POSTER: Enhancing Security and Privacy Control for Voice Assistants Using Speaker Orientation

Shaohu Zhang
North Carolina State University
szhang42@ncsu.edu

Aafaq Sabir
North Carolina State University
asabir2@ncsu.edu

Anupam Das
North Carolina State University
anupam.das@ncsu.edu

ABSTRACT

Home voice assistants (VA) like Amazon Echo and Google Home have gained popularity due to the ease of controlling devices through voice commands. VAs continuously listen to detect the wake word and send the subsequent audio data to the manufacturer-owned cloud service to interpret the commands. However, studies have revealed that imperfect voice recognition can lead to unintentional activations when similar-sounding words are spoken in the background. Existing privacy controls are not effective in preventing such misactivations. Recent studies have shown that the visual gaze plays an important role when interacting with conservation agents such as VAs, and users tend to turn their heads or body toward the VA when invoking it. In this study, we propose a device-free, non-obtrusive acoustic sensing system called *HeadTalk* to thwart the misactivation of VAs. The proposed system leverages the user's head direction information and verifies that a human generates the sound to minimize accidental activations.

1 INTRODUCTION

Voice-controlled speakers like Amazon Echo and Google Home have become increasingly pervasive due to the convenience they provide. Voice assistants (VAs) keep listening to detect the wake command (e.g., "Alexa") and send the subsequent voice command to the manufacturer-owned cloud service for processing to identify actionable commands. However, the always-listening nature of voice assistants gives rise to security and privacy concerns [1]. For example, VAs can misactivate either accidentally due to suboptimal wake-word recognition engine or artificially by manipulating the pronunciation of the wake word [2]. In addition, with more and more devices integrating voice-assistant-like capabilities (e.g., smart TVs), multiple VAs will likely share the same physical space, which can lead to misactivating the wrong VAs. Existing privacy controls for VAs include: usage of different safewords, physical mute button, and access to the command history. However, such privacy controls are not effective as safewords can also lead to misactivations [2]. Furthermore, while users are aware of the ability to review audio logs and mute their smart speaker, Study [3] has shown that users do not use such privacy-enhancing features for multiple reasons. Due to the limitation of existing privacy controls for VAs, more device-free privacy controls are needed.

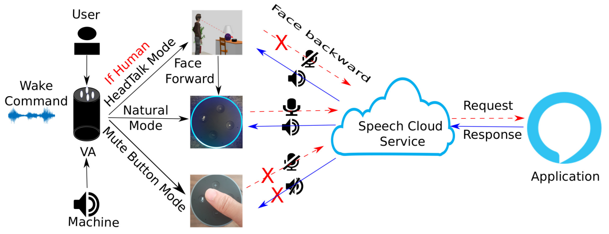


Figure 1: *HeadTalk* privacy control.

Recent studies have shown that visual gaze [4] plays an important role when interacting with VAs. They also found that participants rated the overall user experience to be higher when they could view the VA as opposed to not seeing it as visual cues increased their confidence in the VA's response. Leveraging such insights, we develop a device-free, non-obtrusive acoustic sensing system called *HeadTalk* to thwart the misactivation of VAs in this paper. Figure 1 shows our proposed privacy control for VAs. In addition to the mute button, which fully disables the VA function, users can select *HeadTalk* mode through voice command (e.g., by saying "Alexa, enter HeadTalk mode"). *HeadTalk* only accepts the given wake word when it is spoken facing the VA. Once the wake word is detected while facing forward, the user does not need to continuously face the device for the remaining session. If the user faces backward, the VA will not record and transmit audio data to the cloud service, but the smart speaker will still be functional (e.g., streaming music or news). In this way, we can essentially implement a *soft mute* operation while still enabling the speaker to function.

We are the first to propose a speaker-orientation-based privacy control for VAs while processing the wake word. We show that speech alone can be used as a directional communication channel, in much the same way visual gaze specifies a focus. We implement *HeadTalk* using a commercial off-the-shelf (COTS) speaker and collect data under various real-world settings covering both a lab setting and a real-home setting.

2 SYSTEM OVERVIEW

As shown in Figure 2, *HeadTalk* is comprised of two main components, including *Liveliness Detection* (shown in green color) and *Speaker Orientation Detection* (shown in gray color). The *Prepossessing* block captures the wake command, removes noise, and outputs as *denoised audio*. The *Feature Extraction* block takes the denoised audio as input and extracts features for liveliness detection and

WiSec '23, May 29 - June 01, 2023 , Guildford, Surrey, United Kingdom
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/21/06…$15.00
https://doi.org/10.1145/xxxxxxx.xxxxxxx

speaker orientation detection, respectively. Next, if the speech command is identified to originate from a mechanical speaker, *HeadTalk* will reject the command and remain in 'mute mode'. If it is classified as human speech, then the human speaker's orientation is determined to evaluate whether the human speaker is facing and not facing the VA. If the speech command is identified as facing, *HeadTalk* will accept the command and upload it to the corresponding cloud service for further processing.

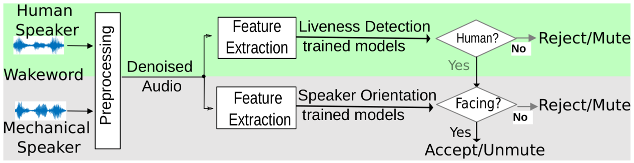


Figure 2: System Overview for *HeadTalk*

Threat Model. Our threat model covers two possible scenarios: 1) accidental misactivation through audio broadcasts (e.g., smart TVs streaming media such as TV shows and news) or people chatting in the background; 2) we also consider a more active attack scenario where the adversary can compromise/control a media device to replay in the same physical location as the VA.

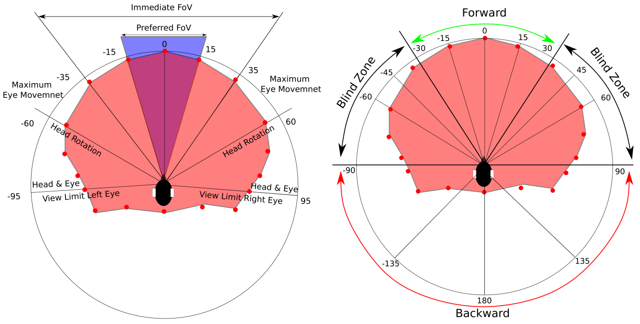
2.1 Speaker Orientation Estimation

Figure 3a shows a person's horizontal Field-of-View (FoV). Considering the human eye or mouth as the centerline, the 15° on both sides of the centerline is considered as the *preferred viewing area* [5], where human vision is most sensitive. 35° on both sides of the centerline is referred to as the *immediate FoV* that represents the maximum angle where both eyes can observe an object simultaneously. Figure 3 shows that a sound source has directivity in its spatial radiation. The power energy is highest when directly facing the device at 0 degree. The incoming acoustic signal will most likely change with changes in the orientation of the sound source. Based on the human FoV and speech directivity, we define the angles within the range of $-30°$ to $30°$ as the forward-facing orientation while the range of $-90°$ to $90°$ as the non-facing orientation. Like 'blind spots' that a driver cannot see without turning his/her head around, we define arcs that are hard to determine when the speaker's head is facing a specific angle. We consider the arc of $-90°$ to $-30°$ and $30°$ to $90°$ as the "blind zone". As a human head can turn as much as $90°$, the speaker can easily turn his/her head toward the facing zone to activate the device.

When a user speaks towards a device, the direct path from the mouth to the device is the loudest and least-distorted, whereas all other reflected signals are delayed, lower power, and more distorted. In addition, the higher frequency acoustic signals are more directional, carrying the most significant amplitude in their emitted direction, while lower frequency components spread out in a more omnidirectional fashion. Therefore, We extract such speech reverberation and directivity features to estimate a speaker's head orientation.

3 EXPERIMENT RESULT

We implement *HeadTalk* using a 6-channel Seeed's ReSpeaker Core V2.0 [6] and record raw audio at 48 kHz. Most recently speech representation learning networks such as wav2vec2 [7] have shown their advantage in speech recognition and speaker recognition. We use the SpeechBrain library and ASVSpoof 2019 physical access dataset to train our wav2vec2 model to distinguish human speakers from mechanical speakers. We use the default ASVSpoof 2019 dataset splits and train the network for 20 epochs. The accuracy was 98.56% (EER 3.36%) and 98.52% (EER 3.90%) for the validation and test dataset. Next, we evaluate performance using our samples of live-human speech (i.e., "Computer" utterance) and samples of replayed audio through a Sony speaker. We use the previously trained model to test the unseen samples and get 84.87% accuracy (EER 16.50%). We, therefore, adopt an incremental learning approach to create a better-generalized classifier, where we split the unseen samples into the following train, validation, and test datasets (20:20:60). After retraining on the 20% new training data, we get 98.61% accuracy (EER 1.76%) and 98.68% accuracy (EER 2.58%) for the validation and test dataset, respectively, with just 10 epochs of training. To detect speaker orientation, we use the implementation of support vector machine (SVM) and select the best complexity parameter for Radial Basis Function (RBF) through grid search. We evaluated *HeadTalk* covering three wake words (e.g., "Hey Assistant!", "Computer" and "Amazon") in two room settings (e.g., lab and home), and showed that it could achieve an average accuracy of 96% to detect the speaker orientation. We believe this simple yet effective head orientation-based privacy control can help consumers better protect sensitive operations carried out by voice assistants. Our proposed approach has the potential to make distributed voice interactions more practical and privacy-preserving.



(a) Human FoV (b) Proposed field of sound

Figure 3: The power distribution of human speech aligns with human FoV. (a) Human FoV; (b) Our proposed field of sound.

REFERENCES

[1] S. Maheshwari, "Hey, Alexa, what can you hear? and what will you do with it?" https://tinyurl.com/hwxtau39, March 2018.
[2] L. Schönherr, M. Golla, T. Eisenhofer, J. Wiele, D. Kolossa, and T. Holz, "Unacceptable, where is my privacy? exploring accidental triggers of smart speakers," *arXiv preprint arXiv:2008.00508*, 2020.
[3] J. Lau, B. Zimmerman, and F. Schaub, "Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers," *Proc. ACM Human-Computer Interaction*, vol. 2, no. CSCW, Nov. 2018.
[4] A. Mhaidli, M. K. Venkatesh, Y. Zou, and F. Schaub, "Listen only when spoken to: Interpersonal communication cues as smart speaker privacy controls," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 2, pp. 251–270, 2020.
[5] "Guidelines on ergonomic criteria for bridge equipment and layout," International Maritime Organization, Tech. Rep., 12 2000.
[6] "Seeed's respeaker core v2.0," https://wiki.seeedstudio.com/ReSpeaker_Core_v2.0/, 2021.
[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.