VELODY: Nonlinear Vibration Challenge-Response for Resilient User Authentication

Jingjie Li
University of Wisconsin–Madison
jingjie.li@wisc.edu

Kassem Fawaz
University of Wisconsin–Madison
kfawaz@wisc.edu

Younghyun Kim
University of Wisconsin–Madison
younghyun.kim@wisc.edu

ABSTRACT

Biometrics have been widely adopted for enhancing user authentication, benefiting usability by exploiting pervasive and collectible unique characteristics from physiological or behavioral traits of human. However, successful attacks on "static" biometrics such as fingerprints have been reported where an adversary acquires users' biometrics stealthily and compromises non-resilient biometrics.

To mitigate the vulnerabilities of static biometrics, we leverage the unique and nonlinear hand-surface vibration response and design a system called VELODY to defend against various attacks including replay and synthesis. The VELODY system relies on two major properties in hand-surface vibration responses: uniqueness, contributed by physiological characteristics of human hands, and nonlinearity, whose complexity prevents attackers from predicting the response to an unseen challenge. VELODY employs a challenge-response protocol. By changing the vibration challenge, the system elicits input-dependent nonlinear "symptoms" and unique spectrotemporal features in the vibration response, stopping both replay and synthesis attacks. Also, a large number of disposable challenge-response pairs can be collected during enrollment passively for daily authentication sessions.

We build a prototype of VELODY with an off-the-shelf vibration speaker and accelerometers to verify its usability and security through a comprehensive user experiment. Our results show that VELODY demonstrates both strong security and long-term consistency with a low equal error rate (EER) of 5.8% against impersonation attack while correctly rejecting all other attacks including replay and synthesis attacks using a very short vibration challenge.

CCS CONCEPTS

• **Security and privacy → Authentication**.

KEYWORDS

Authentication; nonlinear vibration; challenge-response; biometric

ACM Reference Format:
Jingjie Li, Kassem Fawaz, and Younghyun Kim. 2019. VELODY: Nonlinear Vibration Challenge-Response for Resilient User Authentication. In *2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3319535.3354242

CCS '19, November 11–15, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6747-9/19/11…$15.00
https://doi.org/10.1145/3319535.3354242

1 INTRODUCTION

The mass proliferation of "smart" devices has created unprecedented security and privacy concerns to their users. One of the significant security concerns comes from unauthorized entities accessing and controlling user devices. Stronger access control goes a long way towards alleviating security and privacy threats to users and their devices. User authentication, where a user has to prove their identity to a system, is one core mechanism to achieve adequate access control.

Biometric user authentication, which relies on the unique physiological or behavioral traits of the user to verify their identity, has been touted as the solution that meets both security and usability goals. Thanks to its low cognitive burden, it is more attractive to the users who wish to authenticate themselves to their devices without having to memorize a password or use an additional security device.

Several commercial and research solutions have been proposed or deployed to achieve biometric authentication. These solutions range from the traditional approaches such as fingerprints [14] and iris scan [12] to the more advanced modalities such as human touching [6, 32, 34], human speech [31, 43, 44], eye movement patterns [13, 15, 33], electrophysiological measurements [2, 38, 45], and vibration responses [5, 21, 23]. Of these modalities, vibration response has emerged as an attractive method due to its compatibility with commodity devices. Consumer devices, such as smartphones and watches, are commonly equipped with vibration motor, microphone, accelerometer, and gyroscope which can generate and measure the vibrations off the human body.

Typical biometric approaches rely on what we refer to as "static" biometrics. An initial training phase collects physiological or behavioral information from the user, such as a gesture, fingerprint, or voice print. At the authentication phase, the user proves their identity by reusing the same information every time. The problem lies in that **human biometrics are non-resilient** [24, 26, 27, 41]: once the biometric information has been compromised, the user cannot recover. Some biometric methods such as gesture-based vibration [21] can scale to multiple traits corresponding to a specific gesture. Their usability, however, will degrade significantly as a cost due to increased training effort and mental burden.

In this work, we attempt to answer this question: *Is it possible to leverage the strengths of biometric authentication while avoiding its pitfalls?* We answer this question in the affirmative and argue that the key to answering this question is to consider a dynamic view of human biometrics. The human body is a complex and dynamic system that reacts differently to different physical stimuli. If through some training phase, an authenticating service knows the responses

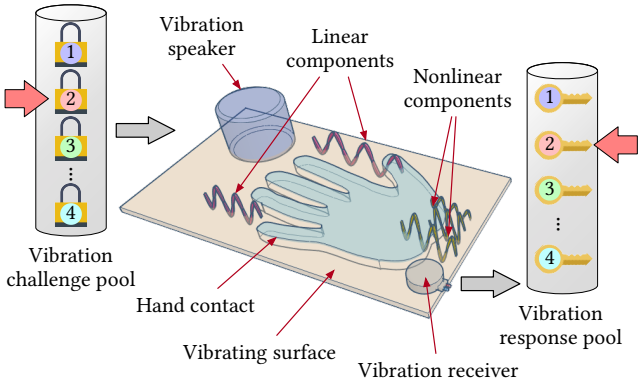


Figure 1: Illustration of VELODY.

to a large set of stimuli, then it can play a new and disposable stimulus at each session. It collects the response and attempts to match it to the previously recorded response. Instead of reusing the same biometric to authenticate the user, an authenticating service can use a new biometric for each authentication session and never use it again. We refer to this model as **challenge-response biometric authentication**. This model is akin to physically unclonable functions (PUFs) that are popular in hardware security [35].

In this paper, we present VELODY, a system that adopts a challenge-response protocol for biometric authentication. It leverages the nonlinear and complex nature of hand-surface vibration. Figure 1 illustrates the use case of VELODY. It has access to a pool of pre-collected challenge-response pairs from a user. The challenge refers to a vibration stimulus to the user's hand through a surface, and the response is the collected vibration. Due to the properties of the user's hand contact, each response is unique per-challenge and per-user. At each authentication session, VELODY plays a disposable challenge and uses a classifier to decide whether the measured response matches the pre-collected one. By design, VELODY is resilient to an attacker replaying previously used biometric information.

To realize VELODY, we have to design two core components: (1) the challenges to play and (2) the classifiers to compare the collected and pre-collected responses.

Challenge design: A challenge is a vibration stimulus that comprises different spectral components. First, to maximize the user-distinguishability as a biometric, a frequency sweep is used to capture the frequency selectivity contributed by the physiological traits in human hands. Second, combinations of sinusoidal waves with random frequencies act as stimuli along with the frequency sweep in disposable challenges to elicit the user-distinct and varying degrees of complicated nonlinearity in vibration responses, including harmonics and intermodulation, which are hard to model and predict for unseen responses.

Response classification: VELODY is a per-user system; a VELODY user does not have access to other users' response data for privacy and security considerations. This requirement constrains VELODY's classifier design as it cannot obtain negative samples from other users. To address this issue, we utilize the one-class k-nearest neighbor (OC-kNN) classifier, which relies on the similarity between inference-time observations and training instances. VELODY trains one classifier for each challenge. We devised a novel mechanism to set the matching threshold of the classifier per-user as to reduce the misclassification rate.

We implement VELODY using off-the-shelf speaker and accelerometer. Our evaluation via 15 individuals shows the following:

- VELODY exhibits a favorable performance in terms of security and usability with an EER at 5.8% evaluated using long-term authentication session against impersonation attacks.
- VELODY can reject 97.3% impersonation samples and 100% replay and synthesis attacks with reasonable effort in passive enrollment and an extremely short 200-ms vibration challenge in one authentication session.
- VELODY's challenge-response design is resilient to variations in the challenge design. Using shorter challenges with fewer spectral stimuli still maintains a satisfactory EER.

2 BACKGROUND ON HAND-SURFACE VIBRATION RESPONSE

In this section, we introduce two properties of hand-surface vibrations that enable the operation of VELODY: user distinguishability and nonlinearity.

2.1 User-Distinct Vibration Response

A human hand exhibits unique physiological features such as geometry, bone shape, bone-muscle ratio, bone density, which have been utilized as a static biometric for a while [3]. These features lead to the human-distinguishable characteristics of acoustic dispersion, absorption, and reflection when a person places his/her hand on a vibration surface. Specifically, the contact area between a hand and the vibration surface affects the reflection and absorption of the surface vibration. Differences in the contact area (due to different hand geometry of different users) contribute to different vibration propagation paths and varying constructive or destructive interferences at different frequencies – leading to frequency-selective vibration responses. Moreover, the differences in hand's damping and acoustic absorption relate to composition, the force and distribution of contact between the hand and surface, contributing to vibration responses that are user-distinct, too [8].

One can naively model the vibration response of a hand using a spring-mass-damper system. Such a model, however, ignores several practical issues, including the multipath-induced frequency selectivity dependent on the hand-surface contact and the nonlinear spectral interactions. As a result, an accurate user-specific model for hand contact interaction is extremely hard to build even by state-of-the-art 3D finite-element (FE) modeling techniques [9, 36].

2.2 Nonlinear Effects in Vibration Response

The second property that VELODY utilizes is the nonlinearity in the vibration responses of the hand-surface system, which is difficult to model and predict [22, 39, 42]. Previous studies have demonstrated that a hand itself, due to its geometry and composition, is a nonlinear medium for acoustic propagation [9, 16].

Here, we show a model of nonlinear acoustics to explain the complexity of vibration responses of the hand-surface system. For a linear system, the output signal S_{out} is a linear combination of

the input signals S_{in} , which can be represented as:

$$S_{out} = A \cdot S_{in}. \tag{1}$$

The complex gain only affects the phase and amplitude of the inputs, and no new frequency component appears in the response of the linear system. In a nonlinear system, however, like the hand-surface system, the response will contain new frequency components. For simplicity, we model the nonlinear response as a power-series of inputs with different gains at each term:

$$S_{out} = \sum_{n=1}^{\infty} A_n \cdot (S_{in})^n. \tag{2}$$

For example, if the input is a single sinusoidal wave at a frequency f_1, different orders of *harmonics* $(n \cdot f_1)$ will appear in the response. For an input composed of two signals, the output of this nonlinear system exhibits *intermodulation*:

$$S_{out} = A_1 \cdot (S_{in,1} + S_{in,2}) + A_2 \cdot (S_{in,1} + S_{in,2})^2 \ldots \tag{3}$$

For example, the second order term in Eq. 3 has a product of signals resulting in new frequency components at $f_1 - f_2$ and $f_1 + f_2$. We can rewrite the second-order term of the output in the equation above as follows.

$$\begin{aligned} S_{out,2} =& a_1^h \sin(2\pi \cdot 2f_1) + a_2^h \sin(2\pi \cdot 2f_2) \\ &+ a_1^m \sin(2\pi(f_1 + f_2)) + a_2^m \sin(2\pi(f_1 - f_2)), \end{aligned} \tag{4}$$

where a_i^h are the gains for harmonics and a_i^m are those for the intermodulation.

The harmonic gains depend on the medium properties and the frequency, while the intermodulation gains depend on several factors including the material coefficients between f_1 and f_2, the amplitudes of both f_1 and f_2, which are sensitive to the structure of vibration medium [22] – the hand-surface system in our case. The system creates more complicated intermodulation interactions for higher order terms which are hard to predict.

Note that this simplified model does not convey the dynamics and component interactions of a nonlinear system as the nonlinear responses are highly input-dependent within the same nonlinear system. The model fails to describe the non-analytic responses like complicated energy exchange between different frequencies as well as temporal dependencies of nonlinear coefficients [39, 42]. Other nonlinear effects include nonlinear attenuation rates at different frequencies depending on the input excitation level [1]. Due to this complex and nonlinear nature of vibration responses in a hand-surface system, precise modeling or prediction of arbitrary responses preserving individual traits is highly implausible. It is very hard to predict the hand response for a previously unobserved input signal, to the best of our knowledge.

2.3 Motivational Example of Hand-Surface Vibration

We take an exemplification approach to motivate the distinct and nonlinear hand-surface vibration. We record the vibration responses of a hand-surface system to provide an intuition about our model. We use a portable vibration speaker (Vib-Tribe Troll Plus) to generate an input vibration and we collect the responses using a contact microphone (BU-27135 accelerometer) from a vibrating copper surface (setup similar to Figure 1).

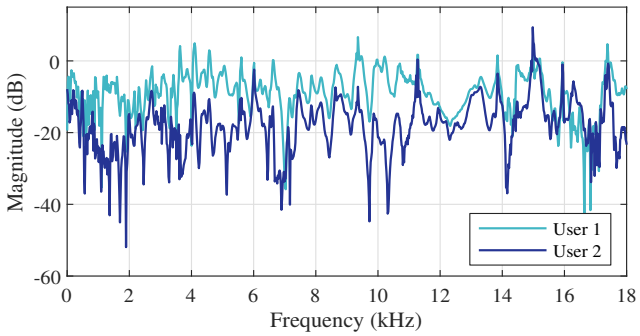
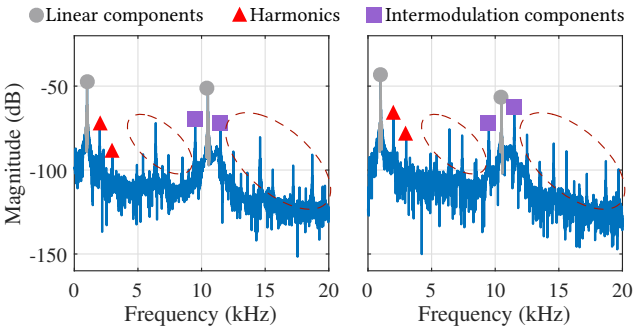


Figure 2: Vibration responses of two difference users.



(a) With a hand on surface (b) Without a hand on surface

Figure 3: Nonlinearity in hand-surface measurement.

User distinguishability: We first examine the user distinguishability of frequency responses. Two users place their hands on the vibration surface with the same gesture (relaxed with spreading fingers). Meanwhile, the vibration speaker plays a sweeping sinusoidal vibration from 0.2 to 18 kHz for a duration of 200 ms. Figure 2 shows the frequency response of the transfer function of each user, illustrating the attenuation at different frequencies. It is evident from the figure that the responses of the two users are easily distinguishable. The transfer function does not capture all sources of nonlinearity like harmonics and intermodulation which result in more distinguishability.

Nonlinearity: To visualize the nonlinearity in hand-surface system, we play two sinusoidal waves at 1 kHz and 10.5 kHz simultaneously. We show the frequency response of the raw recorded signals (not the transfer functions as before) with and without a hand placed on the vibration surface in Figures 3(a) and 3(b), respectively. We mark the major frequencies in grey dots, some representative harmonics in red triangles, and intermodulation components as purple squares. The spectral locations of the newly-generated frequencies match the anticipated harmonics and intermodulation results in both scenarios. The intermodulation components are significant in both cases and even comparable with the major frequencies. Also, the hand exhibits distinguishable modification of nonlinear components as evident from components marked and circled in Figures 3(a) and (b).

The findings above show an intuition that the vibration responses of hand-surface system are distinct between users (Figure 2), and the

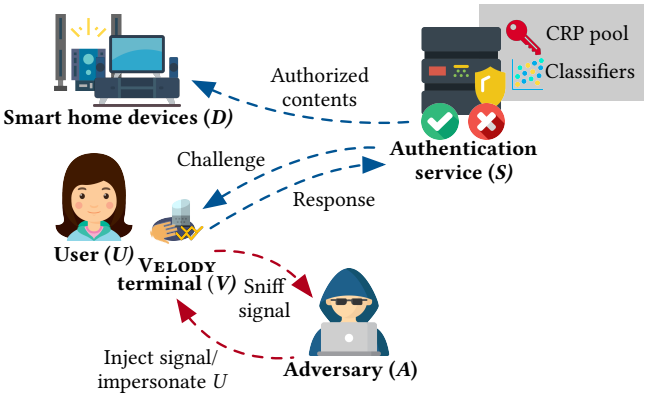


Figure 4: System and threat model.

nonlinear effects are significant (Figures 3), too. Both observations are critical to the design of VELODY.

3 SYSTEM AND THREAT MODELS

In this section, we describe the system and threat models for VELODY.

3.1 System Model

Figure 4 shows an overview of the system model, including the involved parties. We assume a general scenario where VELODY is employed to authenticate a user (U) to use smart devices (D). The authenticator service (S) grants permission for the user (U) to use smart devices (D) and access to authorized contents. The user (U) requests authentication and permission through the VELODY terminal (V), which is associated with an interface consisting a surface, a vibration speaker, and contact microphones. For example, V can simply refer to laptop or a smartphone paired with a smartwatch that has a high bandwidth accelerometer [17]. V generates a vibration signal according to a challenge assigned by S, collects the response, and sends it to S. We assume a secure training phase during which S collects all vibration challenge-response pairs securely for future verification.

For each authentication request, S randomly selects one disposable vibration challenge and sends to V, which collects the hand-surface response. The response is sent back to S to verify the claimed identity U. Note that V may not only verify the identity solely relying on vibration challenge-responses but also on other factors like password in a multi-factor authentication scenario. Once U is verified and authenticated, the requested D will be activated, and the authorized contents, such as a video stream, will be distributed.

Figure 4 depicts the involved parties in our system model as separate entities, just for visualization. There is nothing preventing V, D and S to be part of the same device, such as a laptop, desktop, or even a smartphone.

3.2 Threat Model

The goal of the adversary (A) is to deceive S to grant the access to the victim, U. In addition to the attacker capabilities that have been typically assumed in previous work, such as physical access to the authentication devices, we take one step further and assume that

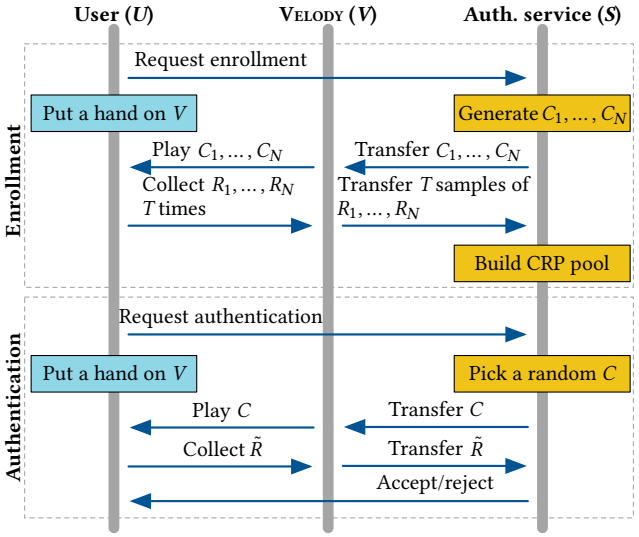


Figure 5: Authentication protocol of VELODY.

the active attacker is able to observe previously used responses and replay raw or synthesized response corresponding to an unknown challenge through a side channel. This side channel could refer to (1) a compromised networking interface between V and S or to (2) the attacker collecting responses through a placed/compromised device in the same environment. In this paper, we assume a strong adversary that is capable of recording the exact challenge-response pairs. By considering a strong adversary model capable of recording and replaying biometric information, we avoid the pitfalls of previous defense approaches. Under this scenario, we consider the following attack scenarios.

- **Zero-effort attack.** In this scenario, A only bypasses the password and tries to authenticate opportunistically by vibrating an empty surface without hand contact using the authentication-time challenge assigned by VELODY.
- **Impersonation.** In this scenario, A has access to V, bypasses other authentication factors like password, claims the identity of U, and places his hand on the vibration surface to impersonate legitimate U using the same gesture.
- **Raw signal replay attack.** In this scenario, A acquires previously-used vibration challenge-responses from U and replays an arbitrary raw response to S during an authentication session through a compromised wireless channel.
- **Synthesis attack.** More advanced than simply replaying raw signal, A attempts to predict the response of a specific challenge by modeling from previously observed responses and inject the synthesized signal in real time. We consider the implementation of multiple synthesis methods in our evaluation.

4 VELODY PROTOCOL AND FRAMEWORK

In this section, we present the design details of VELODY.

4.1 Authentication Protocol

Challenges and responses: VELODY employs a challenge-response protocol as illustrated in Figure 5. At each authentication session, S

sends the user a challenge and receives a response. Only after matching the measured response to the previously recorded response is the user authenticated. Each challenge-response pair (CRP) is disposable; a challenge will not be reused in other authentication sessions.

A vibration challenge (C) is a specially designed acoustic signal played by V. S collects a challenge-specific and user-distinct response for *verifying the user identity*. The n_{th} challenge $C_n = (f_{crp}, f_n^1 ... f_n^M)$ can be characterized by M randomly selected distinct spectral stimuli (sinusoidal waves) appearing at different slots within the entire time period of the challenge, and f_{crp} is the time-varying frequency of a chirp signal. For each challenge C_n, the response is measured T times. R_n, the response to challenge C_n has T elements: R_n^i ($i = 1, 2, \ldots, T$). As explained earlier, each response is a function of the challenge as well as the nonlinearities associated with playing the challenge to the user's hand. The nonlinear dynamics are challenge-dependent and user-specific; each challenge produces a unique response for each user.

Enrollment: The enrollment phase of Velody is initiated when requested by U, or CRPs are depleted. S generates N new random challenges C_1 to C_N that are not previously used for authentication. V plays each C_n with the user's hand placed on the panel and records the corresponding response R_n. This procedure is repeated T times to generate a robust training set. After receiving the responses R_1 to R_N, S trains the classifiers for the new CRPs; Velody trains one classifier for each CRP. We employ one-class k-nearest neighbors (OC-kNN) classifier for verifying the response corresponding to a challenge. During training, a threshold Th_n is computed for each classifier corresponding to each challenge. We assume that the enrollment phase takes place in a secure setting (attacker cannot record/alter the recorded responses).

Authentication: After the enrollment is completed, U can request authentication to S. Upon receiving an authentication request, S randomly chooses a challenge C_n from unused challenge pool, which is sent to V. While U places their hand on the vibration surface, V plays the challenge C_n collects a response $\tilde{R_n}$, which is sent to S. S performs the feature extraction and decision making.

The authentication decision D on $\tilde{R_n}$ corresponding to C_n is described as follows:

$$D = F_n(\tilde{R_n}, R_n, Th_n), \tag{5}$$

where F_n represents the process starting at feature extraction and ending at the OC-kNN-based classification. R_n represents T training responses collected during enrollment; and Th_n is a challenge- and user-specific threshold. The challenge used in the current session, C_n, is disposed to ensure security against replay attack. The detailed decision process is discussed in Section 4.5.

4.2 Framework Overview

The processing framework of Velody is illustrated in Figure 6 including is major stages. The collected responses during the enrollment session (i.e., $R_n^1, R_n^2, \ldots, R_n^T$) and authentication session (i.e., $\tilde{R}$) is synchronized and segmented first; then, filters and normalization are applied on the raw response segments. Velody extracts effective spectrotemporal features from the raw time-domain response. For each CRP on normalized feature vectors, an OC-kNN

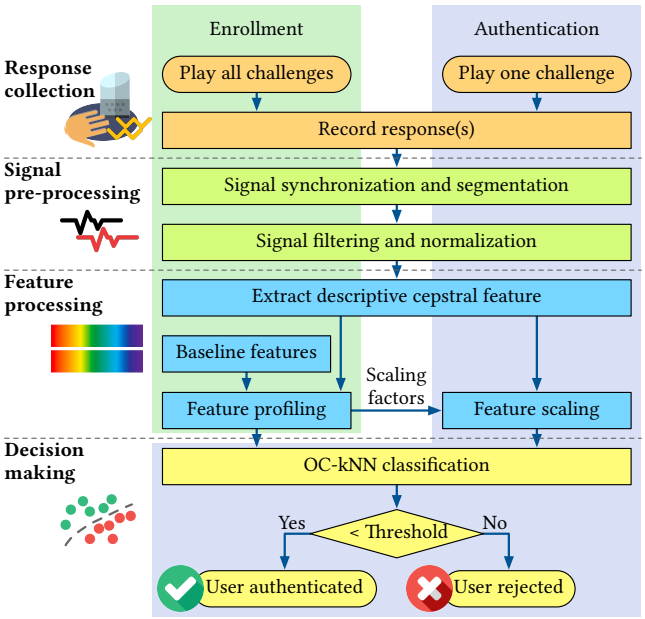


Figure 6: Processing framework of Velody.

classifier is built. An authentication decision is made based on the comparison of the CRP-specific threshold and the OC-kNN distance between observed features of response $\tilde{R}$ and the templates. The advantage of using OC-kNN as a classifier is that training can be conducted per-person, without the need to collecting data from multiple people.

4.3 Vibration Challenge Design

We have two requirements from Velody's vibration CRPs: (1) distinguishability between the users of the system and (2) distinguishability as well as unpredictability from previously observed CRPs. These requirements necessitate the careful design of the challenges.

To meet the first requirement, we adopt a chirp vibration signal (frequency sweep) to capture the frequency selectivity contributed by the physiological characteristics of human hand in a short time. We meet the second requirement by designing each challenge to evoke a unique vibration response each time. The period of entire challenge is divided into several time slots, and in each slot, Velody superimposes a sinusoidal wave at a random frequency onto the chirp instance to make the response unpredictable. The superimposition of the chirp signal with a sinusoidal wave generates complex harmonics and intermodulation interactions of different orders simultaneously, which is practically unpredictable from previously observed CRPs.

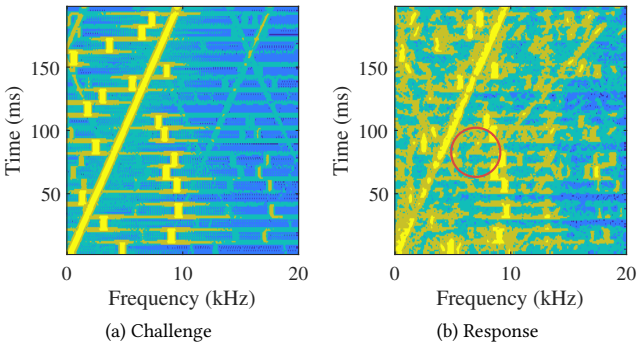
The vibration challenge signal $C_n(t)$ as a function of time t is expressed as:

$$C_n(t) = S_{crp}(t) + \sum_{i=1}^{I} S_{sin,i}(t). \tag{6}$$

The linear chirp signal $S_{crp}(t)$ is constructed by:

$$S_{crp}(t) = A_{crp} \sin\left(2\pi f_{crp}(t)t + \phi_{crp}\right), \tag{7}$$

where A_{crp} and ϕ_{crp} denote the amplitude and phase of the chirp signal, respectively; and $f_{crp}(t)$ is the frequency of the chirp, which



(a) Challenge (b) Response

Figure 7: Comparison of challenge and response spectrograms. The challenge contains the chirp as well as superimposed sinusoidal signals at different frequencies. Some nonlinear components are highlighted in the response.

linear changes from f_b to f_e over time:

$$f_{crp}(t) = \frac{f_e - f_b}{T_{crp}} t + f_b. \tag{8}$$

The random component $S_{sin,i}(t)$ in (6) is defined as:

$$S_{sin,i} = \begin{cases} A_i \sin(2\pi f_i t + \phi_i) & \text{if } (i-1)\frac{T_{crp}}{I} \leq t < i\frac{T_{crp}}{I} \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where A_i is the amplitude of the sinusoidal wave in the i-th time slot, $(i-1)\frac{T_{crp}}{I} \leq t < i\frac{T_{crp}}{I}$; and f_i is the random frequency.

In our prototype, the chirp S_{crp} changes from f_b =0.5 kHz to f_e =10 kHz, in which the vibration speaker generates stable vibration and hand-surface responses preserve useful information for distinguishing different users. The duration T_{crp} is set to 200 ms, short enough to avoid annoying the user during enrollment and authentication. The changeable stiumli of each challenge consist of 20 different sinusoidal waves of random frequencies (i.e., $I = 20$), uniformly distributed over in a range between 0.5 kHz to 10 kHz to ensure diversity of both linear and nonlinear components. The amplitudes of sinusoidal stimuli, A_i, is also randomly determined for challenge diversity.

In Figure 7, we show two spectrograms: one from a challenge and one from its corresponding response. From Figure 7(b), we can clearly observe some nonlinear components, such as the highlighted ones, including harmonics and intermodulation, which are widely spread over a wide frequency range.

4.4 Feature Processing

Signal pre-processing: First, we perform signal alignment and segmentation to minimize bias for feature extraction, resulting from imperfect hardware synchronization. We align the measured response with the challenge by finding the time lag that maximizes the cross-correlation between them. Second, we apply a bandpass filter between 0.3 kHz and 20 kHz to remove external vibration induced by motion. Also, we apply multi-band spectral subtraction to clean the in-band noise due to measurement. Finally, we apply Z-score normalization on each response signal to reduce the variability from gesture inconsistency.

Cepstral feature extraction: The cepstral features are widely adopted for acoustic modeling of music, human speech, and structural damage, etc. which are of complex or nonlinear nature. Intuitively, cepstral coefficients describe the dynamics among the different frequency bands of a signal, including the contribution of linear and nonlinear spectral components. Cepstral coefficients are calculated by applying discrete consine transform (DCT) on the complex logarithm of the Fourier transform of a time-domain signal. A sliding window is used to extract cepstral coefficients over the duration of a signal to model its temporal dynamics.

The Mel-frequency cepstral coefficient (MFCC) is the most frequently used cepstral feature for human speech modeling and recognition since the Mel-scale filter banks are optimized for human speech and perception frequency. Instead of using the Mel-band, Velody applies linearly allocated filter banks before calculating the coefficients. We argue that unlike human speech where high-frequency components contribute less to human perception, the nonlinear vibration responses of Velody are spread more widely across the spectrum. Specifically, the band edges of overlapped filter banks are separated by 0.25 kHz, and we take 40-th order cepstral coefficients at each time window of 10 ms, with a window overlap of 8 ms to capture fine-grained dynamics. Moreover, the delta and delta-delta of the cepstral coefficients are also computed to capture more fine-grained spectral dynamics within a short time frame. A cepstral feature map combines all the cepstral coefficients with the log energy and first/second order delta energies per window.

Statistical feature extraction: Raw cepstral features exhibit inconsistencies brought by several factors such as circuitry randomness, gesture variation, and imperfect signal segmentation. To overcome this issue, we extract statistical features for each coefficient channel. Each coefficient channel is defined as the sequence of the values of cepstral coefficients over signal duration.

Besides mean, variance, entropy, and power, which are standard metrics in characterizing a random variable or its distribution, we adopt other metrics to assess the distribution of cepstral coefficients over the signal period. Skewness measures the degree of symmemtry of left and right parts of a distribution; kurtosis estimates the 'tailedness' of one distribution compared to normal distribution; and crest factor examines the significance of the extreme peak in the distribution [18]. The final feature vector comprises statistical features describing the cepstral, delta-cepstral, and delta-delta-cepstral coefficients as well as log frame energies. This results in a feature vector with 1722 elements per response in this work.

4.5 Classification

Velody is a per-user system; a Velody user does not have access to other users' CRPs for privacy and security considerations. This requirement constrains Velody's classifier design as it cannot obtain negative samples from other users. OC-kNN is an instance-based classifier that relies on the similarity between inference-time observations and training instances. Velody trains one OC-kNN classifier for each CRP; the underlying assumption is that the response to a challenge for a user is different from those to other challenges. It is also different from responses to the same challenge from other users. The authenticator service passes the features from the response to the CRP's OC-kNN that decides whether the response

is valid or not for the played challenge. The two major steps in OC-kNN decision making are distance calculation and threshold comparison.

Distance calculation: Recall that during enrollment, VELODY plays each challenge T times, so that it collects T copies of the response. Each response is associated with a feature vector. For the rest of this discussion, $x^i_{n,j}$ refers to the i_{th} feature of a training response (R^j_n) to the challenge C_n. To keep the notation simple, we use x^i_n instead of $x^i_{n,j}$, except for Eq. 12.

We first normalize each feature to the same scale by min-max normalization for the fairness of the distance-based OC-kNN:

$$\hat{x^i_n} = \frac{x^i_n - min(x^i_n)}{max(x^i_n) - min(x^i_n)}, \tag{10}$$

where the min and max are taken for a feature value over the T responses.

Given an unseen feature vector z_n at the authentication phase, VELODY scales it using the *min* and *max* factors computed during training: $\hat{z^i_n} = \frac{z^i_n - min(x^i_n)}{max(x^i_n) - min(x^i_n)}$. We observe that different features have varying sensitivity to system or gesture randomness. We introduce a weight for each feature so that the more consistent features have higher weights [5]:

$$w^i_n = \frac{max(E(\hat{x^i_n})) - E(\hat{x^i_n})}{\sum_{i=1}^{1722}(max(E(\hat{x^i_n})) - E(\hat{x^i_n}))}. \tag{11}$$

The expectation is taken over the T training samples (responses to a single challenge during enrollment). The min and max are taken over the 1722 features.

The weights are applied to both the training and test instances. The ℓ_1 distance is calculated between the weighted test instance $\hat{z_n}$ and all T training instances $\hat{x}_{n,1\ldots T}$ as:

$$d_j = d(\hat{z_n}, \hat{x}_{n,j}) = \sum_{i=1}^{1722} \left| (\hat{z^i_n} - \hat{x^i_{n,j}}) \cdot w^i_n \right|. \tag{12}$$

The final distance of the test instance to the challenge is calculated by averaging the k smallest d_j values. Comparing the final distance to a threshold Th yields the final classification result.

Threshold estimation: The major obstacle in VELODY's classification is determining a proper Th for each user and each CRP. An ideal Th accepts legitimate samples while rejecting all illegitimate samples. The ℓ_1 distances from the classifier show great diversity among users and CRPs, hence, a fixed threshold for every user and CRP is not ideal. Nevertheless, we notice that distances between training instances and baseline responses collected from vibrating surface without hand contact correlate with those of illegitimate distances ($\rho > 0.5, p = 0.000$) for each user. VELODY utilizes these baseline samples available to every user during enrollment to estimate Th_n corresponding to the n_{th} challenge of one user. VELODY calibrates Th by leave-one-out cross validation based on training and baseline samples. More specifically, one training instance is held out at a fold, and its kNN distance $d_{n,pos}$ as well as distance of baseline samples $d_{n,bl}$ are computed using the rest training instances. Then, the threshold Th_n is determined by

$$Th_n = E(d_{n,pos}) + \alpha \times (E(d_{n,bl}) - E(d_{n,pos})) \tag{13}$$

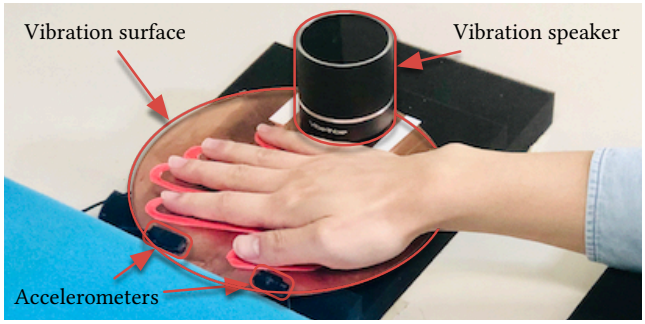


Figure 8: VELODY prototype setup.

where the expectation is taken through all folds and α is a global tuning factor, the usability of which on all CRPs will be evaluated instead of determining thresholds by exhaustive search.

5 PROTOTYPE AND DATA COLLECTION

5.1 Hardware Prototype

A prototype of VELODY is built, as shown in Figure 8. A commercial off-the-shelf vibration speaker Vib-Tribe Troll Plus is used to play challenges. It is attached to a vibration surface, which is an 8-inch copper plate laying on a polymer foam pad. The speaker has an effective frequency range between 80 Hz to 18 kHz and a signal to noise ratio (SNR) of 80 dB. Two contact microphones (BU27135 accelerometer) are attached on two different locations of the vibration surface to measure vibration responses. The BU27135 is an analog accelerometer with a wide effective spectrum and a high sensitivity. Since VELODY relies on the physiological properties of human hand instead of behavioral traits, we fix the gesture for all the users: all users are asked to put the right hand with fingers spread on the vibration surface, where we draw a hand shape for consistent alignment in evaluating impersonation attack. As a proof-of-concept, a PC is used to output all challenges through a built-in sound card and collect responses through a dual-channel USB sound card, sampling at 48 kHz. MATLAB's data acquisition (DAQ) toolbox is used.

We argue transferring challenges and collecting responses can be done remotely via wireless protocol, such as WiFi and Bluetooth. in a real-world use case. The duration of each challenge is set to 200 ms. We generate 100 challenges, and these challenges are kept unchanged for all users for establishing impersonation attacks.

5.2 Data Collection

We recruited 15 subjects with body mass index (BMI) ranging from 17.5 to 29.6 with a median of 22.2. The entire course of data collection took place over one and a half months, during which each participant was involved in three data collection sessions. The first two sessions were performed within one day with a time gap of at least 30 minutes. This was to verify intra-day (short-term) consistency and to establish baselines of consistency. The third session was arranged at least five days after the first two sessions to collect data for verifying inter-day (long-term) consistency.

Each session took about 20 to 30 minutes, including introduction, orientation, surveying, and data collection. After explaining the consent form, having user's agreement and signature and collecting

basic information about the user, each participant was demonstrated with how to interact with VELODY interface and take a good gesture. For each challenge, responses were measured for 15 trials. In between two consecutive trials, the user was asked to remove the hand from the plate and relax to ensure the diversity of the data set. Each trial took 30 seconds, including short intervals of 100 ms between two consecutive challenges. No complicated task or gesture for enrollment or authentication was needed. In a real use case, each authentication session will take only 200 ms, which is short enough to ease user's burden. The user study is approved by the Institutional Review Board (IRB) of our institution.

The total number of collected responses is 67,500 (3 sessions × 15 users × 100 challenges × 15 trails). Additional 15 responses were collected from empty vibration surface for threshold estimation and attack evaluation.

As for impersonation attack, for each user, we consider all other 14 users as active impersonators. Therefore, we use 3×15×100×14 = 63,000 samples for impersonation attack against each user. As for replay attack of raw signal attack, we use responses collected for challenges other than the legit one. For each participant, the number of raw signal replay samples is 99×100 = 9900. For each user, we also conduct benchmarking sessions for evaluating the attack using modeling and synthesis.

6 EVALUATION

In this section, we evaluate the VELODY framework focusing on answering two questions about its usability and security aspects.

Q1: How well does VELODY authenticate legitimate users? The major factor impacting the usability of biometric authentication is its success rate of verifying true users (true positive), which is typically compared against the possibility that an illegitimate user is falsely accepted (false positive), where we adopt responses from other users performing the same gesture while being stimulated by the same challenges as impersonation samples.

More specifically, four detailed usability aspects need to be analyzed to answer *Q1* comprehensively, for which we vary VELODY's configuration like threshold, training set size, and CRP complexity, and interpret results of FNR, FPR, and EER.

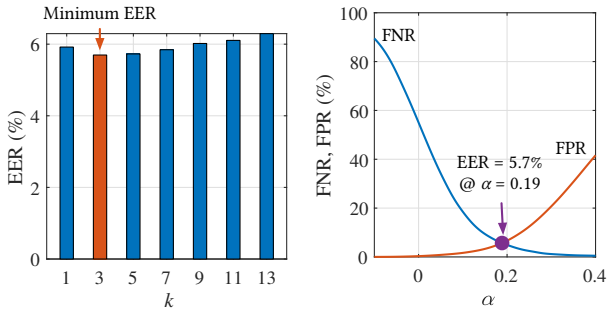
- How sensitive is VELODY to system parameters such as k in OC-kNNs and threshold factor α?
- How consistent is VELODY's accuracy in the long term?
- How much training data do we actually need?
- How scalable are the CRPs of VELODY?

Q2: How robust is VELODY against various attacks? The security evaluation focuses on examining and comparing the attack success rate of zero-effort attack, impersonation attack, raw signal replay attack, and synthesis attacks. The following question will be answered in this regard.

- What is the most effective attack modality, and why?

Evaluation metrics: The major metrics used for quantitatively analyzing the system's usability and security are as follows:

- False negative rate (FNR): The rate of mistakenly rejecting legitimate users, as a function of classification threshold. It is a usability metric.



(a) EER vs. k (b) FNR and impersonation FPR vs. α ($k = 3$)

Figure 9: Authentication performance of intra-day sessions.

- False positive rate (FPR): The ratio of how many illegitimate samples are accepted, as a function of classification threshold. It is a security metric.
- Equal error rate (EER): The rate when FPR equals to FNR for a certain classification threshold. It is a widely adopted metric to assess the overall accuracy and how well usability and security are balanced in an authentcaion system.

6.1 Accuracy of Authenticating Legitimate Users

6.1.1 System parameter baselining using intra-day sessions. One of the challenges in implementing VELODY's classification scheme is tuning the large number of OC-kNN classifiers corresponding to many CRPs with a minimal effort since it is not practical to exhaustively search the optimal configuration for each classifier of every CRP. To this end, we evaluate whether two major parameters, OC-kNN component k and global threshold tuning factor α, are sufficient to achieve a good overall authentication accuracy.

For each user, two separate sessions are used for evaluating system performance. Though physiological characteristics of human hand are relatively consistent, we argue that multiple factors, such as gesture, posture, and contact force, which may not be well controlled by users without concentration across different sessions, may influence the authentication success rate. The system configuration of VELODY should be robust against these variations.

Setup: We use two sessions within one day (intra-day) but 30 minutes apart for all 15 users and 100 CRPs to establish a baseline for authentication accuracy. One session is used as a training set, and another acts as a test set. Each session includes 15 trials for every CRP. For each user, 30 trials of both two sessions from all 14 other impersonators are used as illegitimate samples for the classifier of each CRP. We evaluate $k = 1, 3, 5, \ldots, 13$, which are fixed for both threshold estimation and OC-kNN testing. Tuning factor α is varied from -0.1 to 0.4 with a step of 0.02.

Results: The impact of k in OC-kNNs is evaluated first. A very small k may lead to noisy classification results and unstable performance; on the other hand, if k is too large, it will cause under-fitting and the decision boundary will be overly smoothed. Figure 9(a) shows the average EERs of all the users and classifiers with various k values, which are calculated by finding the crossover of interpolated FNR and FPR data points at varying discrete α. We can

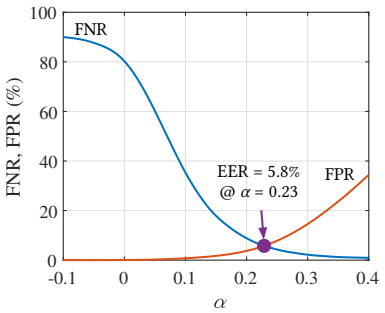
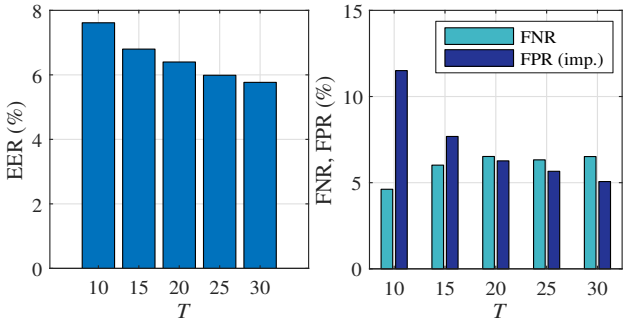


Figure 10: Authentication performance of inter-day sessions.



(a) EER vs. T (b) FNR, impersonation FPR vs. T (a = 0.22)

Figure 11: Authentication performance with different training set sizes.

see that Velody is able to achieve a satisfactory EER below 6.3% within a wide range of k from 1 to 13. The minimum EER of 5.7% is attained when $k = 3$.

The trend of FNR and FPR with varying threshold factor α from -0.1 to 0.4 is shown in Figure 9(b) at an optimal $k = 3$, which is fixed for following experiments. Both FNR and FPR change smoothly and monotonically with α as a larger α accepts more legitimate samples while misclassifying more impersonation samples as well, which is intuitive regarding the distance-based OC-kNN classification. FNR and FPR intersect at $\alpha = 0.19$ when EER is 5.7% (marked with a purple dot in Figure 9(b)). Velody performs satisfactorily within a broader range of α. For example, if $\alpha = 0.14$ is chosen, Velody can reject over 97.1% of attacks while maintaining a FNR at 10.7%.

Hence, we verify that Velody's classification can achieve a good overall authentication accuracy with a large pool of CRPs without tuning parameters in a brute-force manner, and it is capable of handling inter-session variation of intra-day tests.

6.1.2 Long-term consistency evaluation on inter-day sessions. To verify long-term consistency and strengthen our usability argument, we collect the third session, following the same experimental procedure, but five days later than the first two sessions for each user. In daily usage, larger variation in vibration responses may occur due to behavioral changes by different cognitive and physical statuses, which may not be well considered by intra-day experiments.

Setup: We fix k to 3 and use the first two sessions, including $T = 30$ trials as the training set to authenticate the third session, which capture more variation of users due to inter-session behavioral inconsistency, as we observe that using training data collected in a single session for authenticating inter-day trials may not cover this variation perfectly, resulting a higher average EER of 7.9% by training on two individual sessions respectively.

Results: We show the varying FNR and impersonation FPR evaluated on inter-day sessions in Figure 10. We observe similar trend of FNR and FPR compared to intra-day verification results. A low EER of 5.8% can be achieved at $\alpha = 0.23$ (marked with a purple dot in Figure 10), which indicates negligible difference compared to 5.7% from intra-day evaluation. Though the optimal α varies slightly, Velody still achieves low FNR and FPR of 11.8% and 2.7% respectively using an $\alpha = 0.18$, close to the interpolated EER point at $\alpha = 0.19$ of intra-day verification, indicating good consistency.

Therefore, we verify that Velody is robust to system and behavioral variation and attains good long-term consistency with reasonable training effort. We argue that physiological properties of human hand are relatively stable regarding to time despite that physical development process of children or aging may affect the properties [28], which can be addressed by updating the CRP pool.

6.1.3 Impact of training set size. Though Velody employs very short CRPs of 200 ms and almost passive enrollment/authentication sessions without performing complicated tasks, the size of training set influences usability in multiple angles such as duration of enrollment, the computation time for kNN at authentication phase, as well as data storage. To investigate the sensitivity of authentication performance to the number of instances used in training each classifier, we vary the number of training instances and examine accuracy of Velody for each case.

Setup: We prune the training set from $T = 30$ instances of two intra-day sessions to 10 with a step of 5 by trimming those have larger average pairwise ℓ_1 distances to other training instances in the validation phase then test using 15 inter-day trials.

Results: In Figure 11, we show the variation of authentication performance ((a): EER, (b): FNR/FPR) with training set sizes. From the EER plot, we conclude that the performance is generally stable against different T, however, the more legitimate templates we have, the better Velody's overall performance is, as the EER decreases from 7.6% to 5.8% by varying the number of instances from 10 to 30. Also, from Figure 11(b) we see both FPR and FNR do not vary substantially from 15 to 30 at a fixed α of 0.22, meanwhile a smaller size T benefits consistency while sacrificing security slightly.

These findings indicate more training instances do improve system robustness, nevertheless, using fewer training instances around 15 is feasible to achieve comparable authentication performance while saving enrollment time if users keep good consistency, as well as data storage and computation at authentication time.

6.1.4 Scalability of Velody CRP. The CRP pool of Velody can be scaled by changing challenges in different domains like sinusoidal frequencies or complexity in terms of challenge duration and bandwidth of signal. We still anticipate that Velody maintains its performance when a larger CRP pool is deployed for realistic usages with daily authentication activity, which is evaluated here.

Setup: First, for validating the variation in authentication success rate regarding different combinations of sinusoidal stimuli,

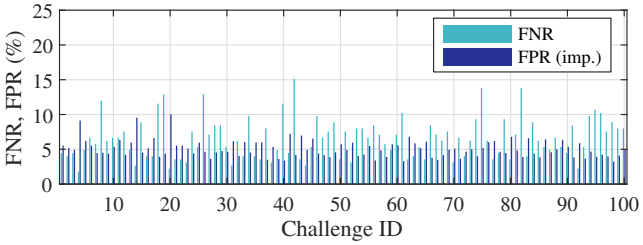


Figure 12: Authentication performance of different CRPs.

we use the inter-session results and demonstrate the individual accuracies of all 100 200 ms-challenge.

Also, based on the same dataset we have, we can emulate the scenario when the challenge complexity is varied by truncating each 200-ms CRP in time domain to 100 ms and 50 ms respectively, starting from $t = 0$ which ensures that responses are not impacted by previous signal. Each truncated challenge-response has a narrower effective chirp bandwidth and fewer sinusoidal stimuli.

Results: The accuracy statistics of different vibration challenges are shown in Figure 12. The performances of vibration challenges of varying combinations of stimuli are quite consistent, and 99% of them have an average FPR lower than 10%. The average FNR per CRP is stable across various challenges, though more variant than FPR, and only a few challenges' (11%) FNRs are higher than 10%.

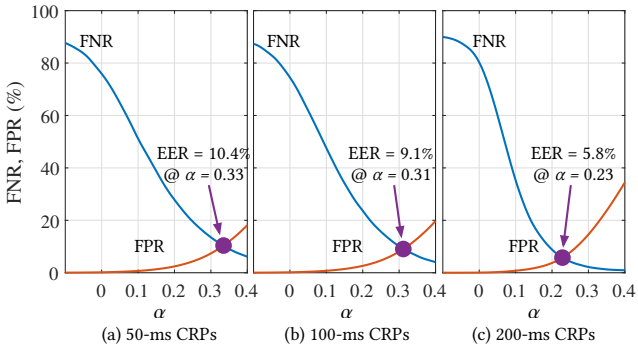
For verifying the efficacy of CRPs with reduced complexity, the threshold tuning factor α is varied for each case, and we also evaluate the FNR with FPR from impersonation attack, whose results are shown in Figure 13. From the plot, we observe that EER only degrades slightly from 5.8% to 9.1% and 10.4% when 200-ms, 100-ms, and 50-ms CRPs are used, respectively. Despite the observation that CRPs with reduced complexity lead to higher FNR while contributing to lower FPR with an α ranging from 0.15 to 0.4, and the thresholds to achieve equal error drift from that using 200-ms CRPs.

Revisiting the findings, we conclude that the design of Velody vibration challenge is scalable and flexible. A user can enlarge the CRP pool by different approaches like updating the spectral stimuli, changing chirp bandwidth, and varying signal duration. Also, the enrollment and authentication time will be saved proportionally using decreased challenge duration with an insignificant penalty in system accuracy. However, Velody also leaves the opportunities for improving the accuracy of different CRP designs by reconfiguring framework parameters such as the duration of sliding window, cepstral filter banks, etc., in feature extraction.

6.2 Robustness against Various Attacks

To answer $Q2$, we set up multiple attack scenarios with varying attacker capabilities and compare the results in respect to usability represented by FNR, whose results are shown in Figure 14.

Setup: The configuration detail of all evaluated attacks is explained below. Note that all classifiers are trained using 30 trials and k is set to 3. (i) *Zero-effort attack:* For evaluating this attack, we collect 15 responses from the vibrating Velody surface without hand contact to attack all 100 classifiers and all 15 users. (ii) *Impersonation attack:* This attack is evaluated with system consistency in previous section. Every classifier is attacked by responses of other 14 impersonators in all 3 sessions. (iii) *Raw signal replay attack:* We consider the worst case scenario that all previous CRPs are overheard For each classifier, one raw response from every other challenge is replayed, resulting 99×100 = 9900 replay attacks per user. (iv) *Synthesis attack:* Based on resources assumed in raw signal replay attack, the adversary is capable of predicting users' responses in real time using different modeling methods.



(a) 50-ms CRPs (b) 100-ms CRPs (c) 200-ms CRPs

Figure 13: Authentication performance using CRPs with various complexities.

Three modeling methods used in synthesis attack are as follows. (a) *Transfer function-based synthesis*: The adversary approximately models the nonlinear vibration system using transfer function. Chirp signal is frequently used for identifying vibration system [29]. The attacker caculates the transfer function from the response of a linear frequency sweep between 0.2 kHz to 18 kHz with a duration of 200 ms, same as a legal challenge. The transfer function is computed by averaging 10 estimates. Two inputs are considered: raw/original challenge templates (TF-O in Figure 14) and responses acquired from the empty vibration surface (TF-E). Using the second input, the attacker focuses on modeling the effect contributed by contact of the user's hand. (b) *Nonlinear system identification-based synthesis*: The attacker adopts cascaded Hammerstein model, which is a well established method to identify nonlineartiy in vibration system [30]. In this method, nonlinear system is modeled as multiple branches of nonlinear static polynomial elements followed by a linear impulse response, which is computed by measurement from an optimized exponential frequency sweep. Similar to transfer function-based synthesis, we compute the Hammerstein model for each user by exciting the hand-surface system with a 0.2 kHz to 18 kHz optimized sweep of 200 ms, and attack 100 times for each user, considering two input sources same as (a) (NI-O, NI-E respectively). (c) *Feature-level synthesis*: Features of an unknown response is predicted by estimating a feature-level mapping between challenge and responses modelled by the least square solution x in $Ax = B$ where A is the feature vector extracted from responses of empty surface and B is that obtained from the corresponding hand-surface vibration response signal. The attack success rate is represented by FPR (FT) in Figure 14.

Results: Comparing various attack success rates in Figure 14, we conclude that impersonation attack is the strongest one. More specifically, when α below 0.8, none of the other modalities succeeds in attacking Velody (0% FPR). We interpret the finding as follows. The failure of zero-effort attack is due to largely different force distributions and linear/nonlinear responses on the surface compared

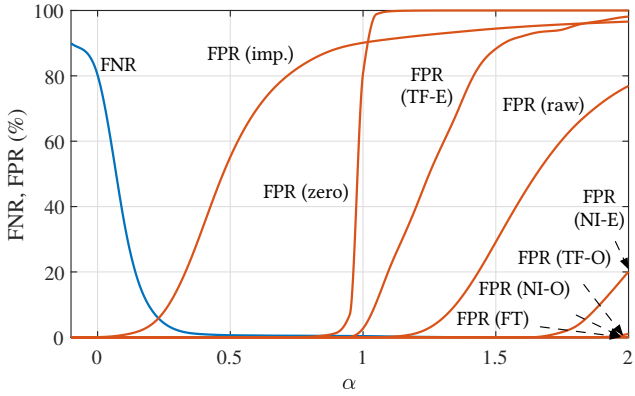


Figure 14: Robustness of Velody against various attacks.

to impersonation. Replaying raw responses is not a feasible attack due to unique spectrotemporal characteristics of randomized stimuli in each challenge. The failure of synthesis methods attributes to the heavy nonlinearity in the vibration response introduced by either circuitry, vibration surface, or hand contact. Also, a 'corrupted' measurement consisting of complicated nonlinear responses of different orders' harmonics and intermodulation even biases estimation by transfer function or Hammerstein model. These findings confirm that precise modeling and prediction in Velody's scenario is very difficult because of multiple factors including non-analytic nonlinearity in real-world measurement. Hence, impersonation is the strongest attack in this case because of similar physical properties between hands and surface contact condition among multiple users.

So far, we have answered all questions post before. To summarize, Velody authenticates legitimate users consistently across time with minimal effort in fine-tuning for many CRPs, minimal authentication effort, and reasonable training effort. Velody's disposable CRPs are scalable for long-term usage. Security under various attacks is also guaranteed as Velody achieves a low EER at 5.8% impersonation attack and stopping 100% of other attacks including replay and synthesis, benefiting from the unique spectrotemporal characteristics of nonlinear vibration responses.

7 DISCUSSION

We have successfully demonstrated Velody's usability and security against various attacks through extensive experiments and analysis. To further improve Velody's practicality and security, the following issues are to be considered.

Deployment in various settings: In this work, we used a vibration speaker as a vibration source and a copper plate as a vibration media. We envision that Velody can be deployed in a variety of settings with a different vibration source and vibration media as long as nonlinearity exists in vibration propagation. It could even be embedded in smart devices, such as laptops and smartwatches. To achieve this vision, a platform-specific challenge generation scheme and evaluation would be required.

Enriching Velody's CRP pool: The most important security attribute of Velody is its non-static and disposable biometric features. Other than the dimensions we discussed in the analysis, such as duration, random frequencies, and bandwidth, more aspects can be tuned to increase the the number of CRPs and improve distinguishability. Examples include the number and composition of spectral stimuli at each window and different gestures made by the user during enrollment and authentication, etc.

Velody's training protocol balances between the effort in generating non-static biometrics and the size of the CRP pool to cover the user's authentication needs. According to a recent user study about daily authentication behavior [25], the average biometric authentication frequency is about 20 times per week for each user. Velody can enroll 100 CRPs, each lasting for 200 ms, within 30 minutes. These CRPs can cover the user's authentication needs for 5 weeks.

Emerging attacks: Although we assumed an attacker with strong capabilities, except obtaining a precisely replicated physical model of the victim's hand, we cannot completely eliminate the possibility of more sophisticated attacks in the future. Existing methods of nonlinear system modeling like Hammerstein model, mostly work only in a constrained and controlled scenario. These methods rely on sufficient measurement, specially designed excitation, etc., for limited objectives, such as assessing the total harmonics distortion, instead of covering all nonlinear dynamics like non-analytic intermodulation. We can also consider neural network-based modeling methods, such as voice or music synthesis. However, they typically require a mature auditory model or sufficient training [10, 37], which require the adversary much more effort and stronger capabilities. We believe these attacks are applicable to Velody's scenario.

8 RELATED WORK

In this section, we revisit previous effort on biometric authetnication, where we both qualitatively and quantitatively compare Velody with the state-of-art to show Velody's contribution.

Traditional biometrics can be categorized into physiological biometrics and behavioral biometrics. Physiological characteristics like fingerprint, hand geometry, iris structure, or physiological signals like electroencephalogram (EEG), electrocardiogram (ECG), and electromyogram (EMG), have been used as biometrics [2, 38, 45]. Behavioral biometric refers to unique characteristics preserved in human dynamics such as gesture dynamics, speech, or gait [7, 11, 32, 34], which are easy to acquire.

In Table 1, we compare several state-of-art biometric authentication systems with Velody. The works are divided by protocols, namely physiological, behavioral, and challenge-response. Note that the biometric-based challenge-response protocol here also relies on physiological properties of users but leveraging unique, passive, and varying responses to different stimuli. Following attributes are listed together: modality, FNR, FPR by impersonation, FPR by replay and synthesis. If the EER between falsely rejecting user samples and accepting impersonator is available, it is reported as FNR and FPR (impersonation) separately.

In Cardiac Scan [20], authors exploit sensing capability of a DC-coupled continuous wave radar to sense unique motion pattern of users' hearts and achieve an EER as low as 4.42%. Note that the FPR (impersonation) reported here originates from zero-effort

Table 1: Comparison among biometric systems. (∗: zero-effort impersonator; †: reduced replay quality; ‡: static user features.)

Work	Protocol	Modality	FNR	FPR		
				Impersonation	Replay	Synthesis
Cardiac Scan [20]	Physiological	Radar-measured heart motion	4.42%	4.42%*	N.A.	N.A.
Wang et al. [40]	Physiological	Heartbeat-induced vibration	2.48%	2.48%*	N.A.	N.A.
BiLock [46]	Behavioral	tooth click sound	<5%	<1.5%	5.6%†	N.A.
BreathPrint [4]	Behavioral	Breathing gesture-induced sound	6%	2%	2%†	N.A.
Taprint [5]	Behavioral	Tapping-induced vibration	1.74%	1.74%	N.A.	N.A.
VibWrite [21]	Behavioral	Vibration response of dynamic gestures	10%	2%	N.A.	N.A.
Sluganovic et al. [33]	Challenge-response‡	Reflective eye movement	6.3%	6.3%	0.06%	N.A.
Brain Password [19]	Challenge-response	Electroencephalogram	2.503%	2.503%	0.789%	N.A.
Velody (this work)	Challenge-response	Vibration response	5.8%	5.8%	0%	0%

impersonators since it is not possible to mimic one's heartbeat. Similar characteristics of heartbeat are utilized in [40] with heartbeat-induced vibration captured by smartphones. The EER is as low as 2.48% against zero-effort impersonator. However, this protocol may not be applicable to defend against replay and synthesis attacks in Velody's threat model where static biometic features may be leaked through a compromised channel.

The authors of [46] harvest unique sound from a tooth click recorded by commodity devices and achieve good consistency and security through a comprehensive user study and evaluation. With an increasing replay distance, the FPR of replay attack decreases to 5.6%. The authors of BreathPrint [4] utilizes distinction in users' breath, and three types of breathing gestures–sniff, normal, and deep breathing are evaluated, whose FNR and FPR are 6% and 2% respectively. Chen et al. [5] designed a system named Taprint that uses vibration induced by finger tapping measured for user authentication, whose EER is as low as 1.74%. Liu et al. [21] leverage the facts that varying user gesture will change the frequency response measured from a vibrating surface and designed a generalizable platform called VibWrite for authenticating users by password, lock pattern, and gesture input. We report the FNR and FPR by using password input. Even under imitation attack when the password is leaked, the FPR is as low as 2%. Though in [4, 46], the authors acknowledge and evaluate the security against replay attack of recorded noisy biometric samples, they are not applicable to Velody's threat model where clean raw responses can be injected directly, since they discover that the efficacy of replay attacks on these biometrics is highly dependent on the quality of replaying.

In terms of protocol, our work is most similar to [19, 33], where the replay attack of raw responses can be stopped by adopting a challenge-response protocol with changing visual stimuli that elicits unique passive reflective eye movement for each challenge and each user. The system achieves an EER of 6.3% against impersonation, and it rejects almost all replay samples. Note that though a challenge-response protocol is used, the security against synthesis attack is guaranteed by the high complexity of synthesizing eye movement because features used to verify user identity from different responses are still static. Hence, this modality may not be suitable for Velody's use case as well. Also, a similar protocol is implemented by using the event-associated electroencephalogram to generate vision-related challenge-response pairs, achieving good accuracy. The cognitive factors involved in the user enrollment, however, restricted the number of responses gathered within a satisfactory time [19]. Velody takes the advantage of the challenge-response protocol and the modality of hand-surface vibration response to achieve robust authentication, where the physiological characteristics of hand and the nonlinearity in hand-surface vibration responses are utilized to generate numerous disposable CRPs for defending against various attacks including raw signal replay and even strong synthesis attacks. Velody attains low error rates while succeeding in rejecting all synthesized samples, too.

9 CONCLUSION

This paper verifies the feasibility of using the nonlinear response from hand-surface system for user authentication, relying on the unique physiological characteristics of human hand with a challenge-response protocol. By building the prototype of Velody and conducting extensive user experiments, we validate several properties of Velody regarding usability and security. First, Velody is able to achieve an EER against impersonation as low as 5.8% in long term, showing a negligible loss with 5.7% using short-term test trials, indicating good temporal permanence. Moreover, this result can be attained with reasonable training effort and negligible authentication time of a 200-ms challenge. Furthermore, we verify the scalability of Velody's disposable CRPs by examining the FNR and FPR of individual challenges and challenges of different complexities. More importantly, Velody succeeds in defending against all replay and synthesis attacks, benefiting from distinct features in each nonlinear response to a unique challenge.

Our findings suggest that Velody's non-static biometrics are robust even when strong attackers present. Nevertheless, to further improve the scalability, more effort should be put to investigate its performance on ubiquitous settings and the design space of CRPs.

ACKNOWLEDGMENTS

This work was supported by the Wisconsin Alumni Research Foundation and NSF under grants CNS-1719336 and CNS-1845469. We also acknowledge the contribution of our anonymous participants.

REFERENCES

[1] Surajudeen Adewusi, Subhash Rakheja, Patrice Marcotte, and Jérôme Boutine. 2010. Vibration transmissibility characteristics of the human hand–arm system under different postures, hand forces and excitation levels. *Journal of Sound and Vibration* 329, 14 (2010), 2953–2971.

[2] Corey Ashby, Amit Bhatia, Francesco Tenore, and Jacob Vogelstein. 2011. Low-cost electroencephalogram (EEG) based authentication. In *Proceedings of IEEE/EMBS Conference on Neural Engineering (NER)*. 442–445.

[3] Silvio Barra, Maria De Marsico, Michele Nappi, Fabio Narducci, and Daniel Riccio. 2019. A hand-based biometric system in visible light for mobile environments. *Information Sciences* 479 (2019), 472–485.

[4] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing acoustics-based user authentication. In *Proceedings of the ACM Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 278–291.

[5] Wenqiang Chen, Lin Chen, Yandao Huang, Xinyu Zhang, Lu Wang, Rukhsana Ruby, and Kaishun Wu. 2019. Taprint: Secure text input for commodity smart wearables. In *Proceedings of the ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*.

[6] Yimin Chen, Jingchao Sun, Rui Zhang, and Yanchao Zhang. 2015. Your song your way: Rhythm-based two-factor authentication for multi-touch mobile devices. In *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*. 2686–2694.

[7] Mohammad Omar Derawi, Claudia Nickel, Patrick Bours, and Christoph Busch. 2010. Unobtrusive user-authentication on mobile phones using biometric gait recognition. In *Proceedings of the IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*. 306–311.

[8] Ren G. Dong, Aaron W. Schopper, Thomas McDowell, Daniel E. Welcome, John Wu, W. Paul Smutz, Christopher M. Warren, and Subhash Rakheja. 2004. Vibration energy absorption (VEA) in human fingers-hand-arm system. *Medical engineering & physics* 26, 6 (2004), 483–492.

[9] Ren G. Dong, John Wu, and Daniel E Welcome. 2005. Recent advances in biodynamics of human hand-arm system. *Industrial health* 43, 3 (2005), 449–471.

[10] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of the International Conference on Machine Learning (ICML)*. 1068–1077.

[11] Anil Jain, Lin Hong, and Sharath Pankanti. 2000. Biometric identification. *ACM Communications* 43, 2 (2000), 90–98.

[12] Nathan D Kalka, Jinyu Zuo, Natalia A Schmid, and Bojan Cukic. 2006. Image quality assessment for iris biometric. In *Biometric technology for human identification III*, Vol. 6202. 62020D.

[13] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. Gazetouchpass: Multimodal authentication using gaze and touch on mobile devices. In *Proceedings of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*. 2156–2164.

[14] Muhammad Khurram Khan, Jiashu Zhang, and Xiaomin Wang. 2008. Chaotic hash-based fingerprint biometric remote user authentication scheme on mobile devices. *Chaos, Solitons & Fractals* 35, 3 (2008), 519–524.

[15] Tomi Kinnunen, Filip Sedlak, and Roman Bednarik. 2010. Towards task-independent person authentication using eye movement signals. In *Proceedings of the ACM Symposium on Eye-Tracking Research & Applications (ETRA)*. 187–190.

[16] Wolfgang Klippel. 2006. Tutorial: Loudspeaker nonlinearities–causes, parameters, symptoms. *Journal of the Audio Engineering Society* 54, 10 (2006), 907–939.

[17] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity bioacoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the ACM Annual Symposium on User Interface Software and Technology (UIST)*. 321–333.

[18] Zhengxiong Li, Zhuolin Yang, Chen Song, Changzhi Li, Zhengyu Peng, and Wenyao Xu. 2018. E-Eye: Hidden electronics recognition through mmwave nonlinear effects. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*. 68–81.

[19] Feng Lin, Kun Woo Cho, Chen Song, Wenyao Xu, and Zhanpeng Jin. 2018. Brain password: A secure and truly cancelable brain biometrics for smart headwear. In *Proceedings of the ACM Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 296–309.

[20] Feng Lin, Chen Song, Yan Zhuang, Wenyao Xu, Changzhi Li, and Kui Ren. 2017. Cardiac scan: A non-contact and continuous heart-based user authentication system. In *Proceedings of the ACM Annual International Conference on Mobile Computing and Networking (MobiCom)*. 315–328.

[21] Jian Liu, Chen Wang, Yingying Chen, and Nitesh Saxena. 2017. VibWrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 73–87.

[22] Peipei Liu and Hoon Sohn. 2017. Development of nonlinear spectral correlation between ultrasonic modulation components. *NDT & E International* 91 (2017), 120–128.

[23] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. 2018. Vocal resonance: Using internal body voice for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 1 (2018), 19.

[24] Jaime Lorenzo-Trueba, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen. 2018. Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. *arXiv preprint arXiv:1803.00860* (2018).

[25] Shrirang Mare, Mary Baker, and Jeremy Gummeson. 2016. A study of authentication in daily life. In *Proceedings of the USENIX Symposium on Usable Privacy and Security (SOUPS)*. 189–206.

[26] Parimarjan Negi, Prafull Sharma, Vivek Jain, and Bahman Bahmani. 2018. K-means++ vs. behavioral biometrics: One loop to rule them all. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.

[27] Saurabh Panjwani and Achintya Prakash. 2014. Crowdsourcing attacks on biometric systems. In *Proceedings of the USENIX Symposium on Usable Privacy and Security (SOUPS)*. 257–269.

[28] Jaebum Park, Nemanja Pažin, Jason Friedman, Vladimir M Zatsiorsky, and Mark L Latash. 2014. Mechanical properties of the human hand digits: Age-related differences. *Clinical Biomechanics* 29, 2 (2014), 129–137.

[29] Alexandre Presas, David Valentin, Eduard Egusquiza, Carme Valero, Mònica Egusquiza, and Matias Bossio. 2017. Accurate determination of the frequency response function of submerged and confined structures by using PZT-patches. *Sensors* 17, 3 (2017), 660.

[30] Marc Rébillat, Romain Hennequin, Etienne Corteel, and Brian FG Katz. 2011. Identification of cascade of Hammerstein models for the description of nonlinearities in vibrating devices. *Journal of Sound and Vibration* 330, 5 (2011), 1018–1038.

[31] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10, 1–3 (2000), 19–41.

[32] Michael Sherman, Gradeigh Clark, Yulong Yang, Shridatt Sugrim, Arttu Modig, Janne Lindqvist, Antti Oulasvirta, and Teemu Roos. 2014. User-generated free-form gestures for authentication: Security and memorability. In *Proceedings of the ACM Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 176–189.

[33] Ivo Sluganovic, Marc Roeschlin, Kasper B Rasmussen, and Ivan Martinovic. 2016. Using reflexive eye movements for fast challenge-response authentication. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 1056–1067.

[34] Yunpeng Song, Zhongmin Cai, and Zhi-Li Zhang. 2017. Multi-touch authentication using hand geometry and behavioral information. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. 357–372.

[35] G Edward Suh and Srinivas Devadas. 2007. Physical unclonable functions for device authentication and secret key generation. In *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*. 9–14.

[36] A Talarico, M Malvezzi, and Domenico Prattichizzo. 2014. Modeling the human touch: A FEM model of the human hand fingertips for haptic application. In *Proceedings of the COMSOL Conference*.

[37] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[38] Shreyas Venugopalan, Felix Juefei-Xu, Benjamin Cowley, and Marios Savvides. 2015. Electromyograph and keystroke dynamics for spoof-resistant biometric authentication. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*. 109–118.

[39] Alexander Voishvillo, Alexander Terekhov, Eugene Czerwinski, and Sergei Alexandrov. 2004. Graphing, interpretation, and comparison of results of loudspeaker nonlinear distortion measurements. *Journal of the Audio Engineering Society* 52, 4 (2004), 332–357.

[40] Lei Wang, Kang Huang, Ke Sun, Wei Wang, Chen Tian, Lei Xie, and Qing Gu. 2018. Unlock with your heart: Heartbeat-based authentication on commercial mobile phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 3 (2018), 140.

[41] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Mohammed Sahidullah, Aleksandr Sizov, Nicholas Evans, and Massimiliano Todisco. 2017. ASVspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing* 11, 4 (2017), 588–604.

[42] Vladimir Yu. Zaitsev, Lev A. Matveev, and Alex Matveyev. 2011. Elastic-wave modulation approach to crack detection: Comparison of conventional modulation and higher-order interactions. *NDT & E International* 44, 1 (2011), 21–31.

[43] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 57–71.

[44] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 1080–1091.

[45] Zhaoyang Zhang, Honggang Wang, Athanasios V Vasilakos, and Hua Fang. 2012. ECG-cryptography and authentication in body area networks. *IEEE Transactions on Information Technology in Biomedicine* 16, 6 (2012), 1070–1078.

[46] Yongpan Zou, Meng Zhao, Zimu Zhou, Jiawei Lin, Mo Li, and Kaishun Wu. 2018. BiLock: User authentication via dental occlusion biometrics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 2, 3 (2018), 152.