Huan Feng *Facebook Inc* **Kassem Fawaz** *University of Wisconsin-Madison* **Kang G. Shin** *University of Michigan*

Editors: Nic Lane and Xia Zhou

WEARABLE TECHNOLOGY BRINGS SECURITY TO ALEXA AND SIRI

Photo, istockphoto.com

Excerpted from "Continuous Authentication for Voice Assistants," from MobiCom '17, *Proceedings of the 23rd Annual ACM International Conference on Mobile Computing and Networking*, with permission. https://dl.acm.org/citation.cfm?id=3117823 © ACM 2017

IT companies are heavily investing in the future of voice assistants. Siri, Cortana, Google Now, Alexa, and Samsung Bixby are already part of our everyday fixtures. Through voice interactions, 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. Further, a voice interface is even more essential for the increasingly prevalent Internet of Things (IoT) devices that lack touch capabilities.

Security concerns, however, have become a major roadblock against the adoption of voice-based interactions. 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 [1, 2], ultrasound [12], wireless signals [3], or through public radio stations [4] without the user's awareness. Existing voice authentication mechanisms, such as Google's "Trusted Voice" and Nuance's "FreeSpeech" (used by banks [5]) fail to provide the security necessary for voice-assistant systems. An attacker can bypass these voice-as-biometric authentication mechanisms by impersonating the user's voice (a feature already enabled by commercial software, such as Adobe Voco) or simply by recording and replaying the user's voice. Even Google warns against its voice authentication feature as being insecure, and some security companies [6] recommend

relinquishing voice interfaces altogether until security issues are resolved. The implications of attacking voice-assistant systems can be frightening, ranging from information theft and financial loss [7] all the way to inflicting physical harm via unauthorized access to smart appliances and vehicles. There is an urgent need to address these security concerns before voice assistants become more prevalent.

VAuth: A WEARABLE SOLUTION

In this article, we describe *VAuth* – a novel wearable system that delivers continuous authentication for voice assistant systems [8]. Designed as a wearable security token, it supports ongoing authentication by introducing an additional channel that provides physical assurance. VAuth collects the body-surface vibrations of a user and continuously matches them to the voice commands received by the voice assistant. This vibration, while not enough to pass as a high-fidelity speech signal, serves as a ground truth of what the user communicated to the voice assistants. This way, VAuth ensures that the voice assistant executes only the commands that originate from the voice of the owner. We implement VAuth using an accelerometer because of its extremely low energy footprint and its ability to collect on-body vibrations (the direct product of the user's speech), which the attacker cannot readily compromise or alter. 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.

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. The first component can be very small in size and can be incorporated easily 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 the 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. Figure 1 illustrates the overall architecture of the system.

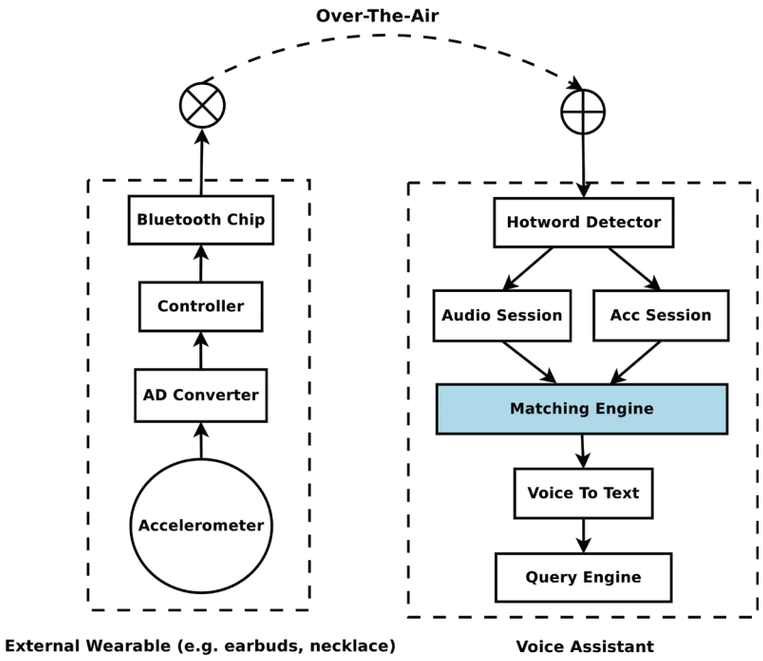

FIGURE 1. The high-level design of VAuth, consisting of the wearable and the voice assistant extension. The wearable component samples the vibration signal from the user's body and sends it to the voice assistant extension, which performs the matching with the microphone signal.

WHY VAuth?

VAuth 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 common 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 allows access for any 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.

In addition, VAuth overcomes a major shortcoming 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 [9]. 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

TABLE 1.

Scenario	Movement	TP(%) 10th percentile	FP(%) 90th percentile
Earbud	Still	100	0.26
	Jogging	91.33	0.33
Eyeglasses	Still	96.66	0.26
	Jogging	96.66	0.34
Necklace	Still	81.66	0.15
	Jogging	93.33	0.15

The detection accuracy of VAuth for the 18 users in the Still and Jogging positions for 30 voice commands. In more than 90% of the cases, the True Positive rate of VAuth is more than 82% (the outlier is the necklace position for two participants with low accelerometer signal energy). Also, in more than 90% of the cases, the False Positive rate was less than 0.34%.

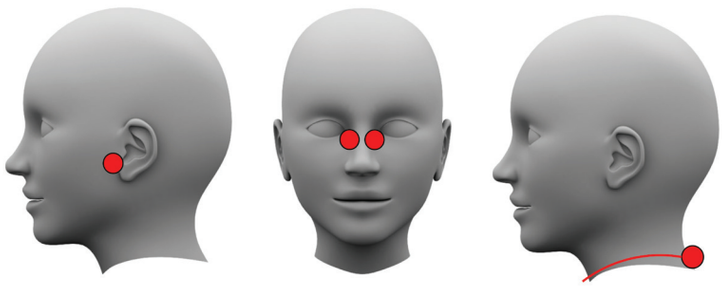

FIGURE 2. The different placement options for VAuth: as part of an earbud (left), as the nosepad of eyeglasses (middle), and on the back of a necklace (right).

TABLE 2.

Scenario	Movement	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

The detection accuracy of VAuth for four different languages. Except for the Korean language which lacks nasal consonants, VAuth performs well over the four languages under different placements.

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 [10]. 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.

MATCHING ENGINE

Sitting at VAuth's core is a matching engine which 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*.

The pre-processing stage filters the artifacts of the low-frequency user movement to the accelerometer signal (such as walking or breathing). After synchronizing both signals, it identifies the energy envelope of the accelerometer signal and overlays it on top of the speech signal. This way, only voice signals that result from the body vibrations are kept. In the segment analysis, VAuth partitions both signals into smaller speech segments. It extracts and compares the glottal cycles of each segment from both signals and keeps only the matching ones. Finally, VAuth employs a machine learning classifier to output the final decision, given the input of a cross-correlation vector between the voice and accelerometer signals. The model was only trained once on English phonemes of a test user; it was directly applicable to all other cases, even languages.

Our prototype performed well in various practical settings. 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. VAuth is shown to yield 97% detection accuracy and close to 0.1% false positives (detailed results available in Table 1). This indicates that most of the commands are correctly authenticated from the first trial and VAuth only matches the command that originates from the owner. It also works out-of-the-box regardless of variation in accents, mobility (still vs. jogging), or even languages (Arabic, Chinese, English, Korean, Persian) – detailed results available in Table 2. VAuth delivers almost perfect detection accuracy, 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 [11]. VAuth also incurs low latency (an average of 300ms) and energy overhead (requiring recharging only once a week).

Most importantly, VAuth delivers a strong security guarantee against various attacks. We tested VAuth under various strict settings and demonstrated that first, when the user is silent, the attacker cannot inject any command to the voice assistant, even when the attacker employs a very loud sound to induce vibrations at the accelerometer chip of VAuth. We also demonstrated that even when the user is actively speaking and generating vibrations (not necessarily to the voice assistant), the attacker still fails to inject voice commands

which sounds almost the same to a human listener. This is because VAuth monitors the instantaneous matching between the vibrations generated by the speaker and sound collected by the voice assistants. Any minor differences of the timing or the tune might trigger an inconsistency. In other words, the attacker needs to emulate exactly what the user is saying to bypass the checking, which invalidates any practical attack scenario. Still, our matching engine yielded a non-zero false positive rate. Even though this suggests some signal might accidentally match the vibrations of the user and leak through the matching engine, but in practice, all of the signals are random noise and do not correspond to comprehensible or meaningful words/sentences.

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 areas. This allows us to incorporate VAuth into wearables that people are already using on a daily basis. Our prototype supports three widely adopted wearable scenarios: earbuds/earphones/headsets, eyeglasses, and necklace/lockets. Figure 2 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 since they conform with widely adopted wearables.

We 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. 70% of the participants are willing to wear at least one of VAuth's configurations to provide security protection. There is no discrepancy in the wearable options among both genders. 75% of the users are willing to pay $10 more for a wearable equipped with this technology, and more than half are willing to pay $25 more.

CONCLUSION

In this article, we have described VAuth, a novel 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, and evaluation with real users under practical settings shows high accuracy and very low false positive rate. We hope the insights envisioned by VAuth can enable better security protection for voice assistants in the future. ■

Huan Feng is a research scientist at Facebook Inc., working on fraud detection and abuse prevention. He earned his PhD in Computer Science and Engineering from the University of Michigan, Ann Arbor, where he developed practical systems that protect mobile systems and users.

Kassem Fawaz is an assistant professor in the ECE department at the University of Wisconsin – Madison. He earned his PhD in Computer Science and Engineering from the University of Michigan. His research interests include the security and privacy of the interactions between users and connected systems.

Kang G. Shin is the Kevin & Nancy O'Connor Professor of Computer Science in the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor. His current research focuses on QoS-sensitive computing and networking as well as on embedded real-time and cyber-physical systems.

REFERENCES

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

[2] 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).

[3] Chaouki Kasmi and Jose Lopes Esteves. 2015. "IEMI threats for information security: Remote command injection on modern smartphones." *IEEE Transactions on Electromagnetic Compatibility*, 57, 6 (2015), 1752–1755.

[4] 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-goescrazy-for-npr-too. (Mar. 2016).

[5] Nuance Communications, Inc. 2018. FreeSpeech biometric voice authentication system. https://www.nuance.com/omni-channel-customer-engagement/security/multi-modal-biometrics/freespeech.html (2018).

[6] Yuval Ben-Itzhak. 2014. "What if smart devices could be hacked with just a voice?" https://web.archive.org/web/20141017174847/http://now.avg.com/voice-hacking-devices (2014).

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

[8] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. "Continuous authentication for voice assistants." *In Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking* (MobiCom '17) https://doi.org/10.1145/3117811.3117823

[9] 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. http://dx.doi.org/10.1007/978-3-319-24177-7_30

[10] Saurabh Panjwani and Achintya Prakash. 2014. "Crowdsourcing attacks on biometric systems." In *Proceedings of 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

[11] Timothy Trippel, Or Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. "WALNUT: Waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks." In *Proceedings of the 2nd IEEE European Symposium on Security and Privacy* (EuroS&P 2017).

[12] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. "Dolphin attack: Inaudible voice commands." In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,* pages 103–117. ACM, 2017.