

# Putting wake words to bed

We speak wake words with systematically varied prosody, but CUIs don't listen

Saul Albert  
Loughborough University  
s.b.albert@lboro.ac.uk

Magnus Hamann  
Loughborough University  
m.hamann@lboro.ac.uk

## ABSTRACT

'Wake words' such as "Alexa" or "Hey Siri", as conversation design elements, mimic the interactionally rich 'summons-answer' sequence in natural conversation, but their function amounts to little more than a button-push: simply activating the interface. In practice, however, users vocally overdesign their wake words with all the detail of a 'real' interactional summons. We hear users uttering wake words with a specific prosody and intonation, as though for a particular recipient in a particular social/pragmatic context. This presents a puzzle for designers of conversational user interfaces (CUIs). Previous research suggests that expert users simplify their talk when interacting with CUIs, but with wake words we observe the opposite. When users do the extra interactional work of varying their wake words in ways that seem 'recipient designed' for a specific other, does that suggest that designers are successfully eliciting natural interaction from users, or is it violating user expectations? Our two case studies highlight how the mismatch between user expectations and the limitations of how wake words are currently implemented can lead to cascades of interactional trouble, especially in multi-party conversations. We argue that designers should find new ways to activate CUIs that align users' expectations with conversational system design.

## CCS CONCEPTS

• Human-centered computing; • Natural language interfaces;

## KEYWORDS

Speech, Conversation Analysis, Voice, VUI, Design

### ACM Reference Format:

Saul Albert and Magnus Hamann. 2021. Putting wake words to bed: We speak wake words with systematically varied prosody, but CUIs don't listen. In *CUI 2021 - 3rd Conference on Conversational User Interfaces (CUI '21)*, July 27–29, 2021, Bilbao (online), Spain. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3469595.3469608>

## 1 INTRODUCTION

The first utterance in a conversation is an important interactional resource for contextualization, for projecting upcoming action, and

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

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

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-8998-3/21/07...\$15.00  
<https://doi.org/10.1145/3469595.3469608>

for establishing the fundamental interactional roles of speaker and recipient [1, 11, 16]. The design of conversational user interfaces (CUIs) that use 'wake words' such as "Alexa", "OK Google" or "Hey Siri" are parasitic on some aspects of these natural interactional openings – specifically the summons-answer 'pre-sequences' [8, 17] that routinely precede and project further talk. Some of the earliest conversation analytic research focused on the importance of the first few utterances in telephone call openings and showed how tiny but systematic variations in, e.g., the production of the first "hello" can provide inferentially rich resources for establishing the interaction [15]. In principle, similar kinds of inference could be drawn from the way users appear to vary the recipient design of their wake words when initiating interactions with CUIs.<sup>1</sup> By 'recipient design' we refer to the interactional practice through which participants in interaction modify their talk for specific recipients and situations [14]. As early interactional studies of CUIs have noted, when users are pursuing a response or rephrasing a prior request, they often produce prosodically marked wake words, [12] despite the fact that most speech recognition systems ignore this extra interactional effort on the part of users [13]. In this paper we present two case studies from a broader systematic analysis of wake words in the openings of interactions with CUIs. The analysis shows how users' interactional efforts to produce recipient designed wake words index a pragmatic mismatch between user expectations and system design [7]. We show an example of how this mismatch can lead to cascades of interactional problems. We conclude that despite some tempting opportunities to harness the users' prosodic cues as a resource for improving voice interfaces, that we should instead look for new, more flexible approaches to initiating interactions with our virtual agents.

## 2 DATA

We worked with Amazon Echo users who recorded ~100 hours of continuous video in their homes, then provided their Alexa 'voice history' logs<sup>2</sup>, enabling us to locate video clips featuring interactions with Alexa. Participants gave informed consent to share pseudonymized recordings in an open data corpus. The data were transcribed using conversation analytic conventions [2] to describe the production quality of talk. We adapted those conventions slightly to include a vertical bar running between the line numbers and the transcript to indicate when the 'wake light' (that indicates that the Amazon Echo device is 'listening' for further talk) is on. The wake light is only transcribed as a 'turn' on its own line where participants treated it as interactionally relevant by, for example,

<sup>1</sup>Although note that in telephone calls, it is a machine-generated sound (a phone ring) that makes the summons and (usually) a human voice that does the answer, whereas wake words function the other way around.

<sup>2</sup>See <https://www.amazon.com/alexa-history-delete-voice-recordings/>

looking towards the device and waiting for the light to come on before proceeding with a command.

### 3 ANALYSIS

In our corpus, variations in the vocal production of the wake words differ in at least three key parameters.

1. *Pitch/prosody*: we found four prosodic variants of the pitch contour on the wake word ‘Alexa’, which we transcribed as follows: falling intonation (Alexa.); flat intonation (Alexa.); slight rise intonation (Alexa<sub>ζ</sub>); and rising intonation (Alexa?)
2. *Intensity and stress*: each syllable of a wake word can be produced with varying intensity and syllabic stress. In our corpus, stress tends to fall on the middle syllable (Alexa). In the transcripts below we only annotate stress/emphasis where it differs from this apparent norm.
3. *Quality*: the phonetic quality of the wake word is sometimes produced with a hoarse or creaky voice, or with elongation or shortening of syllables.

While these differently designed wake words may seem interchangeable since they all initiate ostensibly the ‘same’ action of summoning, their positions within broader sequences of action suggest that these variations are specific and systematically distributed. For example, in Figure 1, Ted uses each variation of the wake word “Alexa” in a different sequential context. When the interaction starts Ted is alone in the room, sitting with his back turned to Alexa. Prior to this extract starting he has set a timer that will trigger Alexa’s alarm.

#### 3.1 Prosodic variations in wake word/summons initiations

Before exploring Ted’s use of wake words any further, we note two key points about his access to Alexa’s responses. Firstly, Ted does not have visual access to the wake light on the Echo, which is sitting on the table behind him, so he cannot see whether his wake words have been ‘heard’ successfully. Secondly, note that when Alexa ‘hears’ a wake word while already producing a sound (such as music, or – as at line 5 – an alarm), it will turn down the volume: a response that is audible to Ted. The four wake words at lines 1, 6, 11 and 17 initiate different projects (turning off the heater, turning off an ongoing alarm, and calling Ann), using three different prosodic designs. The first project of turning off the heater involves two differently designed wake words. The initial iteration at line 1 is produced with a slight rising intonation. This prosodic pattern matches the intonation of the wake word Ted later uses to initiate the project of calling Ann at line 17. In sequential terms, these are both ‘initial’ iterations of the first part of a summons/response pair, so a summons in this sequential position can be thought of as an *initial summons*. The summons at line 6, produced while Alexa is already making an alarm noise is a shorter variation with added stress and falling intonation. In other recordings we have observed similarly designed wake words used to intervene while Alexa is playing music or giving extended responses to another command. These might be called *interventional summons*. Finally, the second iteration of the heater-action-related wake word in line 11 has a more stressed second syllable and flat intonation. Since this type of summons re-does the first part of a summons/response sequence it

might be grouped alongside the kinds of *second summons* that have previously been identified as components of everyday telephone conversations [5, 15] and in other interactional studies of CUI use in everyday life [12].

Without making any claims here for the generalizability of these categories of summons and their associated prosodic designs, the examples clearly show that Ted designs these summons variably, so at least in principle, such variations could be associated with different sequential/pragmatic contexts. Whatever Ted may know about Alexa’s technical capabilities, the fact that he varies the prosody of his wake words while alone in the room with Alexa<sup>3</sup> suggests that aspects of the natural interactional practice of doing a summons are leaching into the design of his wake words. Although the redundancy of this extra interactional work does not constitute a design problem as such, this analysis does provide some insight into Ted’s apparent misunderstanding of how wake words work. In the next extract, a similar analysis also suggests how users’ misunderstandings about wake words can lead to more significant breakdowns in interactional structure.

In Figure 2, we see how apparently unmet user expectations about wake words can lead to cascading interactional problems, especially where multiple collocated CUIs are involved.

#### 3.2 Using the wrong wake word confuses multiple CUIs

Before the interactional problems begin to unfold between Ted and his two virtual assistants, he is facing his desk where his Alexa device is sitting next to his phone. This means he can see Alexa’s wake light and hear the audible beep that Apple’s voice assistant ‘Siri’ makes in response to a wake word. After he first does an initial summons “Alexa<sub>ζ</sub>” in line 1, Ted may be waiting for Siri to respond since despite being in a position to see Alexa’s active wake light, he continues to look towards Siri (i.e., his phone) while doing a second summons “Alexa.” Ted glances up at Alexa’s wake light in line 9 as he says “Ahh wrong one” under his breath, before switching to “He:y Siri.” Just as participants in everyday conversation often use names as address terms to select a next speaker [6], here Ted uses Siri’s wake word to switch between recipients. However, Ted summons Siri while Alexa’s wake light is on – still visibly ‘listening’ for a command. Alexa shows that it successfully recognizes “Hey Siri” as a misdirected summons since it responds in line 14 with the jokey non-sequitur: “I think you’ve got me confused with someone else.” This scripted response must have been designed for a sequence in which the initial wake word “Alexa” is followed by a ‘competitor’ initial wake word such as “Hey Siri” or “OK Google”. Ironically, Alexa misses the chance to do an even smarter, interactionally fitted response by, for example, responding to the command or, even better, simply ignoring wake words that are clearly directed to another virtual assistant. Instead, Alexa’s wisecrack sets off a cascade of misdirected utterances when Siri treats it as a request for information by responding in lines 16-19 with another equally sequentially ill-fitted offer to look up the contents of Alexa’s turn on

<sup>3</sup>In multi-party settings we have seen that variations in the design of wake words and commands to CUIs may be designed to involve overhearers in the interaction by, for example, recruiting their assistance after multiple failures to summon a virtual agent.

1 Ted: Alexa:¿

2 (2.1) ((Text message alert sounds))

3 Ted: hhh (1.8) turn off: (1.0) heater.

4 (2.0)

5 Ale: ((ALARM-----[----->))

6 Ted [Alexa.

7 Ale: -->>((ALARM STOPS))

8 (0.2)

9 Ted: Off.

10 (6.8)

11 Ted Alexa:,

12 (1.6)

13 Ted: <Tu:rn off:> (0.5) heater.

14 (1.2)

15 Ale: Okay. ((heater turns off 0.2 seconds before this response))

16 (1.0)

17 Ted: Alexa:¿

18 (1.1)

19 Ted: Call (0.2) Ann.

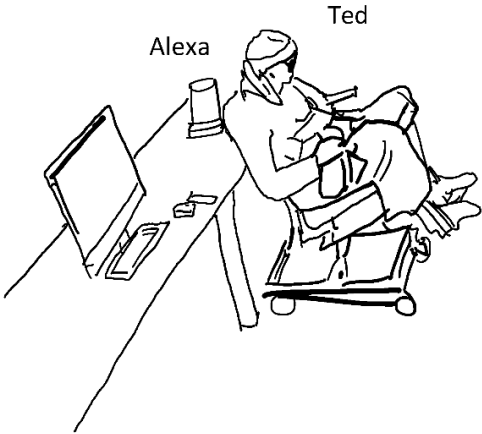


Figure 1: Video available at <http://bit.ly/cui-13-fig1>

the web. At this point Ted gives up on whatever goal had originally motivated his initial summons and lapses into silence.

#### 4 DISCUSSION

These two examples reveal the practical impact of mismatches between CUI user expectations and system design. Ted's use of interactional practices of recipient design such as the prosodic variation in his summons initiations indexes some form of unmet

expectation about the ability of his virtual assistants to respond accordingly. Similarly, his switching between wake words to address two different virtual assistants implies that he assumes his CUIs are able to manage the interactional roles of speaker and recipient in this kind of multi-party interactional setting. The analysis of these case studies suggest there may be fundamental problems with the use of wake words to activate virtual agents, especially in everyday

1	Ted:	Alexa?	
2		(0.5)	
3	Ale:	((Wake light on))	
4		(1.6)	
5	Ted:	Alexa.	
6		(0.4)	
7	Ale:	((Wake light on))	
8		(1.4)	
9	Ted:	*°Ahh wrong one°*	
10		(0.4)	
11	Ted:	He:y Siri,	
12		(1.6)	
13	Sir:	((Siri response [ beep]))	
14	Ale:	I think you've got me confused with someone else.	
15		(0.9)	
16	Sir:	((Siri response beep)) I don't know what that means.	
17	Ted:	Alexa?=-	
18	Sir:	=If you like (.) I could search the web for (0.6) hey Siri I think	
19		you've got me confused with someone else.	

**Figure 2:** Video available at <http://bit.ly/cui-13-fig2>

domestic environments that usually include multiple co-present parties and wake-word-activated devices.

Whereas experimental and survey-based studies have highlighted users' anthropomorphic attitudes to technology [9] and the tendency to overestimate the interactional skills of CUIs [7], this study demonstrates how and why this matters in practice for conversation designers. The interactional practice of summoning another party into conversation achieves far more than just alerting the recipient to await a command. Even without the complexity of face-to-face interactional openings [8], the smallest voice sample from a recipient answering a telephone or the breathiness of their "hello" can be vital to establishing the identities of the caller and the called, the roles of speaker and recipient, and other situated contingencies for the ensuing interaction [1, 14]. When users systematically vary the production of their wake words for different sequential and pragmatic contexts, they are drawing on this rich prosodic repertoire from our natural interactional practices of summoning. When CUI developers borrow from conversational

practices to create interface elements such as wake words, users' unmet expectations about e.g., how they should work in multi-party interactions may cause more interactional problems than they solve.

So what, if anything, should conversation designers do to resolve this mismatch? On the one hand, the systematicity of users' prosodic variations could inspire designers to harness this additional set of interactional 'signals'. Designers could use the regularities of patterns in the prosody of wake words to infer the user's current understanding of the state of a conversation. If systems could detect that a wake word has the prosodic pattern of a 'second summons' or an 'interventional summons', this could provide designers with an opportunity to leverage that information to provide more appropriate CUI responses. Similarly, harnessing more granular details from naturalistic interaction [3] could contribute to related calls to improve CUIs by developing incremental processing pipelines and more sophisticated cognitive models of communication states [2, 5]. On the other hand, the redundancy of users'

interactional efforts and the problematic outcomes of their unmet expectations could point to more serious design flaws with the basic concept of using wake words to initiate interactions with virtual agents.

Looking at our evidence and related interactional studies, we lean towards the latter and suggest putting wake words to bed. Pelikan and Broth's [10] interaction analytic study of conversations with social robots shows that more technologically savvy users tend to reduce the complexity and nuance of their turn structure and vocabulary choice, and related studies suggest that expert users seem to be aware that they are 'talking down' to CUIs [7]. Where users understand the capabilities of CUIs they are better able to simplify their turns to accommodate the known limitations of their virtual recipient. In the case of wake words, however, we see the opposite effect: users overdesign their turns for oblivious virtual recipients. Since the design of wake words draws on the structure of naturalistic summons-response sequences, they can mislead users as to the interactional capabilities of CUIs. We therefore recommend that conversation designers explore alternatives to wake words such as gaze-tracking or other multimodal signals [4] to provide users with more variation, transparency, and control over how they initiate interactions with virtual agents.

## ACKNOWLEDGMENTS

Thanks to Crispin Coombs, Thorsten Gruber, Mark Harrison, Donald Hislop, and Elizabeth Stokoe for their work on the project that produced the data used for this study: Adept at Adaptation: Disability, AI, and Voice Technologies in Social Care Services, supported by a BA/Leverhulme Small Research Grant: SRG19\191529.

## REFERENCES

- [1] Charles Goodwin. 2007. Interactive footing. In *Reporting Talk*, Elizabeth Holt and Rebecca Clift (eds.). Cambridge University Press, Cambridge, 16–46. DOI:<https://doi.org/10.1017/CBO9780511486654.003>
- [2] Alexa Hepburn and Galina B Bolden. 2017. *Transcribing for social research*. Sage, London.
- [3] William Housley, Saul Albert, and Elizabeth Stokoe. 2019. Natural Action Processing. In *Proceedings of the Halfway to the Future Symposium 2019* (HTTF 2019), Association for Computing Machinery, Nottingham, United Kingdom, 1–4. DOI:<https://doi.org/10.1145/3363384.3363478>
- [4] Razan Jaber, Donald McMillan, Jordi Solsona Belenguer, and Barry Brown. 2019. Patterns of gaze in speech agent interaction. In *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19*, ACM Press, Dublin, Ireland, 1–10. DOI:<https://doi.org/10.1145/3342775.3342791>
- [5] Seung-Hee Lee. 2006. Second summonings in Korean telephone conversation openings. *Language in Society*, 35, 02. DOI:<https://doi.org/10.1017/S0047404506060118>
- [6] Gene H Lerner. 2003. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society*, 32, 02, 177–201. DOI:<https://doi.org/10.1017/S004740450332202X>
- [7] Ewa Luger and Abigail Sellen. 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* (CHI '16), Association for Computing Machinery, New York, NY, USA, 5286–5297. DOI:<https://doi.org/10.1145/2858036.2858288>
- [8] Robert J. Moore and Raphael Arar. 2019. *Conversational UX design: A practitioner's guide to the natural conversation framework*. Association for Computing Machinery, New York, NY, USA.
- [9] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. DOI:<https://doi.org/10.1111/0022-4537.00153>
- [10] Hannah R. M. Pelikan and Mathias Broth. 2016. Why That Nao? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, ACM Press. DOI:<https://doi.org/10.1145/2858036.2858478>
- [11] Danielle Pillet-Shore. 2018. How to Begin. *Research on Language and Social Interaction* 51, 3 (July 2018), 213–231. DOI:<https://doi.org/10.1080/08351813.2018.1485224>
- [12] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems - CHI'18*, ACM Press. DOI:<https://doi.org/10.1145/3173574.3174214>
- [13] Stuart Reeves, Martin Porcheron, and Joel Fischer. 2018. "This is not what we wanted": designing for conversation with voice interfaces. *Interactions* 26, 1, 46–51. DOI:<https://doi.org/10.1145/3296699>
- [14] Harvey Sacks. 1995. *Lectures on conversation*. Wiley-Blackwell, London.
- [15] Emanuel A Schegloff. 1968. Sequencing in Conversational Openings. *American Anthropologist* 70, 6, 1075–1095. DOI:<https://doi.org/10.1525/aa.1968.70.6.02a00030>
- [16] Emanuel A Schegloff. 1988. Presequences and indirection: Applying speech act theory to ordinary conversation. *Journal of Pragmatics* 12, 1 (1988), 55–62.
- [17] Emanuel A Schegloff. 2007. *Sequence organization in interaction: Volume 1: A primer in conversation analysis*. Cambridge University Press, Cambridge.