I think because you start off with green so it's kind of familiar" [P04].* Another participant preferred the blue soldier because *"I had the most time with so we bonded" [P10].* Even a short introductory task was enough to foster an increased sense of familiarity. The experience of training with a specific soldier was also mentioned, with one participant preferring the training agent because *"I learned how to shoot with him so that was the first one to actually do stuff" [P16]* Participants often imagined surprisingly nuanced relationships with different agents, for example one participant (who trained with the green soldier) stated that *"you can have this old sentimental attachment to the green but the orange really came through and the blue was pretty solid. They all had their [eh] good points" [P05].*

4.2.2 Perceived Competence. Second to familiarity, perceived competence was an important factor in choosing a preferred agent. The agent giving speech feedback was the most popular, and users often described this agent as more competent or as more likely to listen to commands. For example, when describing the challenges of getting agents to listen, one participant stated that *"you have to do something but your blue [visual feedback] soldiers not listening so you have to argue with him, your orange [behavioural feedback] soldier is lost behind a plant pot" [P20].* Speech feedback was often equated to an agent being responsive and listening, and the absence of speech feedback led to frustrations. When discussing the need to repeat commands, one participant stated that *"I have to say green [visual feedback] multiple times just because I don't know if he heard me" [P03].* When describing the competence of different agents, one participant stated that *"not blue [visual feedback], no. I liked orange [behavioural feedback] he seemed to know what was going on and green [speech feedback] was obedient but I don't think he's leadership material" [P13].* Participants characterised agents based on the feedback they displayed, and speech was often described as competent and/or obedient.

4.2.3 Roleplay and Game Context. The game context, the role of "commander" had an influence on how participants chose to speak and how they related to the toy soldier agents. Participants often enjoyed the experience of commanding in this context, for example when discussing their speech choices one participant stated that they were *"taking the role of commander. I was deliberately trying to phrase things as just a short command" [P19].* Given the technical constraints in speech segmentation and our command-based implementation, speech in short commands also resulted in a better user experience. Participants also enjoyed expressing themselves as part of their commanding role, for example stating that *"it's definitely more enjoyable being able to shout at things and having responses than like typing something".*

4.3 Understanding Feedback Modalities

Participants described the advantages and affordances of the three different feedback modalities. This included the modalities individually (as displayed in the structured and open ended tasks), and in unison (as displayed during the training task).

4.3.1 Eyes-Free Confirmation. Speech was the most popular feedback technique, in particular when participants wanted to visually scan the environment while completing tasks without needing to attend directly to the soldier agents. When discussing the most useful feedback, one participant stated that *"I would say a sound response because you can be looking around the environment to see what there is to do." [P10].* When agents provided speech feedback, participants could attend to different areas in the VE while maintaining an awareness of the location and actions of the agent they had

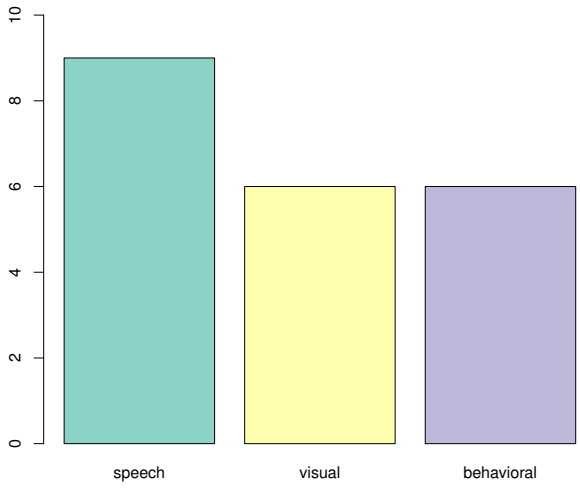

Figure 3: Frequency of Participant Modality Preference

just commanded. In our application, this affordance was especially useful when planning tasks which required attention in different spaces in the VE. One participant stated that *"I didn't look at them I was looking where I wanted them to go... so the audio feedback was very useful" [P14].* The absence of speech feedback often led to confusion when visually attending to different parts of the VE, for example *"there was one time where I told blue [speech feedback] to go somewhere and I wasn't looking at him and he didn't reply to me so I had no idea if he knew what I'd said or not...not getting an audio response made me really confused as to whether or not anything had happened" [P19].* In this case, the participant could not differentiate between a false negative for command execution or simply not hearing the feedback.

4.3.2 Reciprocal Speech. Speech feedback provides information during interaction, but also plays the important role of reciprocating speech input. This resulted in participants perceiving the interactions as more natural, conversational, or enjoyable. The reciprocation of speech is the first step towards conversations, for example one participant stated that *"it just felt like it was like a conversation - I would say something to it and it would talk back. I think there's one that didn't respond at one point and I wasn't sure if it was gonna move or not" [P21].* Even given the relatively limited vocabulary of the toy soldier agents, the speech feedback provided a sense of a dialogue taking place between participants and agents.

4.3.3 Complementary Feedback Modalities. While auditory confirmation was an important feature in the task, participants also discussed errors when speech was used alone. For example, our scene only included one speaking agent at a time but participants highlighted the challenge of differentiating voices if multiple agents occupied the scene. One participant stated that *"I couldn't tell the difference between the voices. There was one point where one said 'yes sir' but I couldn't tell which it was, whereas the health bar [visual feedback] was a lot more obvious" [P15].* Although the speech feedback was implemented with directional audio, the visual cues were explicitly placed at a location within the VE. When more precise feedback was needed, participants were more likely to visually scan the room than listen for audio, for example *"the green bar was also useful if I ever forgot which unit I had selected then I could just look around the room and sometimes see it" [P1].*

4.4 Communication Techniques

When discussing how they approached speaking to the system, participants were varied in how natural or stilted they felt their interactions were. Although many participants described their speech as "natural," there was clear evidence that participants adapted their languages to suit both the game task and the agent partners.

4.4.1 Adapting Language to the Task. Participants often described their speech patterns as natural given the task of commanding toy soldiers. One participant described this as *"it wasn't a conversation it was mostly like me saying commands but it was pretty natural in the sense of the setup ... you don't need to have a conversation when you're commanding soldiers" [P16].* The types of tasks required during the game were easily completed using a simple command structure. Participants described how they simplified their speech in this context, for example *"it's not that it is less natural it's just removing redundant information from the conversation" [P20].*

4.4.2 Adapting Language to the Partner. Participants described how they adapted their speech to better communicate with the toy soldier agents. This was often driven by a desire to simply be understood without having to repeat or clarify commands. For example, one participant stated that *"I was focusing more on speaking clearly than just speaking normally" [P11].* Based on game play

experiences, some participants reduced the number of words in their input and resorted to simpler commands. For example, *"I was just trying to be as clear and precise, sort of procedural about it as possible" [P12].* As participants learned more about the capabilities of the soldiers, they also adapted their language. For example, one participant stated that *"it felt very natural at first until there was the first miscommunication and then it broke down into colour, command" P01.*

The need to adapt utterances based on the application was seen as an error or something that could be improved in future designs. For example, on participant stated that *"our language shouldn't really change for technology technology should adapt for our speech, we shouldn't have to change the way we speak for it to be beneficial" [P7].* Latency often required participants to adapt their manner of speaking, for example *"if possible with voice recognition these days, if it could react faster to what I'm saying and if I could speak a little more informally" [P11].*

5 DISCUSSION

Our evaluation explored how different agent feedback designs affected user experience when introducing speech as a control modality in VEs. Speech has been under-investigated compared to other more common modalities such as gaze and controller based interaction. Our work builds on recent efforts to explore the user experience of speech in VEs [23, 49, 50, 54], focusing in particular on feedback design when commanding agents in VEs.

We found that auditory feedback was preferred above visual and agent behavioural feedback, giving participants the ability to engage with other input modalities, allowing them to scan their environment and move on to other tasks more effectively. However, each feedback technique had strengths and participants felt that the feedback techniques could be used simultaneously. Providing multimodal cues is especially powerful in a VE when users may be attending to complex tasks distributed throughout a large 3D environment.

Participants also clearly altered their speech to the context and task, whilst also adapting because of the machine nature of the partner. Users felt that, although they should not need to change their language to succeed in the interaction, it was necessary to do so.

5.1 Commanding Virtual Agents

Through the statements in the transcripts and the interviews, it is clear that participants commonly utilised commands rather than attempting to use more complex language. Participant comments emphasise that they felt this was driven by the type of task, with commands being seen as a natural way of interacting verbally in this context. They also stated that they felt this command driven approach ensured communicative clarity when interacting with agents. This echoes literature on human machined dialogue, where users are commonly seen to not only use more command based utterances [1], but also adapt their language choices based on this desire to ensure communicative success [4, 16]. People's perceptions of automated partner communicative competence (i.e. their *partner models* [15, 19]) are critical drivers of user language choices in agent interaction, with adaptation on the user side because of automated agents being seen as less competent and flexible dialogue partners [4, 20]. This is likely to inform the command based linguistic choices of the types seen in this work (e.g. [1, 40]). The nature of the task also seems to support this partner model, as the task emphasises the more command based interactions with the agent. This interaction is therefore an example where the interaction task fits well with people's existing preconceptions of agents abilities as communicative partners. It is important for future speech design in VEs, as well as speech applications more widely, to keep this in mind so as to design appropriately for speech based control. It is also important to note that context may significantly influence the types of interactions users feel are appropriate with agents [10]. For instance, a more social task may make users more likely to engage in social conversation with agents. Yet, this may not lead to the same types of language use seen in social conversation with other people. Recent work highlights that when comparing conversation with agents to conversation with other people, users tend to see task-based and command oriented interactions as more characteristic of agent conversation [10], making it fundamentally different to human conversation [10, 40, 42]. Future work could elucidate this debate by exploring the types of patterns of conversation that occur in more social or multimodal interactive contexts with agents in VR.

5.2 Scaffolding Learnability & Discoverability

Similarly, it is important to scaffold the learnability of user input. While discovery-based learning can be successful [38], making the available commands visible to users is a common challenge in speech-based interaction [14, 36, 37]. Recent work on speech agents has shown that strategies to increase discoverability such as making commands available through phrases such as "what can I say?" improve perceived usability compared to agents without these functions [27]. Similar methods could be explored in VE environments when using speech, along with use of multimodal feedback and display to support discoverability. Furthermore, it may be important to scaffold users' understanding of system output depending on the input modalities used. This reflects recent speech and human-artificial intelligence (AI) design guidelines, emphasising the need for appropriate feedback [2, 51, 56] and using multimodal feedback when available [56]. With the introduction of 3D space, however, feedback methods for discoverability and learnability may need to be further evaluated.

6 LIMITATIONS AND FUTURE WORK

Because the aim of the study was for users to be able to directly compare the types of agent based feedback developed, participants were able to interact with all three feedback types simultaneously. This was so as to make the comparison of feedback salient for the post interaction interviews, and allow users to select the agents they interacted with the most in the session. Although beneficial to emphasise comparison, in a real-world interaction agents are unlikely to be categorised into mutually exclusive feedback types. Future work could conduct a more controlled study to compare more quantitatively the effects of using each feedback type in interaction. In future quantitative analysis, we also hope to explore

the effect of feedback type on both the prevalence and the nature of deictic references.

The study focused on one user commanding many agents in a VE interaction, Yet multiple user scenarios are common in VEs. It is therefore important to explore the potential impact of agent feedback design as well as voice control in multi-user multi agent domains. This type of interaction would likely include talk between users and as well as between user and agents. Especially in situations were users were competing, this may lead to significant challenges for voice based interaction with agents, ensuring that the correct command is understood and acted upon. Currently, prominent VE work using voice in VE agent interaction focuses on individual users [55], future studies should look to expand this towards multiple user scenarios.

The study used three conditions that varied in feedback modality, showing that users in the study had a clear preference for auditory feedback. That said, some of the other feedback modalities may have varied in their subtlety by comparison, especially that used in the visual feedback condition. It is therefore important for future work to explore how other types of visual feedback may compare to the conditions within this study.

7 CONCLUSION

Speech is growing in popularity as an input modality, yet is not heavily used in VE interactions. Our novel evaluation explored the role of agent feedback in a speech based VE interaction, asking users to compare their experiences controlling agents that used speech, visual or behavioural feedback mechanisms. We found that, when using speech in VE, auditory feedback was highly preferred as this gave them the ability to multitask and move to other phases of the task. When interacting with the agents using speech, command based language was commonly used. Participants also clearly altered their speech to suit the type of task as well as the nature of the agent as a machine partner, altering their language to make sure they would be understood. Overall our work sheds light on the importance of feedback in speech based VE experiences, identifying that using speech based feedback to match the control modality leads to a positive VE experience.

ACKNOWLEDGMENTS

This research was conducted in part with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 and 18/CRT/6224 at the ADAPT SFI Research Centre and D-REAL Centre for Research Training at University College Dublin. We would like to thank our participants for taking part in this research.

REFERENCES

[1] René Amalberti, Noëlle Carbonell, and Pierre Falzon. 1993. User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies* 38, 4 (1993), 547–566.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. ACM, Glasgow, 1–13.

[3] Richard A. Bolt. 1980. "Put-that-there": Voice and Gesture at the Graphics Interface. *SIGGRAPH Comput. Graph.* 14, 3 (July 1980), 262–270. https://doi.org/10.1145/965105.807503

[4] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (Oct. 2011), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

[5] Grigore Burdea, Paul Richard, and Philippe Coiffet. 1996. Multimodal Virtual Reality: Input-Output Devices, System Integration, and Human Factors. *International Journal of Human-Computer Interaction* 8 (01 1996), 5–. https://doi.org/10.1080/10447319609526138

[6] Alan Chalmers, Kurt Debattista, and Belma Ramic-Brkic. 2009. Towards High-fidelity Multi-sensory Virtual Environments. *Vis. Comput.* 25, 12 (Oct. 2009), 1101–1108. https://doi.org/10.1007/s00371-009-0389-2

[7] H. H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

[8] Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science* 13, 2 (April 1989), 259–294. https://doi.org/10.1016/0364-0213(89)90008-6

[9] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers* 31, 4 (2019), 349–371.

[10] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. *What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi-org.ucd.idm.oclc.org/10.1145/3290605.3300705

[11] Sue VG Cobb, Sarah Nichols, Amanda Ramsey, and John R Wilson. 1999. Virtual reality-induced symptoms and effects (VRISE). *Presence: Teleoperators & Virtual Environments* 8, 2 (1999), 169–186.

[12] P. R. Cohen, M. Dalrymple, D. B. Moran, F. C. Pereira, and J. W. Sullivan. 1989. Synergistic Use of Direct Manipulation and Natural Language. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '89)*. ACM, New York, NY, USA, 227–233. https://doi.org/10.1145/67449.67494

[13] Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. QuickSet: Multimodal Interaction for Distributed Applications. In *Proceedings of the Fifth ACM International Conference on Multimedia* (Seattle, Washington, USA) *(MULTIMEDIA '97)*. ACM, New York, NY, USA, 31–40. https://doi.org/10.1145/266180.266328

[14] Eric Corbett and Astrid Weber. 2016. What can I say?: addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, ACM, Florence, Italy, 72–82.

[15] Benjamin R. Cowan, Holly P. Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue. *International Journal of Human-Computer Studies* 83 (Nov. 2015), 27–42. https://doi.org/10.1016/j.ijhcs.2015.05.008

[16] Benjamin R. Cowan, Philip Doyle, Justin Edwards, Diego Garaialde, Ali Hayes-Brady, Holly P. Branigan, João Cabral, and Leigh Clark. 2019. What's in an Accent?: The Impact of Accented Synthetic Speech on Lexical Choice in Human-machine Dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. ACM, New York, NY, USA, Article 23, 8 pages. https://doi.org/10.1145/3342775.3342786

[17] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, Vienna, Austria, 43.

[18] Nicole N. Craycraft and Sarah Brown-Schmidt. 2018. Compensating for an Inattentive Audience. *Cognitive Science* 42, 5 (2018), 1504–1528. https://doi.org/10.1111/cogs.12614

[19] Philip R Doyle, Leigh Clark, and Benjamin R. Cowan. 2021. What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 244, 14 pages. https://doi.org/10.1145/3411764.3445206

[20] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[21] Justin Edwards, He Liu, Tianyu Zhou, Sandy J. J. Gould, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. 2019. Multitasking with Alexa: How Using Intelligent Personal Assistants Impacts Language-based Primary Task Performance. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3342775.3342785

[22] Horst Eidenberger. 2018. Smell and Touch in the Virtual Jumpcube. *Multimedia Syst.* 24, 6 (Nov. 2018), 695–709. https://doi.org/10.1007/s00530-018-0592-y

[23] Daniel Hepperle, Yannick Weiß, Andreas Siess, and Matthias Wölfel. 2019. 2D, 3D or speech? A case study on which user interface is preferable for what kind of object interaction in immersive virtual reality. *Computers & Graphics* 82 (Aug. 2019), 321–331. https://doi.org/10.1016/j.cag.2019.06.003

[24] Sylvia Irawati, Scott Green, Mark Billinghurst, Andreas Duenser, and Heedong Ko. 2006. An Evaluation of an Augmented Reality Multimodal Interface Using Speech and Paddle Gestures. In *International Conference on Artificial Reality and Telexistence (Advances in Artificial Reality and Tele-Existence)*. Springer Berlin Heidelberg, Berlin, Heidelberg, 272–283.

[25] Ali Israr, Zachary Schwemler, John Mars, and Brian Krainer. 2016. VR360HD: A VR360 & Deg; Player with Enhanced Haptic Feedback. In *Proceedings of the 22Nd ACM Conference on Virtual Reality Software and Technology* (Munich, Germany) *(VRST '16)*. ACM, New York, NY, USA, 183–186. https://doi.org/10.1145/2993369.2993404

[26] Marshall B Jones, Robert S Kennedy, and Kay M Stanney. 2004. Toward systematic control of cybersickness. *Presence: Teleoperators & Virtual Environments* 13, 5 (2004), 589–600.

[27] Philipp Kirschthaler, Martin Porcheron, and Joel E. Fischer. 2020. What Can I Say? Effects of Discoverability in VUIs on Task Performance and User Experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 9, 9 pages. https://doi.org/10.1145/3405755.3406119

[28] Pascal Knierim, Thomas Kosch, Valentin Schwind, Markus Funk, Francisco Kiss, Stefan Schneegass, and Niels Henze. 2017. Tactile Drones - Providing Immersive Tactile Feedback in Virtual Reality Through Quadcopters. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. ACM, New York, NY, USA, 433–436. https://doi.org/10.1145/3027063.3050426

[29] Mike Lambeta, Matt Dridger, Paul White, Jesslyn Janssen, and Ahmad Byagowi. 2016. Haptic Wheelchair. In *ACM SIGGRAPH 2016 Posters* (Anaheim, California) *(SIGGRAPH '16)*. ACM, New York, NY, USA, Article 90, 2 pages. https://doi.org/10.1145/2945078.2945168

[30] Ludovic Le Bigot, Jean-François Rouet, and Eric Jamet. 2007. Effects of Speech- and Text-Based Interaction Modes in Natural Language Human-Computer Dialogue. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49, 6 (Dec. 2007), 1045–1053. https://doi.org/10.1518/001872007X249901

[31] Minkyung Lee and Mark Billinghurst. 2008. A Wizard of Oz Study for an AR Multimodal Interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (Chania, Crete, Greece) *(ICMI '08)*. ACM, New York, NY, USA, 249–256. https://doi.org/10.1145/1452392.1452444

[32] Minkyung Lee, Mark Billinghurst, Woonhyuk Baek, Richard Green, and Woontack Woo. 2013. A Usability Study of Multimodal Input in an Augmented Reality Environment. *Virtual Real.* 17, 4 (Nov. 2013), 293–305. https://doi.org/10.1007/s10055-013-0230-0

[33] V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (Feb 1966), 707.

[34] Nikolas Martelaro, Jaime Teevan, and Shamsi T. Iqbal. 2019. An Exploration of Speech-Based Productivity Support in the Car. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 264, 12 pages. https://doi.org/10.1145/3290605.3300494

[35] Scott McGlashan and Tomas Axling. 1996. A speech interface to virtual environments.

[36] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. 2018. Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, ACM, Barcelona, Spain, 269–276.

[37] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45.

[38] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, ACM, Montreal QC, Canada, 6.

[39] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998. Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech* 41, 3-4 (July 1998), 419–442. https://doi.org/10.1177/002383099804100409

[40] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC, Canada, 640.

[41] Belma Ramic-Brkic and Alan Chalmers. 2010. Virtual smell: Authentic smell diffusion in virtual environments. In *Proceedings of the 7th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*. ACM, Franschhoek, South Africa, 45–52.

[42] Stuart Reeves. 2019. Conversation Considered Harmful?. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 10, 3 pages. https://doi.org/10.1145/3342775.3342796

[43] Holger Regenbrecht, Joerg Hauber, Ralph Schoenfelder, and Andreas Maegerlein. 2005. Virtual Reality Aided Assembly with Directional Vibro-tactile Feedback. In *Proceedings of the 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia* (Dunedin, New Zealand) *(GRAPHITE '05)*. ACM, New York, NY, USA, 381–387. https://doi.org/10.1145/1101389.1101464

[44] Nina Rosa, Wolfgang Hürst, Wouter Vos, and Peter Werkhoven. 2015. The Influence of Visual Cues on Passive Tactile Sensations in a Multimodal Immersive Virtual Environment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. ACM, New York, NY, USA, 327–334. https://doi.org/10.1145/2818346.2820744

[45] Daniel Roth, Peter Kullmann, Gary Bente, Dominik Gall, and Marc Erich Latoschik. 2018. Effects of hybrid and synthetic social gaze in avatar-mediated interactions. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, IEEE, Munich, Germany, 103–108.

[46] Bhuvaneswari Sarupuri, Miriam Luque Chipana, and Robert W Lindeman. 2017. Trigger walking: A low-fatigue travel technique for immersive virtual reality. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, Los Angeles, CA, USA, 227–228.

[47] Stefan Schaffer, Robert Schleicher, and Sebastian Möller. 2015. Modeling input modality choice in mobile graphical and speech interfaces. *International Journal of Human-Computer Studies* 75 (2015), 21–34.

[48] Emily Shaw, Tessa Roper, Tommy Nilsson, Glyn Lawson, Sue V.G. Cobb, and Daniel Miller. 2019. The Heat is On: Exploring User Behaviour in a Multisensory Virtual Environment for Fire Evacuation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 626, 13 pages. https://doi.org/10.1145/3290605.3300856

[49] Jaisie Sin and Cosmin Munteanu. 2020. Let's Go There: Combining Voice and Pointing in VR. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 31, 3 pages. https://doi.org/10.1145/3405755.3406161

[50] Harrison Jesse Smith and Michael Neff. 2018. Communication Behavior in Embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 289:1–289:12. https://doi.org/10.1145/3173574.3173863 event-place: Montreal QC, Canada.

[51] Bernhard Suhm. 2003. Towards best practices for speech user interface design. In *Eighth European Conference on Speech Communication and Technology*. ISCA Archive, Geneva, Switzerland, 2217–2220.

[52] Anastasia Treskunov, Emil Gerhardt, David Nowottnik, Ben Fischer, Laurin Gerhardt, Mitja Säger, and Christian Geiger. 2019. ICAROSmuIti - A VR Test Environment for the Development of Multimodal and Multi-User Interaction Concepts. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC'19)*. ACM, New York, NY, USA, 909–911. https://doi.org/10.1145/3340764.3345379

[53] Edward Tse, Saul Greenberg, and Chia Shen. 2006. GSI Demo: Multiuser Gesture/Speech Interaction over Digital Tables by Wrapping Single User Applications. In *Proceedings of the 8th International Conference on Multimodal Interfaces* (Banff, Alberta, Canada) *(ICMI '06)*. ACM, New York, NY, USA, 76–83. https://doi.org/10.1145/1180995.1181012

[54] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring Virtual Agents for Augmented Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 281:1–281:12. https://doi.org/10.1145/3290605.3300511 event-place: Glasgow, Scotland Uk.

[55] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. *Exploring Virtual Agents for Augmented Reality*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300511

[56] Zhuxiaona Wei and James A Landay. 2018. Evaluating Speech-Based Smart Devices Using New Usability Heuristics. *IEEE Pervasive Computing* 17, 2 (2018), 84–96.

[57] D. Weimer and S. K. Ganapathy. 1989. A Synthetic Visual Environment with Hand Gesturing and Voice Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '89)*. ACM, New York, NY, USA, 235–240. https://doi.org/10.1145/67449.67495