Poster: A 2-FA for Home Voice Assistants using Inaudible Acoustic Signal

Shaohu Zhang
szhang42@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Anupam Das
anupam.das@ncsu.edu
North Carolina State University
Raleigh, NC, USA

ABSTRACT

Voice assistants have been shown to be vulnerable to replay attacks, impersonation attacks and inaudible voice commands. Existing defenses do not provide a practical solution as they either rely on external hardware or work under very constrained settings. We introduce a hand gesture-based authentication system for smart home voice assistants called *HandLock*, which uses built-in microphones and speakers to generate and sense inaudible acoustic signals to detect the presence of a known hand gesture. Our proposed approach can act as a second-factor authentication (2-FA) for performing specific *sensitive* operations like confirming online purchases through voice assistants. The experiments involving 45 participants show that *HandLock* can achieve on average 96.51% true-positive-rate at the expense of 0.82% false-acceptance-rate.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

acoustic sensing, hand gesture, 2-FA, voice assistants

ACM Reference Format:
Shaohu Zhang and Anupam Das. 2022. Poster: A 2-FA for Home Voice Assistants using Inaudible Acoustic Signal. In *The 27th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '21), January 31-February 4, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3447993.3482863

1 INTRODUCTION

Smart home voice assistants (VAs) like Amazon Echo and Google Home thrive on the ability to enable users to interact with devices and services through voice to not only listen to music and flash briefings but also control other smart home appliances. However, the widespread use of VAs also gives rise to both security and privacy concerns due to their always-listening capability [4] and susceptibility to audio-based attacks [6].

One of the major security concerns with current VAs is the limited support for authentication. Other than simple, customizable wake words like "Alexa" or "Hi, Google," there is not much support for authentication in VAs. In recent years, several studies have proposed authenticating users through microphones and speakers embedded in smart devices [1, 8]. However, all of these schemes require the sensing device (i.e., microphones or speakers) to be placed very close to the user's mouth or held by the user, which does not amount to a practical solution for smart home VAs.

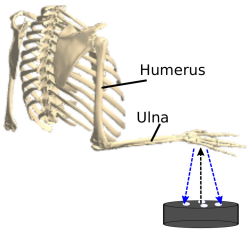

Figure 1: Anatomy of human arm and hand.

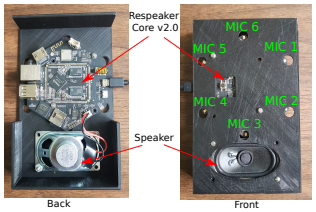

Figure 2: Device setup

In this work, we introduce a gesture-based biometric authentication scheme called *HandLock* that can recognize an authorized user based on his/her hand movement. To this end, *HandLock* emits inaudible acoustic signals and records the reflected signals to identify a user. The underlying hypothesis for *HandLock* is that *it is possible to distinguish different users even if they perform the same hand gesture due to their differing physical biometrics.* Specifically, as shown in the Figure 1, since the length of ulna and humerus of a given user is fixed, the starting and ending positions of the hand stay the same no matter how fast the user moves the hand. Therefore, the speed profile of a given gesture from the same user should remain similar as the speeds of different parts of the hand and limb change proportionally. Theoretically, the phase change that appears in the received acoustic signal is directly proportional to the speed at which a human hand was moved while performing a gesture. By combining our hypothesis with this theoretical result, we make the following observation: the phase shift recorded on the received acoustic signal is significantly different for different users, even if they all perform the *same* hand gesture. Through our evaluations, we see that our approach can be used not only to determine the *physical presence* of a user but also as an effective second-factor authentication method for VAs.

Table 1: Comparison with existing works.

Method	TPR	FAR	Extra Hardware	Device Free
WiID [5]	92.80%	-	WiFi transceiver	Yes
VAuth [2]	≤97%	0.10%	Wearable	No
P2Auth [3]	≤99.55%	2.1%	Wearable	No
HandLock	96.51%	0.82%	No	Yes

To the best of our knowledge, we are the first to propose an acoustic hand gesture-based 2-FA system for VAs. Our approach is *device-free* and *non-obtrusive*. Table 1 highlights a comparison

ACM MobiCom '21, January 31-February 4, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-8342-4/22/01…$15.00
https://doi.org/10.1145/3447993.3482863

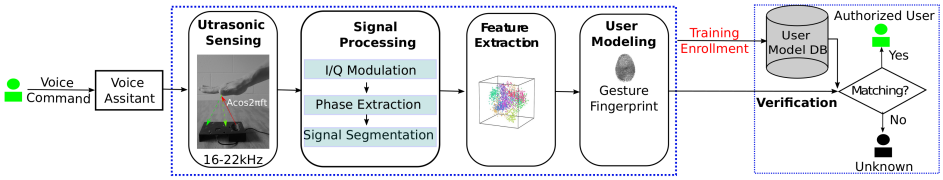

Figure 3: System overview.

with other existing VA authentication systems. *HandLock* achieves similar effectiveness when compared with existing approaches. However, unlike other approaches *HandLock* does not require any additional hardware and operates device-free. Thus, our approach is fully compatible with existing VAs.

2 SYSTEM DESIGN

2.1 System Overview

Figure 3 shows an overview of our proposed system, which consists of five main components: signal sensing, signal processing, feature extraction, user modeling, and verification. In the signal sensing phase, as soon as the VA enters a *sensitive operation* such as confirming an online purchase, the embedded speaker of the VA device prompts the user to perform a hand gesture and starts emitting and recording *inaudible* continuous wave signal. The signal processing phase uses a low pass filter to get the corresponding In-phase (I) and Quadrature (Q) signals. Given that we can extract the acoustic phase shift from the Q signal, which is less susceptible to noise, we use the Q trace alone to extract features. Next, we use these features to train machine learning (ML) classifiers. Lastly, *HandLock* uses the developed ML classifier to distinguish a known user from a set of unknown users.

2.2 Signal Preprocessing

To make our system unobtrusive, *HandLock* uses sound waves with frequencies higher than 16 kHz, which are inaudible to most people and supported by Commercial-Off-The-Shelf (COTS) voice assistant devices. We transmit and record audio signals at 48 kHz. To ensure there is no interference of Doppler shifts caused by two source signals of varying frequencies, we use a frequency interval of 400 Hz. A transmitted signal arrives at the microphone from multiple paths including the structure-borne path via the body of the device, the Light-Of-Sight (LOS) propagation path via the air, and other reflection paths by surrounding objects. Let us assume the phase of the source signal ($A\cos 2\pi ft$) changes by δ due to the Doppler effect caused by a hand movement. Let $2\pi fD(t)/c$ represents the phase delay (i.e., impact of multi-path) caused by the propagation delay of $D(t)/c$, where c is the speed of sound. The recorded inaudible signal will then be $A'\cos(2\pi ft + 2\pi fD(t)/c + \delta)$. Let ϕ represents phase shift $\frac{2\pi fD(t)}{c} + \delta$, then the received signal can be simplified using the equation shown below:

$$A'\cos(2\pi ft + \phi) = I\cos 2\pi ft - Q\sin 2\pi ft \tag{1}$$

The received signal is first multiplied with the transmitted signal $\cos 2\pi ft$ and its phase-shifted version $-\sin 2\pi ft$. We then use a low pass filter (LPF) to filter out frequencies greater than 24 kHz (i.e., the maximum possible frequency at 48 kHz sampling rate) and get the corresponding desired I and Q traces.

$$I = LPF(2A'\cos(2\pi ft + \phi)cos(2\pi ft)) = A'\cos\phi \tag{2}$$

$$Q = LPF(-2A'\cos(2\pi ft + \phi)sin(2\pi ft)) = A'\sin\phi \tag{3}$$

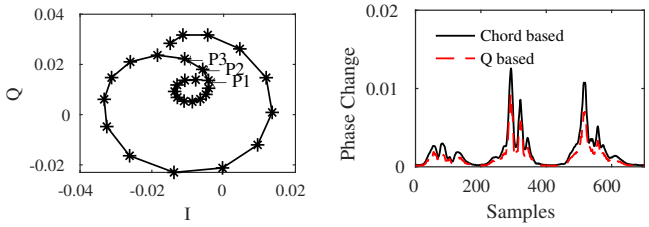

Figure 4: I/Q trace segment of a gesture sample and its phase change based on Chord-based and Q-based approach.

Limitations of prior works. Figure 4a highlights a short time series of IQ traces. Prior work [7] approximates phase (ϕ) by considering small arcs ($\widehat{P_iP_{i+1}}$) formed by two neighbouring IQ points in a circle. Specifically, the chord length ($Chord_i$) is proportional to the angle formed by an arc when it is very small. The length of chord between two neighbouring IQ points is calculated as:

$$\begin{aligned}\widehat{P_iP_{i+1}} &= \sqrt{(I_{i+1}-I_i)^2 + (Q_{i+1}-Q_i)^2} \\ &= 2R\sin(\phi_i/2) \approx R\phi_i\end{aligned} \tag{4}$$

where R is the radius of the circle IQ points form, and ϕ is the central angle of the corresponding chord. Using Taylor's series we know $\sin\phi \approx \phi$, when ϕ is small. Assuming R is constant in a short time interval, $\Delta Chord_i \approx R(\phi_{i+1} - \phi_i)$ as shown in Eq. 5, which is proportional to the phase change. We call this process Chord-based phase extraction.

$$\Delta Chord_i = \|\widehat{P_{i+1}P_{i+2}} - \widehat{P_iP_{i+1}}\| = \|R(\phi_{i+1} - \phi_i)\| \tag{5}$$

$$\begin{aligned}Q_{i+1} - Q_i &= Q_{DC} + A'\sin\phi_{i+1} - (Q_{DC} + A'\sin\phi_i) \\ &= A'(\phi_{i+1} - \phi_i - \frac{\phi_{i+1}^3}{3!} + \frac{\phi_i^3}{3!}) = A'(\phi_{i+1} - \phi_i) \\ &= A'(\frac{2\pi fD(t)}{c} + \delta_{i+1} - \frac{2\pi fD(t)}{c} - \delta_i) \\ &= A'(\delta_{i+1} - \delta_i)\end{aligned} \tag{6}$$

Our approach. As $Q = A'\sin\phi$, the phase shift can be extracted from the Q trace alone as shown in the Eq. 6. Our approach is not dependent on approximating the value of R, and at the same time removes the DC offset and reduces the impact of multipath propagation (eliminating both Q_{DC} and $\frac{2\pi fD(t)}{c}$). Figure 4b contrasts the Chord-based and our Q-based approach of approximating phase change. In our approach, phase change can, therefore, be represented by $\theta_i = \|\delta_{i+1} - \delta_i\| = \|(Q_{i+1} - Q_i)/A'\|$. As A' is constant, θ_i is proportional to $\|Q_{i+1} - Q_i\|$.

2.3 Feature Extraction

The hand speed is proportional to the relative phase change θ_i, which can be derived from Eq. 6. Therefore, we consider θ_i as the relative speed of the hand movement. As the amplitude A' is constant, we weighted two acceleration readings to calculate the average acceleration.

$$Acc_i = \frac{\theta_{i+1} - \theta_i + \frac{\theta_{i+2} - \theta_i}{2}}{2} \quad (7)$$

Feature Vector. We extract temporal and spectral features from both the speed and acceleration time series. First, we compute 8 single-valued features for both speed and acceleration including *Mean*, *Median*, *RMS*, *STD*, *MAD*, *10th percentile*, *90th percentile*, and *median frequency*. To gain more fine-grained insights into the change in speed and acceleration, we split each gesture segment into 20 equal-sized chunks and calculate *RMS*, *STD*, *MAD* and *Mean* from each chuck. To extract the spectral features like power spectral density (PSD) from speed and acceleration, we apply FFT on each time-series data, then perform max-min normalization on the power of all frequencies. Onward, we segment the PSD values into 20 small chunks and calculate the mean value from each chunk. Similarly, we compute the signal auto-correlation coefficient of speed and acceleration, and normalize the auto-correlation coefficient; we then segment the normalized coefficient into 20 chunks and compute the mean value from each chunk. In total, we extract 256 features from both the speed and acceleration time-series data.

3 EXPERIMENTAL RESULTS

As current commercial VAs are not allowed to log raw audio, we implement *HandLock* using a Seeed's ReSpeaker Core V2.0 (shown in Figure 2). The board is equipped with a six microphone array — similar to how microphones are distributed inside an Amazon Echo Dot. We wire it to an external 3W speaker AS07104PO-LW152-R. and use a 3D-printed casing to hold the microphone array and speaker. Next, we obtained the necessary IRB approval to hire 45 participants to interact with our prototype VA located inside a lab space. We randomly split the 45 participants into 39 benign users and 6 attackers. For evaluation purposes we consider five popular gestures: 'Z', 'W', 'X', '✓' and '☆'. Participants performed their hand gestures anywhere in the range of $5 \sim 30$ *cm* from our prototype VA. We collected 60 samples for each of the five gestures for each participant. We randomly select 30 samples out of the 60 samples per gesture from each user as the training set and use the remaining 30 samples as the test set. We randomly label 5 users' data as *negative class* and one out of the remaining 34 users' data as *positive class*.

We adopt Adaptive Synthetic Sampling (ADASYN) method to upsample the positive class instances. We use random forest (RF) learning to perform cross-user (i.e. 34 users) evaluation. Table 2 presents the overall performance of *HandLock* . After balancing the positive class, the false rejection rate (FRR) decreased from 5.49% to 3.49% and F-Score improved from 96.62% to 97.77%, while the average false acceptance rate (FAR) decreased from 1.14% to 0.82%. We also evaluate six attackers who individually mimic the gesture of a given victim for all five gestures. The average FAR for 'Z', 'W', 'X', '✓', and '☆' gesture is 6.2%, 3.6%, 5.14%, 3.62%, and 3.52%, respectively.

Table 2: Perf. of RF with imbalanced/balanced dataset.

Gesture	FRR	FAR	Precision	Recall	F-Score
Z	4.86/3.71	1.35/0.76	98.62/99.26	95.14/96.29	96.85/97.69
W	6.19/3.14	1.05/1.05	98.90/98.95	93.81/96.86	96.29/97.83
X	4.95/4.10	2.14/1.90	97.84/98.17	95.05/95.90	96.43/96.90
✓	6.10/3.14	0.96/0.29	99.00/99.70	93.90/96.86	96.3898.22
☆	5.33/3.33	0.19/0.10	99.80/99.91	94.67/96.67	97.17/98.20
Avg.	5.49/3.49	1.14/0.82	99.82/99.20	94.51/96.51	96.62/97.77

To evaluate the impact of ambient noise, we set up our device one meter away from a TV broadcasting news with a sound pressure of 80 dB. Three participants were asked to perform 20 samples of 'Z' gesture. We observe that the background noise is below 15 kHz. Under this setting the average true positive rate (TPR) is 95.10%. Compared to the reference TPR of 96.29% (see Table 2) we can see that the ambient background noise typically found at homes does not significantly impact our system.

4 CONCLUSION AND FUTURE WORK

In this work, we showcased a new modality of the acoustic signal-based 2-FA system for smart home voice assistants, called *HandLock*, which extracts unique movement characteristics of a user's hand gesture. The result showed that it can achieve an average TPR of 96.51% across 34 users while the FAR is 0.82%. The accuracy might decrease with a larger population. However, in a smart home with limited inhabitants, we believe the approach is still feasible. We believe this simple yet effective 2-FA approach is a first step towards helping consumers better protect *sensitive operations* carried out by VAs. In the future, we plan to extensively evaluate *HandLock* to enable user identification from gestures at longer distances.

ACKNOWLEDGMENTS

This material is based upon work supported in parts by the National Science Foundation under grant number CNS-1849997. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing Acoustics-Based User Authentication. In *MobiSys '17*.
[2] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous Authentication for Voice Assistants. In *MobiCom '17*.
[3] Xiaopeng Li, Fengyao Yan, Fei Zuo, Qiang Zeng, and Lannan Luo. 2019. Touch Well Before Use: Intuitive and Secure Authentication for IoT Devices. In *MobiCom '19*.
[4] Sapna Maheshwari. 2018. *Hey, Alexa, What Can You Hear? And What Will You Do With It?* Retrieved March 25, 2020 from https://www.nytimes.com/2018/03/31/business/media/amazon-google-privacy-digital-assistants.html
[5] Muhammad Shahzad and Shaohu Zhang. 2018. Augmenting User Identification with WiFi Based Gesture Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 134 (Sept. 2018).
[6] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible Voice Commands. In *CCS '17*.
[7] Man Zhou, Qian Wang, Jingxiao Yang, Qi Li, Peipei Jiang, Yanjiao Chen, and Zhibo Wang. 2021. Stealing Your Android Patterns via Acoustic Signals. *IEEE Transactions on Mobile Computing* 20, 4 (2021), 1656–1671.
[8] Yongpan Zou, Meng Zhao, Zimu Zhou, Jiawei Lin, Mo Li, and Kaishun Wu. 2018. BiLock: User Authentication via Dental Occlusion Biometrics. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 152 (Sept. 2018).