Speech-Controlled Smart Speaker for Accurate, Real-Time Health and Care Record Management

Jonathan E. Carrick[1,2], Nina Dethlefs[3], Lisa Greaves[2], Venkata M. V. Gunturi[1], Rameez Raja Kureshi[1], Yongqiang Cheng[4],

[1]School of Computer Science, University of Hull, Hull, HU6 7RX, UK,
[2]Connexin Ltd, K3 Business Park, Hull, HU5 1SN, UK,
[3]Department of Computer Science, Loughborough University, Loughborough, LE11 3TU, UK,
[4]Faculty of Technology: Computing, University of Sunderland, Sunderland, SR6 0DD, UK

Correspondence: j.carrick@hull.ac.uk

Abstract

To help alleviate the pressures felt by care workers, we have begun new research into improving the efficiency of care plan management by advancing recent developments in automatic speech recognition. Our novel approach adapts off-the-shelf tools in a purpose-built application for the speech domain, addressing challenges of accent adaption, real-time processing and speech hallucinations. We augment the speech-recognition scope of Open AI's Whisper model through fine-tuning, reducing word error rates (WERs) from 16.8 to 1.0 on a range of British dialects. Addressing the speech-hallucination side effect of adapting to real-time recognition by enforcing a signal-to-noise ratio threshold and audio stream checks, we achieve a WER of 5.1, compared to 14.9 with Whisper's original model. These ongoing research efforts tackle challenges that are necessary to build the speech-control basis for a custom smart speaker system that is both accurate and timely.

1 Introduction

Health and social care is one of the last major industries to undergo the digital transformation to improve management of information and connectivity (Glaser and Shaw, 2022; Konopik and Blunck, 2023). Reasons include challenges relating to data privacy, tech-literacy and scalability in a highly heterogenuous domain (Aceto et al., 2020). Transformation towards *Healthcare 4.0* is helped by integrating new artificial intelligence technologies into purpose-built smart devices (Wehde, 2019).

Yen et al. (2018) find that, even with the implementation of real-time electronic record management, healthcare administrators spend a quarter of their time on documentation and, due to typing distractions, information is missed. Combined with the job-demanding stresses that care workers experience (Wilberforce et al., 2012) it is clear that there is a need for simplified health care record management to help reduce the burden. This would further benefit those cared for as care resources become more optimised. One way to achieve a quicker, more efficient approach to care record management that is both complete and accurate is through automatic speech recognition (ASR, Ajami, 2016; Alharbi et al., 2021; Malik et al., 2021).

In this paper we focus on recognition of spoken English in the UK. However, typical off-the-shelf ASR models are often trained primarily on American-accented datasets (Vergyri et al., 2010; Mathur et al., 2020) and health and care in the UK is a diverse industry. This includes variations in dialects across the British Isles (MacKenzie et al., 2022), as well as foreign accents from care workers who originate from places such as Eastern Europe, Nigeria, and South Asia, amongst others[1]. Commercial smart speakers, such as Amazon's Alexa, showcase the potential of real-time ASR in a general home assistant setting (Hoy, 2018), and have been used in previous studies to improve well-being in social care (Edwards et al., 2021). However, to the best of our knowledge there is currently no device whose primary function is a smart administrative assistant for health and care workers.

Hence, we have set out to develop a custom-built speaker, starting with new research into the fundamental ASR basis. This paper introduces a novel approach and makes the following key contributions:

[1]https://www.skillsforcare.org.uk/Adult-Social-Care-Workforce-Data/Workforce-intelligence/publications/Topics/Workforce-nationality-and-international-recruitment.aspx

- Fine-tuning an ASR model for greater scope of accent recognition
- Adapting the model and adding voice commands to a real-time recognition pipeline
- Audio processing methods to prevent speech hallucinations caused by background noise and predictive text

2 Smart Speaker Design

We began with a review into different accessible ASR models. In the context of finding the best-suited framework to build and adapt our custom system around, our initial testing of models included wav2vec 2.0 (Baevski et al., 2020) and VOSK[2] with the Kaldi toolkit[3]. Ultimately, we decided to utilise Open AI's Whisper (Radford et al., 2023) model, due to its free, open license, ongoing development in state-of-the-art ASR and ease of adapting to our own needs with Python.

Rather than use an established smart speaker, we develop our own hardware[4] to, first, keep the solution cost-effective for customers in the care sector, who might not need or want a full-fledged commercial system, and second, to keep full control of confidentially sensitive data. While the device itself can run most of the required data management functionality, speech inference runs on a GPU cloud-server. We use sound cues to give audio feedback to the user to confirm that voice commands are understood and functions are carried out.

3 Accent Adaption

Despite Whisper's extensive training, we find that it struggles to generalise to a broad variety of British as well as other foreign accents found in the care sector. Graham and Roll (2024) find a similar bias towards North American over other British accents.

We start by adapting Whisper to better recognise the variations in six different British accents: 'Southern', 'Northern', 'Midlands', 'Scottish', 'Welsh', 'Irish' from the OpenSLR[5] dataset of ∼30 hours of spoken English (Demirsahin et al., 2020). With this dataset, we fine-tune Whisper's `medium.en` model, which balances speed with accuracy, and is the largest model that we can enforce with English-only recognition; the larger models would occasionally incorrectly recognise speech as a different language and attempt to translate. Furthermore, the large model requires twice the VRAM but offers diminishing returns in performance (Radford et al., 2023) and we do not require the additional feature of multi-language ASR.

Our fine-tuning[6] is done with 95% of the data, with the remainder used for validation. By observing the evolution of the word error rate (WER)[7] and validation loss through training, we find that the model begins to plateau half an epoch in and converges in approximately two training epochs, beyond which the model begins to overfit the dataset. Training for 3,072 steps (batch size 16 and evaluation every 256 steps), we achieve a minimum WER of 1.0 at step 2,048, where validation loss is also minimised[8]. This checkpoint defines the fine-tuned model uesd in this study. Table 1 shows the improvement in WERs per accent through fine-tuning. Recognition of all accents surpasses human-level transcription (Amodei et al., 2016; Stolcke and Droppo, 2017; Lippmann, 1997).

4 Dealing with Hallucinations

Off-the-shelf, Whisper requires an audio file uploaded manually in a controlled process. Adapting Whisper to a real-time pipeline presented an unexpected challenge: hallucinations in ASR are defined as 'recognised' text that arises completely independently from what is spoken. While not limited to real-time ASR (Dolev et al., 2024), the phenomenon becomes more apparent in this adaption. Hallucinations are not simply mis-recognised words or phrases, but recognition in the absence of speech. These need to be prevented as hallucinated text, while often common words/phrases, e.g. "Thank you", "Yes", can be unexpected or even harmful (Koenecke et al., 2024). Without automatic mitigation, hallucinations may cause confusion in care records and require additional work to fix, resulting in the opposite of what we aim to achieve with our smart speaker. We find two causes of hallucination in our setting, as detailed below.

[2] https://alphacephei.com/vosk/

[3] https://github.com/kaldi-asr/kaldi

[4] We use a Raspberry Pi (Model 4), 8GB RAM, GPIO speaker & USB speaker, USB microphone, one-button 'keyboard' and touchscreen.

[5] https://www.openslr.org/83/

[6] We follow a similar method to https://huggingface.co/blog/fine-tune-whisper, adapted to our dataset.

[7] We use the WER implementation from https://huggingface.co/spaces/evaluate-metric/wer

[8] Model fine-tuning was done using Viper (https://hpc.wordpress.hull.ac.uk/), taking approximately 70 hours to optimise.

Accent	WER before fine-tuning	WER after fine-tuning	Number of test samples
'Southern'	16.8	0.9	451
'Midlands'	13.9	1.3	25
'Northern'	16.3	0.9	158
'Welsh'	16.4	1.2	148
'Scottish'	17.9	1.4	93
'Irish'	21.0	2.7	19
Weighted average	**16.8**	**1.0**	

Table 1: WERs for Whisper before and after fine-tuning with the OpenSLR dataset. WER scores are rounded as higher precision is not meaningful with these sample sizes. Averages are weighted as proportions of each accent in the test data differ, shown by the number of test samples. True WER scores (maximum precision) were included in calculation of averages, that are then rounded at the end. Class imbalance is due to random sampling and reflects the number of volunteers for each accent during creation of the original dataset.

4.1 Recognising Background Noise

The first cause is due to continually monitoring with a microphone. If audio input, regardless of its nature, is automatically passed to Whisper, the model will try to process it into text, even if nothing has been spoken. In this case, Whisper tries to recognise speech from effective silence, i.e. background noise, and results in speech hallucinations.

The dynamic energy threshold[9] we employ for microphone input is not sufficient in separating clear speech from background noise. Hence, we apply a check in each processing loop before passing the queued audio data to the ASR (Figure 1). A signal-to-noise ratio (SNR) threshold is defined during initialisation and we choose SNR $= 50$, determined empirically by testing in different environments, e.g. quiet room at home, noisy office. Then, for each audio loop, the SNR is calculated as

$$\text{SNR} = \frac{\text{Signal Power}}{\text{Noise Power}} = \frac{\sigma_S^2}{\sigma_N^2}, \qquad (1)$$

where σ^2 is the variance (standard deviation squared) for signal S and noise N, and, if it is greater than the threshold, the audio clip is passed to Whisper. σ_S^2 is calculated for each loop's audio clip. σ_N^2 is calculated upon device startup when the speaker records the background noise level of the current environment. We limit σ_N^2 to the range $0.5\text{–}5\times10^{-6}$, determined empirically, as, too low and any sound will be passed to the ASR as speech, and too high and no speech will be recognised.

[9] https://pypi.org/project/SpeechRecognition/2.1.3/

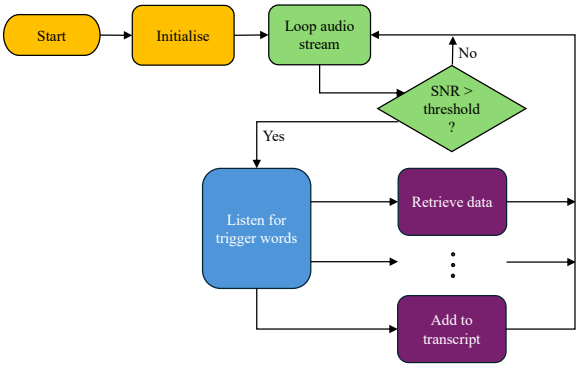
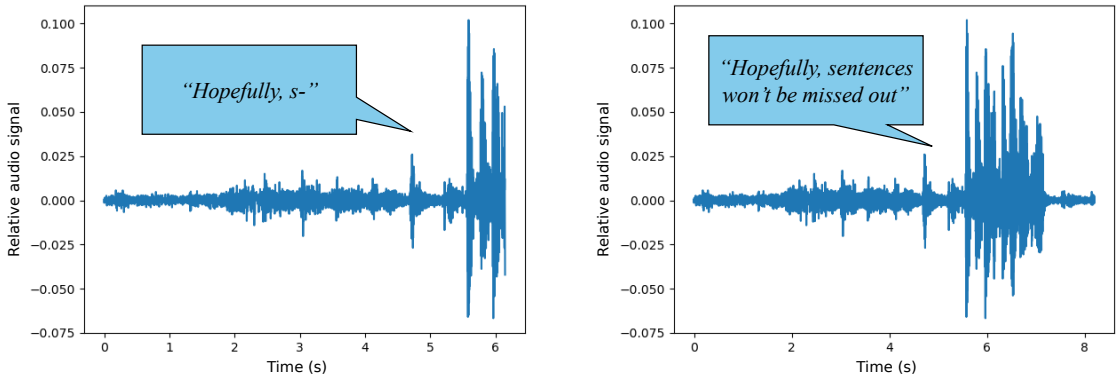


Figure 1: Flow chart of the speaker pipeline. After set up (yellow) and starting the audio loop (green), functions (purple) are evoked via voice commands (blue).

4.2 Record Timeout

LLMs such as Whisper are typically trained on sequences of words (Sutskever et al., 2014; Radford et al., 2023). Therefore, when an initial word is passed to a trained model, it will anticipate the next word/s, based on common sequences it has learned from many hours of training. This learned 'predictive text' means that, if the model considers the speech input to be only part of a phrase, Whisper may automatically output what it thinks the full phrase should be. This form of hallucination occurs when the microphone recording loop times out before a word/phrase is completed. Figure 2 demonstrates this effect with a waveform of speech and its corresponding recognised text, before and after a phrase is completed.

We find a recorded timeout of 2 seconds suitable to balance the trade-off between 'real-timeness' and ASR accuracy. We implement predictive-



(a) Recognised speech: "Hopefully, something will work out"

(b) Recognised speech: "Hopefully, sentences won't be missed out"

Figure 2: Waveforms of speech where the phrase is cut off by a recorded timeout and where the complete phrase is allowed to be fully recorded. Speech bubbles show the true speech recorded and their captions state ASR output.

Real-time setup	WER	WER*
Original model: both hallucination methods	15.4	14.9
Fine-tuned model: without SNR checks	8.7	7.2
Fine-tuned model: without SNR checks, with pauses	185.5	182.4
Fine-tuned model: without audio stream checks	57.8	55.8
Fine-tuned model: both hallucination methods	6.6	5.1
Fine-tuned model: both hallucination methods, with pauses	6.0	5.7

Table 2: WERs for our different real-time tests. 'Both hallucination methods' here means that the SNR threshold and audio stream checks are both in place. The 'original model' is Whisper's `medium.en` model. 'With pauses' means that 3-second pauses were taken after every sentence. WER* denotes the word error rate when we ignore errors due to lexical differences that can still be considered as the recognition having the correct understanding.

hallucination prevention by tracking the audio data that is passed to the ASR. This method ensures that the each new transcript entry[10] is only saved when a full phrase is spoken, with recognised text corresponding only to the processed audio. When data between successive recognitions overlap, we ensure that the current transcript entry is updated with the most recent recognition. New transcript entries are added when at least 3 seconds have passed since the previous recognition *and* the recognition is on all new audio data. For additional robustness, we combine this with a comparison of texts between successive recognitions to check whether the current recognition is a continuation of the previous entry.

5 Real-time Recognition Results

We present the results of real-time recognition with our fine-tuned model and hallucination-prevention methods in Table 2. To test performance of our real-time ASR pipeline, the same script of 332 words (an excerpt from a paper draft) was read for different setups including comparison of our fine-tuned model to the original Whisper model, and with/without our hallucination methods. The reading for each setup was done in the same office meeting room in one take, where some background noise from adjacent rooms was present to help simulate a real environment where our device may be used, and was read by the same speaker who self-identifies as having a 'Northern' accent. For each test, the transcription is compared to the original script and we calculate the WER.

The model's full potential is demonstrated with both hallucination methods reaching a minimum WER across all tests of 6.6. The improvement over the original Whisper model is substantial (down from WER = 15.4), although limited compared to the reduction achieved with fine-tuning (Table 1). We attribute this to the real-time adaption where arbitrarily-segmented audio clips are input automat-

[10]Each 'entry' is a string element in the transcription list.

ically and the test environment, where some levels of background noise were present.

Some of the errors we find in the transcripts are not necessarily inaccuracies, but rather mismatches with the original script. For example, sometimes the model will recognise "*UK*" as "`United Kingdom`", and "*100,000s*" as "`hundreds of thousands`". While we consider the original script as the 'ground truth' for these tests, considering these differences as *correct*, WER reduces to as low as 5.1 with our fine-tuned model. Results for these cases are shown in Table 2 under WER*.

Without the SNR checks in place, there is less reduction in performance (WER = 8.7), however, the crucial importance of including a SNR threshold is demonstrated when 3-second pauses are taken after every sentence. Recognition from silence/background noise results in multiple hallucinations throughout. The generated text during these quiet moments is often gibberish, repeated out multiple times and with no relation to the context of the previous speech, increasing WER to as high as 185.5. In comparison, the same test with speech pauses using both hallucination methods, achieves similar results to the first test: WER = 6.0.

Finally, we test our fine-tuned model in real time without checking the audio stream for repeated recognitions of overlapping data. The WER is again high at 57.8 and results in several instances where a sentence is hallucinated or repeated multiple times in the transcript.

These results highlight that our fine-tuned model is more than twice as effective as Whisper's original model and that hallucination prevention is essential to achieve the lowest WERs possible.

6 Conclusion

We demonstrated that an off-the-shelf Whisper is not well-adapted to a wide range of spoken British accents and that WERs can be reduced substantially through fine-tuning to the set of target varieties. Adapting Whisper as a real-time ASR results in the unexpected side effect of speech hallucinations. This is addressed by enforcing a SNR criterion in each audio clip and tracking audio data passed to the ASR to ensure that recognised text consists of complete and accurate phrases.

Future work will include greater accent scope, integration into health and care plan systems, sophisticated care data querying and monitoring methods, and trigger/alert systems to improve administration efficiency and help identify errors. Upon successful deployment of these features, we will trial our smart speaker in a real care-home environment to gain a better understanding of technological capabilities, user requirements and to maximise the social impact of our specialised speaker system.

Limitations

Our initial fine-tuning of Whisper that is described covers a range of British accents from a single dataset. We would like to expand on this, especially with accents representing the diversity of health and care workers in the UK, but have not yet been able to because of a lack of available datasets with suitable coverage of a variety of accents. Initial testing of ASR performance in real time was done with a single speaker only for our pilot speaker. We are planning to expand this in future.

Ethical Considerations

Ethical reviews, including draft consent forms, have been completed and approved to prepare for user testing. While initial testing will be done with dummy care data, we have plans in place to follow General Data Protection Regulation with the handling of any sensitive information in the case of in-situ health and care environments. As we progress in our development, we will address privacy concerns with secure logins and encryption methods. Measures are being taken for accurate recording of important information, especially with regards to treatments, medicine, etc., following guidelines, e.g. from the British National Formulary[11].

Acknowledgments

This project was supported by an Innovate UK Knowledge Transfer Partnership (KTP/13520). Much support from the University of Hull is gratefully acknowledged, including computing resources, and input and insightful comments from researchers in Health Sciences, the ethics committee and the Dementia Advisory Group. We also thank Connexin Ltd for making this project possible. Our thanks extends to our connections in adult social care for suggestions on user requirements. Finally, many thanks to the organisers of the 15th International Workshop on Spoken Dialogue Systems Technology and our anonymous reviewers.

[11] https://bnf.nice.org.uk/

References

Giuseppe Aceto, Valerio Persico, and Antonio Pescapé. 2020. Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. *Journal of Industrial Information Integration*, 18:100129.

Sima Ajami. 2016. Use of speech-to-text technology for documentation by healthcare providers. *The National medical journal of India*, 29(3):148.

Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9:131858–131876.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. 2020. Open-source multi-speaker corpora of the English accents in the British isles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6532–6541, Marseille, France. European Language Resources Association.

Eyal Dolev, Clemens Lutz, and Noëmi Aepli. 2024. Does whisper understand Swiss German? an automatic, qualitative, and human evaluation. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 28–40, Mexico City, Mexico. Association for Computational Linguistics.

Katie J Edwards, Ray B Jones, Deborah Shenton, Toni Page, Inocencio Maramba, Alison Warren, Fiona Fraser, Tanja Križaj, Tristan Coombe, Hazel Cowls, and Arunangsu Chatterjee. 2021. The use of smart speakers in care home residents: Implementation study. *J Med Internet Res*, 23(12):e26767.

John Glaser and Stanley Shaw. 2022. Digital transformation success: What can health care providers learn from other industries? *Catalyst non-issue content*, 3(2).

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2):025206.

Matthew B. Hoy. 2018. Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1):81–88. PMID: 29327988.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless whisper: Speech-to-text hallucination harms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1672–1681, New York, NY, USA. Association for Computing Machinery.

Jens Konopik and Dominik Blunck. 2023. Development of an evidence-based conceptual model of the health care sector under digital transformation: Integrative review. *J Med Internet Res*, 25:e41512.

Richard P. Lippmann. 1997. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

Laurel MacKenzie, George Bailey, and Danielle Turton. 2022. Towards an updated dialect atlas of british english. *Journal of Linguistic Geography*, 10(1):46–66.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457.

Akhil Mathur, Fahim Kawsar, Nadia Berthouze, and Nicholas D. Lane. 2020. Libri-adapt: a new speech dataset for unsupervised domain adaptation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7439–7443.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Andreas Stolcke and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. In *Interspeech 2017*, pages 137–141.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. 2010. Automatic speech recognition of multiple accented english data. In *Interspeech*, pages 1652–1655.

Mark Wehde. 2019. Healthcare 4.0. *IEEE Engineering Management Review*, 47(3):24–28.

Mark Wilberforce, Sally Jacobs, David Challis, Jill Manthorpe, Martin Stevens, Rowan Jasper, Jose-Luis Fernandez, Caroline Glendinning, Karen Jones, Martin Knapp, Nicola Moran, and Ann Netten. 2012. Revisiting the causes of stress in social work: Sources of job demands, control and support in personalised adult social care. *The British Journal of Social Work*, 44(4):812–830.

Po-Yin Yen, Marjorie Kellye, Marcelo Lopetegui, Abhijoy Saha, Jacqueline Loversidge, Esther M Chipps, Lynn Gallagher-Ford, and Jacalyn Buck. 2018. Nurses' time allocation and multitasking of nursing activities: A time motion study. *AMIA Annu Symp Proc*, 2018:1137–1146.

A Appendix

A.1 Voice Controls and Speaker Functions

Table 3 details some of the main trigger words and their functions in our smart speaker pipeline. Initialisation steps include microphone calibration, defining starting settings, e.g. 'asleep', and beginning to listen in the background. The different functions are only carried out when the appropriate 'trigger' words are located in the recognised text, e.g. "*Activate and retrieve latest medication entry*". In the future we would like to add a more conversational-agent approach, with a text-to-speech output to fulfil the role of 'speaker'.

A.2 Fine-tune Training

Figures 3 and 4 show the fine-tuning evolution of training/evaluation loss and WER respectively.

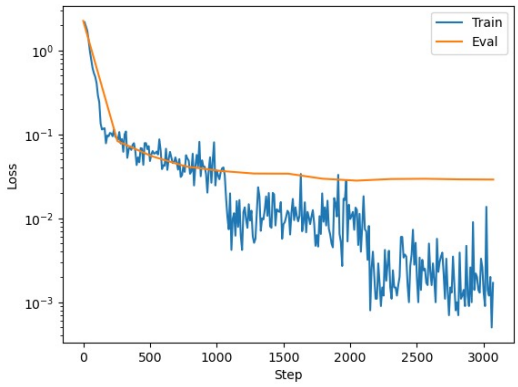


Figure 3: Loss on training and evaluation (test) data. The loss scale is logarithmic to help visualise the difference between loss evolutions. Evaluation loss is minimised at step 2048, beyond which a small degree of overfitting is observed. This approximately coincides with training over two epochs. Evaluation is done every 256 steps due to computational time constraints.

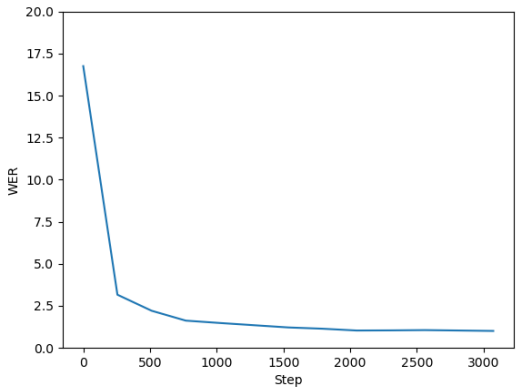


Figure 4: Through fine-tuning, WER on test data reduces and plateaus quickly, minimising at step 2048.

Trigger word	Functionality
Wake word, e.g. *"Activate"*	Wake speaker and unlock all other functionalities
Care record section, e.g. *"Medication"*	Start recording transcript for the given section
"Sign off"	Save transcript to care record with date/time-stamp
"Undo"	Removes most recent transcript addition
"Retrieve"	Return data from care record section
"Restart"	Erase current transcript and sleep the device again

Table 3: The main speaker functions and the trigger words that activate them.