Continuous Authentication for Voice Assistants

Huan Feng*, Kassem Fawaz*, and Kang G. Shin
University of Michigan
Ann Arbor, Michigan 48109
{huanfeng,kmfawaz,kgshin}@umich.edu

ABSTRACT

Voice has become an increasingly popular User Interaction (UI) channel, mainly contributing to the current trend of wearables, smart vehicles, and home automation systems. Voice assistants such as Alexa, Siri, and Google Now, have become our everyday fixtures, especially when/where touch interfaces are inconvenient or even dangerous to use, such as driving or exercising. The open nature of the voice channel makes voice assistants difficult to secure, and hence exposed to various threats as demonstrated by security researchers. To defend against these threats, we present VAuth, the first system that provides *continuous* authentication for voice assistants. VAuth is designed to fit in widely-adopted wearable devices, such as eyeglasses, earphones/buds and necklaces, where it collects the body-surface vibrations of the user and matches it with the speech signal received by the voice assistant's microphone. VAuth guarantees the voice assistant to execute *only* the commands that originate from the voice of the owner. We have evaluated VAuth with 18 users and 30 voice commands and find it to achieve 97% detection accuracy and less than 0.1% false positive rate, regardless of VAuth's position on the body and the user's language, accent or mobility. VAuth successfully thwarts various practical attacks, such as replay attacks, mangled voice attacks, or impersonation attacks. It also incurs low energy and latency overheads and is compatible with most voice assistants.

1 INTRODUCTION

Siri, Cortana, Google Now, and Alexa are becoming our everyday fixtures. Through voice interactions, these and other voice assistants allow us to place phone calls, send messages, check emails, schedule appointments, navigate to destinations, control smart appliances, and perform banking services. In numerous scenarios such as cooking, exercising or driving, voice interaction is preferable to traditional touch interfaces that are inconvenient or even dangerous to use (e.g., while driving). Furthermore, a voice interface is even essential for the increasingly prevalent Internet of Things (IoT) devices that lack touch capabilities [29].

With sound being an open channel, voice as an input mechanism is inherently insecure as it is prone to replay attacks, sensitive to noise, and easy to impersonate. Recent studies have even demonstrated that it is possible to inject voice commands stealthily and remotely with mangled voice [9, 39], wireless signals [17], or through public radio stations [26] without raising the user's attention. Existing voice authentication mechanisms, such as Google's "Trusted Voice" and Nuance's "FreeSpeech", used by banks [33], fail to provide the security features for voice assistant systems. An adversary can bypass these voice-as-biometric authentication mechanisms by impersonating the user's voice (a feature already enabled by commercial software [5]) or simply launching a replay attack. Researchers have already shown that voice morphing techniques can defeat human and machine-based speaker verification systems [30]. Even Google warns against its voice authentication feature as being insecure,[1] and some security companies [6] recommend relinquishing voice interfaces all together until security issues are resolved. The implications of attacking voice-assistant systems can be frightening, ranging from information theft and financial loss [25] all the way to inflicting physical harm via unauthorized access to smart appliances and vehicles.

To defend against these threats, we propose VAuth, a novel system that provides *continuous* authentication for voice assistant systems. Designed as a wearable security token, it supports on-going authentication by introducing an additional channel that provides physical assurance. Specifically, VAuth collects the body-surface vibrations of a user and continuously matches them to the voice commands received by the voice assistant. This way, VAuth ensures that the voice assistant executes *only* the commands that originate from the voice of the owner.

In this paper, we choose to implement VAuth using an accelerometer because of its extremely low energy footprint and its ability to collect on-body vibrations (the direct product of user's speech) which the attacker cannot readily compromise or alter. This, however, does not preclude utilizing other sensors (e.g., specialized microphones), in VAuth, that can achieve a similar functionality and security guarantees as for the accelerometer. VAuth offers the following salient features.

Continuous Authentication. VAuth specifically addresses the problem of continuous authentication of a speaker to a voice-enabled device. Most authentication mechanisms, including all smartphone-specific ones such as passwords, PINs, patterns, and fingerprints, provide security by proving the user's identity *before* establishing a session. They hinge on one underlying assumption: the user retains exclusive control of the device right after the authentication. While such an assumption is natural for touch interfaces, it is unrealistic for the case of voice assistants. Voice

*Co-primary authors.

MobiCom '17, Snowbird, UT, USA

978-1-4503-4916-1/17/10…$15.00
DOI: 10.1145/3117811.3117823

[1] When a user tries to enable Trusted Voice on Nexus devices, Google explicitly warns that it is less secure than password and can be exploited by the attacker with a very similar voice.

allows access for any third party during a communication session, rendering pre-session authentication insufficient. VAuth provides ongoing speaker authentication during an entire session by ensuring that every speech sample recorded by the voice assistant originates from the speaker's throat. Thus, VAuth complements existing mechanisms of initial session authentication and speaker recognition.

Improved Security Features. VAuth overcomes a major security problem of voice biometric authentication schemes: the possibility of the voice biometric information to be leaked or compromised. A voice biometric (akin to a fingerprint) is a long term property of an individual, and compromising it (e.g., through impersonating the owner's voice) renders the voice authentication insecure. Automated speech synthesis engines can construct a model of the owner's voice (thereby impersonating him/her) using a very limited number of his/her voice samples [30]. On the other hand, when losing VAuth for any reason, the user has to just unpair the token and pair a new one.

Many of the existing biometric-based authentication approaches try to reduce time-domain signals to a set of vocal features. Regardless of how descriptive the features are of the speech signal, they still represent a projection of the signal to a reduced-dimension space. Therefore, collisions are bound to happen; two different signals can result in the same feature vector. Such attacks weaken the security guarantees provided by almost all voice-biometric approaches [35]. In contrast, VAuth depends on the instantaneous consistency of the entire signal from the accelerometer and the microphone. Thus, it can identify even minor changes/manipulations of the signal.

Usability. A user can use VAuth out-of-the-box as it does not require any user-specific training, a drastic departure from existing voice biometric mechanisms. It only depends on the instantaneous consistency between the accelerometer and microphone signals. Therefore, VAuth is immune to voice changes over time and different situations, such as sickness (a sore throat) or tiredness – a major limitation of voice biometrics. VAuth provides its security features as long as it touches the user's skin at any position on the facial, throat, and sternum[2] areas. This allows us to incorporate VAuth into wearables that people are already using on a daily basis, such as eyeglasses, Bluetooth earbuds and necklaces/lockets. Our usability survey of 952 individuals revealed that users are willing to accept the different configurations of VAuth when it comes in the forms with which they are already comfortable.

Another issue that affects the usability of VAuth is the quality of the accelerometer and voice signals. In the real world, the microphone and accelerometer do not pick up clean signals, rendering their matching non-trivial. VAuth is equipped with a matching algorithm that can handle the noise and other artifacts in both the accelerometer and voice signals without sacrificing its security features.

We have built a prototype of VAuth using a commodity accelerometer and an off-the-shelf Bluetooth transmitter. Our implementation is integrated into the Google Now system in Android, and could easily extend to other voice-based platforms such as Cortana, Siri, or even phone banking services. VAuth can thus be utilized by individuals in enterprises and organizations with high-security requirements such as financial institutions. To demonstrate the effectiveness of VAuth, we recruited 18 participants and asked each of them to issue 30 different voice commands using VAuth. We repeated the experiments for three wearable scenarios: eyeglasses, earbuds and necklace. We found that VAuth:

- delivers results with 97% detection accuracy and close to 0.1% false positives. This indicates most of the commands are correctly authenticated from the first trial and VAuth only matches the command that originates from the owner;
- works out-of-the-box regardless of variation in accents, mobility (still vs. jogging), or even languages (Arabic, Chinese, English, Korean, Persian);
- effectively thwarts mangled voice attacks and blocks unauthenticated voice commands replayed by an attacker or impersonated by other users; and
- incurs low latency (an average of 300ms) and energy overhead (requiring re-charging only once a week).

The rest of the paper is organized as follows. Sections 2 and 3 discuss the related work and background. Section 4 states the system and threat models and Section 5 details the design and implementation of VAuth. We discuss our matching algorithm in Section 6, and conduct a phonetic-level analysis on the matching algorithm in Section 7. We evaluate VAuth's effectiveness in Section 8. Section 9 discusses different aspects of VAuth, and finally, the paper concludes with Section 10.

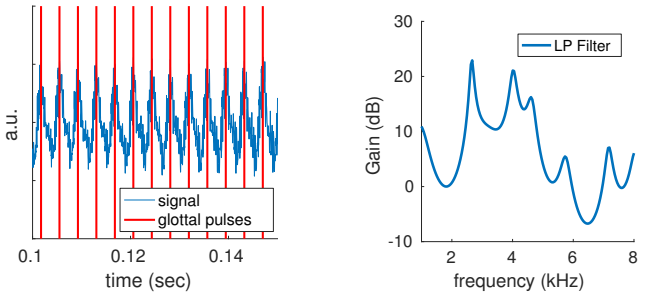
2 RELATED WORK

Smartphone Voice Assistants. Many researchers have studied the security issues of smartphone voice assistants [13, 17, 36, 39]. They have also demonstrated the possibility of injecting commands into voice assistants with electromagnetic signals [17] or with a mangled voice that is incomprehensible to humans [39]. These practical attack scenarios motivate us to build an authentication scheme for voice assistants. Petracca *et al.* [36] proposed a generic protection scheme for audio channels by tracking suspicious information flows. This solution prompts the user and requires manual review for *each* potential voice command. It thus suffers from the habituation and satisficing drawbacks since it interrupts the users from their primary tasks [14].

Voice Authentication. Most voice authentication schemes involve training on the user's voice samples and building a voice biometric [4, 10, 12, 18]. The biometric may depend on the user's vocal features or cultural backgrounds and requires rigorous training to perform well. There is no theoretical guarantee that they provide good security in general. Approaches in this category project the signal to a reduced-dimension space and collisions are thus inherent. In fact, most companies adopt these mechanisms for the usability benefits and claim they are not as secure as passwords or patterns [31]. Moreover, for the particular case of voice assistants, they all are vulnerable to simple replay attacks.

Mobile Sensing. Researchers have studied the potential applications of accelerometers for human behavior analysis [3, 24, 34, 42]. It has been shown possible to infer keyboard strokes [3], smartphone touch inputs [42] or passwords [3, 34] from acceleration

[2]The sternum is the bone that connects the rib cage; it vibrates as a result of the speech.



(a) Source (b) Filter

Figure 1: The source–filter model of human speech production using the vowel {i:} as an example.

information. There are also applications utilizing the correlation between sound and vibrations [20, 28] for health monitoring purposes. Doctors can thus detect voice disorder without actually collecting the user's daily conversations. These studies are very different from ours which focuses on *continuous* voice assistant security.

3 BACKGROUND

We introduce some basic concepts and terminology regarding the generation and processing of human speech, which will be referenced throughout the paper.

3.1 Human Speech Model

The production of human speech is commonly modeled as the combined effect of two separate processes [16]: a voice source (vibration of vocal folds) that generates the original signal and a filter (determined by the resonant properties of vocal tract including the influence of tongue and lips) that further modulates the signal. The output is a shaped spectrum with certain energy peaks, which together map to a specific phoneme (see Fig. 1(b) for the vowel {i:} – the vowel in the word "see"). This process is widely used and referred to as the *source-filter* model.

Fig. 1(a) shows an example of a female speaker pronouncing the vowel {i:}. The time separating each pair of peaks is the length of each glottal pulse (cycle). It also refers to the *instantaneous fundamental frequency* (f_0) variation while the user is speaking, which is the pitch of speaker's voice. The value of f_0 varies between 80 to 333Hz for a human speaker. The glottal cycle length (being the inverse of the fundamental frequency) varies 0.003sec and 0.0125sec. As the human speaker pronounces different phonemes in a particular word, the pitch changes accordingly, which becomes an important feature of speaker recognition. We utilize the fundamental frequency (f_0) as a reference to filter signals that fall outside of the human speech range.

3.2 Speech Recognition and MFCC

The most widely used features for speech recognition tasks are Mel-frequency cepstral coefficients (MFCC) [21], which models the way humans perceive sounds. In particular, these features are computed on short-term windows when the signal is assumed to be stationary. To compute the MFCCs, the speech recognition system computes the short-term Fourier transform of the signal, then scales the frequency axis to the non-linear Mel scale (a set of Mel bands). Then, the Discrete Cosine Transform (DCT) is computed on the log of the power spectrum of each Mel band. This technique works well in speech recognition because it tracks the invariant feature of human speech across different users. However, it also opens the door to potential attacks: by generating mangled voice segments with the same MFCC feature, an attacker can trick the voice assistant into executing specific voice commands without drawing any attention from the user.

4 SYSTEM AND THREAT MODELS

4.1 System Model

VAuth consists of two components. The first is wearable, housing an accelerometer that touches the user's skin at any position on the facial, throat, and sternum areas. This allows us to incorporate it into wearables that people are already using on a daily basis, such as eyeglasses, Bluetooth earbuds and necklaces/lockets. It constitutes a specialized "microphone" that only registers the user's voice via his body, providing a physical security guarantee. The second component is an extended voice assistant that issues voice commands after correlating and verifying both the accelerometer signal from the wearable device and the microphone signal collected by the assistant. VAuth is not only compatible with smartphone voice assistants such as Siri and Google Now, but also applies to voice systems in other domains such as Amazon Alexa and phone-based authentication systems used by banks.

We assume the communications between the two components are encrypted. Attacks to this communication channel are orthogonal to our work. We also assume the wearable device serves as a secure token that the user will not share with others. The latter is known as *security by possession*, which is widely adopted in the security field in the form of authentication rings [40], wristbands [23], or RSA SecurID. Thus, the problem of authenticating the wearable token to the user is orthogonal to VAuth and has been addressed elsewhere [11]. Instead, we focus on the problem of authenticating voice commands, assuming the existence of a trusted wearable device.

4.2 Threat Model

We consider an attacker who is interested in stealing private information or conducting unauthorized operations by exploiting the voice assistant of the target user. Typically, the attacker tries to hijack the voice assistant of the target user and deceive it into executing mal-intended voice commands, such as sending text messages to premium phone numbers or conducting bank transactions. The results can be much more serious considering voice is becoming the most promising UI interface for controlling smart home appliances and vehicles. The adversary mounts the attack by interfering with the audio channel. This does not assume the attacker has to be physically at the same location as the target. It can utilize equipment that can generate a sound on its behalf, such as radio channels or high-gain speakers. Specifically, we consider the following three categories of attack scenarios.

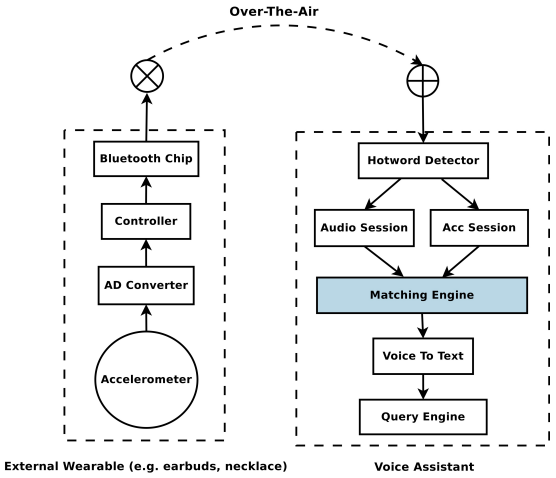


Figure 2: The high-level design of VAuth, consisting of the wearable and the voice assistant extension.

Scenario A – Stealthy Attack. The attacker attempts to inject either inaudible or incomprehensible voice commands through wireless signals [17] or mangled voice commands [9, 39]. This attack is stealthy in the sense that the victim may not even be aware of the on-going threat. It is also preferable to the attacker when the victim has physical control or within proximity of the voice assistant.

Scenario B – Biometric-override Attack. The attacker attempts to inject voice commands [35] by replaying a previously recorded clip of the victim's voice, or by impersonating the victim's voice. This attack can have a very low technical barrier: we found that by simply mimicking the victim's voice, an attacker can bypass the Trusted Voice feature of Google Now within five trials, even when the attacker and the victim are of different genders.

Scenario C – Acoustic Injection Attack. The attacker can be more advanced, trying to generate a voice that has a direct effect on the accelerometer [38]. The intention is to override VAuth's verification channel with high energy vibrations. For example, the attacker can play very loud music which contains embedded patterns of voice commands.

5 VAUTH

We present the high-level design of VAuth, describe our prototype implementation with Google Now, and elaborate on its usability aspects.

5.1 High-Level Overview

VAuth consists of two components: (1) a wearable component, responsible for collecting and uploading the accelerometer data, and (2) a voice assistant extension, responsible for authenticating and launching the voice commands. We chose to employ an accelerometer instead of an additional microphone because the accelerometer does not register voice (vibrations) through the air, thus providing a better security guarantee. The first component easily incorporates into existing wearable products, such as earbuds/earphones/headsets, eyeglasses, or necklaces/lockets.

When a user triggers the voice assistant, for example, by saying "OK, Google" or "Hey, Siri", our voice assistant extension collects accelerometer data from the wearable component, correlates it with signals collected from microphone and issues the command only when there is a match. It is worth noting that the wearable component of VAuth stays in an idle mode (idle connection and no accelerometer sampling) and only wakes up when it receives a trigger from the voice assistant extension. After the command finishes, the wearable component goes back to its idle mode. This helps reduce the energy consumption of VAuth's wearable component by reducing its duty cycle.

Fig. 2 depicts the information flows in VAuth. To reduce the processing burden on the user's device, the matching does not take place on the device (that runs the voice assistant), but rather at the server side. The communication between the wearable component and the voice assistant takes place over Bluetooth BR/EDR [7]. Bluetooth Classic is an attractive choice as a communication channel, since it has a relatively high data rate (up to 2Mbps), is energy-efficient, and enables secure communication through its pairing procedure. Upon losing the wearable component of VAuth, the user has to just unpair it with the voice assistant. This unpairing prevents an attacker from misusing a stolen wearable component to access the user's voice assistants.

5.2 Prototype

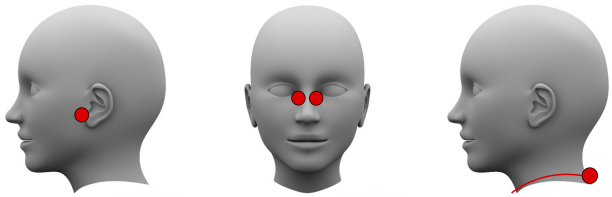
To build the wearable component, we use a Knowles BU-27135 miniature accelerometer with the dimension of only 7.92×5.59×2.28mm so that it can easily fit in any wearable design. The accelerometer uses only the z-axis and has an *analog* bandwidth of 11kHz, enough to capture the bandwidth of a speech signal, as opposed to the accelerometers available in commercial wearables with typical bandwidth of 200Hz. We utilize an external Bluetooth transmitter that provides Analog-to-Digital Conversion (ADC) and Bluetooth transmission capabilities to the voice assistant extension. To reduce energy consumption, VAuth starts streaming the accelerometer signal only upon request from the voice assistant. Our prototype communicates the microphone and accelerometer signals to a Matlab-based server which performs the matching and returns the result to the voice assistant. Fig. 3 depicts our wireless prototype standalone and attached to a pair of eyeglasses.

Our system is integrated with the Google Now voice assistant to enable voice command authentication. VAuth starts execution immediately after the start of a voice session (right after "OK Google" is recognized). It blocks the voice assistant's command execution after the voice session ends until the matching result becomes available. If the matching fails, VAuth kills the voice session. To achieve its functionality, VAuth intercepts both the HotwordDetector and the QueryEngine to establish the required control flow.

We implement our voice assistant extension as a standalone user-level service. It is responsible for retrieving the accelerometer signals from the wearable device and sending both accelerometer and microphone to our Matlab-based server for analysis. The user-level service provides two RPC methods, start and end, which are triggered by the events generated when the hotword "OK Google"

(a) Wireless (b) Eyeglasses

Figure 3: Our prototype of VAuth, featuring the accelerometer chip and Bluetooth transmitter, (a) compared to US quarter coin and (b) attached to a pair of eyeglasses belonging to one of the authors.



(a) Earbuds (b) Eyeglasses (c) Necklace

Figure 4: The wearable scenarios supported by VAuth.

is detected, and when the query (command) gets executed, respectively. The first event can be observed by filtering the Android system logs, and we intercept the second by overriding the Android IPC mechanisms, by filtering the Intents sent by Google Now.

Note that the above modifications and interceptions (in the Android framework) are necessary only because Google Now is closed source. The incorporation of VAuth is straightforward in the cases when developers try to build/extend their voice assistant.

5.3 Usability

VAuth requires the user to wear a security-assisting device. Since security has always been a secondary concern for users [41], we design VAuth in a way such that it can be easily embedded into existing wearable products that the users are already comfortable with in their daily lives. Our prototype supports three widely-adopted wearable scenarios: earbuds/earphones/headsets, eyeglasses, and necklace/lockets. Fig. 4 shows the positions of the accelerometer in each scenario. We select these areas because they have consistent contact with the user's body. While VAuth performs well on all facial areas, shoulders and the sternal surface, we only focus on the three positions shown in Fig. 4 since they conform with widely-adopted wearables.

We have conducted a survey to evaluate the users' acceptance of the different configurations of VAuth with 952 participants using Amazon Mechanical Turk. We restricted the respondent pool to those from the US with previous experience with voice assistants;

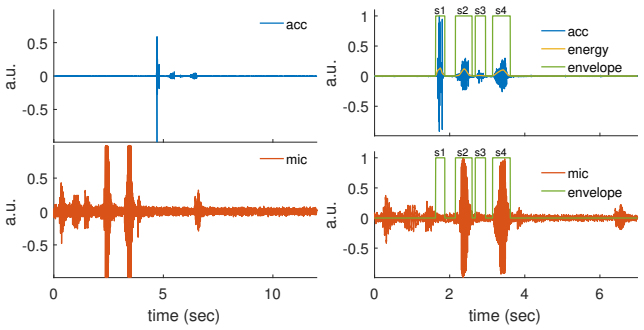
Figure 5: A breakdown of respondents' wearability preference by security concern and daily wearables. *Dangerous* and *Safe* refer to participants' attitudes towards the attacks to voice assistants after they've been informed; the *Dangerous* category is further split according to the wearables that people are already wearing on a daily basis; *Yes* and *No* refer to whether participants are willing to use VAuth in at least one of three settings we provided.

58% of them reported using a voice assistant at least once a week. We follow the USE questionnaire methodology [22] to measure the usability aspects of VAuth. We use a 7-point Likert scale (ranging from Strongly Disagree to Strongly Agree) to assess the user's satisfaction with a certain aspect or configuration of VAuth. We pose the questions in the form of how much the respondent agrees with a certain statement, such as: *I am willing to wear a necklace that contains the voice assistant securing technology.* Below, we report a favorable result as the portion of respondents who answered a question with a score higher than 4 (5,6,7) on the 7-point scale. Next to each result, we report the portion of those surveyed, between brackets, who answered the question with a score higher than 5 (6 or 7).

We first asked the respondents about their opinion regarding the security of voice assistants. Initially, 86% (63%) of the respondents indicate that they think the voice assistants are secure. After being primed about the security risks by iterating the attacks presented in Section 4, the respondents' perceptions shifted considerably. 71% (51%) of the respondents indicate that attacks to voice assistants are dangerous, and 75%(52%) specified that they would take steps to mitigate the threats. This suggests that serious efforts remain to be made to raise the public awareness of the security concerns of voice assistants.

Then, we study the perception of using VAuth from individuals who are already aware of the security problems of voice assistants. We ask the participants about their preferences for wearing VAuth in any of the three configurations of Fig. 4. We have the following takeaways from the analysis of survey responses.

- 70%(47%) of the participants are willing to wear at least one of VAuth's configurations to provide security protection.



(a) Raw input signals (b) Energy envelopes

Figure 6: Pre-processing stage of VAuth's matching.

Specifically, 48% (29%) of the respondents favored the earbuds/earphone/headset option, 38% (23%) favored the eyeglasses option and 35% (19%) favored the necklace/locket option.

- As expected, the findings fit the respondents' wearables in their daily lives. 71% of the respondents who wear earbuds on a daily basis favored that option for VAuth, 60% for eyeglasses and 63% for the necklace option.
- There is no discrepancy in the wearable options among both genders. The gender distribution of each wearable option followed the same gender distribution of the whole respondent set.
- 75% of the users are willing to pay $10 more for a wearable equipped with this technology while more than half are willing to pay $25 more.

Fig. 5 presents a breakdown of the major findings from our usability survey. These results demonstrate that users are willing to accept the different configurations of VAuth, when they are well-educated about the privacy/security threats and when VAuth comes in the form with which they are already comfortable.

6 MATCHING ALGORITHM

To enable the useful scenarios we described in the previous section, the matching algorithm of VAuth (highlighted in Fig. 2) needs to be (1) resilient to the different placements of the accelerometer, while being (2) sensitive enough to identify even minor discrepancies of potentially modified microphone signals.

The matching algorithm of VAuth takes as input the speech and vibration signals along with their corresponding sampling frequencies. It outputs a decision value indicating whether there is a match between the two signals as well as a "cleaned" speech signal in case of a match. VAuth performs the matching in three stages: *pre-processing*, *speech segments analysis*, and *matching decision*. In what follows, we elaborate on VAuth's matching algorithm using a running example of a male speaker recording the two words: "cup" and "luck" with a short pause between them. The speech signal is sampled by an accelerometer from the lowest point on the sternum at 64kHz and recorded from a built-in laptop (HP workstation) microphone at a sampling frequency of 44.1kHz, 50cm away from the speaker.

6.1 Pre-processing

First, VAuth applies a highpass filter which removes all the artifacts of the low-frequency movement to the accelerometer signal (such as walking or breathing) and negates the effect of other environmental vibrations (such as the user being on a subway platform). We use 100Hz as a cutoff threshold because humans cannot generate more than 100 mechanical movements per second. VAuth then re-samples both accelerometer and microphone signals to the same sampling rate while applying a low-pass filter to prevent aliasing.

Second, VAuth normalizes the magnitude of both signals to have a maximum magnitude of unity, which necessitates removal of the spikes in the signals. Otherwise, the lower-energy components referring to the actual speech will not be recovered. The matching algorithm computes a running average of the signal's energy and enforces a cut-off threshold, keeping only the signals with energy level within the moving average plus six standard deviation levels.

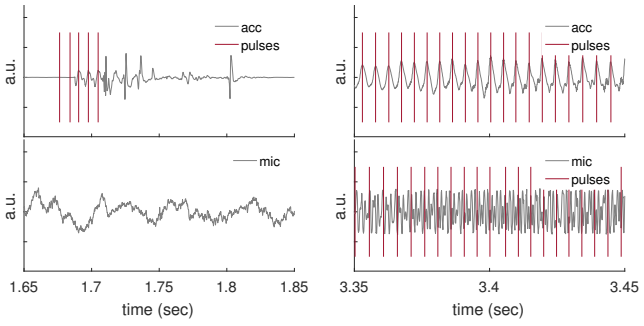
After normalizing the signal magnitude, as shown in the top plot of Fig. 6(b), VAuth aligns both signals by finding the time shift that results in the maximum cross correlation of both signals. Note that VAuth does not utilize more sophisticated alignment algorithms such as Dynamic Time Warping (DTW), since they remove timing information critical to the signal's pitch and require a higher processing load. Fig. 6(b) shows both accelerometer and microphone signals aligned and normalized.

The next pre-processing step includes identification of the energy envelope of the accelerometer signal. VAuth identifies the parts of the signal that have a significant signal-to-noise ratio (SNR). These are the "bumps" of the signal's energy as shown in the top plot of Fig. 6(b). This results in four energy segments of the accelerometer signal of Fig. 6(b). The thresholds for energy detection depend on the average noise level (due to ADC chip's sampling and quantization) when the user is silent.

Finally, VAuth applies the accelerometer envelope to the microphone signal so that it removes all parts from the microphone signal that did not result from body vibrations, as shown in the bottom plot of Fig. 6(b). This is the first real step towards providing the security guarantees. In most cases, it avoids attacks on voice assistant systems when the user is not actively speaking. Inadvertently, it improves the accuracy of the voice recognition by removing background noise and sounds from the speech signals that could not have been generated by the user.

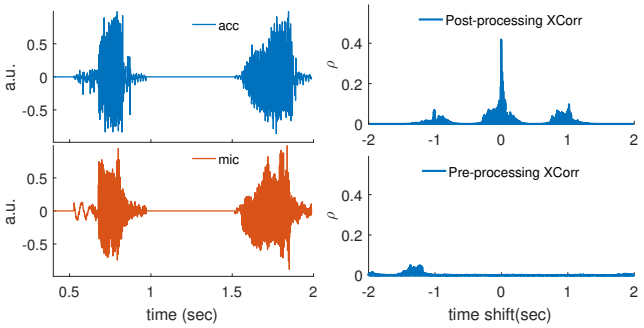
6.2 Per-Segment Analysis

Once it identifies high-energy segments of the accelerometer signal, VAuth starts a segment-by-segment matching. For each segment, VAuth normalizes the signal magnitude to unity to remove the effect of other segments, such as the effect of the segment *s1* in Fig. 6(b). VAuth then applies the approach of Boersma [8] to extract the glottal cycles from each segment. The approach relies on the identification of periodic patterns in the signal as the local maxima of the auto-correlation function of the signal. Thus, each segment is associated with a series of glottal pulses as shown in Fig. 7. VAuth uses information about the segment and the corresponding glottal pulses to filter out the segments that do not correspond to human speech. Specifically, we keep only those segments with the average



(a) Non-matching segment *s1* (b) matching segment *s4*

Figure 7: Per-segment analysis stage of VAuth.



(a) Output signals (b) Cross correlation

Figure 8: Matching decision stage of VAuth's matching.

fundamental frequency falling into the human speech range ([80Hz, 333 Hz]).

Also, if the average relative distance between glottal pulse sequence between the accelerometer and microphone segments is higher than 25%, then VAuth removes the segment from both signals. This refers to the case of interfered speech (e.g., attacker trying to inject speech); the instantaneous pitch variations should be similar between the accelerometer and microphone [27] in the absence of external interference. For example, it is evident that the pitch information is very different between the accelerometer and microphone of Fig. 7(a).

After performing all the above filtering steps, VAuth does a final verification step by running a normalized cross correlation between the accelerometer and microphone segments. If the maximum correlation coefficient falls inside the range [-0.25,0.25], then the segments are discarded. We use this range as a conservative way of specifying that the segments do not match (correlation coefficient close to zero). The correlation is a costlier operation but is a known metric for signal similarity that takes into consideration all the information of the time-domain signals. For example, the segment "s4" depicted in Fig. 7(b) shows matching pitch information and a maximum cross-correlation coefficient of 0.52.

6.3 Matching Decision

After the segment-based analysis finishes, only the "surviving" segments comprise the final accelerometer and microphone signals. In Fig. 8(a), only the segments "s2" and "s4" correspond to matching speech components. It is evident from the bottom plot that the microphone signal has two significant components referring to each word.

The final step is to produce a matching decision. VAuth measures the similarity between the two signals by using the normalized cross-correlation, as shown in the top plot of Fig. 8(b). VAuth cannot just perform the cross-correlation on the input signals before cleaning. Before cleaning the signal, the cross-correlation results do not have any real indication of signal similarity. Consider the lower plot of Fig. 8(b), which corresponds to the cross-correlation performed on the original input signals of Fig. 6(a). As evident from the plot, the cross-correlation shows absolutely no similarity between the two signals, even though they describe the same speech sample.

Instead of manually constructing rules that map the cross-correlation vector to a *matching* or *non-matching* decision, we opted to utilize a machine learning-based classifier to increase the accuracy of VAuth's matching. Below, we elaborate on the three components of VAuth's classifier: the feature set, the machine learning algorithm and the training set.

Feature Set. The feature vector comprises the normalized cross-correlation values ($h(t)$) of the final accelerometer and microphone signals. However, we need to ensure that the structure of the feature vector is uniform across all matching tasks. To populate the feature vector, we identify the maximum value of $h(t)$, and then uniformly sample 500 points to the left and another 500 to the right of the maximum. We end up with a feature vector containing 1001 values, centered at the maximum value of the normalized cross-correlation.

Classifier. We opted to use SVM as the classifier thanks to its ability to deduce linear relations between the cross-correlation values that define the feature vector. We utilize Weka [15] to train an SVM using the Sequential Minimal Optimization (SMO) algorithm [37]. The SMO algorithm uses a logistic calibrator with neither standardization nor normalization to train the SVM. The SVM utilizes a polynomial kernel with the degree equal to 1. We use the trained model in our prototype to perform the online classification.

Training Set. We recorded (more on that in Section 7) all 44 English phonemes (24 vowels and 20 consonants) from one of the authors at the lower sternum position using both the accelerometer and microphone. Hence, we have 44 accelerometer and microphone pair of recordings corresponding for each English phoneme. To generate the training set, we ran VAuth's matching over all 44 × 44 accelerometer and microphone recordings to generate 1936 initial feature vectors and labeled them accordingly.

Here, it is critical to specify that VAuth's classifier is trained *offline*, only *once* and *only* using a single training set. The classifier is thus agnostic of the user, position on the body and language. In our user study and rest of the evaluation of Section 8, this (same) classifier is used to perform all the matching. To use VAuth, the user *need not* perform any initial training.

Table 1: The IPA chart of English phonetics.

Vowel	Examples	Consonants	Examples
ʌ	CUP, LUCK	b	BAD, LAB
ɑː	ARM, FATHER	d	DID, LADY
æ	CAT, BLACK	f	FIND, IF
e	MET, BED	g	GIVE, FLAG
ə	AWAY, CINEMA	h	HOW, HELLO
ɜːr	TURN, LEARN	j	YES, YELLOW
ɪ	HIT, SITTING	k	CAT, BACK
iː	SEE, HEAT	l	LEG, LITTLE
ɒ	HOT, ROCK	m	MAN, LEMON
ɔː	CALL, FOUR	n	NO, TEN
ʊ	PUT, COULD	ŋ	SING, FINGER
uː	BLUE, FOOD	p	PET, MAP
ɑɪ	FIVE, EYE	r	RED, TRY
ɑʊ	NOW, OUT	s	SUN, MISS
eɪ	SAY, EIGHT	ʃ	SHE, CRASH
oʊ	GO, HOME	t	TEA, GETTING
ɔɪ	BOY, JOIN	tʃ	CHECK, CHURCH
eər	WHERE, AIR	θ	THINK, BOTH
ɪər	NEAR, HERE	ð	THIS, MOTHER
ʊər	PURE, TOURIST	v	VOICE, FIVE
-	-	w	WET, WINDOW
-	-	z	ZOO, LAZY
-	-	ʒ	PLEASURE, VISION
-	-	dʒ	JUST, LARGE

After computing the matching result, VAuth passes the final (cleaned and normalized) microphone signal to the voice assistant system to execute the speech recognition and other functionality.

7 PHONETIC-LEVEL ANALYSIS

We evaluate the effectiveness of our matching algorithm on phonetic-level matchings/authentications. The International Phonetic Alphabet (IPA) standardizes the representation of sounds of oral languages based on the Latin alphabet. While the number of words in a language, and therefore the sentences, can be uncountable, the number of phonemes in the English language are limited to 44 vowels and consonants. By definition, any English word or sentence, as spoken by a human, is necessarily a combination of those phonemes [19]. Our phonetic-level evaluation represents a baseline of VAuth's operation.

We study if VAuth can correctly match the English phoneme between the accelerometer and microphone (true positives), and whether it mistakenly matches phoneme samples from accelerometer to other phoneme samples from the microphone (false positives). We recruited two speakers, a male and a female, to record the 44 examples listed in Table 1. Each example comprises two words, separated by a brief pause, both representing a particular phoneme. We asked the speaker to say both words, not just the phoneme, as it is easier for the speaker to pronounce the phoneme in the context of a word. Both speakers were wearing VAuth, with the accelerometer taped to the sternum.

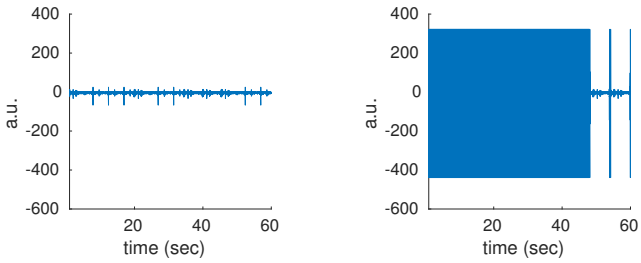
7.1 Accelerometer Energy & Recognition

Phonemes originate from a possibly different part of the chest-mouth-nasal area. In what follows, we show that each phoneme results in vibrations that the accelerometer chip of VAuth can register, but does not retain enough acoustic features to substitute a microphone speech signal for the purpose of voice recognition. This explains our rationale for employing the matching-based approach.

All phonemes register vibrations, with the minimum relative energy (accelerometer relative to microphone) coming from the ɔɪ (the pronunciation of "oy" in "boy") phoneme of the male speaker.

Table 2: The detection accuracy for the English phonemes.

microphone	accelerometer	TP (%)	FP (%)
consonants	consonants	90	0.2
consonants	vowels	-	1.0
vowels	consonants	-	0.2
vowels	vowels	100	1.7
all	all	94	0.7



(a) Low energy and white noise (b) High energy and periodic noise

Figure 9: Examples of tested noise signals.

While the accelerometer chip senses considerable energy from the chest vibrations, it cannot substitute for the microphone. To confirm this, we passed the recorded and cleaned accelerometer samples of all phonemes for both speakers to the Nuance Automatic Speaker Recognition (ASR) API [32]. The state-of-the-art ASR engine fails to identify the actual spoken words. In particular, for about half of the phonemes for both speakers, the ASR fails to return any result. For the rest of the phonemes, Nuance API returns three suggestions for each accelerometer sample. The majority of these results do not match any of the spoken words. In only three cases for consonants phonemes for both speakers, the API returns a result that matches at least one of the spoken words.

The above indicates that existing ASR engines cannot interpret the often low-fidelity accelerometer samples, but it does not indicate that ASR engines cannot be retrofitted to recognize samples with higher accuracy. This will, however, require significant changes to deploying and training these systems. On the other hand, VAuth is an entirely client-side solution that requires no changes to the ASR engine or the voice assistant system.

7.2 Phonemes Detection Accuracy

We then evaluate the accuracy of detecting each phoneme for each speaker as well as the false positive across phonemes and speakers. In particular, we run VAuth to match each accelerometer sample (88 samples — corresponding to each phoneme and speaker) to all the collected microphone samples; each accelerometer sample must match only one microphone sample. Table 2 shows the matching results.

First, we match the consonant phonemes across the two speakers as evident in the first row. The true positive rate exceeds 90%, showing that VAuth can correctly match the vast majority of consonant phonemes. We also report the false positive rate which indicates the instances where VAuth matches an accelerometer sample to the

inappropriate microphone sample. As shown in the same figure, the false positive rate is very low. Using the Hoeffding bound, it is easy to show (we do not provide this proof because of space constraints) that the false positive rate of a sentence comprising more phonemes decays exponentially fast in the number of phonemes. As such, the false positive rate of phonemes matching is only a (loose) upper bound of the whole sentence matching. Our evaluation in Section 8 further supports this property of VAuth. We also show that this non-zero false positive rate does not constitute a viable attack vector.

Moreover, a phoneme contains speaker-independent features. VAuth overcomes these similar features to successfully distinguish the same phoneme across the two speakers. The fourth row of Table 2 shows comparable results when attempting to match the vowel phonemes for both speakers.

The second and third rows complete the picture of phoneme matching, showing the matching results of the vowel to the consonant phonemes for both speakers. Both rows do not contain true positive values as there are no phoneme matches. Finally, the fifth row shows results of matching all the accelerometer samples to all the microphone samples. The true positive rate is 93%, meaning that VAuth correctly matched 82 accelerometer samples matched to their microphone counterparts. Moreover, the false positive rate was only 0.6%.

7.3 Idle Detection Accuracy

Last but not least, we evaluate another notion of false positives: VAuth mistakenly matches external speech to a silent user. We record idle (the user not actively speaking) segments from VAuth's accelerometer and attempt to match them to the recorded phonemes of both participants. We considered two types of idle segments: the first contains no energy from speech or other movements (Fig. 9(a)), while the other contains significant abrupt motion of the accelerometer resulting in recordings with high energy spikes (similar to the spike of Fig. 6(a)). We also constructed a high energy noise signal with periodic patterns as shown in Fig. 9(b).

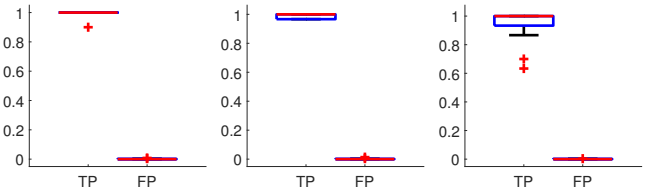
We executed VAuth over the different idle segments and microphone samples and recorded the false matching decisions. In all of the experiments, we did not observe any occurrence of a false matching of an idle accelerometer signal to any phoneme from the microphone for both speakers. As recorded phonemes are representative of all possible sounds comprising the English language, we can be confident that the false positive rate of VAuth is zero in practice for silent users.

8 EVALUATION

We now evaluate the efficacy of VAuth in identifying common voice assistant commands, under different scenarios and for different speakers. VAuth is shown to achieve a matching accuracy exceeding 95% (True Positives, TPs) regardless of its position on the body, user accents, mobility patterns, or even across different languages except for the Korean language — which we will describe later. Since the TP rate is high, the False Negative rate is very low (less than 5%), meaning that the user need not repeat the command to get a positive match in the absence of an attacker. Moreover, we elaborate on the security properties of VAuth, demonstrating its effectiveness in thwarting various attacks. Finally, we report the delay and energy consumption of our wearable prototypes.

Table 3: The list of commands to evaluate VAuth.

Command	Command
1. How old is Neil deGrasse Tyson?	**16.** Remind me to buy coffee at 7am from Starbucks
2. What does colloquial mean?	**17.** What is my schedule for tomorrow?
3. What time is it now in Tokyo?	**18.** Where's my Amazon package?
4. Search for professional photography tips	**19.** Make a note: update my router firmware
5. Show me pictures of the Leaning Tower of Pisa	**20.** Find Florence Ion's phone number
6. Do I need an umbrella today? What's the weather like?	**21.** Show me my bills due this week
7. What is the Google stock price?	**22.** Show me my last messages.
8. What's 135 divided by 7.5?	**23.** Call Jon Smith on speakerphone
9. Search Tumblr for cat pictures	**24.** Text Susie great job on that feature yesterday
10. Open greenbot.com	**25.** Where is the nearest sushi restaurant?
11. Take a picture	**26.** Show me restaurants near my hotel
12. Open Spotify	**27.** Play some music
13. Turn on Bluetooth	**28.** What's this song?
14. What's the tip for 123 dollars?	**29.** Did the Giants win today?
15. Set an alarm for 6:30 am	**30.** How do you say good night in Japanese?



(a) earbuds (b) eyeglasses (c) necklace

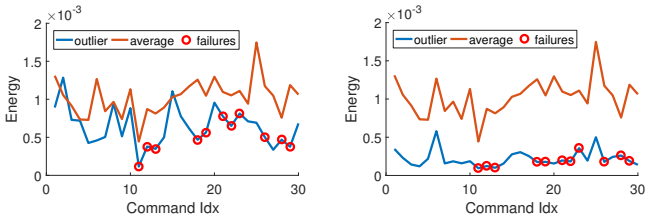
Figure 10: The detection accuracy of VAuth for the 18 users in the still position.

8.1 User Study

To support the conclusions derived from our model, we conducted a detailed user study of the VAuth prototype with 18 users and under six different scenarios. We tested how VAuth performs at three positions, each corresponding to a different form of wearable (Fig. 4) eyeglasses, earbuds, and necklace. At each position, we tested two cases, asking the user to either stand still or jog. In each scenario, we asked the participants to speak 30 phrases/commands (listed in Table 3). These phrases represent common commands issued to the "Google Now" voice assistant. In what follows, we report VAuth's detection accuracy (TPs) and false positives (FPs) when doing a pairwise matching of the commands for each participant. We collected no personally identifiable information from the individuals, and the data collection was limited to our set of commands and posed no privacy risk to the participants. As such, we obtained non-regulated status from the IRB of our institution.

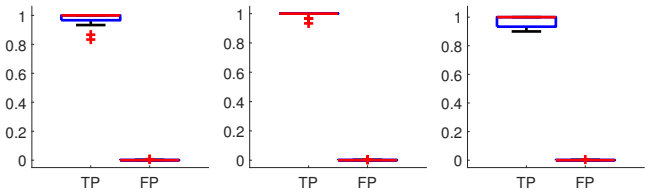
Still. VAuth delivers an overall detection accuracy rate (TPs) that is very close to 100% (97% on average). This indicates most of the commands are correctly authenticated from the first trial, and VAuth does not introduce a usability burden to the user. The false positive rate is 0.09% on average, suggesting that very few signals will leak through our filtering. Note that the non-zero false positive rate does not constitute a viable attack vector.

These false positive events occurred because after our matching algorithm removes all non-matching segments from both signals, the remaining segments of the microphone signal and accelerometer



(a) User A (b) User B

Figure 11: The energy levels of the outlier users (in Fig. 10(c)) compared to average users. The circles represent commands of the outlier users that VAuth fails to match.



(a) earbuds (b) eyeglasses (c) necklace

Figure 12: The detection accuracy of VAuth for the 18 users in the moving position.

signal accidentally match. In fact, what was eventually received by the voice assistants contains no semantic information and sounds very like random noise. The voice recognition system (Voice-to-Text) fails to pick them up as sensible voice commands. Fig. 10 shows the overall distribution of detection results for each scenario.

VAuth presents with a median detection accuracy of 100% in two wearable scenarios, eyeglasses and earbuds, but has two significant outliers in the case of the necklace. We investigated the commands that VAuth fails to recognize and found they happen when there are significant energy dips in the voice level. Fig. 11 reports the energy levels of the voice sessions for the two outlier users compared to the average across users. This suggests both participants had a lower (than average) voice when performing the experiments which did not generate enough energy to achieve the authentication.

Mobility. We asked the participants to repeat the experiments at each position while jogging. Our algorithm successfully filters the disturbances introduced by moving, breathing and VAuth's match accuracy remains unaffected (see Fig. 12). In fact, we noticed in certain cases, such as for the two outliers observed in our previous experiments, the results are even better. We studied the difference between their samples in the two scenarios and found both accelerometer and microphone received significantly higher energy in the jogging scenario even after we filtered out the signals introduced by movement. One explanation is that users are aware of the disturbance introduced by jogging and try to use louder voice to compensate. This observation is consistent across most of our participants, not just limited to the two outliers.

Table 4: The detection accuracy of VAuth for the 4 different languages.

Scenario	Language	TP (%)	FP (%)
earbuds	Arabic	100	0.1
	Chinese	100	0
	Korean	100	0
	Persian	96.7	0.1
eyeglasses	Arabic	100	0
	Chinese	96.7	0
	Korean	76.7	0
	Persian	96.7	0
necklace	Arabic	100	0
	Chinese	96.7	0
	Korean	96.7	0
	Persian	100	0

Table 5: The protections offered by VAuth.

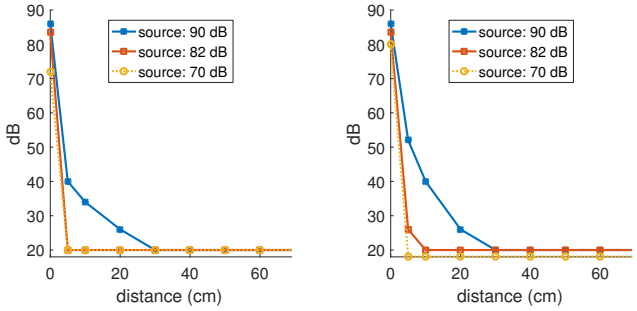
Scenario	Adversary	Example	Silent User	Speaking User
A	Stealthy	mangled voice, wireless-based	✔	✔
B	Biometric Override	replay, user impersonation	✔	✔
C	Acoustic Injection	direct communication, loud voice	distance cut-off	distance cut-off

Language. We translated the list of 30 commands into four other languages — Arabic, Chinese, Korean and Persian — and recruited four native speakers of these languages. We asked the participants to place and use VAuth at the same three positions. As shown in Table 4, VAuth performs surprisingly well, even though the VAuth prototype was trained on English phonemes (Section 6.3). VAuth delivers detection accuracy of 97% (100% in some cases), except for one case, with the user speaking Korean when wearing eyeglasses. The Korean language lacks nasal consonants and thus does not generate enough vibrations through the nasal bone [43].

8.2 Security Properties

In Section 4, we listed three types of adversaries against which we aim to protect the voice assistants. Table 5 lists the protections offered by VAuth, when the user is silent or actively speaking, for each attack scenario.

Silent User. When the user is silent, VAuth completely prevents any unauthorized access to the voice assistant. In Section 7.3, we evaluate the false positive rate of VAuth mistakenly classifying noise while the user is silent for all English phonemes. VAuth is shown to have a zero false positive rate. When the user is silent, the adversary *cannot inject* any command for the voice assistant, especially for scenarios A and B of Section 4. There might be an exception, however, for scenario C; an adversary can employ a very loud sound to induce vibrations at the accelerometer chip of VAuth. Note that, since the accelerometer only senses vibrations at the z-axis, the attacker must make the extra effort to direct the sound wave perpendicular to the accelerometer sensing surface. Next, we will show that beyond a cut-off distance of 30cm, very loud



(a) Exposed accelerometer (b) Covered accelerometer

Figure 13: The magnitude of the sensed over-the-air vibrations by the accelerometer as a function of the distance between the sound source and the accelerometer.

sounds (directed at the z-axis of the accelerometer) do not induce accelerometer vibrations. Therefore, to attack VAuth, an adversary has to play a very loud sound within less than an arm's length from the user's body — which is highly improbable.

We conduct experiments on two scenarios: the first with VAuth exposed and the second with VAuth covered with cotton clothing. Fig. 13 reports how the accelerometer chip of VAuth reacts to over-the-air sound signals of different magnitudes at different distances. In each of these scenarios, we played a white noise at three sound levels: 2x, 4x and 8x the conversation level at 70dB, 82dB and 90dB, respectively [2]. The noise is directed perpendicularly to the sensing surface of the accelerometer. Fig. 13(a) shows the recorded magnitude of the accelerometer signal as a function of the distance between the sound source and VAuth when it is exposed. As evident from the plots, there is a cut-off distance of 30cm, where VAuth's accelerometer cannot sense even the loudest of the three sound sources. Beyond the cut-off distance, the magnitude of the recorded signal is the same as that in a silent scenario. This indicates that an adversary cannot inject commands with a high sound level beyond some cut-off distance. These results are consistent with the case of VAuth covered with cotton, as shown in Fig. 13(b).

Speaking User. The adversary may try to launch an attack on the voice assistant when the user is actively speaking. As we will show, VAuth matches signals in their entirety, thus can detect even minor discrepancies of the injected voice command compared to the authentic voice session. First, we show how VAuth can successfully thwart the stealthy attacks of scenario A. Vaidya *et al.* [9, 39] presented an attack that exploits the gap between voice recognition system and human voice perception. It constructs mangled voice segments that match the MFCC features of an injected voice command. An ASR engine can recognize the command, but not the human listener. This and similar attacks rely on performing a search in the MFCC algorithm parameter space to find voice commands that satisfy the above feature.

Fig. 14 shows the evaluation flow. For each of the recorded command of the previous section, we extract the MFCCs for the full signal and use them to reconstruct the voice signal. We vary the number Mel filter bands between 15 and 30. At 15 Mel filter bands,

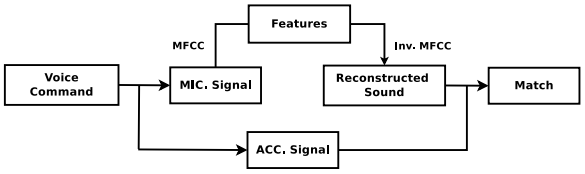


Figure 14: The flow of the mangled voice analysis.

the reconstructed voice command is similar to what is reported in existing attacks [39]. At 30 Mel filter bands, the reconstructed voice command is very close to the original; it shares the same MFCCs and is easily identifiable when played back. Finally, we execute VAuth over the reconstructed voice segment and the original accelerometer sample to test for a match.

In all cases, while the original microphone signal matches accelerometer signals near perfectly as indicated before, the reconstructed sound failed to match the accelerometer signal in 99% of the evaluated cases. Of 3,240 comparisons (2 Mel filter band lengths per command, 90 commands per user and 18 users), the reconstructed sound matched only a handful of accelerometer samples, and only in cases where we used 30 Mel filter bands. Indeed, those sound segments were very close to the original sound segment that corresponds to the matched accelerometer samples and are not stealthy at all. To constitute a practical attack, the mangled voice segment is not supposed to originate from the sound the user is speaking, let alone preserving discernible acoustic features. This demonstrates that VAuth matches the signals in their entirety, even subtle changes which are indistinguishable to human ears can result in discrepancies of our matching/authentication results.

In scenario B, an attacker also fails to overcome VAuth's protection. We indicated earlier in Section 7.2 and in this section that VAuth successfully distinguishes the phonemes and commands of the same user. This indicates 'sounds like' the user does not help the attacker much — it must inject the same command the user is saying using the same voice. Moreover, even if the user is speaking and the adversary is replaying another sound clip of the same user, VAuth can differentiate between the microphone and accelerometer samples and stop the attack.

Finally, the matching algorithm of VAuth might result in some false positives (albeit very low). Even though these false positives indicate the remaining segments after the "per-segment" stage of VAuth match, useful information in these accelerometer and microphone signals have already been filtered by our algorithm. What gets delivered to the voice assistant is practically random noise. Note that VAuth could use a more stringent classifier tuned to force the false positive rate to be 0. Such a classifier comes at the cost of usability but could be preferable in high-security situations.

8.3 Delay and Energy

We measure the delay experienced at the voice assistant and the energy consumption of the wearable component, using our prototype. As shown in Fig. 2, VAuth incurs delay only during the matching phase: when VAuth uploads the accelerometer and microphone signals to the remote service and waits for a response. According to our evaluation over the same list of 30 commands, we found that a

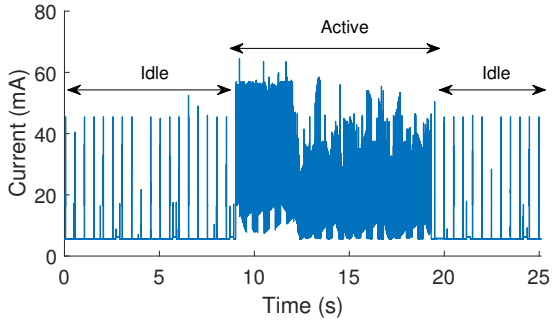


Figure 15: Current levels of the prototype in the idle and active states.

successful match takes 300–830ms, with an average of 364ms, while an unsuccessful match takes 230–760ms, with an average of 319ms. The response time increases proportionally with the length of the commands. Still, matching a command containing 30 words still takes less than 1 second. We expect the delay to decrease further with a more optimized server implementation.

When the wearable component transmits accelerometer signals, it switches between an *idle* state that keeps the connection alive and an *active* state to transmit the data. We connected our prototype to the Monsoon power monitor and recorded the current levels of the prototype in these two states when powered by a fixed voltage (4V). Fig. 15 illustrates the changes of the current levels when our prototype switches from idle to active and then back to idle. In the active state, our prototype consumes as much as 31mA, while in the idle state, it only consumes an average of 6mA. Most of the energy is used to keep the Bluetooth connection and transmit data (in the active state) — the energy consumed by the accelerometer sensor is almost negligible (3 μA).

Assuming the user always keeps the wearable open at daytime and sends 100 voice commands per day (10 seconds per voice command). Our prototype consumes 6.3mA on average. This might even be an over-estimation since 90% of the users issue voice commands at most once per day according to our survey. A typical 500mAh Li-Ion battery used by wearables (comparable to a US quarter coin) can power our prototype for around a week. 80% of the participants in our usability survey think they have no problem with recharging the wearable on a weekly basis. We conducted all the analyses on our prototype which directly utilizes off-the-shelf hardware chips without any optimization, assuming that VAuth is provided as a standalone wearable. If incorporated into an existing wearable device, VAuth will only introduce an additional energy overhead of less than 10mAh per day (to power the additional accelerometer).

9 DISCUSSION

VAuth requires the user to wear an additional device that is in continuous contact with his/her body. This requirement restricts our design options to embedding VAuth within wearables that maintain constant contact with the user's body. In this paper, we presented a proof-of-concept implementation of VAuth where we investigated three possible wearable scenarios: eyeglasses, earbuds, and necklaces. With more engineering effort, we can expand the use of VAuth to cover more wearables, such as watches and wristbands.

VAuth's energy consumption is another issue that affects its usability and practicality. In Section 8, our prototype was shown to deliver a week worth of battery life under a usage scenario of 100 commands per day. This overhead constitutes the energy consumption of the entire prototype of VAuth, most of which is spent on keeping the Bluetooth connection alive (between the matching and wearable components of VAuth). Integrating VAuth within existing Bluetooth-equipped wearables (such as a Bluetooth earbud or smartwatch) can limit its additional energy overhead to that needed to collect and communicate the vibrations from the user's body to the matching component. The latter only takes place when the user is issuing a voice command.

In its core, VAuth consists of the matching and wearable components. The wearable component collects the speech signal as it traverses the user's body (not injected through the air or another medium) and communicates them to the matching engine. In this paper, we presented a proof-of-concept implementation of the wearable component. This implementation, however, does not preclude other realizations of VAuth, as long as they meet both requirements: securely collect and communicate speech from the body to the matching component. We have used an accelerometer which captures vibrations through contact with the skin, as opposed to an on-body microphone, for instance, which captures voice through the air making it susceptible to voice injection. Even with a reasonably tuned sensitivity threshold, an attacker could always use higher power (louder voice) inject speech signals over the air for the on-body microphone. On the other hand, the human skin reflects most of the energy (95% to 99%) of the voice signal [1]; the accelerometer will only collect 1% to 5% of the over-air voice signal. We evaluated this property of the accelerometer in Section 8.2.

10 CONCLUSION

In this paper, we have proposed VAuth, a system that provides continuous authentication for voice assistants. We demonstrated that even though the accelerometer information collected from the facial/neck/chest surfaces might be weak, it contains enough information to correlate it with the data received via microphone. VAuth provides extra physical assurance for voice assistant users and is an effective measure against various attack scenarios. It avoids the pitfalls of existing voice authentication mechanisms. Our evaluation with real users under practical settings shows high accuracy and very low false positive rate, highlighting the effectiveness of VAuth. In future, we would like to explore more configurations of VAuth that will promote wider real-world deployment and adoption.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers, and our shepherd Marco Gruteser, for their constructive suggestions. They would also like to thank Timothy Trippel for his help in building VAuth's prototype and Mariam Fawaz for her comments on the earlier drafts of this manuscript. The work reported in this paper was supported in part by NSF under Grants CNS-1505785 and CNS-1646130.

REFERENCES

[1] Eugene Ackerman and Fujio Oda. 1962. *Acoustic absorption coefficients of human body surfaces.* Technical Report MRL-TDR-62-36. Pennsylvania State University, University Park.

[2] IAC Acoustics. 2017. Comparative Examples Of Noise Levels. http://www.industrialnoisecontrol.com/comparative-noise-examples.htm. (2017). Accessed: 2017-03-16.

[3] Adam J Aviv, Benjamin Sapp, Matt Blaze, and Jonathan M Smith. 2012. Practicality of accelerometer side channels on smartphones. In *Proceedings of the 28th Annual Computer Security Applications Conference*. ACM, 41–50.

[4] Mossab Baloul, Estelle Cherrier, and Christophe Rosenberger. 2012. Challenge-based speaker recognition for mobile authentication. In *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*. IEEE, 1–7.

[5] BBC News. 2016. Adobe Voco 'Photoshop-for-voice' causes concern. http://www.bbc.com/news/technology-37899902. (2016). Accessed: 2017-07-26.

[6] Yuval Ben-Itzhak. 2014. What if smart devices could be hacked with just a voice? http://now.avg.com/voice-hacking-devices/. (Sep. 2014).

[7] Bluetooth SIG. 2014. *Specification of the Bluetooth System.* Version 4.2. https://www.bluetooth.org/en-us/specification/adopted-specifications.

[8] Paul Boersma. 1993. ACCURATE SHORT-TERM ANALYSIS OF THE FUNDAMENTAL FREQUENCY AND THE HARMONICS-TO-NOISE RATIO OF A SAMPLED SOUND. *Institute of Phonetic Sciences - University of Amsterdam* 17 (1993), 97–110.

[9] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, Austin, TX, 513–530. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini

[10] Cory Cornelius, Zachary Marois, Jacob Sorber, Ron Peterson, Shrirang Mare, and David Kotz. 2014. Vocal resonance as a passive biometric. (2014).

[11] Cory Cornelius, Ronald Peterson, Joseph Skinner, Ryan Halter, and David Kotz. 2014. A Wearable System That Knows Who Wears It. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '14)*. ACM, New York, NY, USA, 55–67. DOI: http://dx.doi.org/10.1145/2594368.2594369

[12] Amitava Das, Ohil K Manyam, Makarand Tapaswi, and Veeresh Taranalli. 2008. Multilingual spoken-password based user authentication in emerging economies using cellular phone networks. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE, 5–8.

[13] Wenrui Diao, Xiangyu Liu, Zhe Zhou, and Kehuan Zhang. 2014. Your Voice Assistant is Mine: How to Abuse Speakers to Steal Information and Control Your Phone. In *Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices (SPSM '14)*. ACM, New York, NY, USA, 63–74. DOI: http://dx.doi.org/10.1145/2666620.2666623

[14] Adrienne Porter Felt, Serge Egelman, Matthew Finifter, Devdatta Akhawe, and David Wagner. 2012. How to Ask for Permission. In *Proceedings of the 7th USENIX Conference on Hot Topics in Security (HotSec'12)*. USENIX Association, Berkeley, CA, USA, 7–7. http://dl.acm.org/citation.cfm?id=2372387.2372394

[15] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. DOI: http://dx.doi.org/10.1145/1656274.1656278

[16] Haskins Laboratories. 1999. The Acoustic Theory of Speech Production: the source-filter model. http://www.haskins.yale.edu/featured/heads/mmsp/acoustic.html. (1999).

[17] Chaouki Kasmi and Jose Lopes Esteves. 2015. IEMI Threats for Information Security: Remote Command Injection on Modern Smartphones. *Electromagnetic Compatibility, IEEE Transactions on* 57, 6 (2015), 1752–1755.

[18] Max Kunz, Klaus Kasper, Herbert Reininger, Manuel Möbius, and Jonathan Ohms. 2011. Continuous Speaker Verification in Realtime.. In *BIOSIG*. 79–88.

[19] Mark Liberman. 2016. Linguistics 001 – Sound Structure of Language. http://www.ling.upenn.edu/courses/ling001/phonology.html. (2016). Accessed: 2016-05-23.

[20] Yu-An S Lien, Carolyn R Calabrese, Carolyn M Michener, Elizabeth Heller Murray, Jarrad H Van Stan, Daryush D Mehta, Robert E Hillman, J Pieter Noordzij, and Cara E Stepp. 2015. Voice Relative Fundamental Frequency Via Neck-Skin Acceleration in Individuals With Voice Disorders. *Journal of Speech, Language, and Hearing Research* 58, 5 (2015), 1482–1487.

[21] Mumtaj Begam Lindasalwa Muda and I. Elamvazuthi. 2010. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal Of Computing* 2, 3 (2010), 138–143.

[22] A. M. Lund. 2001. Measuring usability with the USE questionnaire. *Usability Interface* 8 (2001), 3–6. Issue 2.

[23] Shrirang Mare, Andrés Molina Markham, Cory Cornelius, Ronald Peterson, and David Kotz. 2014. ZEBRA: zero-effort bilateral recurring authentication. In *Security and Privacy (SP), 2014 IEEE Symposium on*. IEEE, 705–720.

[24] Philip Marquardt, Arunabh Verma, Henry Carter, and Patrick Traynor. 2011. (sp) iPhone: decoding vibrations from nearby keyboards using mobile phone accelerometers. In *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 551–562.

[25] Arnold Martin and Greenhalgh Hugo. 2016. Banking biometrics: hacking into your account is easier than you think. https://www.ft.com/content/959b64fe-9f66-11e6-891e-abe238dee8e2. (2016). Accessed: 2017-07-26.

[26] Rachel Martin. 2016. Listen Up: Your AI Assistant Goes Crazy For NPR Too. http://www.npr.org/2016/03/06/469383361/listen-up-your-ai-assistant-goes-crazy-for-npr-too. (Mar. 2016).

[27] D. D. Mehta, J. H. Van Stan, and R. E. Hillman. 2016. Relationships Between Vocal Function Measures Derived from an Acoustic Microphone and a Subglottal Neck-Surface Accelerometer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 4 (April 2016), 659–668. DOI: http://dx.doi.org/10.1109/TASLP.2016.2516647

[28] Daryush D Mehta, Matías Zañartu, Shengran W Feng, Harold A Cheyne, and Robert E Hillman. 2012. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *Biomedical Engineering, IEEE Transactions on* 59, 11 (2012), 3090–3096.

[29] Rachel Metz. 2014. Voice Recognition for the Internet of Things. https://www.technologyreview.com/s/531936/voice-recognition-for-the-internet-of-things/. (Oct. 2014).

[30] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. 2015. *All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines*. Springer International Publishing, Cham, 599–621. DOI: http://dx.doi.org/10.1007/978-3-319-24177-7_30

[31] Lisa Myers. 2004. An Exploration of Voice Biometrics. https://www.sans.org/reading-room/whitepapers/authentication/exploration-voice-biometrics-1436. (April 2004).

[32] Nuance Cloud Services. 2013. HTTP Services 1.0 Programmer's Guide. https://developer.nuance.com/public/Help/HttpInterface/HTTP_web_services_for_NCS_clients_1.0_programmer_s_guide.pdf. (Dec. 2013).

[33] Nuance Communications, Inc. 2017. Authentication via conversation. http://www.nuance.com/for-business/customer-service-solutions/voice-biometrics/freespeech/index.htm. (2017). Accessed: 2017-07-26.

[34] Emmanuel Owusu, Jun Han, Sauvik Das, Adrian Perrig, and Joy Zhang. 2012. ACCessory: password inference using accelerometers on smartphones. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*. ACM, 9.

[35] Saurabh Panjwani and Achintya Prakash. 2014. Crowdsourcing Attacks on Biometric Systems. In *Tenth Symposium on Usable Privacy and Security, SOUPS 2014, Menlo Park, CA, USA, July 9-11, 2014*. 257–269. https://www.usenix.org/conference/soups2014/proceedings/presentation/panjwani

[36] Giuseppe Petracca, Yuqiong Sun, Trent Jaeger, and Ahmad Atamli. 2015. AuDroid: Preventing Attacks on Audio Channels in Mobile Devices. In *Proceedings of the 31st Annual Computer Security Applications Conference (ACSAC 2015)*. ACM, New York, NY, USA, 181–190. DOI: http://dx.doi.org/10.1145/2818000.2818005

[37] John Platt. 1998. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, In Advances in Kernel Methods - Support Vector Learning. (January 1998). https://www.microsoft.com/en-us/research/publication/fast-training-of-support-vector-machines-using-sequential-minimal-optimization/

[38] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. WALNUT: Waging Doubt on the Integrity of MEMS Accelerometers with Acoustic Injection Attacks. In *In Proceedings of the 2nd IEEE European Symposium on Security and Privacy (EuroS&P 2017). To appear.*

[39] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. 2015. Cocaine noodles: exploiting the gap between human and machine speech recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*.

[40] Tam Vu, Akash Baid, Simon Gao, Marco Gruteser, Richard Howard, Janne Lindqvist, Predrag Spasojevic, and Jeffrey Walling. 2012. Distinguishing Users with Capacitive Touch Communication. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12)*. ACM, New York, NY, USA, 197–208. DOI: http://dx.doi.org/10.1145/2348543.2348569

[41] Ryan West. 2008. The Psychology of Security. *Commun. ACM* 51, 4 (April 2008), 34–40. DOI: http://dx.doi.org/10.1145/1330311.1330320

[42] Zhi Xu, Kun Bai, and Sencun Zhu. 2012. Taplogger: Inferring user inputs on smartphone touchscreens using on-board motion sensors. In *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks*. ACM, 113–124.

[43] Kenji Yoshida. 2008. Phonetic implementation of Korean denasalization and its variation related to prosody. *IULC Working Papers* 8, 1 (2008).