Safe Audio AI Services in Smart Buildings

Jennifer Williams, Vahid Yazdanpanah, and Sebastian Stein
{j.williams,v.yazdanpanah,s.stein}@soton.ac.uk
University of Southampton, United Kingdom

ABSTRACT

Audio AI services present an opportunity to conceptualise smart buildings in a new light. Microphones can capture fine-grained audio information that can be used for determining how many people are inside of a building, where they are, and what kinds of activities are taking place. This information can feed into smart resource management systems or it could be used for assistive technologies. Generally speaking, audio is regarded as a less intrusive type of information collection than video surveillance, but significant issues of privacy and security persist with audio capture. Such issues warrant a serious discussion about how safe it is to use audio-capture in smart buildings for AI decision-making. This position paper initiates a discussion of research directions for the safety of audio services related to three key areas: data degradation strategies, dynamic customisation of tools, and privacy-aware technologies. In each area, we identify key challenges and highlight solution concepts with the potential to address the issue.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → *Adversarial learning*; *Speech recognition*; • **Computer systems organization** → Embedded software.

KEYWORDS

audio, accessibility, occupancy, privacy, smart buildings

ACM Reference Format:
Jennifer Williams, Vahid Yazdanpanah, and Sebastian Stein. 2022. Safe Audio AI Services in Smart Buildings. In *The 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22), November 9–10, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3563357.3564076

1 INTRODUCTION

A variety of opportunities arise from incorporating artificial intelligence (AI) services based on speech and audio into smart buildings, as such services have the potential to improve experiences of individuals and groups. One class of applications incorporates audio-capture for intelligent technologies that require information such as monitoring of room occupancy or movement and activities of people throughout a building in order to optimise energy resource consumption, e.g., by regulating heating, ventilation or lighting [10]. Another class of applications involve using audio capture to provide assistive services, for example for deaf and hard of hearing populations (DHH), such as audio scene analysis, talker localization, alerts for audible events, or new wearable devices that work in synergy with microphones inside smart buildings [18].

Technologies that involve audio capture can give rise to a range of issues relating to *safety*. We understand and refer to safety of audio AI systems in a broad sense. This includes the reliability of data degradation strategies (detailed in Section 2.1), their fairness in the way they treat sensitive information (detailed in Section 2.2), and awareness of privacy (detailed in Section 2.3). Safety is important particularly from those who are concerned about relinquishing control of their privacy or security in the surrounding environment alongside the perception of continuous surveillance. Oversight from GDPR [25] highlights idealised protections, though some of these protections are the responsibility of technology developers. Providing audio AI services in smart buildings implies installing multiple microphones throughout a building. While different audio applications have differing levels of autonomous decision making, they have in common the need to address consumer concerns of safety. Safety issues in this context may range from actual or perceived violations of privacy and security, to problematic decisions made by an audio AI system [9].

Against this background, this position paper explores the safety issues of audio AI for three key areas: (1) performance reliability from data degradation, (2) the role of personal and dynamic customization, and (3) privacy-aware technologies. We look at each of these areas in terms of two safety-related trust relationships. In each case, we identify open research challenges, define key components of the challenge, and propose some useful strategies for understanding how the challenge could be overcome through research or by taking advantage of currently available solutions.

2 CONCEPTUAL ANALYSIS: SAFETY TRUST RELATIONSHIPS

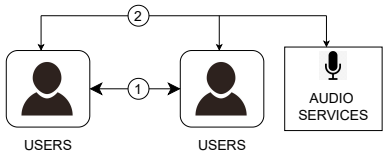

Figure 1: Conceptual diagram showing two forms of trust for audio AI services: (1) interpersonal trust between users and their agreement to participate in the service together, and (2) the trust between users and the audio AI services.

We introduce two overarching relationships in Figure 1 that are important for AI safety, and these relationships will form the

BuildSys '22, November 9–10, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9890-9/22/11…$15.00
https://doi.org/10.1145/3563357.3564076

foundation of our discussion of safety and trust throughout this paper. The first relationship involves how audio AI services may affect safety, privacy, and comfort between users (interpersonal). Some audio AI services capture audio from bystanders who are not directly involved in a service, but whose presence and activities may contribute audio data to a service meant for another. Examples of this relationship can be understood from audio scene analysis, audio event detection, or closed-captions that provides acoustic information to the DHH communities [18]. Interpersonal trust is important because bystanders may inadvertently participate in the co-construction of acoustic scenes and this is a recurring issue in many AI systems [28]. The second relationship from Figure 1 involves the trustworthiness of the AI service (i.e., the trust between the service and the users). This type of trust describes the reliability and fairness of AI-based decision making. In this relationship, users must be willing participants and relinquish their audio data while also feeling safe with regard to how their data will be processed and used. Another part of this relationship is whether users have an ability to opt out of participating in the audio service for a particular context (e.g., microphones installed in restrooms to support users with physical conditions but unnecessary for others).

3 RESEARCH DIRECTIONS AND CHALLENGES

3.1 Reliability: Data Degradation Strategies

Consumer trust is dependent upon a series of factors, and in this section we examine the technical effectiveness and reliability of algorithmic solutions. Installing microphones throughout smart buildings can make some people uneasy. Many people have a reasonable expectation of privacy when it comes to how their smart technologies store, process, and transmit audio data [5]. Recent work in audio-based activity recognition has shown that people begin to trust audio privacy more if monitored speech has been rendered unintelligible to humans through an audio degradation technique called *scrambling* [13]. Some types of scrambling or noise may make audio content unintelligible to humans, but does not neutralise information in the audio signal for an AI system. Another audio degradation technique involves injecting different types of noise into data in order to prevent salient features (such as identity, spoken content, or emotion) from being detected, while retaining enough information to perform the target task.

Research Direction 1. ***Reliable audio AI services require privacy-preserving techniques that do not impede the AI service, and this can be met through the creation of methods that perform data degradation and verified through adversarial tasks.*** *Let x be the inputs to an AI system $\mathbb{S}$ that is optimised to output a decision y (for example, a classification prediction problem). Let x also be the inputs into an AI counter-system $\mathbb{A}$ such that the counter-system is optimised to perform poorly on a separate task. The counter system may act as an adversary that attempts to gain access to private information, for example a separately trained speaker recognition system that attempts to recover individual identity from audio. The aim is to develop a method $\mathbb{N}$ that creates a noise n to be applied to the input x in order to transform x into x' so that it satisfies optimal performance in system $\mathbb{S}$ and degraded performance in system $\mathbb{A}$.*

Data degradation is an open research area and this line of enquiry can expose a class of algorithmic vulnerabilities called *adversarial attacks* [22]. Expanding on previous work in the area of audio scrambling [13], we suggest increased research to explore adversarial AI components that check whether speech content and individual identity could be reconstructed from degraded or scrambled audio. The purpose of the adversarial component (which we have defined as system $\mathbb{A}$) is to enhance the trust partnership between AI and consumers. Another form of data degradation involves a signal-emitting device (system $\mathbb{N}$ from our problem description) that can be installed into a physical space (e.g., an office or meeting room). The device emits a signal (audible or subsonic) that conceals human conversation, such that anyone outside of the physical space cannot eavesdrop [1, 7, 12]. In this scenario, the system $\mathbb{N}$ emits a noise n to transform the input speech so that an automatic speech recognition (ASR) system (or human listener), as system $\mathbb{A}$, performs poorly on the task of word recognition. The noise-emitting system $\mathbb{N}$ serves to strengthen both types of trust (interpersonal and human-AI). Once users understand that they can trust the functionality of system $\mathbb{N}$, they can focus their efforts on building interpersonal trust through private conversations and co-participation in the AI service.

3.2 Personal and Dynamic Customization

Attitudes differ when it comes to the notion of perceived or actual surveillance by microphones and audio AI services. Solutions for audio AI must account for different people having different preferences and therefore audio AI services must be dynamic and allow for customization. Privacy concerns with audio AI are so strong that some people admit to modifying how they behave or speak when they are around audio-enabled devices like personal voice assistants [24]. While people in the DHH community, for example, can benefit from audio AI services that are always listening, this does effectively create a scenario of continuous surveillance. An exception could be related to audio AI for the purpose of serving as an assistive technology. It was found by Profita et al. that bystanders whose data is collected in audio capture are more forgiving of their privacy concerns if they believe that an audio device is an assistive device meant to help others [19].

Research Direction 2. ***The need for fairness in audio AI services can be met with selective masking techniques that target pre-defined sensitive information to allow for dynamic customisation and opt-in strategies.*** *Let $\mathbb{S}$ be an AI system, with input values y that is designed to perform a service task (e.g., monitor the acoustic landscape and provide an alert when a specific sound is audible). Let $\mathbb{M}$ be an AI system that takes audio features x as input and provides y to $\mathbb{S}$ such that the two systems are composed $\mathbb{M} \circ \mathbb{S}$. The aim is to design and develop $\mathbb{M}$ to intelligently mask specific portions of an audio signal that contain sensitive information (e.g., features of voice identity or sensitive phrases spoken in the background)*

There are two general types of speech privacy: voice anonymity and content masking. Methods that achieve anonymity change the acoustic properties of voice so that an individual's identity (including age, gender, accent, etc.) is unrecognisable from audio data by human listeners and AI algorithms [15]. On the other hand, speech content masking involves methods that remove or conceal content from audio such as sensitive spoken words or acoustic events

[27]. Incorporating speech privacy into smart buildings could be achieved by composing masking techniques with audio AI services. In such a system composition, raw audio data is processed to conceal sensitive information, and then the output of that system is passed on to the AI service (with sensitive content removed).

The challenge of fairness that we describe in this paper could be approached through composing systems of privacy and systems of audio AI services. Successfully overcoming this challenge will require innovative work in that direction, and to some extent is already being explored by the speech community through the development of spoofing countermeasure systems to protect voice biometric systems [26]. As mentioned earlier, views of privacy may shift for some people based on the communication context and purpose of audio capture [2, 3, 19]. Deciding which types of information should be concealed is a matter of additional research, in part to gain an understanding of public perceptions of privacy and also to understand how information masking affects performance of assistive audio AI services. This challenge highlights both the importance of interpersonal trust relationships among users as well as the importance of trust between users and audio AI services.

Masking strategies address an element of fairness for individuals who may have the option to participate (or not) in audio AI services. However another important dimension comes from audio AI services that are designed to provide assistance to people, as assistive technologies. People who rely on audio AI services as a form of assistive technology do not necessarily have the option to opt out of services, so these circumstances warrant careful consideration of safety. Examples of audio-based assistive technologies that could be implemented in a smart building include sound event detection for hearing impaired, wireless acoustic sensing networks based on microphones distributed throughout an area [11], localization of sound (i.e., automatically determining where a sound originated from), reducing or removing noise from acoustic environments [18], summaries of audio scenes [23], and wearable audio services for DHH [16]. Individuals who rely on audio AI services in smart buildings may not have a choice to control how they opt in and opt out of services, if they require particular services. Similarly, bystanders whose information is inadvertently captured by assistive audio AI may not be at leisure to opt out without causing a disruption to the assistive technology. Addressing this will improve safety and contribute to interpersonal trust as well as human-AI trust.

3.3 Privacy-Aware Technologies

Maintaining user trust from audio AI services in smart buildings requires joining together technical solutions from *reliability* and *fairness* while also delivering continued trust after gaining users' consensus. In this section, we introduce several engineering strategies that can help to assure user safety.

Research Direction 3. ***Safer audio AI services can be developed with technologies that assure privacy-by-design.*** *Let $\mathbb{S}$ be an audio AI system and let x be the audio signal input into the system. The system $\mathbb{S}$ includes components $\mathbb{C}$ (such as hardware, additional signal processing technologies, secure enclaves, etc). The aim is to design $\mathbb{C}$ such that it constrains how $\mathbb{S}$ encapsulates information x internally and distributes it externally.*

Similar to the noise-emitting device described earlier, another signal-based privacy technique involves purposefully rendering a microphone unable to capture audio. When done in a controlled manner, this can facilitate customised or dynamic audio capture wherein all of the audio can be blocked from a microphone. Subsonic signals (acoustic signals below the hearing threshold) can be harnessed to control when and where information is captured by a microphone. This is referred to as *microphone jamming* [21]. Jamming can be used for temporarily rendering a microphone inoperable for a period of time, for example during an especially sensitive meeting. Since jamming can be performed with subsonic signals, the jam does not interfere with the human experience of the acoustic landscape. Additional research in this area could explore how to target specific microphones among many inside of a building, as well as jamming signal reach and strength.

Another privacy-aware technology involves the optimisation of AI algorithms (including audio-based neural networks) to enable them to run directly on a chip or micro-controller. This approach is referred to as *edge computing*. With edge computing, audio data is inherently prevented from being physically recorded or stored, and all of the AI processing occurs directly on a chip equipped with a microphone. Edge computing solutions are often divorced from connectivity (such as WiFi or Bluetooth), so this approach makes it difficult to hack or attack microphones [4]. In a case when particular microphones must be linked or transmit information, edge computing can be designed to allow transmission of specific information, such as a number indicating occupancy levels, whether specific people are in the room (if they have opted in to voice identity), or if significant acoustic events have been detected in an audio scene [14, 23]. Example edge computing solutions include distillation of neural networks with Tensorflow Lite Micro [6]. An example free commercial off-the-shelf (COTS) toolkit that supports multiple types of microchips is also available is called Edge Impulse [8]. The ONNX [17] libraries allow multiple different types of neural network toolkits to interface with accelerators and microcontrollers, enabling algorithms to be optimised for edge computing. This enables localised audio AI services wherein data need not be transmitted to a central server: for example using voice biometrics to unlock a secure door or room.

4 DISCUSSION

We have presented three research directions related to safety of audio AI services in the context of smart buildings. For each area, we described challenges and potential solutions that could be explored further. All of the challenges to safety are at the intersection of individual privacy and security. Our research directions are summarised in Figure 2. In **Research Direction 1**, we described an external coordinator that "adds noise" to audio data. In **Research Direction 2**, we described a composite system that works in sequence to filter specific targeted information from the raw audio data. And finally in **Research Direction 3**, we described a system that does not require coordination with other components because it "self-organises" privacy by design.

We described how audio data can be misused, for example stealing voice biometric markers to create deepfakes. But the potential harms for consumers extend beyond deepfakes. The presence of

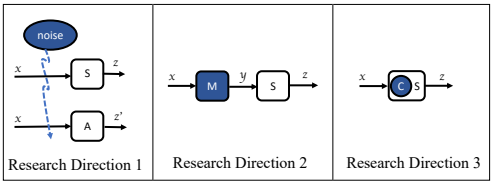

Figure 2: Visualization of research directions (1), (2), and (3) with proposed research components (dark colouring).

microphones inside smart buildings can invoke feelings of unease due to privacy concerns and possible surveillance of conversations or personal matters. Two overarching types of trust are at work, (1) interpersonal trust among users, and (2) trust between users and AI systems. In the course of preserving individual privacy (and reducing potential or perceived harms) it is important that audio AI systems continue to perform with high reliability. On behalf of AI systems, a mistake or mis-identification could range from benign and annoying to critical for life and safety (e.g., acoustic scene analysis that alerts DHH individuals of dangers, or someone calling out for emergency help).

Audio services are generally under-explored in AI research, and the safety and harms are not fully understood. Privacy-preserving anonymisation techniques and content masking techniques are active research areas for the audio processing community. As we have described, however, there is no one-size solution to managing consumer safety. Trust can be developed though a methodology of co-design that establishes human-AI partnerships [20] as well as human-human partnerships. As much as humans are consumers of AI, they are also co-creators of AI because they supply the data used in training and inference, and ultimately decision-making.

Our proposed research agenda supports advances and provides provides a timely opportunity to exploit the limits of software on chip, by ensuring that speech cannot be recorded or stored during audio capture in smart buildings. Educating consumers about audio AI safety will empower them in other domains where audio capture is increasingly used, such as digital voice assistants inside the home, smart watches worn by others, and informational robots.

ACKNOWLEDGMENTS

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the Trustworthy Autonomous Systems Hub (EP/V00784X/1) and a Turing AI Acceleration Fellowship on Citizen-Centric AI Systems (EP/V022067/1).

REFERENCES

[1] Masato Akagi and Yoshihiro Irie. 2012. Privacy Protection for Speech Based on Concepts of Auditory Scene Analysis. *Proceedings of INTERNOISE* (2012), 485.
[2] Deeksha Anniappa and Yoohwan Kim. 2021. Security and Privacy Issues with Virtual Private Voice Assistants. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. 0702–0708.
[3] Tom Backstrom, Andreas Nautsch, Karla Markert, and Ingo Siegert. 2021. How to collect speech data with human rights in mind. In *Proceedings of 2021 ISCA Symposium on Security and Privacy in Speech Communication*. 86–88.
[4] Sourav Bhattacharya, Dionysis Manousakas, Alberto Gil CP Ramos, Stylianos I Venieris, Nicholas D Lane, and Cecilia Mascolo. 2020. Countering Acoustic Adversarial Attacks in Microphone-Equipped Smart Home Devices. *ACM Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–24.
[5] Jeremy A Blumenthal, Meera Adya, and Jacqueline Mogle. 2009. The Multiple Dimensions of Privacy: Testing Lay 'Expectations of Privacy'. *University of Pennsylvania Journal of Constitutional Law* 11, 2 (2009), 331.
[6] Robert David, Jared Duke, Advait Jain, Vijay Janapa Reddi, Nat Jeffries, Jian Li, Nick Kreeger, Ian Nappier, Meghna Natraj, Tiezhen Wang, et al. 2021. TensorFlow Lite Micro: Embedded Machine Learning for TinyML Systems. *Proceedings of Machine Learning and Systems* 3 (2021), 800–811.
[7] Jacob Donley, Christian Ritz, and W Bastiaan Kleijn. 2016. Improving Speech Privacy in Personal Sound Zones. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 311–315.
[8] Edge Impulse. 2022. https://www.edgeimpulse.com/. [Accessed June-2022].
[9] Leah Findlater, Steven Goodman, Yuhang Zhao, Shiri Azenkot, and Margot J. Hanley. 2020. Fairness issues in AI systems that augment sensory abilities. *ACM SIGACCESS Accessibility and Computing* 125 (2020), 8.
[10] Shabnam Ghaffarzadegan, Attila Reiss, Mirko Ruhs, Robert Duerichen, and Zhe Feng. 2017. Occupancy Detection in Commercial and Residential Environments Using Audio Signal. In *INTERSPEECH*. 3802–3806.
[11] Gee Yeun Kim, Seung-Su Shin, Jin Young Kim, and Hyoung-Gook Kim. 2018. Sound event detection and haptic vibration based home monitoring assistant system for the deaf and hard-of-hearing. In *Proceedings of the 2018 Workshop on Multimedia for Accessible Human Computer Interface*. 1–7.
[12] Kazuhiro Kondo and Hiroki Sakurai. 2014. Gender-Dependent Babble Maskers Created From Multi-Speaker Speech for Speech Privacy Protection. In *10th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 251–254.
[13] Dawei Liang, Wenting Song, and Edison Thomaz. 2020. Characterizing the Effect of Audio Degradation on Privacy Perception And Inference Performance in Audio-Based Human Activity Recognition. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–10.
[14] Daniyal Liaqat, Salaar Liaqat, Jun Lin Chen, Tina Sedaghat, Moshe Gabel, Frank Rudzicz, and Eyal de Lara. 2021. Coughwatch: Real-World Cough Detection using Smartwatches. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8333–8337.
[15] Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, and Els Kindt et al. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language* 58 (2019), 441–480.
[16] Adedayo O Olaosun and Olawale Ogundiran. 2013. Assistive Technology For Hearing and Speech Disorders. *Assistive Technology* 3, 17 (2013).
[17] Open Neural Network Exchange (ONNX). 2022. https://onnx.ai/index.html. [Accessed June-2022].
[18] Yi-Hao Peng, Ming-Wei Hsu, Paul Taele, Ting-Yu Lin, Po-En Lai, and Leon Hsu et al. 2018. SpeechBubbles: Enhancing Captioning Experiences for Deaf and Hard-of-Hearing People in Group Conversations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, 293.
[19] Halley Profita, Reem Albaghli, Leah Findlater, Paul Jaeger, and Shaun K Kane. 2016. The AT effect: how disability affects the perceived social acceptability of head-mounted display use. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4884–4895.
[20] Sarvapali D Ramchurn, Sebastian Stein, and Nicholas R Jennings. 2021. Trustworthy Human-AI Partnerships. *iScience* 24, 8 (2021), 102891.
[21] Ke Sun, Chen Chen, and Xinyu Zhang. 2020. "Alexa, stop spying on me!" Speech Privacy Protection Against Voice Assistants. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 298–311.
[22] Naoya Takahashi, Shota Inoue, and Yuki Mitsufuji. 2021. Adversarial Attacks on Audio Source Separation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 521–525.
[23] João E da R Tavares, Tefferson D da R Guterres, and Jorge LV Barbosa. 2019. Apollo APA: Towards a model to care of people with hearing impairment in smart environments. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*. 281–288.
[24] M Vimalkumar, Sujeet Kumar Sharma, Jang Bahadur Singh, and Yogesh K Dwivedi. 2021. 'Okay Google, what about my privacy?': User's Privacy Perceptions and Acceptance of Voice Bbased Digital Assistants. *Computers in Human Behavior* 120 (2021), 106763.
[25] Paul Voigt and Axel Von dem Bussche. 2017. The EU general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.
[26] Xin Wang and Junichi Yamagishi. 2021. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. In *Proc. Interspeech 2021*. 4259–4263.
[27] Jennifer Williams, Junichi Yamagishi, Paul-Gauthier Noé, Cassia Valentini-Botinhao, and Jean-François Bonastre. 2021. Revisiting Speech Content Privacy. In *1st ISCA Symposium of the Security & Privacy in Speech Communication*. 42–46.
[28] Efstathios Zavvos, Enrico H. Gerding, Vahid Yazdanpanah, Carsten Maple, Sebastian Stein, and m.c. schraefel. 2021. Privacy and Trust in the Internet of Vehicles. *IEEE Transactions on Intelligent Transportation Systems* (2021), 1–16.