Securely Capturing People's Interactions with Voice Assistants at Home: A Bespoke Tool for Ethical Data Collection

Angus Addlesee
Heriot-Watt University
Edinburgh
a.addlesee@hw.ac.uk

Abstract

Speech production is nuanced and unique to every individual, but today's Spoken Dialogue Systems (SDSs) are trained to use general speech patterns to successfully improve performance on various evaluation metrics. However, these patterns do not apply to certain user groups - often the very people that can benefit the most from SDSs. For example, people with dementia produce more disfluent speech than the general population. In order to evaluate systems with specific user groups in mind, and to guide the design of such systems to deliver maximum benefit to these users, data must be collected securely. In this short paper we present CVR-SI, a bespoke tool for ethical data collection. Designed for the healthcare domain, we argue that it should also be used in more general settings. We detail how off-the-shelf solutions fail to ensure that sensitive data remains secure and private. We then describe the ethical design and security features of our device, with a full guide on how to build both the hardware and software components of CVR-SI. Our design ensures inclusivity to all researchers in this field, particularly those who are not hardware experts. This guarantees everyone can collect appropriate data for human evaluation *ethically*, *securely*, and in a timely manner.

1 Introduction

Data collection is vital if we are to create more *natural* and more *accessible* spoken dialogue systems (SDSs) embedded within voice assistants and social robots (MacWhinney et al., 2004; Yu and Deng, 2016; Devlin et al., 2018; Williams et al., 2022). As these technologies are applied with admirable goals in the healthcare domain, general voice datasets lose the ability to accurately reflect the end-user. For example, speech production changes as cognition declines; people use more prepositions, slow their speech rate, pause more frequently mid-sentence, and pause for longer durations as dementia progresses (Boschi et al., 2017; Slegers et al., 2018; Zhu et al., 2018; Nasreen et al., 2019; Luz et al., 2021). We can refine evaluation metrics endlessly, but a system's practical benefit to the end-user remains unknown without data representing that specific user group.

It is critical that this data is collected *ethically* and *securely* as vulnerable user groups are particularly common in the healthcare domain. Issues around consent have been explored as individuals develop cognitive impairments, but identifiable information will still be captured and this is a concern (Haider and Luz, 2019; Addlesee and Albert, 2020). Data privacy does not just affect people with cognitive impairments however, people affected by sight loss can unwittingly reveal sensitive information (Ramil Brick et al., 2021; Baker et al., 2021), as will individuals conversing during a GP consultation (Ryan et al., 2019).

Off-the-shelf devices are not secure. If used, all sensitive data that is captured will be fully accessible to anyone if the device is lost or stolen. Very few audio recorders even exist with this capability due to copyrighting of encrypted audio codecs (Chege, 2019), and the ones that do exist are expensive and not applicable or adaptable for ethical data collection (see Table 1 in which we have included the Philips DPM8000 for comparison). This is a serious risk that should not be overlooked when seeking ethical approval. In this short paper we will detail a bespoke device, called CVR-SI, with *ethics* and *data security* at the core of its design.

2 Previous Work

A data capture device, called CVR, was used to collect similar data in a less-sensitive domain (Porcheron et al., 2018). This device was used to collect family interactions with Amazon Alexa devices within participants homes over a period of one-month. While we would argue that the CVR would have certainly captured personally identifiable information, this risk is heightened in our

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 25 - 30
December 7, 2022 ©2022 Association for Computational Linguistics

Desired Features	DPM	CVR	CUSCO	CVR-SI
Captures audio	✓	✓	✓	✓
Clearly indicates when 'on' to user	✗	✓	✗	✓
Clearly indicates when 'recording' to user	✗	✓	✗	✓
User can easily stop the device listening	✗	✓	✗	✓
Data is securely stored	✓	✗	✓	✓
Data is encrypted in real-time	✓	✗	✓	✓
Recording uses wake-word detection	✗	✓	✗	✓
Adequate Storage Capacity	✗	✓	✓	✓

Table 1: A list of desired system features with indicators of their presence within each device.

domain of interest, that is healthcare.

A security-focused data capture device, called CUSCO (Addlesee and Albert, 2020), was created for sensitive in-person data collections like medical conversations. Participants would interact or complete a task with the researchers in attendance at all times. Therefore, this device does not face the same challenges as a long-term device that cannot be monitored or controlled mid-study. CUSCO does implement real-time data encryption however, a critical feature that ensures no data can be accessed even if the device is stolen *during* recording.

With advice from the creators of the CVR (Conditional Voice Recorder), we used their work as a starting point. Hence our device's name: CVR-SI (Conditional Voice Recorder for Sensitive Information). We then adapted the data security features of CUSCO and integrated them to create CVR-SI. In Table 1 you can see which of our desired features the CVR, CUSCO, and Philips DPM8000 devices are missing. For example, the user must be able to easily stop the device from 'listening' while a health worker is visiting.

CVR-SI has been ethically approved for use by Heriot-Watt University's Ethics Committee and has been successfully used within vulnerable participant's homes. In the following sections we will describe the device's software, explain the security features, detail exactly how to construct the CVR-SI, and highlight components that tackle ethical issues[1]. The final CVR-SI can be seen in Figure 1.

3 Device Software

3.1 Wake-Word Detection

As mentioned above, we used the CVR (Porcheron et al., 2018) as the starting point of our CVR-SI device. We therefore started with Snowboy's wake-word detection, trained to detect "Alexa", by Kitt AI (Kitt-AI, 2020). For security reasons, this wake-word detection must take place on-device and cannot use a cloud service (Cho et al., 2018; Bolton et al., 2021; Singh et al., 2021). This ensures all data remains offline and cannot be intercepted. Additionally, as the CVR-SI does not need to connect to home wifi, the setup is simple and non-invasive.

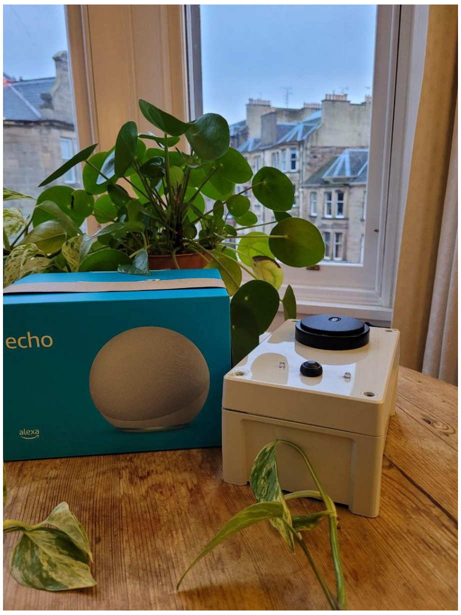

Figure 1: The fully built CVR-SI device with the accompanying Alexa voice assistant.

Another popular on-device Snowboy alternative is called Porcupine (Picovoice, 2022). It is more recent and their benchmark[2] suggested that it would noticeably outperform Snowboy. We explored this with both system's "Alexa" models at different activation sensitivities and with utterances containing various phrases similar to the target wake-word (other wake-words are available).

We want the CVR-SI to activate more often than the actual Alexa voice assistant, capturing instances where Alexa fails to listen to the user's utterance. In order to test this we prepared some phrases that are similar to "Alexa" (for example: "Lexa", "a Lexus", and "Alexis"), and some that are less-similar (for example: "My Lexus", "election", and "a lexeme"). We set up Porcupine and Snowboy with identical microphones and ran them simultaneously at the same distance from the test user. Each test phrase was spoken within a sentence at a range of different sensitivities. We found that both models performed indistinguishably. We do not dispute Porcupine's benchmark results and suggest referring to them for a more detailed and rigorous evaluation. We simply

[1] A full writable .img of CVR-SI and a list of specific hardware component URLs can be found here for reproducibility: https://github.com/AddleseeHQ/CVR-SI

[2] https://github.com/Picovoice/wake-word-benchmark

conclude that switching to Porcupine would not impact the CVR-SI's overall performance enough *practically* to warrant carrying out the potentially troublesome task.

3.2 Audio Buffer

As mentioned, we want the CVR-SI to capture all failed interactions with Alexa, and this includes failed wake-word detection. The original CVR stored a 60-second buffer of audio for this reason, assuming that failed interaction would be followed by another interaction attempt. It was found that users would repeat their utterance, clearly enunciating and stripping disfluencies from their speech (Porcheron et al., 2018). We kept this buffering feature as it is particularly important in the healthcare domain. For example, we can discover whether people with dementia learn to clean their speech of disfluencies in the same manner. Storing a constant buffer of audio is a security concern as people are certainly going to utter personally identifiable information in their own home at some point. This highlights the need for *real-time* encryption.

3.3 Data Security

Data security is imperative to avoid ethical and legal ramifications following a data breach (Romanosky et al., 2014; Labrecque et al., 2021; Masuch et al., 2021). These concerns are magnified when collecting data with vulnerable participants (Kavanaugh et al., 2006; Nordentoft and Kappel, 2011; McReynolds et al., 2017). We therefore reproduced the data security focused design of CUSCO (Addlesee and Albert, 2020) by using an audited, open-source, disk encryption software called Veracrypt (Knight, 2017). Data is encrypted in real time and can only be accessed with a generated key. This ensures the security of the entire corpus during collection, transport, exchange, and storage. The CVR-SI can therefore be handled by multiple parties without any of them being able to access collected data.

4 Device Hardware

We created several prototypes of the CVR-SI device, and then built this device at scale (20 units) as seen in Figure 2. Various hardware design decisions were made to mitigate ethical concerns[3].

Figure 2: All of the materials laid out to build 20 CVR-SI devices with accompanying Alexa assistants.

4.1 Raspberry Pi and Storage

Each CVR-SI uses a Raspberry Pi 3 Model B+ as its foundation. We made this decision based upon the CVR-SI performance requirements. Wake-word detection needs to run over audio continuously as the buffer and stored audio is encrypted live. The software runs smoothly on the Raspberry Pi 3 Model B+, so the additional cost to upgrade to a higher model was deemed redundant.

A microSD card is needed to run the software and store the corpus. We initially used a 16Gb microSD card, but this was not sufficient due to our deliberate over-capturing discussed above. Some participants placed the CVR-SI next to their TV or radio, which frequently activated the device. We therefore upgraded to a 256Gb version of the software (the only difference being the capacity of the encrypted drive), and this is sufficient for 1-month collections. Both the 16Gb and 256Gb versions of the software will be made available.

4.2 Microphone

As the purpose of the CVR-SI is to capture audio, a suitable microphone is required. We selected three off-the-shelf microphones at varying price points, and we tested the audio recording quality. We set up all three microphones in two different rooms. These were placed right next to each other and run simultaneously to avoid any external factors like background noise. We walked around the room while talking to investigate how each microphone handled audio input from various distances and orientations. We also spoke at varying volumes and while facing away from the microphones to test different user setups. Some participants may speak

[3] The full construction manual with component links, tool specifications, and circuit diagrams can be found here: https://github.com/AddleseeHQ/CVR-SI

more quietly (Maslan et al., 2011), so this was a vital deciding factor. We found that the cheapest microphone had a background crackle at all times (we tested multiple, so this was not a defect). This crackle made it very difficult to hear what was being said at long distances, and low volumes. It was therefore discounted as an option. The other two microphones were similar as the utterances could always be heard. The most expensive microphone had many interesting features, including a bidirectional mode for example. These features were not useful in this omnidirectional setting, so we selected the mid-range microphone due to cost.

4.3 Peripherals for Ethical Design

In order to support a few design features that we considered ethically necessary, LEDs and a button are required (Pearl, 2016; Abdi et al., 2019). One green LED lights to clearly indicate when the CVR-SI is on and *listening*. One red LED lights to clearly indicate when currently *recording*. The button stops the device recording and listening when pressed, and then reactivates the device to listen once pressed again - indicated by the green LED. This feature can be used when family members are visiting, health workers are in the house, or simply if the participant is having a conversation that they don't want to be captured.

The communication between the Raspberry Pi and the peripherals is achieved through a circuit board that has to be soldered. We designed the circuit with suitable resistors to protect the LEDs and button, ensuring they do not burn out in-use. All of the circuitry and Raspberry Pi is housed within a simple container with holes drilled into it for the lights, button, and microphone cable. A soldering iron, drill, drill bits (matching the LED and button sizes), and glue (to attach the microphone securely) is needed to build the CVR-SI. Please follow the links and guide on GitHub for step-by-step guidance and the circuit diagram. The device build process can be seen in Figure 3.

5 Findings from Use in Practice

In this short paper we have detailed both the software and hardware of CVR-SI, a data capture device with both *data security* and *ethics* at the core of its design. The CVR-SI has been ethically approved and used to capture interactions between people with dementia and Alexa voice assistants in their own home. We have already learned a great deal from real-world deployment, for example the microSD storage upgrade described in Section 4.1.

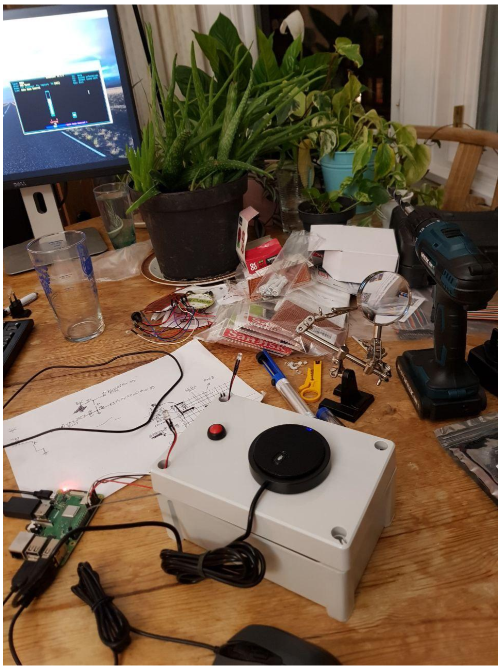

Figure 3: The CVR-SI mid-construction.

Participants have reported using the device's button to stop the CVR-SI capturing audio when family or health-workers are visiting, indicating that this feature is desired for privacy and that the LEDs are clear. One participant noted that they used the button at times they felt "big brother was listening". This is an understandable feeling that is generally felt with smart speakers (Lau et al., 2018), indicating again that the LEDs are clear and the button is a necessary device feature for participant comfort.

Although data analysis is yet to be complete, initial observations have revealed instances in which the Alexa does not activate when the user says the wake-word. The buffer has therefore proven to be a useful feature, driving the need for live encryption. Recordings have been clear and do not skip, demonstrating the sufficient capabilities of both the microphone and Raspberry Pi.

Finally, participants have described how they have used the Alexa device in their day-to-day lives. People with dementia have been able to reawaken their love for music, set reminders to take medication or walk their dogs, get help with their crosswords, and even find new recipes to help get involved with family mealtimes. Voice assistants can clearly have a positive impact, so we hope our work will accelerate voice accessibility research.

Ethical and Societal Implications

The next generation of voice assistants need to be more naturally interactive and accessible for everyone, especially as SDSs are increasingly applied in the healthcare setting. In order to make informed design decisions and effectively evaluate new dialogue systems with specific user groups in mind, potentially sensitive data *must* be collected. Off-the-shelf audio recorders are not secure and cannot be ethically approved for use, creating a barrier to complete crucial research.

This work will not only enable us to design dementia-friendly assistants and social robots in the future. We hope other researchers use the CVR-SI to make a positive impact with similar goals in mind, and in more general settings to ensure data privacy.

Acknowledgements

This work was funded by The Data Lab, Wallscope, and the EU H2020 program under grant agreement no. 871245 (SPRING)[4].

This device couldn't be used to collect valuable data without the generosity of participant's time - thank you all. I would also like to thank CENSIS for providing many of the Raspberry Pi's, Amazon Science for providing the Alexa devices, and Spotify for providing Spotify Premium to participants.

Finally - I would like to thank Derek Davis for facilitating the purchase of the device components, and Antero Duarte for his electronics advice.

References

Noura Abdi, Kopo M Ramokapane, and Jose M Such. 2019. More than smart speakers: security and privacy perceptions of smart home personal assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 451–466.

Angus Addlesee and Pierre Albert. 2020. Ethically collecting multi-modal spontaneous conversations with people that have cognitive impairments. *LREC Workshop on Legal and Ethical Issues*.

Katie Baker, Amit Parekh, Adrien Fabre, Angus Addlesee, Ruben Kruiper, and Oliver Lemon. 2021. The spoon is in the sink: Assisting visually impaired people in the kitchen. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 32–39.

Tom Bolton, Tooska Dargahi, Sana Belguith, Mabrook S Al-Rakhami, and Ali Hassan Sodhro. 2021. On the security and privacy challenges of virtual assistants. *Sensors*, 21(7):2312.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. 2017. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:269.

Isaac Chege. 2019. Best encrypted voice recorder.

Geumhwan Cho, Jusop Choi, Hyoungshick Kim, Sangwon Hyun, and Jungwoo Ryoo. 2018. Threat modeling and analysis of voice assistant applications. In *International Workshop on Information Security Applications*, pages 197–209. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fasih Haider and Saturnino Luz. 2019. A system for real-time privacy preserving data collection for ambient assisted living. In *INTERSPEECH*, pages 2374–2375.

Karen Kavanaugh, Teresa T Moro, Teresa Savage, and Ramkrishna Mehendale. 2006. Enacting a theory of caring to recruit and retain vulnerable participants for sensitive research. *Research in nursing & health*, 29(3):244–252.

Kitt-AI. 2020. Snowboy hotword detection. `https://github.com/Kitt-AI/snowboy`. Online; accessed 08 May 2022.

G Knight. 2017. Encrypt data using veracrypt.

Lauren I Labrecque, Ereni Markos, Kunal Swani, and Priscilla Peña. 2021. When data security goes wrong: Examining the impact of stress, social contract violation, and data type on consumer coping responses following a data breach. *Journal of Business Research*, 135:559–571.

Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–31.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.

Brian MacWhinney, Steven Bird, Christopher Cieri, and Craig Martell. 2004. Talkbank: Building an open unified multimodal database of communicative interaction.

Jonathan Maslan, Xiaoyan Leng, Catherine Rees, David Blalock, and Susan G Butler. 2011. Maximum phonation time in healthy older adults. *Journal of Voice*, 25(6):709–713.

[4] `https://spring-h2020.eu/`

Kristin Masuch, Maike Greve, and Simon Trang. 2021. What to do after a data breach? examining apology and compensation as response strategies for health service providers. *Electronic Markets*, 31(4):829–848.

Emily McReynolds, Sarah Hubbard, Timothy Lau, Aditya Saraf, Maya Cakmak, and Franziska Roesner. 2017. Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5197–5207.

Shamila Nasreen, Matthew Purver, and Julian Hough. 2019. A corpus study on questions, responses and misunderstanding signals in conversations with alzheimer's patients. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue-Full Papers. SEMDIAL, London, United Kingdom (Sep 2019), http://semdial.org/anthology/Z19-Nasreen semdial*, volume 13.

Helle Merete Nordentoft and Nanna Kappel. 2011. Vulnerable participants in health research: Methodological and ethical challenges. *Journal of Social Work Practice*, 25(3):365–376.

Cathy Pearl. 2016. *Designing voice user interfaces: Principles of conversational experiences*. " O'Reilly Media, Inc.".

Picovoice. 2022. Porcupine on-device wake word detection. https://github.com/Picovoice/Porcupine. Online; accessed 08 May 2022.

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*, pages 1–12.

Elisa Ramil Brick, Vanesa Caballero Alonso, Conor O'Brien, Sheron Tong, Emilie Tavernier, Amit Parekh, Angus Addlesee, and Oliver Lemon. 2021. Am i allergic to this? assisting sight impaired people in the kitchen. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 92–102.

Sasha Romanosky, David Hoffman, and Alessandro Acquisti. 2014. Empirical analysis of data breach litigation. *Journal of Empirical Legal Studies*, 11(1):74–104.

Padhraig Ryan, Saturnino Luz, Pierre Albert, Carl Vogel, Charles Normand, and Glyn Elwyn. 2019. Using artificial intelligence to assess clinicians' communication skills. *Bmj*, 364.

Abhishek Singh, Rituraj Kabra, Rahul Kumar, Manjunath Belgod Lokanath, Reetika Gupta, and Sumit Kumar Shekhar. 2021. On-device system for device directed speech detection for improving human computer interaction. *IEEE Access*, 9:131758–131766.

Antoine Slegers, Renee-Pier Filiou, Maxime Montembeault, and Simona Maria Brambati. 2018. Connected speech features from picture description in alzheimer's disease: A systematic review. *Journal of Alzheimer's Disease*, 65(2):519–542.

Louise R Williams, Myzoon Ali, Kathryn VandenBerg, Linda J Williams, Masahiro Abo, Frank Becker, Audrey Bowen, Caitlin Brandenburg, Caterina Breitenstein, Stefanie Bruehl, et al. 2022. Utilising a systematic review-based approach to create a database of individual participant data for meta-and network meta-analyses: the release database of aphasia after stroke. *Aphasiology*, 36(4):513–533.

Dong Yu and Li Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2018. Detecting cognitive impairments by agreeing on interpretations of linguistic features. *arXiv preprint arXiv:1808.06570*.