Bäckström, Tom; Das, Sneha; Perez Zarazaga, Pablo; Fischer, Johannes; Findling, Rainhard; Sigg, Stephan; Nguyen, Le

Intuitive Privacy from Acoustic Reach: A Case for Networked Voice User-Interfaces

Published in:
Proceedings of the 1st ISCA Symposium on Security and Privacy in Speech Communication

DOI:
10.21437/SPSC.2021-12

Published: 01/11/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Intuitive Privacy from Acoustic Reach: A Case for Networked Voice User-Interfaces

Tom Bäckström[1], Sneha Das[1], Pablo Pérez Zarazaga[1], Johannes Fischer[2], Rainhard Dieter Findling[3], Stephan Sigg[3] and Le Ngu Nguyen[3]

[1]Dept Signal Processing and Acoustics, Aalto University, Finland
[2] Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany
[3]Dept Communications and Networking, Aalto University, Finland

first.lastname@{aalto.fi, iis.fraunhofer.de}

Abstract

The effect that advances in voice interface technologies have on privacy has not yet received the attention it deserves. Systems in which multiple devices collaborate to provide a unified user-interface amplify those worries about privacy. We discuss ethical implications of voice enabled devices on privacy in typical scenarios at home, office, in a car and in the public. From our findings, it follows that the reach of voice can be exploited as a feature to intuitively define the extent of privacy. In particular, the acoustic reach of speech signals can serve as a feature for designing privacy-gentle voice user-interfaces which are intuitive to use. We argue that this approach poses reasonable technological requirements and establishes a natural experience of privacy which confirms intuitive perception.

Index Terms: voice user interfaces, privacy and security, ethics in engineering, distributed speech processing

1. Introduction

Voice user interface technology has recently become fashionable with the advent of personal digital assistants (PDAs) such as Siri[1], Alexa[2] and Cortana[3]. They provide an intuitive way of interacting with devices and services via voice. Traditional user-interfaces could thereby be improved or replaced with a voice operated interface [1]. Current solutions are primarily single-device approaches, often with the support of a cloud server. However, collaborative multi-device approaches could improve signal quality and increase flexibility (and thereby usability) in user interface design [2, 3].

Such audio-capable device networks could provide a transparent acoustical front-end for all voice-operated services [2]. We aim to employ the orchestration of all devices with microphones to collaborate for services such as hands-free telephony and personal digital assistants. A usual objective in distributed speech recognition is to recognize speech within a distributed network of audio-capable devices without a modular acoustic front-end [4]. In contrast, we propose a standardized and modular acoustic front-end for distributed networks of audio-capable devices. This enables interoperability between manufacturers such that 1. any voice-service could use all devices, 2. multiple service-providers could co-exist and collaborate, and 3. interaction between multiple devices of a single service-provider would be straightforward. With most current solutions, two co-located devices of the same brand would both respond to a voice command. Such behavior is undesirable and can be avoided if the two devices collaborate.

Parts of this work have been supported by the SPEAKER project (FKZ 01MK20011A), funded by the German Federal Ministry for Economic Affairs and Energy.

[1]https://www.apple.com/ios/siri/

[2]https://developer.amazon.com/alexa

[3]https://www.microsoft.com/en-us/windows/cortana

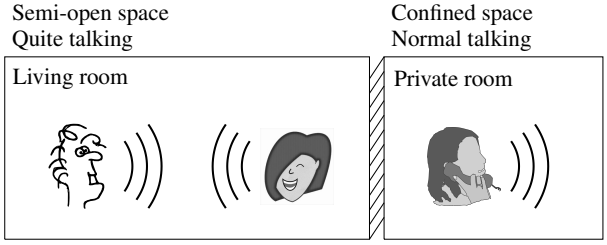


Figure 1: *Acoustic range of speech implies the extent of privacy. Semi-public spaces necessitate quiet voice to improve privacy.*

Collaboration between devices demands a open standards, but also raises questions on security, privacy, and authorization. Which devices are allowed to collaborate and share acoustic data? How to verify that devices do not violate privacy? Voice user-interfaces should be designed such that they are aware of users' expectation of privacy and, in particular, should attempt to provide the desired level of privacy. For example, family members might regularly share experiences, but they also need their private domain. Our contributions are:

1. An exploration of use-cases (Sec. 2) which indicates that people consciously and accurately adjust speaking volume and content to enforce different privacy levels (cf. figure 1). We argue that user-interfaces adhering to such a notion of acoustic reach are intuitive to use.
2. We show that voice user-interfaces for which privacy and collaboration is defined by the acoustic reach require only modest extensions to existing technologies (Sec. 3).
3. Our threat analysis (Sec. 4) indicates that acoustic reach can provide reasonable privacy.

We intend to bring attention to the ethical implications of speech processing technologies. We argue that it is not feasible to design objective measurements to determine ethical boundaries, because such measurements would, by design, cross those ethical boundaries. We therefore follow the tradition of the social sciences and approach the problem through observations within narratives of four typical use cases and discuss the implications these scenarios have on technological solutions. Future work may design systems such that usability studies on their performance do not pose ethically unsustainable risks.

2. Use Cases

We discuss expectations towards audio privacy with four characteristic use cases: at home, in an office, in a car, and in a public space. For each use-case we highlight characteristic expectations and user behavior with regard to privacy.

2.1. Home

Scenario: Consider a typical family of three; parents Jim and Jane, as well as their teenage daughter Jill. They live together in relative harmony and share stories about their experiences over dinner. After dinner, Jill typically retreats to her room, closes the door behind her and talks to her boyfriend on the phone. In the meanwhile, Jim and Jane frequently discuss private topics in hushed voices in the living-room (Fig. 1). When away from home, all three chat as well as share pictures and videos with each other on social media.

Analysis: This family home features several layers of privacy. At dinner, they talk about their experiences more openly than they would if outsiders were present. They also have a desire to share with each other. However, when Jill closes the door of her room, it is not just a sign that she wants to be left in private, but it is also an acoustic barrier.

In the living-room, the parents have a private discussion, but since it does not have doors, they speak with lowered volume. The reduced volume makes it harder to eavesdrop and concurrently advertises to a possible passer-by (Jill) that the conversation is private.

In both cases, the acoustical barriers thus impede eavesdropping and concurrently communicate the desire for privacy. The three however also have a joint desire to share certain things. When they are not together they compensate for the absence by chatting as well as sharing photos and videos.

2.2. Office

Scenario: Colleagues Jake and Jonathan are about to meet their customer Julia, all three under a non-disclosure agreement. Before the meeting, Jake and Jonathan have a talk to agree on their position on some key questions about their project with Julia. In the meeting, after the sales-pitch of Jake and Jonathan, Julia realizes that she needs additional information from her assistant Jester, so they set up a teleconference. After the teleconference, Julia signs the contract with Jake and Jonathan.

Analysis: Different layers of privacy are involved in this use case. Prior to the meeting, Jake and Jonathan discuss company secrets and tactics which cannot be revealed to Julia. Still, the three have mutual interests to discuss. Jester is, in the meeting, an outsider who receives opt-in permission to join through a teleconference, but he is dismissed as soon as the necessary information is received.

We have explicit legal boundaries on privacy based on a non-disclosure agreement. Even colleagues of Jake and Jonathan, would be contractually excluded from accessing the conversation records. Still, it could be useful to have shared voice records of the meeting for legal purposes. For the guest, Julia, to set up a teleconference with her assistant Jester, is often an awkward experience. Julia does not generally have access rights to the meeting room teleconference equipment, nor would she have experience in using it, whereby she has no option but to share her assistant's private contact information with Jake and Jonathan for them to set up the connection.

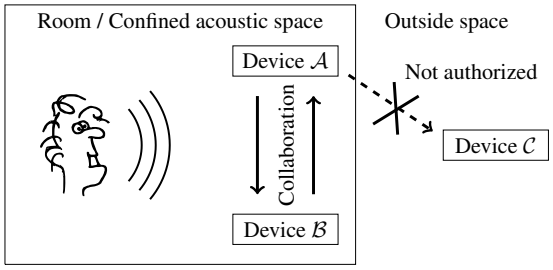


Figure 2: *Illustration of the extent of privacy as defined by the acoustic reach of a speech signal.*

2.3. Car

Scenario: For meetings with his clients, Jeremy has to drive several hours a week, and for efficient use of time, he often talks with colleagues and clients on the hands-free while driving. Twice a week, he drives his daughter Josephine to her hobbies. Occasionally, he also drives with his friend Jacob. When his wife Jessica calls him in the car, he has the habit of starting the conversation as "Hey honey, I'm here with Jacob/Josephine..." depending on who is in the car. Sometimes Jessica also borrows Jeremy's car to run some errands.

Analysis: The car setting introduces novel features of privacy: 1. It is a flexible environment where a range of people can travel in it. 2. those present in the car will (and potentially people outside the car can) overhear any conversation. When Jeremy receives a call, social conventions require him to make clear who can overhear the conversation. When using a hands-free device utilizing the car loudspeakers, everyone in the car can overhear both sides of the conversation. In contrast to the car, in an office or at home one would often move to a private location if others are present when receiving a call. The driver has arguably little control over the level of privacy, except for controlling the content of the conversation itself.

2.4. Public Space

Scenario: Jennifer, Jasmine and Jordan are childhood friends, who like to meet in a cafeteria. They know each other well, which leads to them sharing secrets with each other that they would not reveal otherwise When sharing such secrets they tend to huddle together and reduce their voices to a whisper.

Analysis: Jennifer, Jasmine and Jordan share certain private information which even their corresponding life partners would not hear. Still, they would be unlikely to share all their secrets amongst each other. It is this exact combination of persons and the context that determines which secrets can be revealed. If with others, the level of privacy would be different.

The cafeteria is a public space that might not have any clear, physical boundaries to the acoustic reach of speech. A person might inadvertently overhear parts of a conversation ongoing at the next table. This is why the three come closer and only whisper certain secrets in an effort to avoid accidental eavesdroppers. A cafeteria with plenty of background noises, including competing talkers and generic music, might thereby come to be a benefit to private discussions by masking speech from eavesdroppers.

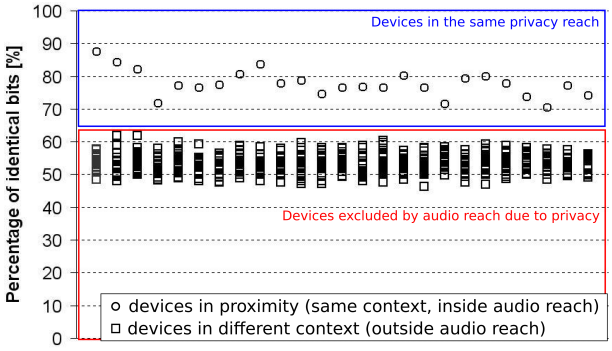


Figure 3: *Devices in proximity achieve higher similarity in audio fingerprints than devices which are outside audio reach.*

3. Privacy Defined by Acoustic Reach

People have an intuitive understanding of the level of privacy in their interactions. Certain situations might become uncomfortable when the normal sphere of privacy is not available, for example, when there is no background noise in a cafeteria, or when receiving a hands-free phone call in a car with passengers on board. This intuition of audio privacy should be taken into account when designing voice interfaces [5].

People have a good guess of who can hear their voice, and adjust the volume of speech and their message accordingly. The acoustic reach of a signal is thus an indicator for sufficient access authorization. In terms of a voice-interface, this concept points to a design where the acoustic reach of a signal defines which entities are allowed access to that signal. For example, in Fig. 2, device $\mathcal{A}$ may observe and analyze speech. It may not, however, transmit that speech signal to a device $\mathcal{C}$ outside the room, without prior consent (opt-in) [6].

For multiple devices $\mathcal{A}$ and $\mathcal{B}$ co-present in a room it would be reasonable to collaborate. Technically, audio-based implicit access rules can be implemented based on ad-hoc spontaneous encrypted communication using fuzzy cryptography based protocols exploiting audio fingerprints [7–10] (Fig. 3).

In this regard, acoustic reach should be defined according to human performance. If a human in the same location as the device would be able to overhear a speech signal, then the device should be allowed to access that speech signal as well. The technological requirement for preserving privacy is thus a method which determines whether two devices are within the same acoustic space. This task has been addressed prior works, e.g. by acoustic handshake and audio watermarking [11–13].

4. Threat Model

In an environment with collaborating speech interface devices, all devices shall gain access to all sensed speech, independently of which device sensed it. Such environments are not necessarily limited in an intuitive way, e.g. by an implicitly assumed acoustic space of households or company rooms. Hence, if no limit is imposed on when, what, and how those devices share information, they necessarily share all sensed speech independently on whether they are in the same acoustic space or not.

Attackers would thereby only need to inject a single device into the environment, which would then share and receive speech from other devices. Possible attackers include people owning devices connected to such an environment. They also include any party being able to illegitimately gain access to such devices [14, 15]. They further include device manufacturers, resellers, and maintainers.

Regarding the impact of attacks, note that devices transmitting speech sensed in the environment to external services should be considered a privacy breach, irrespective whether the transmission is legitimate or not [16–18]. Examples range from family members who overhear conversations, over cybercriminals, to government players. A mitigation to the above threats could be to impose limits on when, how, and what speech devices share in such environments. Utilizing the acoustic reach to impose such limits could be one way of creating feasible and at the same time intuitive limits. When devices only share sensed speech if they are in the same acoustic space, eavesdropping with a single device is limited to audio that is sensed in the corresponding acoustic space.

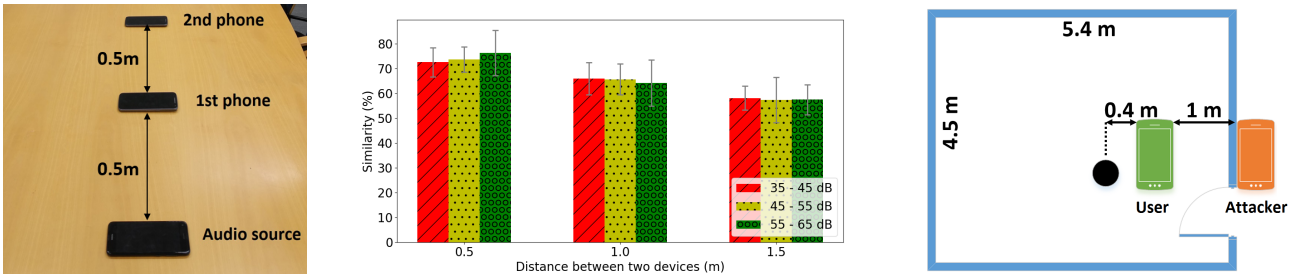
5. Applications

Privacy has received little attention with respect to voice user interfaces so far. However, it is an important aspect of those [5]. In contrast to voice user interfaces, with many modern consumer products (e.g. messaging services) privacy has already become a central selling point. Furthermore, the European Union has adopted the General Data Protection Regulation (GDPR), which has impacted and will continue to have an impact on voice user interfaces [19]. In fact, the European Data Protection Board has recently published guidelines for handling privacy in virtual voice assistants [20]. Similarly, conventional phones have well-defined legal frameworks which specify the extent of privacy required [21]. Privacy is also well-studied in the different other fields, including e.g. law, robotics, and psychology [22–24].

The main observation of our current work is that people have an intuitive understanding of who can hear their speech, whereby they adjust their message and volume of speech accordingly. Voice user interfaces should thus allow collaboration between devices only when they are within the same acoustic space, or when the user specifically enabled further collaboration between devices in an opt-in manner.

The technical requirements for this are well manageable. Firstly, there is a need for methods to determine whether two devices are in the same acoustic space, hence can hear the same signal – without revealing sensitive information in the process, such as previous approaches comprising audio fingerprinting and fuzzy cryptography [7]. Secondly, to access information about conversations afterwards, keeping a log of access rights would be beneficial. For instance, decentralized consensus via blockchain or central server-based solutions would be feasible. With these system features, seamless device collaboration over wireless systems could be enabled, while at the same time maintaining a reasonable level of privacy.

Collaboration between devices, in turn, is useful in many ways. For example, distributed sensor network methods can be used to improve signal quality to obtain high-fidelity even with a number of low-quality sensors [25]. Moreover, by enabling hardware of different manufacturers to collaborate in a secure and privacy-adhering way, dedicated hardware for every service is not required and resources can be used more efficiently. This could potentially make vendor lock-in more difficult, facilitate a lively competition, and allow users to better control and monitor the level of privacy with their voice interfaces.



(a) *Devices to simulate overhearing in a meeting.*

(b) *The similarity of audio fingerprints decreases with increasing distance between devices*

(c) *An attacker observing voice pairing outside the room.*

Figure 4: *Layout of devices in the experimental settings as well as observed fingerprint similarity.*

6. Ethics

Similarly to the outlined privacy concerns, engineering ethics has not yet received much attention within the speech processing community. A central challenge is that the publication tradition in engineering fields requires quantifiable and objective measurements. Ethics, however, can not be measured objectively. Moreover, experiments for measuring ethical issues would likely also be unethical by design.

Consequently, in this paper we utilize considerations of probable scenarios in the form of use-cases as the publication format for ethical discussions. We thereby outline potential consequences of using this technology if the corresponding privacy and ethical problems are not addressed, and propose design-concepts which might facilitate resolving those.

The fact that ethical discussions are typically not explicitly discussed, when proposing design-concepts in our community, is undoubtedly evident also in our humble contribution. With this work we thus want to encourage other researchers and engineers to challenge our approach as well as to point out improvements and challenges. We especially encourage readers to propose improvements not only to technology, but also to the format of publication within the ethical realm.

7. Experimental Evaluation

We demonstrate the effect of distance between audio sources and audio interfaces on the privacy defined by acoustic reach with the experiments. Further experiments are presented in [9].

In the first experiment, we simulate multiple mobile devices overhearing a meeting in a small office with an arrangement of mobile devices acting as audio sources and audio interfaces (see Figure 4a). We put the mobile device that acts as the audio source on a fixed position on a table in the small office, where it continuously broadcasts an audio recording. The device acting as the first audio interface is positioned on the same table in $0.5\,\mathrm{m}$ distance to the audio source. The device acting as the second audio interface is also positioned on the same table, but in distance d to the audio source, with $d = \{1\,\mathrm{m}, 1.5\,\mathrm{m}, 2\,\mathrm{m}\}$. We control the sound level of the audio source in the range of $35-45$, $45-55$, and $55-65$ dB, which corresponds to the range of verbal conversation loudness. For processing the audio sensed on the audio interfaces we apply the techniques of [7], which aims to derive a similar audio fingerprint on both devices based on the sensed audio (see Figure 4a). The Hamming-distance-based similarity of those extracted audio fingerprints is shown in Figure 4b. We thereby observe that the similarity of the audio fingerprints consistently decreases with an increasing inter-device distance, independently of the level of verbal conversation loudness. We thus conclude that it is feasible to use audio fingerprinting (e.g. fuzzy cryptography schemes) to allow shared communication only between devices that are within a certain audio proximity, hence within a certain acoustic reach.

In the second experiment, we address the problem of pairing two smart devices for allowing secure connection between them using only verbal commands. An application scenario of this would be to establish a secure connection between a mobile phone and a smart television for holding a presentation in a meeting room. We also consider an attacker which is outside the room in which the devices get paired, with either an opened or closed door (see Figure 4c). In the experimental setup we use one mobile device in the room to continuously broadcast an audio recording, and two more devices in the room to observe the audio and derive audio fingerprints. The audio broadcast and the extraction of audio fingerprints from audio sensed on both devices are the same as in experiment 1. When both audio interfaces are within 1 m distance of each other in the room we observe the corresponding audio fingerprint to be in the range of 68.2%. In contrast, the device of the attacker outside the room senses audio of which the extracted fingerprint only shows a similarity of 52.3% in case the door is openend, and 45.7% when the door is closed. In both cases, the similarity is lower than with the devices intended to be paired in the room. From this we can conclude that the fingerprint similarity can be utilized to distinguishing benign and malicious devices during pairing, given that the acoustic distance is higher for malicious devices than for benign ones – hence, that malicious devices are further away.

8. Conclusion

In this paper, we explored several use-cases in which users enforce different privacy levels through adjusting their speaking volume. We proposed that designing voice user-interfaces based on the acoustic reach requires only modest modification to existing technologies. Specifically, such functionality can be implemented through a comparison of microphone signals between devices, where the comparison is implemented in a way which does not reveal the content of sounds, but only allows estimation of proximity. Our threat analysis indicates that acoustic proximity can provide acceptable security and privacy. We performed two experiments to verify our proposed approach in realistic scenarios.

9. References

[1] C Pearl, *Designing Voice User Interfaces: Principles of Conversational Experiences*, O'Reilly Media, Inc., 2016.

[2] T Bäckström, F Ghido, and J Fischer, "Blind recovery of perceptual models in distributed speech and audio coding," in *Proc. Interspeech*, 2016, pp. 2483–2487.

[3] Y Jia, Y Luo, Y Lin, and I Kozintsev, "Distributed microphone arrays for digital home and office," in *Proc. ICASSP*, 2006, vol. 5.

[4] ETSI, *ES 201 108 standard – Distributed speech recognition (V1.1.3)*, 2003.

[5] B Schneier, "Stop trying to fix the user," *IEEE Security & Privacy*, vol. 14, no. 5, pp. 96–96, 2016.

[6] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, IEEE, version 2 edition, 2017.

[7] D Schürmann and S Sigg, "Secure communication based on ambient audio," *IEEE Trans. Mobile Comput.*, vol. 12, no. 2, pp. 358–370, Feb 2013.

[8] S Sigg, D Schuermann, and Y Ji, "Pintext: A framework for secure communication based on context," in *Proceedings of the Eighth Annual International ICST Conference on Mobile and Ubiquitous Systems:Computing, Networking and Services (MobiQuitous 2011)*, 2011.

[9] Pablo Pérez Zarazaga, Tom Bäckström, and Stephan Sigg, "Acoustic fingerprints for access management in ad-hoc sensor networks," *IEEE Access*, vol. 8, pp. 166083–166094, 2020.

[10] Mikhail Fomichev, Flor Álvarez, Daniel Steinmetzer, Paul Gardner-Stephen, and Matthias Hollick, "Survey and systematization of secure device pairing," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 517–550, 2017.

[11] S Sigg, M Budde, Y Ji, and M Beigl, "Entropy of audio fingerprints for unobtrusive device authentication," in *International and Interdisciplinary Conference on Modeling and Using Context*. Springer, 2011, pp. 296–299.

[12] W Berchtold, "Secure communication protocol for a low-bandwidth audio channel," in *Proc. EUSIPCO*, 2017.

[13] Y Lin and W H Abdulla, *Audio Watermark*, Springer, 2015.

[14] K Angrishi, "Turning Internet of Things (IoT) into Internet of Vulnerabilities (IoV) : IoT botnets," *CoRR*, vol. abs/1702.03681, 2017.

[15] T Yu, V Sekar, S Seshan, Y Agarwal, and C Xu, "Handling a trillion (unfixable) flaws on a billion devices: Rethinking network security for the internet-of-things," in *Proceedings of the 14th ACM Workshop on Hot Topics in Networks*, New York, NY, USA, 2015, HotNets-XIV, pp. 5:1–5:7, ACM.

[16] M Crocco, M Cristani, A Trucco, and V Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 52:1–52:46, Feb. 2016.

[17] Y Guo and M Hazas, "Localising speech, footsteps and other sounds using resource-constrained devices," in *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, April 2011, pp. 330–341.

[18] T D Räty, "Survey on contemporary remote surveillance systems for public safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 5, pp. 493–515, 9 2010.

[19] European Parliament, *Directive 95/46/EC General Data Protection Regulation*, 2016.

[20] European Data Protection Board, *Guidelines 02/2021 on virtual voice assistants, Version 2.0*, 2021.

[21] European Parliament, "31996G1104 Council Resolution of 17 january 1995 on the lawful interception of telecommunications," *Official Journal*, vol. C 329, pp. 0001–0006, 1995.

[22] P Gewirtz, "Privacy and speech," *The Supreme Court Review*, vol. 2001, pp. 139–199, 2001.

[23] M K Lee, K P Tang, J Forlizzi, and S Kiesler, "Understanding users' perception of privacy in human-robot interaction," in *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 2011, pp. 181–182.

[24] D M Pedersen, "Psychological functions of privacy," *Journal of Environmental Psychology*, vol. 17, no. 2, pp. 147–156, 1997.

[25] Sneha Das and Tom Bäckström, "Enhancement by postfiltering for speech and audio coding in ad hoc sensor networks," *JASA Express Letters*, vol. 1, no. 1, pp. 015206, 2021.