NAuth: Secure Face-to-Face Device Authentication via Nonlinearity

Xinyan Zhou[12*], Xiaoyu Ji[12*], Chen Yan[1], Jiangyi Deng[1], Wenyuan Xu[12†]
[1]USSLab, the College of Electrical Engineering, Zhejiang University
[2]Alibaba-Zhejiang University Joint Institute of Frontier Technologies
xinyanzhou,xji,yanchen,jiangyideng,wyxu@zju.edu.cn

***Abstract*—With the increasing prevalence of mobile devices, face-to-face device-to-device (D2D) communication has been applied to a variety of daily scenarios such as mobile payment and short distance file transfer. In D2D communications, a critical security problem is verifying the legitimacy of devices when they share no secrets in advance. Previous research addressed the problem with device authentication and pairing schemes based on user intervention or exploiting physical properties of the radio or acoustic channels. However, a remaining challenge is to secure face-to-face D2D communication even in the middle of a crowd, within which an attacker may hide. In this paper, we present `NAuth`, a nonlinearity-enhanced, location-sensitive authentication mechanism for such communication. Especially, we target at the secure authentication within a limited range such as 20 cm, which is the common case for face-to-face scenarios. `NAuth` contains a *verification scheme* based on the nonlinear distortion of speaker-microphone systems and a location-based *validation model*. The verification scheme guarantees device authentication consistency by extracting acoustic nonlinearity patterns (ANP) while the validation model ensures device legitimacy by measuring the time difference of arrival (TDOA) at two microphones. We analyze the security of `NAuth` theoretically and evaluate its performance experimentally. Results show that `NAuth` can verify the device legitimacy in the presence of nearby attackers.**

I. Introduction

Mobile devices are becoming increasingly prevalent in our daily life. With this growing trend, face-to-face Device-to-Device (D2D) communication has emerged and involves a pair of devices nearby to communicate directly, e.g., face-to-face mobile payment [1] and short distance file transfer.

In D2D communications, typically two devices share no secrets in advance, and it is important to ensure that they are indeed communicating with each other even if many other devices are around. Taking the mobile payment in Fig. 1 as an example, the payer device should authenticate the legitimacy of the payee device (cashing machine), under the risk of nearby attackers (fake cashing machines). Typically, standard protocols such as Bluetooth ask the payee to input a "code" provided by the payer, thereby ensuring the authentication of the payee. Such an approach mandates user intervention and the security cannot be guaranteed [2], [3].

To eliminate such levels of user intervention, alternative solutions are proposed for device authentication in D2D communications. One type of approaches is to extract a shared authentication key from the physical environment, such as radio channels [4]–[10] and human body [11], [12]. An alternative approach relies on verifying the identity of the device that is embedded in hardware or environment, e.g., Xie et al. [13] use acoustic channel response as a consistent identity.

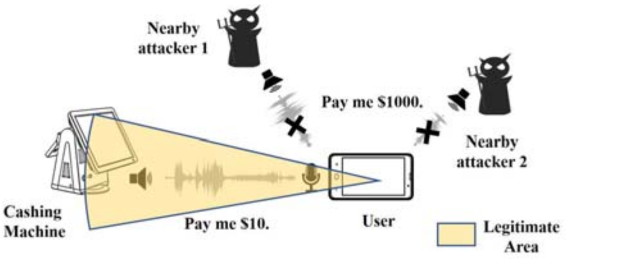


Fig. 1. A `NAuth`-based mobile payment scenario. `NAuth` can authenticate the legitimate device and detect nearby hidden attackers.

Although the aforementioned approaches improve the convenience by reducing user intervention, they may be bypassed by an attacker located close to the device to be authenticated [3]. In Fig. 1, for example, the fake cashing machine can impersonate the genuine one and have money transferred by the payer device. Such a threat is made possible because of the low location-sensitivity of the medium for key extraction, i.e., devices nearby may extract similar keys from the radio channel or the acoustic channel.

In this paper, we focus on the device authentication problem in face-to-face D2D communicationin the presence of nearby attackers. We propose `NAuth`, a nonlinearity-enhanced, location-sensitive authentication mechanism for secure authentication. The key insight of `NAuth` is to utilize the nonlinear distortions for authentication. Nonlinear distortions are essentially fine-grained and location-sensitive because they are combinations of multiple frequency harmonics. In particular, we extract the acoustic nonlinear distortions of the speaker-microphone system (SMS), which is common for current mobile devices. Moreover, a location-based security model is designed to shrink the legitimate area and to decrease the chance of attacks. The high-resolution nonlinearity feature works together with the location-based security model to eliminate any attacks within the legitimate area, and hereafter we name the two components the nonlinearity-based *verification scheme* and the location-based *validation model*.

The design of `NAuth` needs to explore the following questions. Firstly, *can nonlinear distortions be utilized for device authentication?* The basic requirement for device authentication is that the nonlinear distortion should be unique and device dependent. Moreover, the nonlinear distortion characteristic should be hard to imitate, otherwise the attacker can replay the signals easily. Secondly, *how to extract sufficient*

∗ Xinyan Zhou and Xiaoyu Ji are co-first authors.
† Corresponding faculty author.

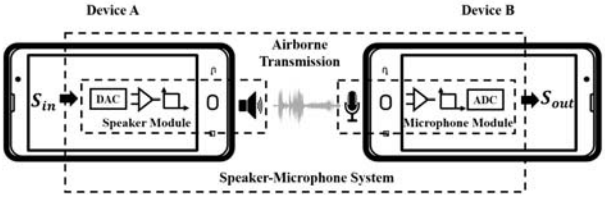


Fig. 2. A speaker-microphone system.

nonlinear distortion characteristics for device authentication? Even if nonlinear distortion can be used for device authentication, it is unknown whether and how it can be applied for real D2D applications. Last but not the least, *how to guarantee the extracted nonlinear distortion characteristics come from the legitimate device?* If the source legitimacy cannot be guaranteed, the extracted characteristics are thus invalid.

To tackle the questions above, we first explore the nonlinear distortions for speaker-microphone systems and validate the fact that nonlinear distortions are both device and location dependent, which are essential for device authentication. We derive unique patterns, i.e., the *acoustic nonlinearity patterns* (ANP), with an elaborately designed amplitude modulation (AM) signal. The verification scheme verifies device consistency during the authentication process. Besides, we design a lightweight location-based model to validate the source location by measuring the time difference of arrival (TDOA) at two microphones. NAuth can be utilized in various application scenarios, including mobile payment, data transmission, etc.

We summarize our main contributions as follows:

- We propose and validate that nonlinearity can be used as a fine-grained feature for device authentication with a speaker-microphone system.
- We design NAuth, a secure and location-sensitive device authentication mechanism for face-to-face D2D communications built on a nonlinearity-based verification scheme and a location-based validation model.
- We evaluate the performance and analyze the security of NAuth. Theoretical and experimental results prove the efficiency and security of our mechanism.

II. Nonlinearity of Speaker-Microphone Systems For Device Authentication

A. Nonlinearity in Speaker-Microphone System

Microphones and speakers are transducers that convert signals between acoustic and electrical states. For the purpose of user experience, stereo effect and noise cancelling, most smart devices (iPhone, Echo, etc.) are built with two or more modules of both microphones and speakers. For example, even early versions of smartphones (e.g., iPhone 5) have three microphones and two speakers [14]. Multiple signal processing circuits are utilized in microphone and speaker modules. Taking microphone module as an example, the converted electrical signals are processed by multiple stages of amplifiers and low-pass filters (LPF) before being sampled by the analog-to-digital converter (ADC).

For a *speaker-microphone System (SMS)*, the signal goes through three stages in the speaker-microphone channel in sequence—a speaker module, airborne transmission, and a microphone module, as shown in Fig. 2. Ideally, one can expect the speaker-microphone system to be linear, which means for a given input signal S_{in} at the speaker module, the output S_{out} at the microphone module is

$$S_{out} = AS_{in} \tag{1}$$

where A is the amplification factor.

However, real speaker-microphone systems are nonlinear because the signal processing circuits are made of nonlinear electronic components, e.g. transistors and the transducers are nonlinear [15], [16]. In general, a nonlinear system can be modeled as the following polynomial equation:

$$S_{out} = a_0 + a_1 S_{in} + a_2 S_{in}^2 + a_3 S_{in}^3 + \ldots = \sum_{i=0}^{\infty} a_i S_{in}^i \tag{2}$$

where a_i is the corresponding polynomial coefficient.

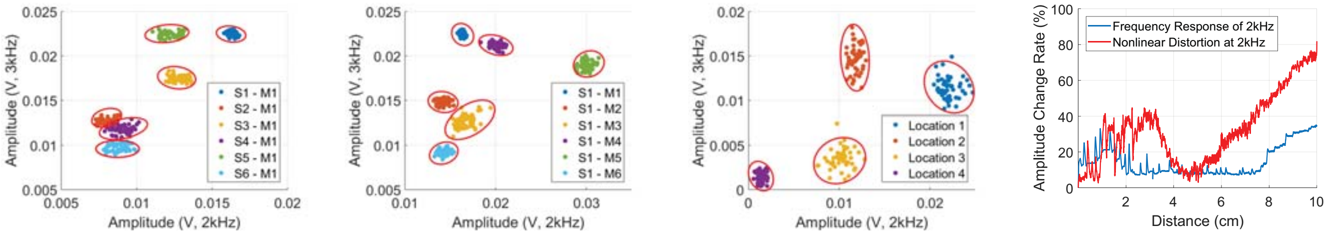
Speaker-Microphone System Nonlinearity. Besides the linear component $a_1 S_{in}$ in Eq. (2), S_{out} contains nonlinear distortions including a DC signal a_0 and $\{a_i S_{in}^i\}(i > 1)$, which are exponents of the input. Nonlinearity can deteriorate the output signals and has unexpected consequences. Despite the manufacturers' efforts in designing linear electronic components especially within the commonly used $100Hz$ to $10kHz$ frequency range, nonlinear distortion is still a common phenomenon among microphone and speaker modules.

B. Distinct Nonlinearity of Devices

The speaker-microphone system demonstrates inevitable nonlinearity and one can formulate the relationship between S_{out} and S_{in} by a vector, named the *nonlinear coefficient vector* $V = [a_0, a_1, a_2, \ldots, a_n]$. Essentially, V is determined by the physical structures of the nonlinear components, i.e., the CMOS chips [17] in both speaker and microphone modules (the nonlinerity casued by the air is ignored). As a result, the nonlinearity of a SMS varies among devices. Moreover, the nonlinearity can be easily observed and quantified because the nonlinear distortions are at different frequency bands from the original input signals. For example, let $S_{in} = \sin(2\pi f_0 t)$, the output of the speaker-microphone system S_{out} would have $2f_0, 3f_0, \ldots, nf_0$ frequency harmonics. Therefore, it is feasible for us to utilize the nonlinearity to identify a device (either the speaker or the microphone) in the speaker-microphone system. In this paper, we authenticate a device (Device A in Fig. 2) by looking at the nonlinearity of the signals received on the microphone side of Device B.

C. Feasibility Validation and Results

Experimentally, we validate the feasibility and effectiveness of speaker-microphone nonlinearity for device authentication. We experiment on 6 stand-alone microphone modules and 6 speakers, both of which are of the same model, and the details are shown in Fig. 8(a). The parameters of the speakers are 8Ω and $0.5W$, and each microphone module consists of a MEMS microphone chip ADMP401 [18], an impedance converter and an output amplifier. We stimulate the speakers with a $1kHz$ tone of $1.5Vpp$ from a function generator and collect the output signals of the microphone modules with a Keysight U2541A data acquisition card (DAQ) sampling at



(a) Speaker dependence.

(b) Microphone dependence.

(c) Location dependence.

(d) Location sensitivity comparison.

Fig. 3. The amplitudes of harmonics on $2kHz$ and $3kHz$ ($f_b = 1kHz$) when sending signals (a) from 6 stand-alone speakers to the same microphone, (b) from the same speaker to 6 different microphones, (c) with the same SMS at 4 locations. (d) The amplitude change rates for nonlinear distortion and fundamental frequency response at different speaker-microphone distances.

$100kHz$. We conduct the experiment in a quiet meeting room, and the distance between the microphones and speakers is $3cm$. In the following, we examine whether the nonlinearity is device dependent and location dependent, which are basic requirements for location-sensitive device authentication.

1) Device Dependence: We examine the nonlinearity behaviors of both the speaker and the microphone modules separately. a) We utilize 6 different speaker modules to stimulate an identical microphone, and b) we use the same speaker to stimulate 6 microphone modules under the above settings and record the frequency response at the microphone(s) side. For each SMS, we collect 50 $10ms$-long samples of the microphone output, perform Fast Fourier Transform (FFT) analysis on the 50 samples, and extract the amplitudes of the $2nd$ and $3rd$ harmonics, i.e., at $2kHz$ and $3kHz$ for simplicity (higher harmonics can also be utilized).

The results are shown in Fig. 3(a) and Fig. 3(b). Generally, samples from the same speaker-microphone system show well-marked clustering characteristic, and samples from different SMSes can be easily separated in a two-dimensional plane (i.e., 2kHz Amplitude as X-Axis and 3kHz Amplitude as Y-Axis). This confirms our assumption that both speakers and microphones share the nonlinearity-specific properties. Though the samples from speaker 2 and 4 (S2-M1 and S4-M1) in Fig. 3(a) partly overlap, they can be distinguished in a higher dimensional space by including more harmonics.

2) Location Dependence: To validate the location-dependence of nonlinearity, we transmit a $1kHz$ tone from the built-in speaker of a Huawei P10 Plus smartphone and record the frequency responses with an iPhone 6s at 4 different distances in a line. The two devices are lying on a table, with one's bottom speaker opposite to the bottom microphone of another. The distance from the Huawei smartphone (speaker) to the iPhone (microphone) is $1cm$, $3cm$, $5cm$ and $8cm$ respectively by moving the Huawei smartphone (we use commercial smartphones here because it is convenient to get moved). We extract the $2kHz$ and $3kHz$ harmonics at the microphone side. The results are shown in Fig. 3(c). We can find that at different distances, the nonlinear distortions are also clearly clustered. Moreover, larger distance results in weaker harmonics, nevertheless, they can be classified by involving more dimensions, i.e., higher harmonics.

3) Location Sensitivity: One may argue that the fundamental frequency response of a speaker-microphone system can also be utilized for authentication, i.e., by measuring the frequency response at $2kHz$ with a $2kHz$ input [13]. We demonstrate the advantage of nonlinearity-based approach over fundamental-frequency-based in terms of location-sensitivity, which enhances security. With the same setup in the location dependence experiment, we extract both amplitudes of 1) the harmonic signal at $2kHz$ stimulated by a $1kHz$ signal, and 2) the $2kHz$ fundamental frequency with a $2kHz$ input, and gradually increase the distance between the two devices from $0cm$ to $10cm$.

We formally define the *Amplitude Change Rate* of frequency f at distance d as:

$$\Upsilon(d)_f = \frac{|A(d)_f - A(d_0)_f|}{A(d_0)_f} * 100\% \qquad (3)$$

where $A(d)_f$ is the amplitude of frequency f at distance d and in our case $d_0 = 0cm$. The results are shown in Fig. 3(d). Compared with the fundamental frequency response, the $\Upsilon(d)_f$ of nonlinear distortion is higher, which coincides the location-sensitive property, and thus can be more secure.

D. Summary

The nonlinearity of speaker-microphone systems is demonstrated to be speaker/microphone specific as well as location-sensitive, which make it a natural candidature for device authentication. In Sec. IV, we elaborate the design details of using nonlinearity for device authentication.

III. Threat Model and Assumptions

NAuth is designed to secure the face-to-face D2D communication through location-sensitive authentication. The threat model involves two parties namely Alice and Bob that need to authenticate each other and an attacker Eve. For simplicity, we only consider the case that Alice authenticates Bob. Eve's purpose is to make Alice believe she is Bob while Alice and Bob share no common secrets in advance.

Without loss of generality, we have the following assumptions for Alice and Bob:

- Alice and Bob are physically close to each other, namely within $50cm$ or closer. The distance can vary across D2D application scenarios, e.g., mobile payment (within $20cm$) or secret information sharing (within $50cm$).
- Alice is for sure that any device within a restricted "legitimate area" is trustworthy and she can control the orientation of her device to guarantee Bob is in the

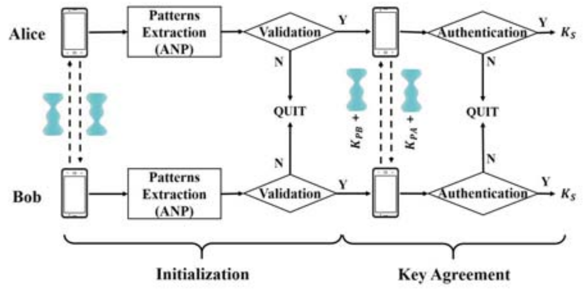


Fig. 4. A NAuth-based key establishment procedure.

"legitimate area", as illustrated in Fig. 1. The design details of this "legitimate area" can be referred to Sec. IV.
- Both parties' devices have speakers and microphones. The party who initiates an authentication, e.g. Alice here, should have two microphones at least.
- Both parties' devices should be relatively stationary during the authentication process.

For Eve, she has the following capabilities and assumptions:

- Eve is free to move anywhere around Alice and Bob. She can even hide her attack equipment in the pocket or under a book. However, neither Eve or her equipment can be between Alice and Bob in face-to-face scenarios.
- Eve is able to capture and inject signals at any stage of the authentication process, and thereby launch replay or man-in-the-middle attacks.
- Eve may be aware of the authentication mechanism.

From the above threat model, a successful authentication relies on three important assumptions: 1) the consistency of the authenticated device can be guaranteed, 2) there is a "legitimate area", and 3) the area is reliable to differentiate attackers such as Eve. In the next section, we elaborate the location-based validation model and the nonlinearity-based verification scheme which satisfy the above two requirements.

IV. Design

In this section, we first provide a system overview of NAuth and then describe our design in detail.

A. System Overview with Key Establishment as an Example

NAuth is a location-sensitive device authentication mechanism built on two key components: a *verification scheme* based on the nonlinearity of speaker-microphone systems and a location-based *validation model*. They mainly address two challenges respectively:

1) How to authenticate a device from the sound it generates?
2) Is the received sound generated by a legitimate device?

To illustrate, we give an example of a secure key establishment process implemented with NAuth. As shown in Fig. 4, Alice and Bob are two legitimate users who need to establish a session key between their devices. The process consists of two steps: initialization and key agreement. To initialize, the two devices send acoustic *authentication signals* to each other and extract nonlinearity patterns from the sounds they receive. Besides, they independently verify the legitimacy of received sounds with the location-based validation model. After that, they can exchange their public keys K_{PA}, K_{PB} via acoustic signals and derive the same session key K_s while constantly sending *declaration signals* to verify the consistency of nonlinearity patterns and validate the source legitimacy.

TABLE I
Notations

Parameter	Description
f_c	The frequency of the carrier in the AM signal.
f_b	The frequency of the baseband in the AM signal.
f_{LPF}	The cut-off frequency of the low-pass filter.
A	The signal amplitude.
$c_{(i,j)}$	The corresponding coefficient for $sin(2\pi j f_b t)$ after the trigonometric expansion of S_{in}^i.
ANP	The acoustic nonlinearity patterns.
L_2, L_1	The distances from the source device to the bottom microphone and the top microphone.
L_m	The distance between the top and bottom microphones.
c	The speed of sound, $340m/s$.
$L_{shoulder}$	The width of the user's shoulder.
D_{u2m}	The distance between the user and the bottom microphone.

In the following we only focus on the verification scheme and the validation model. The acoustic communication borrows existing schemes such as Dolphin [19] and the secret key exchange can be achieved by the Diffie-Hellman protocol [20]. We first summarize all the notations in Tab. I for clear presentation.

B. Nonlinearity-Based Verification Scheme

The nonlinearity-based verification scheme extracts **acoustic nonlinearity patterns** (ANP) from an authentication signal to verify an identity.

Recall that in Eq. (2), we denote the nonlinear coefficient vector $V = [a_0, a_1, \ldots, a_i, \ldots]$ to describe the relationship of the input and the output signals. a_i is the gain of the $i-th$ harmonic and is observable in the frequency domain of the output signal. Intuitively, one can use V for nonlinearity pattern extraction. However, calculating a_i directly is difficult because harmonic at a certain frequency is a combination of sub-frequencies. For example, the harmonic frequency $3kHz$ can be from $1kHz$ and $1.5kHz$.

1) Acoustic Nonlinearity Patterns (ANP): Considering an input signal $S_{in} = sin2\pi f_0 t$, the new frequency components in S_{out} contain $\{f_0, 2f_0, \ldots, nf_0\}, n \in N^+$. Despite of the ambient noise, the amplitudes of these new frequency components are linear combination of the nonlinear coefficient vector, which can be presented as:

$$A(nf_0) = \sum_{i=1}^{\infty} A^i a_i c_{(i,n)} \tag{4}$$

where $A(nf_0)$ is the amplitude of the nf_0, A^i is the signal gain and $c_{(in)}$ is a constant determined by the input signal, which can be calculated by trigonometric expansion. For example, if $S_{in} = \sin(2\pi 1000t)$, $nf_0 = \{1kHz, 2kHz, \ldots, nkHz\}, n \in N^+$, and for $i = 3$, we have $c_{(3,1)} = 0.75$, $c_{(3,2)} = 0$ and $c_{(3,3)} = -0.25$. Based on this, the amplitudes of harmonics can be an alternative nonlinearity pattern. Thus, we define the acoustic nonlinearity patterns (ANP) as:

$$ANP = \ [\ A(f_{nd})\ ] \tag{5}$$

For the same input as above, $ANP = [A(1kHz), A(2kHz), \ldots, A(nkHz)]$. Extracting ANP is feasible for mobile devices while they only need to apply FFT and extract the amplitudes of new frequency components after nonlinear distortions from authentication signals. In order to derive a fine-grained ANP, we elaborately design the authentication signal in the following.

2) Extracting ANP by AM Modulation: Normally, nonlinear distortions of signals at frequencies within $100Hz$ to $10kHz$ are elaborately relieved by manufacturers for the purpose of user experience. To investigate, we measure the frequency response of a microphone module (the same as the one used in Sec. II) with a $1kHz$ input signal and show the normalized amplitudes in Fig. 5(a) and it demonstrates weak harmonics at frequencies like $2kHz$, $3kHz$ and etc. On the other hand, signals above $10kHz$ can only produce limited harmonics due to the low-pass filters (with a cutoff frequency at $20kHz$). As a result, we modulate a baseband signal upon a carrier signal whose frequency is far above than $10kHz$ (e.g., at $20kHz$) to produce significant harmonics, as is used in [15]. In NAuth, we exploit amplitude modulation (AM) and notate the carrier and baseband frequencies f_c and f_b, the AM signal is presented as:

$$S_{in} = A_{f_c} sin(2\pi f_c t)(1 + A_{f_b} sin(2\pi f_b t)) \tag{6}$$

where A_{f_c} and A_{f_b} are the amplitudes of the carrier and baseband signals. Considering the nonlinear relationship of S_{in} and S_{out} in Eq. (2), when the input signal is an AM signal, the new frequency components for the $i-th$ exponent in S_{out} are:

$$f_{S_{in}^i} = \begin{cases} kf_c,\ mf_c \pm nf_b, & k \in \{1,3,\ldots,i\}, \quad i\ is\ odd \\ & m \in \{1,3,\ldots,i\} \\ & n \in \{1,2,\ldots,i\} \\ kf_b,\ mf_c \pm nf_b, & k \in \{1,2,\ldots,i\}, \quad i\ is\ even \\ & m \in \{2,4\ldots,i\} \\ & n \in \{0,1,2,\ldots,i\} \end{cases} \tag{7}$$

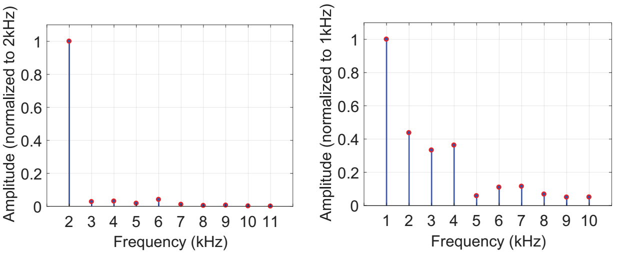
where $f_{S_{in}^i}$ is the frequency components in S_{in}^i.

From Eq. (7), we can find that there are abundant harmonic components whose frequencies can be below $20kHz$. To name a few, kf_b and $(f_c - nf_b)$ can produce frequencies less than $20kHz$. The ANP is actually extracted from those new low-frequencies ($< 20kHz$) composed of harmonic components of other frequencies under AM modulation.

To validate, we send an AM signal with f_c=$20kHz$ and f_b=$1kHz$ to a microphone module. The frequency response is shown in Fig. 5(b) with the same setting as in Fig. 5(a). Compared to Fig. 5(a), the $1kHz$-generated harmonics at $2kHz$, $3kHz$ and even $8kHz$ demonstrate strong power than non-modulated input signal. Therefore, an AM signal can enhance the nonlinear distortions which is beneficial to the ANP extraction.

3) AM Parameters: In the modulation process, three parameters, i.e., f_c, f_b and modulation depth, should be carefully selected to improve the effectiveness of ANP extraction.

f_c and f_b. The maximum value of f_c is constrained by the sampling rate of DAC in the speaker module. Based on the Nyquist sampling theorem, f_c should be less than the half of the sampling rate. Furthermore, to capture enough nonlinear distortions, intuitively f_b should be as small as possible while f_c should be the opposite. Due to the constraint of low-pass filters, we should have:



(a) From a $1kHz$ signal.

(b) From an AM signal.

Fig. 5. The frequency response of (a) a signal at $1kHz$ and its harmonics and (b) the nonlinear distortions of an AM signal with $f_c = 20kHz$ and $f_b = 1kHz$. Higher harmonics can be extracted when the signal is modulated.

$$\begin{aligned} N_{fh} \cdot f_b &\le f_{LPF} \\ f_c - f_b &\le f_{LPF} \end{aligned} \tag{8}$$

where N_{fh} is the space of available harmonics, and typically we prefer a larger N_{fh} to extract efficient nonlinear patterns. The second condition should be satisfied because $f_c - f_b$ also contributes relatively strong harmonics than others. Typically, the sampling rates of devices are higher than $44.1kHZ$, and in our implementation, we select f_c=$20kHz$ and f_b=$1kHz$ empirically.

Modulation depth. Modulation depth is defined as the ratio of the baseband amplitude to the carrier amplitude, i.e., A_{f_b}/A_{f_c}, which impacts the strength of nonlinear distortions. As revealed by Zhang et al. [15], modulation depth should be set to 100% to achieve the best nonlinear distortion and we refer to this setting in our work.

4) Device Verification: To verify the authentication consistency of a device, NAuth requires devices to send declaration signals proactively. We exploit the Euclidean distance to determine whether two ANPs are consistent, specifically, the distance (d) between ANP_1 and ANP_2 is defined as:

$$d_{12} = \sqrt{\sum_{i=1}^{N} (ANP_1(i) - ANP_2(i))^2} \tag{9}$$

where N is the dimension of ANP. If d is smaller than a predefined threshold σ, one can accept the authentication consistence, otherwise a new authentication should be performed. In NAuth, we set $\sigma = 10$ and explain it in Sec. VI.

C. Location-Based Validation Model

The nonlinearity-based verification scheme can ensure authentication consistency. However, it cannot confirm the legitimacy of the device. If an attacker (Eve) sends the authentication signal before a legitimate device (Bob) does, the attacker (Eve) can impersonate Bob because Alice fails to differentiate them. Therefore, the location-based validation model is introduced to distinguish legitimate devices.

1) TDOA-based Validation: The location-based validation model utilizes two microphones embedded in devices. Typically, devices like smartphones and intelligent speakers are designed with at least two microphones to support various applications. We notice that when we record acoustic signals with both microphones, there is always a time difference,

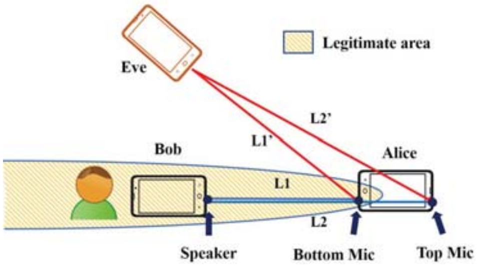


Fig. 6. A device (Bob) inside the legitimate area shows higher TDOA at the two microphones of the authenticator device (Alice) than a device (Eve) outside the legitimate area because $(L_2 - L_1) > (L_2^{'} - L_1^{'})$.

a.k.a. TDOA (time difference of arrival) because the distances between the signal source and two microphones are different.

An illustration of the location-based validation model is shown in Fig. 6. Signals from the speaker at the legitimate sender side are assumed to be along the connecting line of the two microphones at the receiver side. By measuring the TDOA at two microphones, we can approximately estimate the location of the signal source.

$$TDOA = \frac{L_2 - L_1}{c} = \frac{L_m}{c} \tag{10}$$

where L_1 and L_2 are the distances from the source device to the bottom and top microphones respectively, c is the speed of sound and L_m is the distance between the two microphones. Both L_m and c are constants.

2) Legitimate Area: According to Eq. (10), the TDOA of an attacker (Eve) is smaller if she is not located on the connecting line of two microphones because $L_2^{'} - L_1^{'} < L_2 - L_1$, where $L_2^{'}$ and $L_1^{'}$ are distances from Eve's speaker to the two microphones. Therefore, $TDOA_{Eve}$ is smaller than $\frac{L_m}{c}$.

NAuth measures the TDOA by comparing the signal arrival time at two microphones, and the precision of the TDOA is constrained by the device sampling rate f_{sp}, i.e., the measurement of TDOA may have a maximum error of $1/f_{sp}$. Taking the accuracy error into consideration, NAuth validates the source device as a legitimate device if its TDOA satisfies the following constraint:

$$TDOA \geq \frac{L_m}{c} - \frac{1}{f_{sp}} \tag{11}$$

Therefore, points (P) in the legitimate area satisfy:

$$|PM_{top}| - |PM_{bottom}| \geq (L_m - \frac{c}{f_{sp}}) \tag{12}$$

where M_{top} and M_{bottom} are the top and the bottom microphones of the receiver. The boundary of the legitimate area is the equality condition of Eq. (12), which is the left branch of a hyperbola[1] with bottom and top microphones as the two foci. Thus, the device with TDOA satisfying Eq. (11), i.e., located inside the left branch of the hyperbola (the shaded area in Fig. 6), is considered as a legitimate device.

3) User Experience of Legitimacy Validation: The size of the legitimate area is a tradeoff between user experience and security. If the area is too small, legitimate users need to put two devices on a strict straight line to pass the validation, which is hard for users. On the contrary, a bigger area may leave space for attackers.

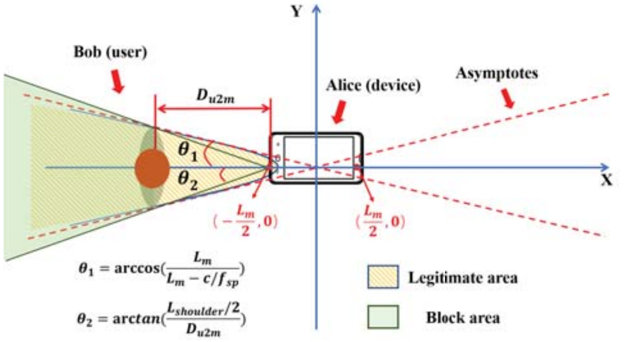


Fig. 7. The boundary of the legitimate area can be approximated to the red dashed asymptotes of the hyperbola. A user (Bob) can essentially block all the legitimate area behind him if $\theta_2 \geq \theta_1$.

Based on the analysis above, the size of the legitimate area is covered by the left branch of a hyperbola. To quantify the legitimate area, we introduce a Cartesian coordinate such that the origin is the center of two microphones and the x-axis is the main axis. We have the bottom microphone as $F_1 = (-\frac{L_m}{2}, 0)$ and the top microphone as $F_2 = (\frac{L_m}{2}, 0)$. With two fixed foci, the hyperbola approaches two asymptotes (red dash lines in Fig. 7) and the shape of the hyperbola is bounded by its asymptotes.

As shown in Fig. 7, we can approximately consider the legitimate area to be within the two asymptotes (the shaded area), and θ_1 is the tolerance of the speaker-to-microphone angle. With basic geometric knowledge, we have:

$$\theta_1 = arccos(1 - \frac{c}{f_{sp}L_m}). \tag{13}$$

where c is the speed of sound (approximately $340m/s$), L_m is a constant related to the device size and f_{sp} is higher than $44.1kHz$ for most of devices. Taking a mobile device with $L_m = 14cm$ as an example (L_m is $13cm$ for iPhone 6s and $15cm$ for iPhone 6s Plus), θ_1 is $19.1°$. Therefore, the tolerance range for the speaker-to-microphone angle is $[-\theta_1, \theta_1]$, which is $[-19.1°, 19.1°]$ in this case. We believe this range is big enough for users when we require them to put the speaker and the microphone on a straight line.

V. Security Analysis

In this section, we analyze the security of NAuth from the following perspectives.

- Can attackers bypass the location-based validation model and impersonate a legitimate user?
- Can attackers deceive the verification scheme and launch replay or man-in-the-middle attacks?

A. Security of the Location-Based Validation Model

If the attacker is outside the legitimate area, she cannot satisfy the requirement imposed by Eq. (11). Even if she exploits multiple speakers, it is extremely difficult to beamform a sound at one microphone without getting it received by another one close by, e.g., $15cm$ away for a smartphone.

[1] A hyperbola is a set of points (P) that have a constant absolute difference between $|PF_1|$ and $|PF_2|$, where F_1 and F_2 are two fixed points (the foci).

Therefore, a more threatening scenario is when the attacker is inside the legitimate area. Since attackers cannot locate between the legitimate users due to the risk of getting visually exposed, we only consider the situation that attackers are behind a legitimate user. In this scenario, the user interaction is considered. Since users are required to manually align the devices, naturally we can assume that they sit or stand behind their devices. Users can block all line-of-sight transmissions of acoustic signals behind them because very few acoustic energies can penetrate through the human body. We highlight the block area in Fig. 7, and the boundary of the block area is the line that connects the bottom microphone and the user's shoulder. The angle between the boundary and the x-axis is:

$$\theta_2 = arctan(\frac{0.5 * L_{shoulder}}{D_{u2m}}) \quad (14)$$

where $L_{shoulder}$ is the width of user's shoulder and D_{u2m} is the distance between the user and the bottom microphone. When $\theta_2 \geq \theta_1$, the user can block all attackers behind her even if the attacker is located in the legitimate area. With Eq. (13) and Eq. (14), we have:

$$D_{u2m} \leq \frac{0.5 * L_{shoulder}}{tan\theta_1} \quad (15)$$

Given $\theta_1 = 19.1°$ and consider a shoulder width of $36cm$ (an average for adult females), we derive $D_{u2m} \leq 51.98cm$, which is typical for face-to-face scenarios. Thus, the location-based validation model is efficient to detect attackers outside and even inside the legitimate area behind users.

B. Replay Attack

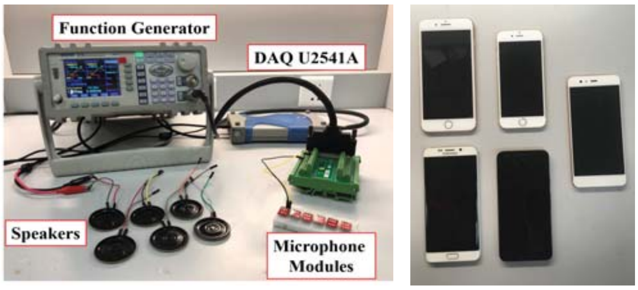
An attacker can receive acoustic signals sent by Alice and Bob, thus can also extract the ANPs. By doing so, she may replay an elaborately designed authentication signal with the previously extracted ANPs. However, as discussed in Sec. II, nonlinearity is location dependent, thus the attacker cannot extract ANPs same as the legitimate users'. Another reason is that the acoustic signal attenuates exponentially during the propagation, thus the attacker at different locations receives different input signals from the legitimate users. Moreover, even if the attacker goes back to the same location after Alice leaves, she still cannot obtain same features as Bob's because nonlinearity is device dependent.

C. Man-in-the-Middle Attack

An attacker may launch man-in-the-middle attacks by impersonating both Alice and Bob at the same time. To do this, the attacker should simultaneously pair with both legitimate devices successfully, which means that she can only locate in the overlapped legitimate areas for both users, i.e., between them. Thus, such attacks are unfeasible.

VI. Evaluation

In this section, we conduct extensive experiments to evaluate the efficiency of NAuth. For the nonlinearity-based verification scheme, we emulate a mobile payment scenario under different experiment settings. Besides, we evaluate the location-based validation model with two smartphones. The experiment setups and tested devices are summarized in Tab. II.



(a) Equipment and stand-alone modules. (b) Tested smartphones.

Fig. 8. Experiment settings for (a) feasibility validation and (b) evaluation.

TABLE II
Summary of Experiment Setups

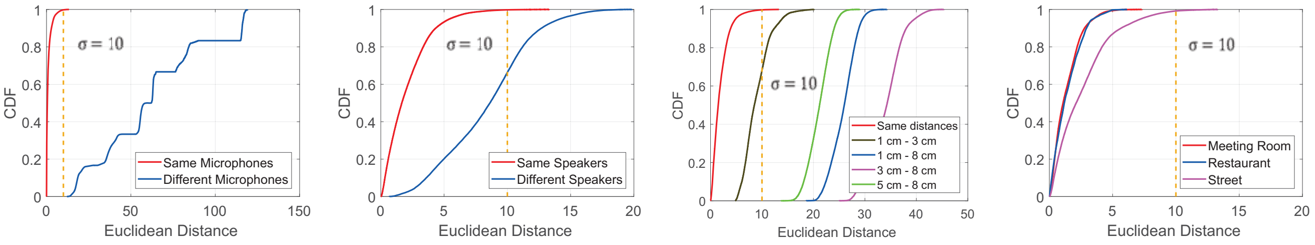
	Model	L_m (mm)	f_{sp} (kHz)	$2\theta_1$
Devices	Apple iPhone 8P	148	48	35.6°
	Apple iPhone 6S	130	48	38°
	Samsung Galaxy S6 Edge	149	48	35.5°
	Google Nexus 5X	140	48	36.6°
	Huawei P10 Plus	145	44.1	37.5°
Scenarios	(1) Quiet meeting room; (2) restaurant; and (3) the street.			
Distances	1cm, 3cm, 5cm and 8cm.			

A. Efficiency of the Verification Scheme

We evaluate the verification scheme with a mobile payment scenario shown in Fig. 1, where a smartphone (the receiver) needs to authenticate a cashing machine (the sender). We envision that NAuth can be applied to various types of acoustic D2D communications, including the trending ultrasonic communication. Therefore, we emulate 4 ultrasound-capable cashing machines with 4 ultrasonic speakers and choose the authentication signal to be an AM signal with $f_c = 20kHz$ and $f_b = 1kHz$. The signals are received on 4 smartphones (iPhone 8P, iPhone 6S, Galaxy S6 Edge and Nexus 5X) shown in Fig. 8(b). For each sender-receiver pair (SMS) in each setting, we collect 300 sets of ANPs and compare the Euclidean distances. We consider four settings that may affect the performance—different receivers, senders, distances and noise levels, which correspond to four assumptions: 1) different customers at the same store, 2) a customer at different stores, 3) a customer pays multiple times at the same store, and 4) payments are performed under different background noises. We investigate the four settings separately in the following.

1) The Impact of Receivers: We send authentication signals from the same speaker and utilize four smartphones as receivers respectively at a distance of $3cm$. We calculate the Euclidean distances of ANPs from the same SMSes ($d(i,i)$, $i \in [1,4]$) and different SMSes ($d(i,j)$, $i,j \in [1,4]$ & $i \neq j$). We show the CDF of Euclidean distances of both cases in Fig. 9(a). The Euclidean distances between the same SMSes are significantly smaller than between different SMSes.

2) The Impact of Senders: Similarly, we send authentication signals from four speakers and receive with the same iPhone 6S. Results in Fig. 9(b) show that over 95% of Euclidean distances from the same SMSes are smaller than 5 while it is 14 for different SMSes. We expect the Euclidean distances between different senders to be bigger if different models of ultrasonic speakers are used.



(a) Varying receivers.

(b) Varying senders.

(c) Varying distances.

(d) Varying noise levels.

Fig. 9. The CDF of Euclidean distances (a) between the same and different receivers; (b) between the same and different senders; (c) between the same and different distances; (d) between the same SMSes under different noise levels.

3) The Impact of Distances: We send from the same speaker to four smartphones at four distances: 1cm, 3cm, 5cm and 8cm. We show the ANP Euclidean distances of the same SMSes at the same and different distances in Fig. 9(c). The results reveal that distance can significantly affect the ANPs. For ANPs at the same distance, the Euclidean distances are smaller than 10, while they increase significantly even if the device moves slightly. The experiment results also indicate that a bigger movement of devices does not necessarily represent a higher Euclidean distance. Thus, we suggest that users do not move the device during the NAuth authentication.

4) The Impact of Noise Levels: We conduct experiments on the same SMS at three places including a quiet meeting room, a restaurant and the street. The average noise levels at the three places are 38.8, 58.2 and $73.7dB$ SPL. As shown in Fig. 9(d), the ANP Euclidean distances on the street is higher than in the meeting room and restaurant, therefore the ambient noise can interfere with the ANPs. Nevertheless, the ANP Euclidean distances are no more than 10 for all scenarios, which indicates that the efficiency will not be affected.

Summary. Experiment results in Fig. 9 show that the Euclidean distances of ANPs from the same SMSes are generally smaller than 10 even in noisy environments, while different SMSes have significantly higher ANP Euclidean distances. Thus, we can set the threshold σ of Euclidean distances to 10 for practical device authentication in NAuth.

B. Efficiency of the Validation Model

We record acoustic signals with a Huawei P10 Plus and measure the TDOA at the top and bottom microphones separated by $145mm$. According to Eq. (11), the TDOA for legitimate devices should be higher than $0.381ms$, which takes approximately 17 sample points at $44.1kHz$.

1) Legitimacy Validation: We send a $500Hz$ tone with an iPhone 6S at 196 locations around the receiver as illustrated in Fig. 10. We mark a red dot on the locations that pass the validation model, and mark a cross for those that fail. We mark the bottom and top microphones of the receiver with diamonds and plot the theoretical legitimate area (37.5°) with yellow shadow. Experiment results show that the passed locations concentrate in a small area, which approximates the theoretical legitimate area. With a second experiment, most false negatives (crosses in the legitimate area) can be eliminated.

To investigate the overall efficiency when multiple attempts are possible, we select 9 locations from the legitimate area, boundary line and rejection area. At each location, we measure the TDOA for 200 times and calculate a pass rate. The results reported in Fig. 11 demonstrate that devices in the legitimate area can pass the validation model easily while it is hard on the boundary and almost impossible in the rejection area.

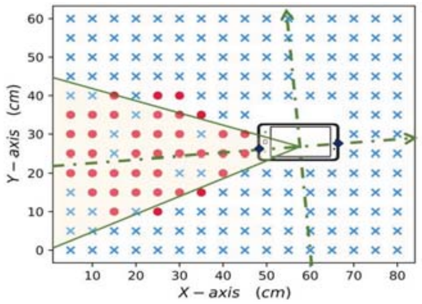


Fig. 10. Results of the validation model running on a Huawei P10 Plus receiver (with two marked microphones near coordinates (50,25) and (65,30)) tested with the speaker of an iPhone 6S at 196 locations around it. The passed and rejected locations are marked with red dots and blue crosses.

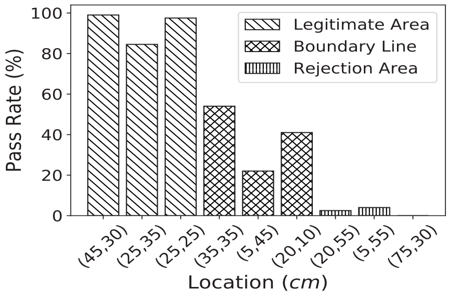


Fig. 11. Rates of passing the validation model at 9 locations in the legitimate area, boundary line and rejection area.

2) User Intervention: We evaluate the efficiency when a user blocks the sound from an attacker behind her but in the legitimate area. A user with a $36cm$ shoulder width sits between the receiver (a Huawei P10 Plus) and an attacker (an iPhone 6S). We place the attacker at 5 locations (listed in Tab. III) in the legitimate area and calculate the pass rates with and without the user as an obstacle. Results show that with the user as an obstacle, the pass rates drop significantly, therefore the user intervention in NAuth is sufficient to prevent attackers in the legitimate area from passing the validation model.

C. Discussion

Time Overhead. NAuth requires users to send authentication and declaration signals and can impose time overhead on the D2D communication. Considering a frequency resolution of $100Hz$ in performing the Fourier Transform (as in our

TABLE III

PASS RATES IN THE LEGITIMATE AREA WITH AND WITHOUT A USER

Locations (cm)	(15,20)	(10,30)	(10,25)	(5,30)	(5,25)
Pass Rate (%) w/o User	91.5	89	85	88	96
Pass Rate (%) w/ User	8.5	0.5	0	0.5	0

experiment), a $10ms$ sample is required. If we average the ANPs of 5 authentication signals for initialization and use 5 declaration signals for verification, the total time overhead is $100ms$, which is acceptable for most application scenarios.

Ambient Noise. Random ambient noises can interfere with the ANP and may affect its consistency over time. However, the application scenarios for NAuth can generally be finished in 1 second, e.g., mobile payment and key establishment, therefore the impact of the ambient noises can be limited.

Device Requirement. The authentication-initiating device should have two microphones in order to measure the TDOA. NAuth is inapplicable to devices with only one microphone.

User Requirement. NAuth requires users to put the devices into the legitimate area and hold them still during the authentication process, which sometimes might be tricky.

VII. RELATED WORK

Extensive research has been proposed for establishing secure D2D communications mainly from three perspectives—proximity, hardware fingerprint, and covert channel.

The proximity-based approaches extract symmetric keys from properties of the wireless channels such as RSS (received signal strength) [4]–[6] and CSI (channel state information) [7]–[10]. Compared with the RSS-based mechanisms, CSI-based ones are more efficient because they can derive finer-grained physical layer information, e.g., the channel response from multiple subcarriers of Orthogonal Frequency-Division Multiplexing (OFDM). However, such methods rely on dedicated hardware (Intel 5300 Wi-Fi card) and cannot be widely implemented on mobile devices.

A number of studies have shown that mobile devices can be fingerprinted with inherent hardware modules including accelerometers [21], microphones [22] and speakers [23]. Although these hardware fingerprints are inimitable, undeniable and stable, the authentication requires prior extraction of features and trained classifiers, thus they are inapplicable to D2D communication when no secret is shared in advance. Xie et al. [13] proposed a key establishment mechanism based on the acoustic channel response of devices. Their methods assume the attackers to be outside a certain range and may not suffice to detect hidden attackers nearby in face-to-face D2D communications.

Roeschlin et al. [11] and Chang et al. [24] exploit secure body channels for key establishment. However, extra hardware like electrodes and on-body sensors are required while NAuth only relies on built-in microphones and speakers.

VIII. CONCLUSION

We propose NAuth, a nonlinearity-enhanced, location-sensitive authentication mechanism for secure face-to-face D2D communication. NAuth consists of two main components: a nonlinearity-based verification scheme and a location-based validation model. We extract acoustic nonlinear patterns (ANP) to verify device consistency in the verification scheme and measure the TDOA at two microphones to guarantee device legitimacy in the validation model. Theoretical analysis and experiment results demonstrate NAuth can authenticate devices efficiently in the presence of nearby attackers.

IX. ACKNOWLEDGMENT

This work is supported by NSFC Grant 61702451, ZJNSF Grant LGG19F020020 and Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

REFERENCES

[1] J. Roberts, “Different types of mobile payments explained,” https://www.mobiletransaction.org/different-types-of-mobile-payments/, 2018.

[2] Y. Jiang, Z. Li, and J. Wang, “Ptrack: Enhancing the applicability of pedestrian tracking with wearables,” in *in Proceedings of the ICDCS*, 2017.

[3] Y. Liu and Z. Li, “aleak: Privacy leakage through context-free wearable side-channel,” in *in Proceedings of the INFOCOM*, 2018.

[4] Y. Luo, L. Pu, Z. Peng, and Z. Shi, “Rss-based secret key generation in underwater acoustic networks: advantages, challenges, and performance improvements.” *IEEE Communications Magazine*, vol. 54, no. 2, pp. 32–38, 2016.

[5] H. Liu, J. Yang, Y. Wang, Y. J. Chen, and C. E. Koksal, “Group secret key generation via received signal strength: Protocols, achievable rates, and implementation,” *IEEE TMC*, vol. 13, no. 12, pp. 2820–2835, 2014.

[6] T. Wang, Y. Liu, and A. V. Vasilakos, “Survey on channel reciprocity based key establishment techniques for wireless systems,” *Wireless Networks*, vol. 21, no. 6, pp. 1835–1846, 2015.

[7] W. Xi, X.-Y. Li, C. Qian, J. Han, S. Tang, J. Zhao, and K. Zhao, “Keep: Fast secret key extraction protocol for d2d communication,” in *Proceedings of the IEEE IWQoS*, 2014.

[8] Y. Liu, S. C. Draper, and A. M. Sayeed, “Exploiting channel diversity in secret key generation from multipath fading randomness,” *IEEE TIFS*, vol. 7, no. 5, pp. 1484–1497, 2012.

[9] W. Xi, C. Qian, J. Han, K. Zhao, S. Zhong, X.-Y. Li, and J. Zhao, “Instant and robust authentication and key agreement among mobile devices,” in *in Proceedings of the CCS*, 2016.

[10] H. Liu, Y. Wang, J. Yang, and Y. Chen, “Fast and practical secret key extraction by exploiting channel response,” in *in Proceedings of the INFOCOM*, 2013.

[11] M. Roeschlin, I. Martinovic, and K. B. Rasmussen, “Device pairing at the touch of an electrode,” in *Proceedings of the NDSS*, 2018.

[12] Z. Li, M. Li, P. Mohapatra, J. Han, and S. Chen, “itype: Using eye gaze to enhance typing privacy,” in *in Proceedings of the INFOCOM*, 2017.

[13] P. Xie, J. Feng, Z. Cao, and J. Wang, “Genewave: Fast authentication and key agreement on commodity mobile devices,” in *in Proceedings of the ICNP, 2017.*

[14] Apple, “Test the microphones on your device,” https://support.apple.com/en-us/HT203792, 2018.

[15] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, “Dolphinattack: Inaudible voice commands,” in *in Proceedings of the CCS, 2017.*

[16] N. Roy, H. Hassanieh, and R. Roy Choudhury, “Backdoor: Making microphones hear inaudible sounds,” in *in Proceedings of the MobiSys, 2017.*

[17] K. Koli and K. A. Halonen, *CMOS current amplifiers: speed versus nonlinearity*. Springer Science & Business Media, 2002.

[18] Analog Devices, “Admp401: Omnidirectional microphone with bottom port and analog output,” 2013.

[19] Q. Wang, K. Ren, M. Zhou, T. Lei, D. Koutsonikolas, and L. Su, “Messages behind the sound: real-time hidden acoustic signal capture with smartphones,” in *in Proceedings of the MobiCom, 2016.*

[20] W. Diffie and M. Hellman, “New directions in cryptography,” *IEEE Transactions on Information Theory*, vol. 22, no. 6, pp. 644–654, 1976.

[21] S. Dey, N. Roy, W. Xu, R. R. Choudhury, and S. Nelakuditi, “Accelprint: Imperfections of accelerometers make smartphones trackable.” in *in Proceedings of the NDSS*, 2014.

[22] Z. Zhou, W. Diao, X. Liu, and K. Zhang, “Acoustic fingerprinting revisited: Generate stable device id stealthily with inaudible sound,” in *in Proceedings of the CCS*, 2014.

[23] A. Das, N. Borisov, and M. Caesar, “Do you hear what i hear?: Fingerprinting smart devices through embedded acoustic components,” in *in Proceedings of the CCS, 2014.*

[24] S.-Y. Chang, Y.-C. Hu, H. Anderson, T. Fu, and E. Y. Huang, “Body area network security: Robust key establishment using human body channel.” in *HealthSec*, 2012, pp. 5–5.