Speaker Orientation-Aware Privacy Control to Thwart Misactivation of Voice Assistants

Shaohu Zhang, Aafaq Sabir, Anupam Das
North Carolina State University, Raleigh, NC, USA
{szhang42, asabir2, anupam.das}@ncsu.edu

***Abstract*—Smart home voice assistants (VAs) such as Amazon Echo and Google Home have become popular because of the convenience they provide through voice commands. VAs continuously listen to detect the wake command and send the subsequent audio data to the manufacturer-owned cloud service for processing to identify actionable commands. However, research has shown that VAs are prone to replay attack and accidental activations when the wake words are spoken in the background (either by a human or played through a mechanical speaker). Existing privacy controls are not effective in preventing such misactivations. This raises privacy and security concerns for the users as their conversations can be recorded and relayed to the cloud without their knowledge.**

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 paper, 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. Our extensive evaluation shows that *HeadTalk* can accurately infer a speaker's head orientation with an average accuracy of 96.14% and distinguish human voice from a mechanical speaker with an equal error rate of 2.58%. We also conduct a user interaction study to assess how users perceive our proposed approach compared to existing privacy controls. Our results suggest that *HeadTalk* can not only enhance the security and privacy controls for VAs but do so in a usable way without requiring any additional hardware.

***Index Terms*—Voice assistant, Privacy control, Signal processing**

I. Introduction

In recent years, voice-controlled speakers (also known as voice assistants) like Amazon Echo and Google Home have become increasingly pervasive due to the convenience they provide, including searching the web, streaming music/news, online shopping, and controlling home appliances through voice command. In order to render all of these services, 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 [47]. For example, VAs can misactivate either accidentally due to suboptimal wake-word recognition engine [26], [57] or maliciously through replay attacks [39], [70]. Lastly, 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 *safe words* (i.e., specific words to enter and exit the 'privacy mode') [57], voice recognition, physical mute button, and access to the command history. However, such privacy controls are not effective as safewords can also lead to misactivations [26], [57], and voice recognition is known to be vulnerable to replay attacks [39], [70]. Furthermore, while users are aware of the ability to review audio logs and mute their smart speaker, Lau et al. [41] have shown that users do not use such privacy-enhancing features for multiple reasons [41]. Lau et al. [41] found that users have an incomplete mental model of these privacy controls and at times do not fully trust the manufacturer to faithfully implement the privacy controls, leading some users to even unplug the device. The study also highlighted that many users dislike the current mute button setting as it is not device-free and disables all functionalities. Thus, more device-free privacy controls are needed.

Recent studies have shown that visual gaze [43], [50] plays an important role when interacting with VAs. Lee et al. [43] have observed that participants tend to stare at the device or turn their bodies towards the device when interacting with a VA. 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 study. We analyze if microphones can accurately infer the speaker orientation and thereby associate addressability with voice commands, allowing VAs to provide an *additional privacy control* where they can record and transmit audio data only when they detect the presence of a human speaker facing them from a distance. *HeadTalk* is fully compatible with existing VAs using only their built-in microphones to sense the head orientation of a human speaker and can run locally on existing VA hardware. Additionally, we demonstrate that *HeadTalk* can effectively distinguish human sources from mechanical speakers using machine learning models. 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

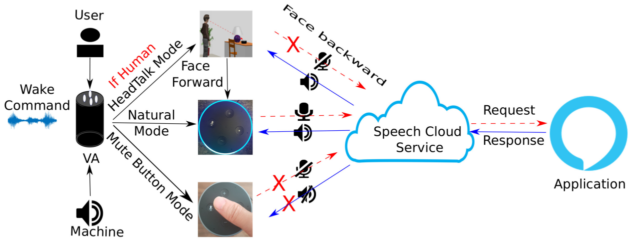


Fig. 1: *HeadTalk* privacy control. VA usually runs in normal mode to detect the presence of the wake word. Mute-button mode loses its normal functionality. *HeadTalk* mode detects the presence of the wake word, and if spoken while facing the VA, it continues to operate in normal mode, otherwise it mutes the microphones.

VA (i.e., the user is forward-facing the device; we will, later on, define what forward-facing means). 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.

There are multiple challenges in making *HeadTalk* effective. First, the user may face the device at various angles and from various distances. Second, the devices may be surrounded by objects and accept commands of varying loudness. Lastly, the system should be temporally stable and generalize across different users. In order to evaluate and showcase the effectiveness of *HeadTalk*, we implement *HeadTalk* using a commercial off-the-shelf (COTS) speaker (e.g., ReSpeaker Core v2.0 [4]) and comprehensively evaluate its performance over a wide variety of real-world settings. Additionally, we collect data using two other microphone arrays (e.g., ReSpeaker 4-channel microphone array [5] and UMA-8 USB microphone array [7]) and use open-source data [13] to demonstrate generalizability. Lastly, we conduct a user study to determine to what extent users prefer *HeadTalk* over existing privacy controls. In summary, we make the following contributions:

- We propose a novel privacy control for VAs to thwart misactivations using built-in microphones. To the best of our knowledge, 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. We perform comprehensive feasibility analysis, studying the impact of various angles, distances, surrounding objects, ambient noise, loudness, and wakeword. Furthermore, we analyze to what extent our approach generalizes across different users and settings. We have also open-sourced our code base. [1]
- We also conduct a user study with 20 participants to access the usability of our prototype VA and highlight their experience and expectations using our proposed head orientation-based privacy control.

II. Related Work

In this section, we will provide an overview of the state-of-the-art related works in the context of alternative privacy controls for VAs and microphone-based speaker orientation estimation techniques. We also highlight the major differences of our proposed approach compared to existing works.

Voice-based Replay Attacks. Voice-based authentication systems leverage unique human voice characteristics such as pronunciation, accent, and physical characteristics of the vocal tract, to recognize a user [49]. However, studies have shown that voice-based authentication systems are vulnerable to impersonation [35], [36] and replay attacks [39], [70]. For instance, Kinnunen et al. [39] reported that the equal error rate (EER) of voice authentication systems could increase anywhere from 1.76% to 31.46% under replay attacks. To counter replay attacks, Speaker-Sonar [44] detects human liveness against remote attackers by continuously emitting ultrasonic sounds (above 18kHz). However, it is vulnerable to replay attacks if the attacker can get hold of the recorded audio. In addition, the author stated that constantly emitting ultrasonic sounds could be disturbing to pets and people with health-related issues. CaField [71] utilizes the sound field feature to distinguish replay speech. However, the sound field is limited to a short distance (i.e., $\leq$ 0.5m). Void [12] detects replay attacks using the differences in spectral power between live-human voices and voices replayed through speakers. Another approach is to use additional sensory hardware such as wearable devices [30], [45], or even leverage wireless signals [55], [61] to capture human biometrics. For example, VAuth [30] collects the body-surface vibrations of a user via a wearable motion sensor and correlates the data with the speech signal recorded by the voice assistant's microphone to achieve continuous authentication.

Privacy Controls for Smart Speakers. Smart speakers have several types of privacy controls: wake word, mute button, and deleting command history. Smart speakers keep listening for an activation/wake keyword (e.g., "Alexa"). Audio is first processed locally until the wake keyword is recognized, and then subsequent audio is recorded and transmitted to the cloud to extract voice commands. Many smart speakers (e.g., Google Home and Amazon Echo) equipped with a physical mute button enable users to deactivate the microphones manually. Users can also delete the audio logs through the companion mobile app [8], website [33] or through voice command [58]. However, studies [26], [50] have shown that current privacy controls for smarter speakers are limited. The activation keyword-based approach is susceptible to false activation [26], [50]. Moreover, users usually do not review or delete their audio history stored by the device's manufacturer, and many do not even know that such options exist [14], [41]. Similarly, the

[1] https://github.com/zhangshaohu/HeadTalk

mute button is rarely used [14], [41]. To protect audio privacy without disrupting VA functionalities, others have proposed using extra hardware such as camera [48], [50] and ultrasound transceiver [20], [64]. Mhaidli et al. have explored the feasibility to activate a VA using *gaze direction* by integrating a depth-camera to recognize a user's head orientation [50]. Sun et al. [64] built a prototype called MicShield, using an ultrasound transceiver to emit ultrasonic audio to interfere with smart speaker's microphones to prevent them from detecting conversations. Neither of these approaches is practical as both approaches rely on using specialized hardware currently not compatible with VAs. Furthermore, depth-camera creates additional privacy concerns due to the presence of a camera, and emitting ultrasonic audio may create uncomfortable noise.

Microphone-based Speaker Orientation Estimation. Prior works have shown the feasibility of using multiple large microphone arrays distributed across a room to estimate the direction of audio speakers [10], [11], [17], [59], [65]. Several works have shown that it is possible to reduce the number of microphones to predict the speaker's orientation [21], [53], [60]. Some prior works [59], [65] leverage existing algorithms used for sound localization and derive coefficients (e.g., Cross-power Spectrum Phase) as a feature vector to train a machine learning model. However, the training data is typically collected through a loudspeaker instead of a human speaker. Muller et al. [53] use a smaller circular microphone array with eight microphones to collect audio played through a loudspeaker, rotated in steps of 30°, but only distinguish whether the loudspeaker is facing in the direction of the microphone, and do not distinguish a human speaker from a mechanical speaker.

The two most closely related works for estimating speaker orientation are Soundr [72], and DoV [13]. Soundr [72] leverages a deep learning (CNN-LSTM) approach using a large dataset collected from real human users to detect speaker orientation with an error bound of 34.3°, but without defining what angles constitute a forward-facing and non-forward-facing direction. Ahuja et al. [13] further analyze sound propagation to determine speaker orientation and achieve an average accuracy of 90% in detecting forward- and backward-facing speakers.

Comparison with Prior Work. In the context of detecting replay attacks, we can not only achieve better accuracy but do so from a longer distance. For example, CaField [71] is limited to a distance of $\leq$ 0.5 m, and Void [12] covers at most 2.6 m, whereas as our approach work for as far as 5 m.

When detecting speaker orientation, previous works first select different angles as the facing direction but do not provide a concrete definition of forward-facing and backward-facing orientation. *HeadTalk* leverages the insight that speaker orientation is aligned with sound propagation and the human field of view (FoV) and defines the angle range of $[-30°, 30°]$ as the forward-facing and $[-90°, 90°]$ as the backward-facing (details discussed in Section III-A). Second, Ahuja et al. [13] use the main feature of GCC-PHAT [40] to detect the speaker's orientation, while *HeadTalk* utilizes SRP-PHAT [25], in reverberant and noisy environments, and improves over 3% accuracy in both normal and cross-environment settings. Third, *HeadTalk* filters mechanical audio sources from real humans and then identifies the speaker's head orientation to thwart the misactivation of VAs. Furthermore, we comprehensively analyze the feasibility of our proposed approach in many real-world settings, including the impact of environmental noise, temporal stability, the impact of the device model, the number of microphones, and surrounding objects, where existing work lacks comprehensive feasibility analysis. Our approach is also *device-free* and *non-obtrusive* without requiring any additional hardware and is fully compatible with existing VAs, unlike other approaches [30], [55], [69], [75].

Given Ahuja et al. [13] has open-sourced their data, we can perform a direct comparison. First, we compare both approaches using the data provided by Ahuja et al. [13]. We extract features to train on all data (across people, wake words, rooms, device placements, distance, and spoken angle) from one session and test on data from another session. Under "forward-facing" definition, our system achieves 94.20% accuracy (F1-score 94.19%) while Ahuja et al. [13] achieve 92.0% accuracy (F1-score 91%). For our dataset, we achieved an average accuracy of 96.14% (F1-score 96.24%). To the best of our knowledge, our system has the best performance for detecting human speaker orientation.

III. Proposed System: HeadTalk

System Overview. Figure 2 provides an overview of our design, where *HeadTalk* is comprised of two main components, including *Liveliness Detection* (shown in green color and discussed in Section III-A) and *Speaker Orientation Detection* (shown in gray color and discussed in Section III-B). The *Prepossessing* block captures the wake command, removes noise, and outputs a time series data, which we will refer to as *denoised audio*. In order to remove low-frequency and high-frequency components generated from the surrounding environment, we adopted the fifth-order Butterworth band-pass filter to keep the audio within the frequency range of $100 \sim 16000$ Hz. The *Feature Extraction* block takes the denoised audio as input and extracts features for liveliness detection and 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.

Threat Model. Our threat model considers replaying wake words to misactivate VAs. In reality, such replays can happen either accidentally (e.g., a smart TV speaker saying the wake word) or maliciously, where the adversary can compromise/control a media device in the same physical location as the VA; for example, a PC or a smartphone,

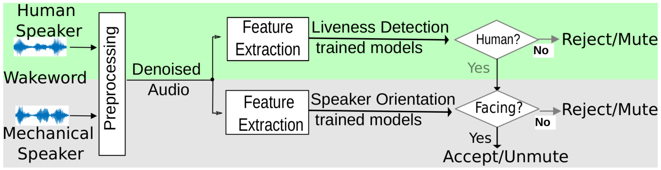


Fig. 2: System Overview for *HeadTalk*

and can capture/record the voice commands spoken by the authorized user and can later replay the commands using the compromised device's speaker. However, the adversary does not have physical access to the VA or its surrounding environment to set up specialized hardware to better spoof human voice (i.e., the attacker can not enter the home and physically set up additional hardware). In both cases, we aim to reduce unwanted misactivation by preventing non-human speech from triggering VAs and only enabling humans to trigger VAs when directly facing them.

A. *Human vs. Mechanical Speaker*

Replay attacks are straightforward to execute, and it only requires a high-quality microphone and loudspeaker for recording and replaying the audio to emulate a real user. A recent study has shown that voice-based authentication systems' equal error rate (EER) can increase from 1.76% to 31.46% under replay attacks [39]. However, recent studies [12], [66] have shown that hardware-level idiosyncrasies may be observable in the audio generated through a loudspeaker. We, therefore, explore extracting acoustic features to distinguish whether a human or mechanical speaker generates a wake word.

Figure 3 shows the spectral power characteristics of the utterance (here utterance refers to saying a wake word) of "Computer" spoken by a live human and then also replayed through Sony SRS-X5 high-end speaker [6] and Samsung Galaxy S21 Ultra smartphone, respectively. We normalize the audio amplitude between -1 and 1. The plots clearly show that the human voice (as shown in Figure 3a) has high-frequency responses above 4 kHz while the replayed voice has fewer such high-frequency responses. Most of the spectral magnitude in human voice lies in the frequency range between 200 Hz and 4 kHz, which shows an exponential power decay at around 4 kHz frequency. However, the frequency magnitude distribution of replayed audio shows more uniformity above 4 kHz. We can use such unique characteristics of the low-frequency and high-frequency responses between a live human and a replayed speaker to determine whether the audio source is from a live human or not.

Most recently speech representation learning networks such as wav2vec2 [15] have shown their advantage in speech recognition [73] and speaker recognition [67] over existing approaches like i-vector [31], x-vector [63], and ECAPA-TDNN [22]. These representation learning models have also been applied in other domains including speech anonymization [28], language detection [62], and emotion identification [52]. We adopt the wav2vec2 model to distinguish human speech from mechanical speakers, which takes the downsampled 16 kHz speech normalized to zero mean and unit variance as input. We use the BASE wav2vec2 structure, in which the convolutional layer has kernel sizes of 128 and 16 groups. The model input dimension is 768 with the inner dimension 3,072 [15]. We will evaluate the effectiveness of the liveliness detection system using the ASVspoof 2019 dataset [66] and our dataset in Section IV-A1.

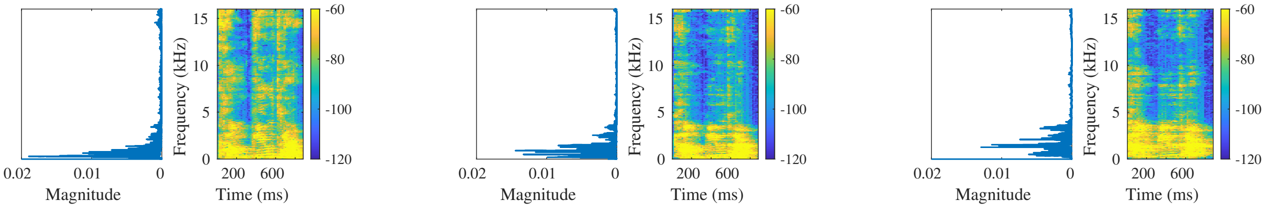
B. *Speaker Orientation Estimation*

1) Intuition: Figure 4a 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* [1], [10], 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. 60° is the maximum focus limit for both eyes in horizontal FoV without head rotation. With head rotation, human eyes can cover around 95° on both sides of the centerline.

Forward vs. Backward Speech. Several studies have looked at detecting people's ability to judge the location of a sound source using only auditory sensors. Researchers have also looked at whether humans can perceive a speaker's head orientation by only listening to audio sources. Neuhoff et al. [54] asked 15 undergraduate students to listen to a loudspeaker broadcasting white noise at six different facing angles (5°, 10°, 15°, 20°, 25°, 30°). The experiment results showed that listeners could roughly sense the loudspeaker's orientation. Kato et al. [38] investigated whether humans can successfully perceive a human speaker's facing angle. Twelve blindfolded listeners were tested on their ability to sense the facing angle of a male speaker who spoke a sentence at 8 different angles with an interval of 45°. The average angle recognition error was 23.5° in the horizontal visual plane. These results show that listeners can perceive a speaker's head orientation by simply listening to the generated audio signal.

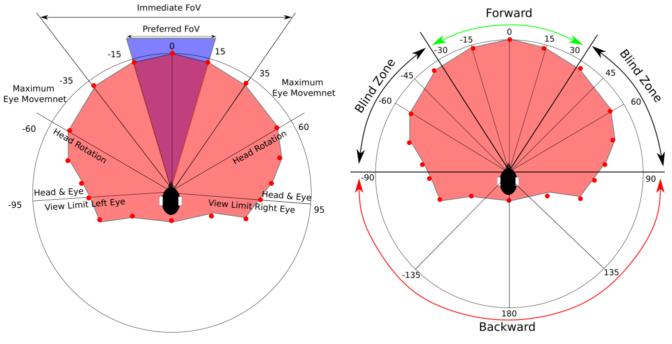
However, defining what angles constitute forward-facing versus backward-facing is still challenging. Muller et al. [53] evaluate four angle values (i.e., 2.5°, 5°, 7.5°, and 10°) as the threshold of determining forward-facing for a loudspeaker. The result shows that 7.5° leads to a near 80% detection rate. However, the sound source is limited to only one loudspeaker without considering the loudspeaker's FoV; moreover, loudspeakers may not be representative of a real human speaker. Ahuja et al. [13] evaluate three ranges of angles as potentially forward-facing: 1) Directly Facing: only 0° angle; 2) Forward Facing: ±45° and 0° arc, and 3) Mouth Line-of-Sight: ±90°, ±45°, and 0° arc. Their classifier-based approach achieves 93.1%, 92%, and 87.3% for the three facing definitions, respectively. In addition, Soundr [72] reported that the average orientation estimates could have errors as large as 34.3°. Thus, it is challenging to set a hard boundary to consider all angles as either "facing" or "non-facing" orientation.

HeadTalk is inspired by interpersonal communication cues that humans exhibit while interacting with a VA [43], [50]. As shown in Figure 4, a sound source has directivity in its



(a) Live-human **(b)** Loudspeaker (Sony SRS-X5) **(c)** Smartphone (Samsung Galaxy S21)

Fig. 3: The utterance ”Computer” generated by a human speaker, a high-end Sony loudspeaker, and a Samsung Galaxy S21 Ultra phone. We see high-frequency responses above 4 kHz from live human speech while the replayed audio has fewer such high-frequency responses.



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

Fig. 4: The power distribution of human speech aligns with human FoV. (a) Human FoV in the horizontal plane (reproduced from MSC/Circ.982 20 December 2000 [1] and [10]); (b) Our proposed field of sound for detecting facing and non-facing speaker orientation.

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.

2) Insights from Speech Propagation: Reflection of sound waves from the surfaces in a small room (height, width, and length dimensions of approximately 17 meters or less) can lead to reverberation [3]. Given a room with volume V, measured in m^3 and the average sound absorption coefficient α, Eyring equation [27] estimates the reverberation time (T in seconds) in small rooms as follow: $T = kV/(S * ln(1-\alpha))$, where k is a constant and S is the reflection surface area in m^2. The speech reverberating signal ($y(t)$) can be modeled as a linear convolution of the speech signal ($x(t)$) and a room impulse response ($h(t)$) [74] as shown below (τ is the signal delay):

$$y(t) = \sum_{\tau=0}^{T} h(\tau)x(t-\tau) = h(t) * x(t) \tag{1}$$

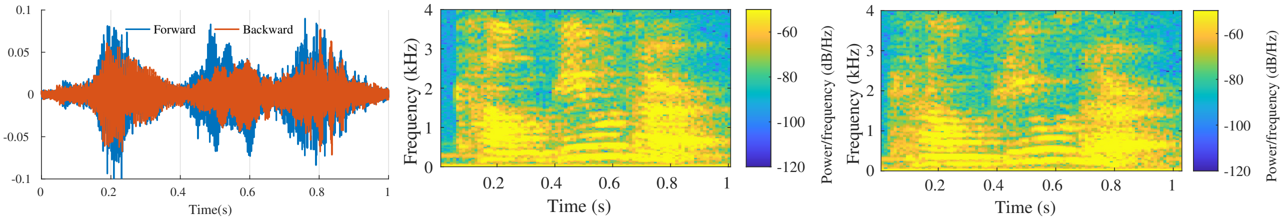
Insight 1: Speech reverberation differs with the speaker orientation. When a user speaks towards a device, the direct path from the mouth to the device is the loudest and least-distorted, whereas all other reflected signals (scattered from various surfaces in the environment) are delayed, lower power, and more distorted. However, the room impulse response $h(t)$ (shown in Eq. 1) changes with the speaker orientation. Figure 5a shows the raw acoustic signal when uttered in both forward direction and backward direction. We can see that the signal has a higher magnitude in the forward direction compared to the backward direction (shown in Figure 5b and 5c). Therefore, we can use the different speech reverberation characteristics in various angles to distinguish the forward and backward directions.

Insight 2: The perceived distribution of human speech frequency varies by angle. In any generated human speech, 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. We refer to this sound characteristic as speech directivity [51], which manifests as a characteristic imbalance between high-band and low-band frequency signals. There is less distortion between the high and low-frequency components when facing the VA, whereas there is significantly more distortion between the high and low-frequency components when not facing the VA. Figure 5b-5c illustratively show the normalized frequency spectral distribution while facing forward and backward towards the VA.

3) Feature Extraction: We extract the following features to estimate a speaker’s head orientation.

Speech Reverberation. Steered Response Power with Phase Transform (SRP-PHAT) [25] and Generalized Cross Correlation with Phase Transform (GCC-PHAT) [40], have been applied to estimate the time delay of arrival (TDoA) between pairs of microphones in speaker localization domain [42], [68]. Compared to GCC-PHAT approach, SRP-PHAT is more robust in reverberant and noisy environments in speaker localization [24], [32], [68]. We are the first to apply SRP feature to speaker orientation detection by modeling the unique delay pattern observed in the forward and backward directions.

SRP-PHAT. The received audio consists of directed propagation and reverberation. Recall the reverberation model of Equation 1, $y(t)$ is the received speech reverberating signal. A microphone array of n microphones have its corresponding



(a) Raw speech waveform

(b) Forward speech

(c) Backward speech

Fig. 5: Utterance of "Computer" spoken in same loudness in 0° direction and 180° direction. The spectrum was normalized.

steering delays $(\delta_1, \delta_2, ...\delta_n)$, the SRP algorithm calculates the delay and sum beamformer in a n microphone acoustic array system and makes them aligned in time, and then sums all these time aligned signals below equation [23].

$$Y(t, \delta_1, \delta_2, ...\delta_n) = \sum_{i=1}^{n} y_i(t - \delta_i) \tag{2}$$

In the frequency domain, the output of an m-element, filter-and-sum beamformer can be presented in Equation 3.

$$Y(\omega, \delta_1, \delta_2, ...\delta_n) = \sum_{i=1}^{n} G_{m_i}(\omega) X_{m_i}(\omega) e^{-j\omega\delta_i} \tag{3}$$

where $X_{m_i}(\omega)$ is the Fourier Transforms of the microphone signals, and $G_{m_i}(\omega)$ is the Fourier transforms of temporal filters for microphone m_i. When locating the location of a sound source, the power of the steered response mostly like reaches a maximum. The SRP can be expressed as the output power of a filter-and-sum beamformer and is defined as follows.

$$P(\delta_1, \delta_2, ...\delta_n) = \int_{-\infty}^{+\infty} Y(\omega, \delta_1, ..., \delta_n) Y'(\omega, \delta_1, ..., \delta_n) \tag{4}$$

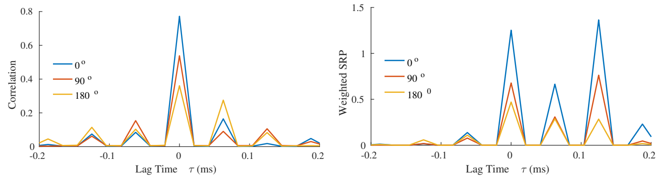
As GCC-PHAT measures the cross correlation for each pair of microphones to estimate the delay, the above SRP function can be expressed as a sum of GCCs for the different microphone pairs at the time-lag corresponding to their TDOA. Considering the GCC-PHAT of a microphone pair m_i and m_j, S_{m_i} and S_{m_j} represent the inverse Fourier transform of the estimated cross-power spectral density for frequencies f. In particular, R_{m_i,m_j} calculates the cross-correlation for a time interval centered at the time instant t_0, which has prominent peaks at a delay of τ (as illustrated in Figure 6a).

$$R_{m_i,m_j}(t) = \int_{-\infty}^{+\infty} \frac{S_{m_i}(f) * S_{m_j}(f)}{|S_{m_i}(f)||S_{m_j}(f)|} e^{j2\pi ft} \tag{5}$$

The weighted SRP-PHAT sums the GCC of all pairs of microphones in the microphone array.

$$P(x) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} R_{m_i,m_j} \tag{6}$$

We calculate SRP-PATH based features for our machine-learning model. VAs typically contain microphone arrays, as shown in Figure 7, where a specific distance separates the microphones. As the distance (d) between orthogonal microphones are 8.5 cm, 9 cm, and 6.5 cm for D1, D2, and D3 microphone arrays, respectively, the maximum delay between two microphones is the distance divided by the speed of sound (i.e., $c = 340m/s$). The number of delayed samples can be represented as $N = d * f/c$. Given that f is 48 kHz, we select the SRP within ±0.25 ms ($0.25 \times 0.001 \times 48000 \times 2 + 1 = 25$ samples), ±0.27 ms (13 × 2+1=27 samples), ±0.2 ms (10 × 2+1=21 samples) for D1, D2 and D3, respectively. Figure 6b plots the SRP, which shows that the smaller the angle with a VA, the higher the power value. It shows each SRP has 3 $\sim$ 4 high peaks due to reverberation. We rank the top three peak values as one feature.



(a) GCC-PHAT

(b) SRP-PHAT

Fig. 6: (a) Signal cross correlations between Mic1 and Mic2 for D3 and (b) the weighted SRP in the direction of 0°, 90°, and 180°.

In addition, we use GCCs and TDoA of all selected pairs of microphones (e.g., for a 4-channel microphone array we compute $\binom{4}{2}$ cross-correlations) as another feature vector (e.g., for D2 there are $6 \times 27 + 6 = 168$ values). Figure 6a plots the GCCs between microphone pair Mic1 and Mic2 on D3, which shows that the smaller angle with the VA has a higher peak value at delay time $\tau = 0$ while the larger angle has a higher peak value at delay time $\tau = \pm 0.0625$ ms or $\tau = \pm 0.125$ ms. We also compute different statistical summaries of SRP and GCCs values, including kurtosis, skewness, maximum, absolute deviation (MAD), and standard deviation.

Speech Directivity. As human speech mainly lies in the frequency range of 100 $\sim$ 4000 Hz (most usable voice frequencies), we thus divide the speech frequency band into the frequency range 100 $\sim$ 400 Hz as low-band and 500 $\sim$ 4000 Hz as high-band. We calculate the mean magnitude of low-band and high-band frequency and get the high-low band ratio (HLBR). We then divide the low-band frequency into 20 smaller chunks and calculate the mean, RMS and standard deviation of each frequency chunk.

IV. Evaluation

In this section, we perform a comprehensive analysis of *HeadTalk* under various settings to evaluate its accuracy, stability, and system-level performance. First, we establish the effectiveness of our model by evaluating the accuracy of detecting human speakers from mechanical speakers (§IV-A1), followed by determining the speaker's head orientation (§IV-A2) and the impact of training set size (§IV-B1). We examine the sensitivity of our system by analyzing the impact of the following factors: distance between speaker and VA (§IV-B2), different wake words (§IV-B3), different devices (§IV-B4), different environments (§IV-B5), number of microphones used (§IV-B6), VA placement (§IV-B7), cross-room settings (§IV-B8), temporal stability (§IV-B9), ambient noise (§IV-B10), seating and standing up (§IV-B11), loudness (§IV-B12), surrounding objects (§IV-B13), cross-users settings (§IV-B14), and runtime (§IV-B15).

Experimental Setup. We implement *HeadTalk* using three off-the-shelf microphone arrays, including a miniDSP 8-channel UMA USB microphone array V2.0 (D1) [7], a 6-channel Seeed's ReSpeaker Core V2.0 (D2) [4], and a 4-channel Seeed Respeaker USB Microphone Array (D3) [5]. Table I summarizes the device specifications. We configured all three devices to record raw audio at 48 kHz. Figure 7 shows the microphone placement on the different boards. Figure 8 and 9 highlights the device setup in a lab and living room environment, respectively. The lab space consists of a 280 square foot ($20' \times 14'$) office room with ten-foot dropped ceilings. The living room is a part of a 2-bedroom apartment with the following dimensions $33' \times 10' \times 8'$. The lab space is shown in Figure 8 which emulates a smart home living room equipped with a table, sofa, desktop computer, smart TV, smart lights, motion sensors, and smart cameras. The default noise level in the lab setting was 33 dB (SPL). The devices were placed on a near-wall study table (74 cm from the ground) in location A as shown in Figure 8. The home setting is shown in Figure 9, and it is exposed to more diverse ambient noise from various household devices such as a refrigerator and microwave as well as external noise from cars passing by (the apartment is located near a two-lane street). The default noise level in the home setting was 43 dB (SPL). The devices were placed on a near-window TV shelf (83 cm from the ground).

TABLE I: Prototype Devices

No	Device	Channels	Description
D1	UMA-8 USB mic array V2.0	7	XMOS XVF3000 chip
D2	Seeed Respeaker V2.0	6	1.5GHz RAM with 1GB RAM
D3	Seeed Respeaker USB Mic Array	4	XMOS XVF-3000 chip

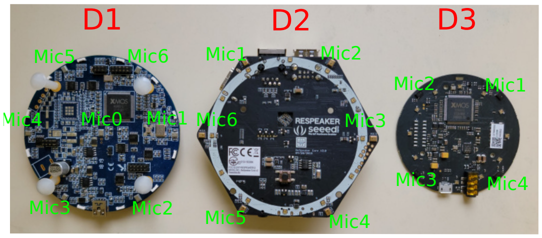


Fig. 7: Configuration of microphones across different prototype devices D1, D2 and D3. D2 is equipped with a 6-microphone array, similar to how microphones are distributed inside an Amazon Echo Dot [9]. D2 takes on average 156 ms and 527 ms to detect the liveliness and speaker orientation.

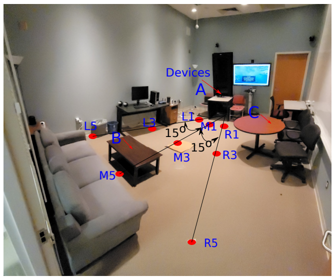


Fig. 8: Lab setup.

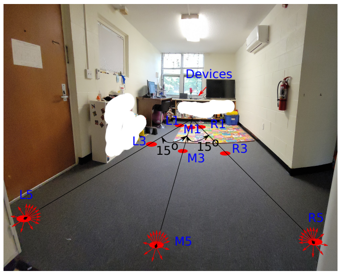


Fig. 9: Home setup.

Data Collection Process. Data was collected using three prototypes VAs (as shown in Figure 7) from January 2021 to November 2021 (across multiple sessions). We collect data at three different distances: 1, 3, and 5 meters, and from three radial directions including $-15°$ (labeled as $L1$, $L3$ and $L5$), $0°$ (labeled as $M1$, $M3$, and $M5$) and $+15°$ (labeled as $R1$, $R3$, and $R5$) as illustrated in Figure 8 and 9. At each grid intersection (i.e., red circles), 14 different angles spanning $360°$ were marked on a cardboard and placed on the floor (visible in Figure 8 at location $M3$). In each data collection session, a given wake word (e.g., "Computer") was spoken at the loudness of 70 dB twice at each angle and then rotated to the neighboring angle in the clockwise direction and the whole process was repeated. Thus, for any grid intersection (i.e., fixed radial direction and distance), we collected $14 \times 2 = 28$ samples per wake word in each data collection session. We collected data for three wake words: "Hey Assistant!", "Computer", and "Amazon". We selected the "Hey Assistant!" phrase as a wake word as this was also used by Ahuja et al. [13]. This way, we could leverage and compare against their dataset. "Computer" and "Amazon" wake words were used as they are common wake words for Amazon Alexa [2].

Datasets. To study the feasibility of our proposed system, we consider many variables that can affect the accuracy of speaker orientation, including robustness across devices, distance, wake words, time, rooms, device placement, noise and loudness. We cover the following different setups:

i. **3 devices**: UMA-8 USB mic array, Seeed's ReSpeaker Core v2.0, and ReSpeaker USB mic array;
ii. **3 wake words**: "Hey Assistant!", "Computer" and "Amazon";
iii. **3 different time frame**: day, week and month;
iv. **2 rooms**: lab and home;
v. **3 device placements**: Location A, B and C (shown in Figure 8);
vi. **3 distances**: 1 meter, 3 meters and 5 meters;
vii. **14 angles**: 0°, +15°, -15°, +30°, -30°, +45°, -45°, +60°, -60°, +90°, -90°, +135°, -135°, 180°.

All audio data is sampled at 48 kHz. We describe the different datasets collected in Table II.

A. Machine Learning Model Selection

We use Dataset-1 and Dataset-2 to evaluate the overall model performance. Also, by default, the utterance "Computer" and device $D2$ are used for our evaluations, unless stated otherwise. We use only four microphones from D1 (i.e., combinations of {Mic2, Mic3, Mic5, and Mic6}) and D2 (i.e., combinations of {Mic1, Mic2, Mic4, and Mic5}) in order to make our findings comparable with D3 (which has only four microphones) as well as to reduce computation time. We will show the impact of the number of microphones in Section IV-B6.

For liveliness detection, we adopt the wav2vec2 learning network [15] to distinguish mechanical speakers from human speakers. For detecting speaker orientation, we compare the performance of four classifiers including Random Forest (RF), Decision Tree (DT), Support Vector Model (SVM), and k-nearest neighbors (kNN). We use the Bagging algorithm for the RF classifier. We test different numbers of trees ranging from 10 to 500 and empirically settle on the number of trees as 200. For DT learning, we select the maximum number of splits as 5. For kNN learning, we select the number of neighbors as 3. To generate each single binary classification model in SVM, we use the implementation of SVDE with 10-fold cross validation in LIBSVM [18] and select the best complexity parameter for Radial Basis Function (RBF) through grid search. As for labeling samples as facing versus non-facing we label 0°, ±15°, ±30°, and ±45° as forward-facing direction while ±60°, ±90°, ±135°, and 180° are labeled as non-facing direction. We later on show how models can be further refined by selectively filtering certain angles from the training set.

As for evaluation metrics we use well-known metrics like True-positive rate (TPR), False-acceptance-rate (FAR), False-rejection-rate (TRR), precision, recall, and F1-Score. We perform a *cross-session* evaluation to determine performance, where we select one session's data as the training set, use the remaining session as the test set, and report the average across the two sessions. We compared the F1-Score of four classifiers for detecting speaker orientation. SVM exhibited the best average F1-Score across both the lab and home settings. We, therefore, use the SVM model for all further evaluations.

1) Distinguish Human vs. Mechanical Speaker: We use the SpeechBrain library [56] and ASVSpoof 2019 physical access dataset [66] to train our wav2vec2 model to distinguish human speakers from mechanical speakers. We use the default ASVSpoof 2019 [66] 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 Dataset-1 and Dataset-2. Dataset-1 provides samples of live-human speech, whereas Dataset-2 provides samples of replayed audio through a Sony speaker. Thus, we have a total of $504 \times 2 \times 2 = 2,016$ samples. We use the previously trained model to test the unseen 2,016 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 2,016 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.

TABLE II: Dataset Summary.

Dataset	Settings	Samples
Dataset-1	2 rooms, 3 devices, 3 utterances, 9 locations, 14 angles, 2 samples, 2 sessions.	$2 \times 3 \times 3 \times 9 \times 14 \times 2 \times 2 = 9072$
Dataset-2 (Replay)	Sony loudspeaker: 2 utterances ("Computer" and "Hey assistant'"), 9 locations, 14 angles each position, 2 repetition, 2 session.	$2 \times 9 \times 14 \times 2 \times 2 = 1008$
Dataset-3 (Temporal)	"Computer" utterance, 3 locations ($M1$, $M3$ and $M5$), 14 angles each position, 2 session, 2 repetition, 2 temporal (week and month).	$3 \times 14 \times 2 \times 2 \times 2 = 336$
Dataset-4 (Ambient)	"Computer" utterance, 2 ambient noise (white noise and TV series), 3 distance ($M1$, $M3$ and $M5$), 14 angles each position, 1 session, 2 repetition.	$2 \times 3 \times 14 \times 2 = 168$
Dataset-5 (Sitting)	"Computer" utterance, 3 locations ($M1$, $M3$ and $M5$), 14 angles each position, 1 session, 2 repetition.	$3 \times 14 \times 2 = 84$
Dataset-6 (Loudness)	"Computer" utterance, 3 locations ($M1$, $M3$ and $M5$), 14 angles each position, 1 session, 2 repetition, 2 loudness (60 and 80 dB).	$3 \times 14 \times 2 \times 2 = 168$
Dataset-7 (Nearby)	"Computer" utterance, 3 locations ($M1$, $M3$ and $M5$), 14 angles each position, 1 session, 2 repetition, 3 settings.	$3 \times 14 \times 2 \times 3 = 252$
Dataset-8 (Multi-user)	10 participants (4 male, 6 female, mean age 20) [13], 9 locations, 8 spoken angles, 2 repetition.	$10 \times 9 \times 8 \times 2 = 1440$

2) Determine Facing and Non-facing Orientation: As we discussed in Section III-B1, we define the angle range of −30° to 30° as the facing direction. Our objective is to achieve a relatively higher TPR (and lower FRR) for facing direction and a lower FAR for non-facing direction. We collected two additional angles to verify our approach, including 75° and −75° for the "Computer" utterance collected by D2 in the lab setting. We train the data from one session and test on data from the other session. We next perform cross-session evaluation and calculate the average performance while considering training data belonging to different arcs:

- Definition-1: 0°, ±15°, ±30°, and ±45° are considered as facing direction; ±60°, ±75°, ±90°, ±135°, and 180° are considered as non-facing direction.
- Definition-2: 0°, ±15°, and ±30° are considered as facing direction; ±60°, ±75°, ±90°, ±135°, and 180° are considered as non-facing direction.
- Definition-3: 0°, ±15°, and ±30° are considered as facing direction; ±75°, ±90°, ±135°, and 180° are considered as non-facing direction.
- Definition-4: 0°, ±15°, and ±30° are considered as facing direction; ±90°, ±135°, and 180° are considered as non-facing direction.

Table III summarizes the performance of the four facing and non-facing definitions. Definition-4 achieves the best perfor-

TABLE III: Accuracy for different definitions of facing and non-facing orientation for the utterance of "Computer".

Definition	Accuracy	FRR	FAR	Precision	Recall	F1-Score
Definition-1	92.54	10.32	5.25	93.04	89.68	91.33
Definition-2	96.43	8.33	0.93	98.20	91.67	94.81
Definition-3	96.99	3.89	2.38	96.71	96.11	96.39
Definition-4	96.95	3.33	2.78	97.28	96.67	96.96

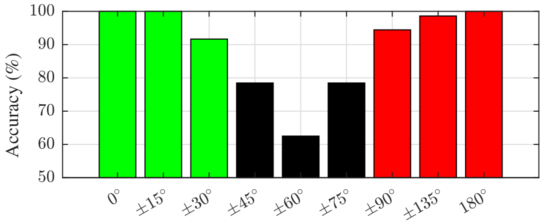


Fig. 10: Detecting speaker orientation at different angles.

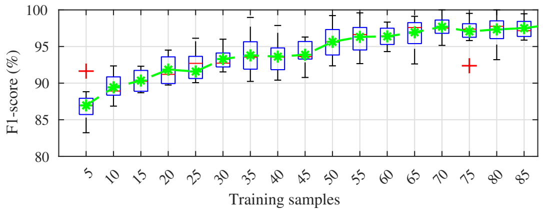


Fig. 11: Impact of training set size on accuracy.

mance with an overall accuracy of 96.95%, while the FRR and FAR are as low as 3.33% and 2.78%, respectively. We also use the trained model based on Definition-4 to test borderline angles such as ±45°, ±60°, and ±75°. Figure 10 shows the accuracy of the facing orientation (marked with green color), borderline orientation (marked with black color), and non-facing orientation (marked with red color). The accuracy for most angles is above 90% excluding the angles of borderline orientation (i.e., ±45° and ±60° and ±75°). The main reason is that these borderline angles form a *soft boundary* between facing and non-facing orientation and thus cause confusion for the classifier. As discussed in Section III-B1, a speaker's immediate field of view is ±35°, closely matching what our classifier can predict with over 90% accuracy. Thus, we consider the direction of Definition-4 for the remaining evaluation in speaker orientation detection. We use these two trained models for sensitivity analysis. Note that our system achieves an average accuracy of 96.14% (F1-Score = 96.24%) in detecting speaker orientation across all devices, utterances, and environments using Dataset-1.

B. Sensitivity Analysis

1) Varying Training Size: To make *HeadTalk* user-friendly, the enrollment effort for a new user is a critical factor. We consider the performance of the classifier in the presence of limited training samples. For this experiment, we vary the training set size (N) for each class from 5 to 100 with an interval of 5 samples and test the remaining samples. We randomly select N samples of the "Computer" utterance for each class collected through D2 under the lab setting and repeat the training and testing process 10 times to calculate the average accuracy. Figure 11 shows the box-plot and mean (shown in green color) of the 10 F1-Score in increasing training set size. The result shows that as the training set size increases, the accuracy also rises. However, we see that with only 20 samples per class, the average F1-Score goes over 92%. Thus, not a significant amount of training samples is required.

2) Impact of Distance: To evaluate the impact of distance between the speaker and microphone, we use the trained model from Section IV-A2 to test against samples of varying distances using Dataset-1 (note that this dataset included samples from all three distances 1 m, 3 m and 5 m). We obtain 36 accuracy values (2 session × 3 devices × 2 rooms × 3 wake words). The average accuracy is 98.38 ± 2.41%, 97.50 ± 4.90%, 92.55 ± 7.19% for the distance of 1 m, 3 m and 5 m, respectively. The result shows that head orientation detection decreases as the distance between the speaker and microphone increases; however, the performance still remains above 92% at a 5-meter distance.

3) Impact of Wake Word: To evaluate the impact of wake words, we plot the F1-Score of each wake word for each session in all three devices across two rooms. In total, we obtain 12 F1-Score values (2 sessions × 3 devices × 2 rooms) for each wake word. Figure 12 shows the box plot for the F1-Score. The average F1-Score is 95.92%, 96.40%, and 96.39% for "Hey Assistant!", "Computer", and "Amazon", respectively. The result shows there are no significant differences across the three wake words.

4) Impact of Devices: We also obtain 12 F1-Score values (2 sessions × 3 wake words × 2 rooms) for each device. We found the average F1-Score to be 97.47%, 96.26%, 94.99% for D1, D2 and D3, respectively. The box plots shown in Figure 13 illustrate that D1 has the best performance. As the distance between each pair of microphones increases, the device can also better sense lower frequencies. Both D1 and D2 have better average performance compared to D3. D1 has slightly higher accuracy than D2. The reason is that the voice recorded by D1 has less noise and thus has a higher signal-to-noise ratio (SNR) compared to D2. For example, we measured the SNR for one session worth of data (using the "Computer" wake word) recorded simultaneously by both D1 and D2. We found the SNR to be 25.09 dB for D1, whereas it was 24.25 dB for D2. As D1 does not support programming on the board, we did not select D1 as our default device.

5) Impact of Environment: Figure 14 presents the box plot for the two rooms, which includes 18 F1-Score values (2 sessions × 3 wake words × 3 devices) for each environment. The average F1-Score is 98.08% and 94.39% for the lab and home setting, respectively. The result shows that the lab setting has a better performance, primarily due to the reduced ambient noise levels in the lab (measured at 33 dB) compared to home (measured at 43 dB). Moreover, the presence of more furniture in homes affects sound transmission paths, resulting in more intricate reverberation. However, even in a typical home setting, we can see an F1-Score of greater than 94%. We further study the impact of surrounding objects in Section IV-B13.

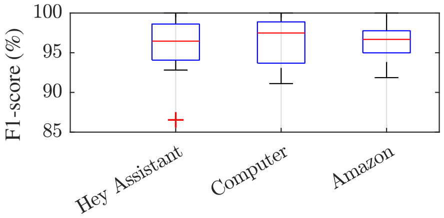


Fig. 12: F1-Score for different wake words.

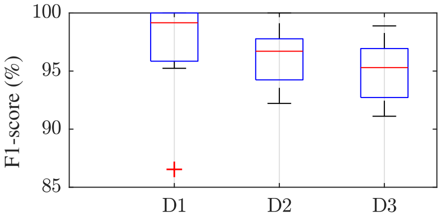


Fig. 13: F1-Score for different devices.

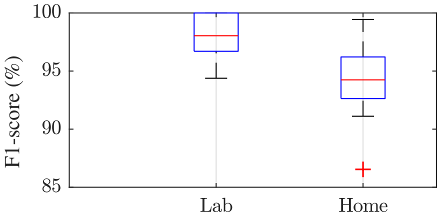


Fig. 14: F1-Score for lab and home setting.

TABLE IV: Performance for different combinations of mics.

No.	Channels	Accuracy	Precision	Recall	F1-Score
2	[1 2]	95.70	95.60	95.83	95.71
3	[1 2 5]	95.83	94.60	97.22	95.90
4	[1 2 4 5]	96.67	96.77	96.67	96.70
5	[1 2 3 4 5]	98.61	100	97.22	98.59
6	[1 2 3 4 5 6]	97.22	97.23	97.22	97.22

6) Impact of the Number of Microphones: To evaluate how the performance is impacted by the number of microphones used, we select N out of the six microphone data of D2 in the lab setting. We select the microphones in an order that results in the greatest distance among them. The greater the distance between two microphones, the longer the delay between them, which can emulate the perceived hearing differences across human ears. We, therefore, select pairs that result in the greatest distance between them. Table IV summarizes the performance. We see that the performance increases with the increasing number of channels and the 5-channel microphones have the best overall performance. After that, the performance decreases with the increasing number of channels.

7) Impact of Device Placement: To evaluate the impact of the device placement, we place D2 on a coffee table (labeled as B in Figure 8, at the height of 45 cm from the ground) and on a work table (labeled as C in Figure 8, at the height of 75 cm from the ground). We collected data across two sessions for the "Computer" wake word at a distance of 3 m along the direction of 0°. We then tested the model that was trained on using data from location A. *HeadTalk* achieves an average accuracy of 97.50% and 91.25% when the device is placed at locations B and C, respectively. While we see a slight drop in accuracy in location C (compared to 96.95% when trained and tested on samples at location A), the performance is still over 90% when trained and tested on samples from different locations within a large room.

8) Cross-environment Performance: To evaluate how the location between the microphone and human speaker impacts accuracy, we train in one environment (e.g., data collected in the home setting) and test in another environment (e.g., data collected in the lab setting). *HeadTalk* achieves an average accuracy of 77.73% (78.20% F1-Score) when trained and tested on samples collected from the home and lab settings, respectively, and vice versa. However, suppose we train on data from one session under both the lab and home setting and test on data from the other session (and vice versa). In that case, the average accuracy is 96.90% (97.09% F1-Score), 95.62% (95.70% F1-Score), and 95.02% (95.70% F1-Score) for "Hey, Assistant!", "Computer", and "Amazon" wake word, respectively. This shows that the model can quickly adapt to new settings and achieve similar accuracy as the normal training setting (96.14%).

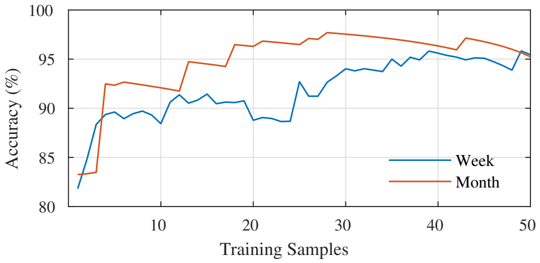


Fig. 15: Temporal accuracy with incremental learning.

9) Temporal Stability: The motivation behind evaluating the accuracy of *HeadTalk* using training and testing samples from different days is that in the real world, the user will provide training samples only on the first day when setting up *HeadTalk* , and then *HeadTalk* should be able to effectively detect orientation on subsequent days. We collected additional data after one week and one month (Dataset-3 in Table II). We use the model trained in Section IV-A and test against Dataset-3. *HeadTalk* achieves 81.25% and 83.19% accuracy when testing against data that is one-week and one-month old, respectively. However, we can adopt an incremental learning approach and reuse high-confidence test samples (i.e., $\geq$ 80%) as training data and rebuild the model periodically. Figure 15 shows that the accuracy improves to 92% and 90% after adding just 10 new training samples from the one-week and one-month dataset, respectively. The accuracy is around 95% after adding 40 new training samples.

10) Impact of Ambient Noise: We use Dataset-4 to evaluate the impact of two types of background noise on a model trained with samples that have no intentional ambient noise. We generate two types of ambient noise: white noise and a TV playing a popular series (which includes various types of sounds typically encountered in a home, such as people chatting, people laughing, people walking, door opening and closing, etc.) in the background at 45 dB (SPL). With white noise, the accuracy is 89%, while the accuracy is 83.33% when the TV is playing. We can see that the background ambient noise typically found in homes degrades performance. Recall with no ambient noise (i.e., default noise at 33 dB), the average accuracy is 98.08% in the lab setting (see details in §IV-B5).

11) Impact of Sitting and Standing: A user may sit on a chair or sofa while interacting with a VA. As our models are trained on data collected while standing up, we use Dataset-

5 to evaluate whether a model trained on data while the speaker is standing up impacts detection when the speaker sits down. We found the accuracy to be 93.33% when trained on standing up and tested on sitting down. Thus, we see that sitting down does not significantly impact detecting the speaker's orientation.

12) Impact of Speech Loudness: We use Dataset-6 to evaluate the impact of speech loudness. Our original model was trained on data collected at 70 dB SPL. The accuracy is 93.33% when tested with samples collected at 60 dB, while the accuracy is 95.83% when tested with samples collected at 80 dB. This shows that increased loudness improves accuracy. At higher loudness, the signal is stronger, and as a result, the signal characteristics for the facing and non-facing orientations are more prominent.

13) Impact of Surrounding Objects: We use Dataset-7 to evaluate the impact of nearby objects. For the partial block scenario, the speaker orientation detection accuracy is 95.83%. However, the accuracy drops to 70% when the device is entirely blocked. The reason is that surrounding objects impact the reverberation paths, making the VA hear the voice like a speech coming from the backward direction. After raising the device height to 14.8 cm, the accuracy improved to 95%, which is close to the overall accuracy (i.e., 96.95%). *HeadTalk* can be mounted higher similar to how the new generation of Amazon Echo devices (e.g., Echo Plus) is produced to reduce the impact of nearby objects. However, VAs are typically placed in an open area easily visible to users.

14) Cross-Users Setting: In a smart home setting, multiple people may want to access the VA without requiring any training effort. To evaluate the feasibility of multi-user settings, we use the dataset provided by Ahuja et al. [13] (i.e., Dataset-8 in Table II) to evaluate the accuracy cross different users. As there are no ±15° and ±30° angles in the dataset provided by Ahuja et al. [13], we consider 0° and ±45° as the facing angles, and the remaining five angles (i.e., ±90°,±135° and 180°) as the backward direction for this experiment. The dataset size is imbalanced due to the unequal size of facing angles and non-facing angles. The facing orientation samples are the minority class, while the non-facing orientation samples are the majority class. Therefore, we need to up-sample the facing angles data to achieve equal class representation. We test Synthetic Minority Over-sampling (SMOTE) [19] and Adaptive Synthetic Sampling (ADASYN) [37] methods which are two popular up-sampling approaches. We selected ADASYN as our up-sampling approach for its superior performance. We then do 10-fold cross-validation, that is, use 9 person's data as training and test against the remaining one person. Figure 16 shows the accuracy for all participants. The average accuracy is 88.66% (F1-Score 85.09%).

15) Run-time Performance: We first ran *HeadTalk* on a personal computer equipped with an Intel i7-2600 3.40 GHz processor and 16 GB RAM. We needed one channel of audio data to detect liveliness and 4-channel audio data to detect speaker orientation. It took on average 42 ms and 136 ms to detect the liveliness and speaker orientation, respectively. We then run it on our prototype VA (i.e., D2 in Figure 7), which is equipped with quad-core ARM Cortex A7 (1.5 GHz) and 1 GB RAM. It took 527 ms to detect speaker orientation. The latest VA models have higher computing power. For example, Google Nest Audio is equipped with a high-performance machine learning hardware engine (quad-core A53 1.8 GHz CPU [34]). Given that *HeadTalk* only processes, the wake command and VAs usually wait for a few seconds before responding, commodity VAs should fulfill our runtime requirement.

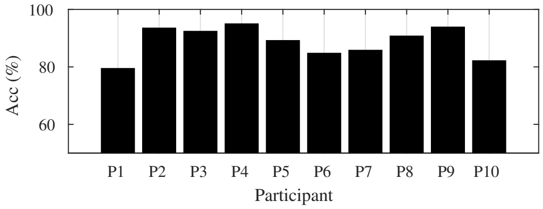
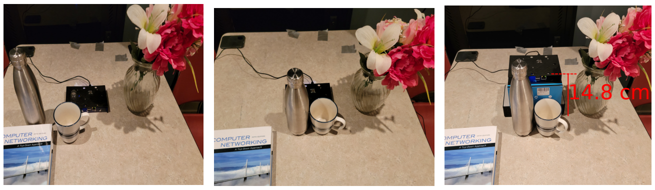


Fig. 16: TPR across different users.



(a) Partially blocked **(b)** Fully-blocked **(c)** Raised

Fig. 17: Setup for testing the impact of surrounding objects.

V. User Study and Feedback

Existing literature has shown that 20 participants are sufficient to conduct usability studies [29], [46]. Thus, we recruited 20 participants (14 male and 6 female graduate students) to interact with our VA prototype in the lab and were paid $10 Amazon gift cards for their participation (for 30 minutes of participation). Our study was IRB approved. Each participant interacted with the VA in three locations (M1, M3, and M5; each location with five forward-facing angles and five backward-facing angles) plus one freestyle speech activity where the participant can either sit or walk anywhere at any angle (10 samples). For each location, the participant spoke a given utterance (e.g., "Computer") once at each angle and then rotated to the neighboring angle. If it is facing, the application will say, "How can I help you?"; if not, the application will say, "Sorry, I didn't hear you."

Each participant in our study completed a post-study survey, where they were asked about their *experience* and *expectations* using our proposed head orientation-based privacy control. We also asked participants questions about comparing our approach with the existing privacy control approach, including pressing the mute button, deleting command history, and unplugging the device. Table V list the basic questions asked. Participants also answered the SUS (System Usability Scale) questionnaire [16], which is a well-known standard for measuring the usability of software systems and consists of 10 standard usability questions, each with five possible answers

TABLE V: Summary of post-study survey and responses.

Question	Response (count)
How many home voice assistants do you have at home?	0 (5), 1 (12), 2 (2), above 2 (1)
How often do you face the VA when you are interacting with the VA (if you have one)?	N/A* (5), Very less (1), Less (4), Often (6), Very often(4)
How easy was it to use *HandTalk* compared with existing privacy controls?	Extremely easy (10), Somewhat easy (9), Neither easy nor difficult (0), Somewhat difficult (1), Extremely difficult (0)
Would you deploy *HeadTalk* on your voice assistant?	Definitely yes (7), Probably yes (7), Might or might not (5), Probably not (0), Definitely not (1)
Compare *HeadTalk* with the existing privacy control.	Much Better (9), Somewhat better(5), About the same (5), Somewhat worse (0), Much worse (1)

* Participants did not own any VA (question skipped)

(5-point Likert scale, where 1 represents strong disagreement and 5 represents strong agreement).

Takeaways. 66.67% (10/15) participants who own at least one VA recall staring at the device when interacting with a VA. 95% (19/20) participants consider *HeadTalk* to be either extremely easy or somewhat easy to use as shown in Table V. 70% (14/20) participants said they would probably or definitely deploy *HeadTalk* on their own VAs. Around 70% of participants felt *HeadTalk* is better (i.e., either somewhat or much better) than using the existing privacy controls. Our survey also asked participants for comments about the pros and cons of our approach. In general, the participants felt that the system was easy to use and could prevent unnecessary activation of VA. There were some concerns about inconsistency with distance and some angles. Next, we compare *HeadTalk* with other alternatives using SUS scores. A SUS score of above 68 is typically considered above average, and anything below 68 is below average [16]. Based on the participants' responses, the 95% confidence interval of SUS score for *HeadTalk* is 77.38 ± 6.26, while the SUS score for the existing privacy control (i.e., physical mute button.) is 74.75 ± 8.12. In general, participants found our approach more usable than the existing privacy control.

Following are some interesting responses from participants.

> P1: It was a new concept to me but I like the idea. Hopefully it'll be possible to implement in VA devices in the future, for more privacy and convenience!
>
> P20: It is an on demand solution for voice privacy: I can choose whether to make the VA to react, instead of other solutions like mute button that I have to toggle beforehand, or delete history afterwards.
>
> P9: I like this orientation feature. I have had moments where my existing speaker responds when not talking. It would be nice to explore orientation of just the head. Sometime I may face the speaker but look down.
>
> P8: It is a nice concept, but learning what angels trigger it whereas what do might need some getting used to. For instance, a lot of people use these smart systems in their kitchens and might want to give a command just turning a bit towards it and not leave their task at hand.

VI. Limitations

There are a few limitations to our work. First, we tried our best to maintain the exact angle while collecting data. However, some human errors may exist. The soft boundary concept between facing and non-facing orientation can potentially mitigate the impact of human error. Alternatively, a VR headset could provide more accurate angle reference in data collection [72]. Second, accuracy decreased when objects were surrounding the VA. However, as we have shown, if the VA's height is higher than the surrounding objects, it is not an issue. Third, we assume speakers tend to face the device for voice interaction. As voice interfaces are designed to be hands-free, users can interact with VA while doing other activities, such as exercise or cooking. Thus, in some scenarios, our proposed approach may not be a good fit, e.g., providing a voice command while lying on the sofa. Also, users may have difficulty locating the VA in a dark room if no visual indicators are present on the VA. However, VAs typically emit lights to help users locate the device in darkness. Our participants also list some scenarios where *HeadTalk* may have limited usability. The participants recruited in our user study were all graduate students, which might have introduced unwanted bias in the result; nevertheless, we believe that our findings still hold for a tech-savvy population. Furthermore, our liveliness detection system relies on the fact that most smart home devices are not able to simulate similar high-frequency responses as live humans; however, going forward, as device capability improves and as audio technology advances, our proposed technique may not be able to distinguish between human and mechanical speakers as effectively. Lastly, our analysis does not cover the impact of moving speakers.

VII. Conclusion

In this paper, we present a new privacy control for VAs, called *HeadTalk* that requires no additional hardware. We make novel contributions by investigating the application of context awareness for smart speaker privacy controls, specifically identifying if a voice command is issued by a human speaker and then examining if the speaker intentionally triggered the speaker by facing the device. *HeadTalk* extracts sound propagation characteristics for facing and non-facing speech. We extensively evaluated *HeadTalk* using an extensive data set, covering three wake words recorded by three VA devices, covering two room settings, and showed that it could achieve an average accuracy of 96.14% 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.

Acknowledgement

We thank our anonymous reviewers for their valuable feedback. This material is based upon work supported in parts by a Meta research gift.

REFERENCES

[1] "Guidelines on ergonomic criteria for bridge equipment and layout," International Maritime Organization, Tech. Rep., 12 2000.

[2] "How to change your wake word," www.amazon.com/b?ie=UTF8&node=21341305011, 2021.

[3] "Reflection, refraction, and diffraction," 2021. [Online]. Available: https://www.physicsclassroom.com/class/sound/Lesson-3/Reflection,-Refraction,-and-Diffraction

[4] "Seeed's respeaker core v2.0," https://wiki.seeedstudio.com/ReSpeaker_Core_v2.0/, 2021.

[5] "Seeed's respeaker microphone array v2.0," https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/, 2021.

[6] "Sony srs-x5 speaker," https://www.sony.com/electronics/support/speakers-wireless-speakers/srs-x5, 2021.

[7] "Uma-8 usb microphone array v2.0," https://www.minidsp.com/products/usb-audio-interface/uma-8-microphone-array, 2021.

[8] "Delete some or all of your voice history in the Alexa app," https://www.amazon.com/gp/help/customer/display.html?nodeId=GVLEU55A9NDZRWBU, 2023.

[9] "Teardown tuesday: Amazon echo dot v2," https://www.allaboutcircuits.com/news/teardown-tuesday-amazon-echo-dot-v2/, 2023.

[10] A. Abad, D. Macho, C. Segura, J. Hernando, and C. Nadeu, "Effect of head orientation on the speaker localization performance in smart-room environment," in *9th European Conference on Speech Communication and Technology*, 2005.

[11] A. Abad, C. Segura, C. Nadeu, and J. Hernando, "Audio-based approaches to head orientation estimation in a smart-room," in *8th Annual Conference of the International Speech Communication Association*, 2007.

[12] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *Proceedings of the 29th USENIX Security Symposium (USENIX Security)*, 2020, pp. 2685–2702.

[13] K. Ahuja, A. Kong, M. Goel, and C. Harrison, "Direction-of-voice (dov) estimation for intuitive speech interaction with smart devices ecosystems," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 1121–1131.

[14] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, "Music, search, and iot: How people (really) use voice assistants," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 26, no. 3, pp. 1–28, 2019.

[15] 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.

[16] J. Brooke, "SUS: a 'quick and dirty' usability scale," *Usability evaluation in industry*, p. 189, 1996.

[17] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *9th European Conference on Speech Communication and Technology*, 2005.

[18] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[20] Y. Chen, H. Li, S.-Y. Teng, S. Nagels, Z. Li, P. Lopes, B. Y. Zhao, and H. Zheng, "Wearable microphone jamming," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*, 2020, p. 1–12.

[21] H. Dai, A. X. Liu, Z. Li, W. Wang, F. Zhang, and C. Dong, "Recognizing driver talking direction in running vehicles with a smartphone," in *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 2019, pp. 10–18.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[23] J. H. DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*. Brown University, 2000.

[24] H. Do and H. F. Silverman, "Srp-phat methods of locating simultaneous multiple talkers using a frame of microphone array data," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 125–128.

[25] H. Do, H. F. Silverman, and Y. Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1. IEEE, 2007, pp. I–121.

[26] D. J. Dubois, R. Kolcun, A. M. Mandalari, M. T. Paracha, D. Choffnes, and H. Haddadi, "When speakers are all ears: Characterizing misactivations of iot smart speakers," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 4, pp. 255–276, 2020.

[27] C. F. Eyring, "Reverberation time in "dead" rooms," *The Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 217–241, 1930.

[28] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[29] L. Faulkner, "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing," *Behavior Research Methods, Instruments, & Computers*, vol. 35, pp. 379–383, 2003.

[30] H. Feng, K. Fawaz, and K. G. Shin, "Continuous authentication for voice assistants," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking (MobiCom)*, 2017, pp. 343–355.

[31] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *12th annual conference of the international speech communication association*, 2011.

[32] M. D. Gillette and H. F. Silverman, "A linear closed-form algorithm for source localization from time-differences of arrival," *IEEE Signal Processing Letters*, vol. 15, pp. 1–4, 2008.

[33] "Manage audio recordings in your web & app activity," 2020.

[34] "An inside look at nest audio," https://store.google.com/us/product/nest_audio_specs?hl=en-US, 2021.

[35] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[36] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry." in *Interspeech*, 2013, pp. 930–934.

[37] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 2008, pp. 1322–1328.

[38] H. Kato, H. Takemoto, R. Nishimura, and P. Mokhtari, "Spatial acoustic cues for the auditory perception of speaker's facing direction," in *In Proc. of 20th International Congress on Acoustics, ICA 2010*, 2010.

[39] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech 2017*, 2017, pp. 2–6.

[40] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[41] 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.

[42] R. Lee, M.-S. Kang, B.-H. Kim, K.-H. Park, S. Q. Lee, and H.-M. Park, "Sound source localization based on gcc-phat with diffuseness mask in noisy and reverberant environments," *IEEE Access*, vol. 8, pp. 7373–7382, 2020.

[43] S. Lee, M. Cho, and S. Lee, "What if conversational agents became invisible? Comparing users' mental models according to physical entity of ai speaker," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–24, 2020.

[44] Y. Lee, Y. Zhao, J. Zeng, K. Lee, N. Zhang, F. H. Shezan, Y. Tian, K. Chen, and X. Wang, "Using sonar for liveness detection to protect smart speakers against remote attackers," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–28, 2020.

[45] X. Lei, G.-H. Tu, A. X. Liu, C.-Y. Li, and T. Xie, "The insecurity of home digital voice assistants-vulnerabilities, attacks and countermeasures," in *Proceedings of the 6th IEEE Conference on Communications and Network Security (CNS)*, 2018, pp. 1–9.

[46] R. Macefield, "How to specify the participant group size for usability studies: a practitioner's guide," *Journal of usability studies*, vol. 5, no. 1, pp. 34–45, 2009.

[47] S. Maheshwari, "Hey, Alexa, what can you hear? and what will you do with it?" https://www.nytimes.com/2018/03/31/business/media/amazon-google-privacy-digital-assistants.html, March 2018.

[48] D. McMillan, B. Brown, I. Kawaguchi, R. Jaber, J. Solsona Belenguer, and H. Kuzuoka, "Designing with gaze: Tama–a gaze activated smart-speaker," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–26, 2019.

[49] W. Meng, D. S. Wong, S. Furnell, and J. Zhou, "Surveying the development of biometric user authentication on mobile phones," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1268–1293, 2014.

[50] 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.

[51] B. B. Monson, E. J. Hunter, and B. H. Story, "Horizontal directivity of low-and high-frequency energy in speech and singing," *The Journal of the Acoustical Society of America*, vol. 132, no. 1, pp. 433–441, 2012.

[52] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.

[53] M. Müller, S. van de Par, and J. Bitzer, "Head-orientation-based device selection: Are you talking to me?" in *Speech Communication; 12. ITG Symposium*. VDE, 2016, pp. 1–5.

[54] J. G. Neuhoff, M.-A. Rodstrom, and T. Vaidya, "The audible facing angle," *Acoustics Research Letters Online*, vol. 2, no. 4, pp. 109–114, 2001.

[55] S. Pradhan, W. Sun, G. Baig, and L. Qiu, "Combating replay attacks against voice assistants," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 3, no. 3, pp. 1–26, 2019.

[56] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[57] 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.

[58] E. H. Schwartz, "Amazon alexa unveils command to delete voice recordings," https://voicebot.ai/2019/05/29/amazon-alexa-delete-voice-recordings-command/, 2020.

[59] C. Segura, A. Abad, J. Hernando, and C. Nadeu, "Speaker orientation estimation based on hybridation of GCC-PHAT and HLBR," in *9th Annual Conference of the International Speech Communication Association*, 2008.

[60] C. Segura and F. J. Hernando Pericás, "GCC-PHAT based head orientation estimation," in *13th Annual Conference of International Speech Communication Association*, 2012, pp. 1–4.

[61] M. Shahzad and S. Zhang, "Augmenting user identification with wifi based gesture recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 2, no. 3, p. 134, 2018.

[62] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, 2018, pp. 105–111.

[63] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[64] K. Sun, C. Chen, and X. Zhang, ""Alexa, stop spying on me!" Speech privacy protection against voice assistants," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 298–311.

[65] R. Takashima, T. Takiguchi, and Y. Ariki, "Estimation of talker's head orientation based on discrimination of the shape of cross-power spectrum phase coefficients," in *13th Annual Conference of the International Speech Communication Association*, 2012.

[66] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[67] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7967–7971.

[68] J. Velasco, C. J. Martin-Arguedas, J. Macias-Guarasa, D. Pizarro, and M. Mazo, "Proposal and validation of an analytical generative model of srp-phat power maps in reverberant scenarios," *Signal Processing*, vol. 119, pp. 209–228, 2016.

[69] Q. Wang, X. Lin, M. Zhou, Y. Chen, C. Wang, Q. Li, and X. Luo, "Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones," in *Proceedings of the 2019 IEEE International Conference on Computer Communications (INFOCOM)*, 2019, pp. 2062–2070.

[70] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Proceedings of the IEEE 2011 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 4, 2011, pp. 1708–1713.

[71] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019, pp. 1215–1229.

[72] J. Yang, G. Banerjee, V. Gupta, M. S. Lam, and J. A. Landay, "Soundr: Head position and orientation prediction using a microphone array," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[73] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Transfer ability of monolingual wav2vec2. 0 for low-resource speech recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–6.

[74] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[75] L. Zhang, S. Tan, J. Yang, and Y. Chen, "Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016, pp. 1080–1091.