Voiceprint-based Access Control for Wireless Insulin Pump Systems

Bin Hao, Xiali Hei and Yazhou Tu
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, LA 70503, USA
{bin.hao, xiali.hei,yazhou.tu1}@louisiana.edu

Xiaojiang Du and Jie Wu
Dept. of CIS
Temple University
Philadelphia, PA 19912, USA
dxj@ieee.org, jiewu@temple.edu

***Abstract*—Insulin pumps have been widely used by patients with diabetes. Insulin pump systems adopt wireless channel with few cryptographic mechanisms, which makes them vulnerable to many attacks. In this paper, we focus on the wireless channel between Carelink USB and insulin pump on which the attackers can launch message eavesdropping and/or therapy manipulation attacks, which may put the patient in a life-threatening situation. Some prior solutions such as certificate-based or token-based schemes need either complicated key management or additional devices. We propose a novel voiceprint-based access control scheme comprising anti-replay speaker verification and voiceprint-based key agreement to secure the channel between the Carelink USB and insulin pump. Our scheme does not need permanent key sharing or additional devices. The anti-replay speaker verification adopts cascaded fusion of speaker verification and anti-replay countermeasure to ensure the insulin pump can be accessed by Carelink USB only after the legitimate user passes the identity verification. The evaluation on ASVspoof 2017 datasets shows that our scheme achieves a 4.02% Equal Error Rate (EER) with the existence of replay impostors. Besides, our scheme uses energy-difference-based voiceprint extraction and secure multi-party computing to generate a common cryptography (temporary) key between the Carelink USB and insulin pump, which can be used to encrypt the subsequent communication, and protect the insulin pump from eavesdropping and therapy manipulation attacks. By appropriately setting the similarity threshold of voiceprints, our key agreement scheme allows the insulin pump to establish a secure channel only with the device in its close proximity.**

***Index Terms*—wireless insulin pump; voiceprint; access control; voice anti-replay; speaker verification; acoustic channel**

I. Introduction

As of 2015, there were an estimated 30.3 million people of all ages in the U.S., or 9.4% of the population, suffering from diabetes [1]. People with type 1 diabetes, about 5% of diabetic patients, need insulin pumps. Insulin pump systems adopt wireless channels with few cryptographic mechanisms, making them vulnerable to many attacks. Patients that use such devices may be in potential security threats. Using off-the-shelf USB device, Radcliffe [3] was able to intercept wireless signals (glucose data) sent between the glucose sensor and the management device and cause blood glucose management devices to display inaccurate readings. Jack [2] was able to capture data transmitted from the computer, control the operations of the insulin pump such as delivering fatal doses (300 unit of insulin) to a diabetic patient. Li et al. [7] reverse-engineered the communication protocol used among continuous glucose monitoring (CGM) and insulin delivery systems and showed that both passive attacks such as message eavesdropping and active attacks such as impersonation can be launched. Marin et al. [8] extended Li's attacks by fully reverse-engineering the wireless communication protocol among all the peripherals of the insulin pump system. The authors carried out replay attacks, message injection attacks, and privacy attacks on the insulin pump system compromising both the safety and privacy of the patient. It is critical to design security mechanisms to mitigate the threats in wireless insulin pump systems.

In this paper, we focus on the wireless channel between the Carelink USB and insulin pump, which is critical in insulin pump systems. Over this channel, the attackers can launch message eavesdropping and/or therapy manipulation (such as remote dosage setting) attacks, which may put the patient in a life-threatening situation. Some prior solutions such as certificate-based or token-based schemes need either complicated key management or additional devices. Marin et al. [8] proposed an AES-MAC based cryptographic solution. The sharing of two symmetric keys and their management make the system complicated. Authors of [9], [10] proposed authentication schemes using additional external devices. Patient infusion pattern based access control scheme (PIPAC) [4] can resist two overdose attacks but assumes the patients glucose levels can only be modified manually, which does not hold in a closed-loop control system.

We propose a novel voiceprint-based access control scheme to mitigate the above attacks over the channel between the Carelink USB and insulin pump (also called "the two devices"). Our scheme does not need permanent key sharing or additional devices except needing two audio sensors to be embedded in the two devices, respectively. When the Carelink USB wants to request data or modify the dosage setting of the insulin pump, the pump needs a target user to grant her or his permission to the Carelink USB by speaking random passphrases. After successful speaker verification, the two devices can start to construct a secure communication channel.

The core ideas of our scheme rely on 1) the anti-replay speaker verification adopting cascaded fusion of speaker verification and anti-replay countermeasure to ensure the insulin pump is accessed by the Carelink USB only after the legitimate

user (not a replay impostor) passes the identity verification; 2) energy-difference-based voiceprint extraction and secure multi-party computing to generate a common cryptography (temporary) key between the two devices. The scheme does not need to share common secret or key, and utilizes an acoustic channel in addition to the wireless channel as a source of proximity declaration. The two devices transmit voiceprints, extracted from audio passphrases, to each other using secure multi-party computing, which leaks no information about the voiceprints to public. The similarity of the two voiceprints is then checked by computing such as Hamming Distance in the two devices, respectively. Then the two devices (if pass check) generate a common secret/key, which can be used to be (or to generate) a session key utilized to encrypt and/or append MAC to the messages exchanged, which can resist attacks such as message eavesdropping and remote dosage setting. Besides, our scheme uses Gaussian mixture model (GMM) to implement anti-replay speaker verification, which only needs to store the target user's model and achieves low equal error rate (EER, a threshold value when false acceptance rate equals false rejection rate) to secure the patient's safety (the lower the EER the higher the accuracy of the verification).

Our contributions are as follows:

- We propose a novel voiceprint-based access control scheme comprising anti-replay speaker verification and voiceprint-based key agreement to secure the channel between the Carelink USB and insulin pump.
- We implement an anti-replay speaker verification scheme using cascaded fusion of speaker verification and anti-replay countermeasure.
- We design a voiceprint-based key agreement between the two devices, and demonstrate its feasibility by experiments. The key agreement ensures the insulin pump will establish a secure channel only with a device in close proximity by setting the similarity threshold of voiceprint to some appropriate value such as 80%.
- We conduct an experimental evaluation on ASVspoof 2017 datasets, the results of which show that the speaker verification scheme achieves a 4.02% EER with the existence of replay impostors.

The remainder of this paper is organized as follows: In Section II, we discuss the related work. Section III describes the system and attacker model. We present our voiceprint (acoustic channel) based access control scheme for wireless insulin pumps in Section IV, and make security analysis in Section V. In Section VI, we describe the experimental results. We make overhead analysis and emergency handling in Section VII, and conclude the paper in Section VIII.

II. Related Work

A. Medical Device Authentication

Some proposals have been provided to add authentication schemes to medical devices. Li et al. [7] showed different types of attacks on an insulin pump using reverse-engineering technology, and proposed a cryptographic solution (rolling code) and body-coupled communication to protect the wireless links. Marin et al. [8] extended their attacks by fully reverse-engineering the wireless communication protocol in the insulin pump system, and proposed an AES-MAC based cryptographic solution which needs sharing of two symmetric keys. Authors of [9], [10] proposed authentication schemes using additional external devices, which may be forgotten, lost or stolen, and could potentially disclose a patient's status. Authors of [5], [6], [13], [14] proposed various access control schemes for wireless medical devices. These works based on general well-studied radio signal channels can be attacked by remote attackers who have sound knowledge of the radio propagation patterns. Our scheme utilizes an acoustic channel as a source of proximity declaration to establish secure communication between unacquainted devices in proximity.

B. Anti-replay Voice Authentication

Biometric identification systems such as face and voice recognition are widely used by healthcare providers. Biometric systems are susceptible to spoofing attacks which use methods such as artifact, mutilations, and replay to achieve impersonation or concealment. For speaker or voice authentication, the spoofing attacks comprise impersonation, voice conversion, speech synthesis, and replay [19]. In this paper, we focus on anti-replay voice authentication. Some countermeasures (CMs) have been proposed to mitigate replay attacks. Commercial voice authentication systems such as Nuance usually use challenge-response based methods and require users' explicit cooperation (repeating a closed set of sentences). Acoustic feature based methods come from the observation that design of spoofing countermeasures should focus on the search for discriminative features rather than the design of complex classifiers [20], [21], [32]. This principle is also adopted in our work. There also exist methods which leverage the response of the human speech production system to external stimuli to implement liveness detection. Zhang et al. propose VoiceGesture which is smartphone based and achieves high accuracy in detecting live users while not requiring additional cumbersome operations from users [15].

C. Secure Channel Establishment

There exist solutions that establish secure communication between (two) devices without any prior trust [11], [16], [18]. Roeschlin et al. [18] proposed a device pairing protocol based on the idea that two devices are permitted to bootstrap a secure channel if both of them are held by the same person. Schürmann et al. [16] proposed a scheme to establish a secure channel between unacquainted devices conditioned on similar ambient audio patterns. The protocol uses ambient audio fingerprints to generate a common cryptographic key between two devices in proximity, and explores error correcting codes to account for noise in the fingerprints. Our scheme utilizes a method similar to [16] to extract voiceprint from the target speaker's voice (not only the ambient audio) but uses secure multi-party computing to bootstrap a secure channel. Based on ultrasonic distance bounding, Rasmussen et al. [11] proposed a

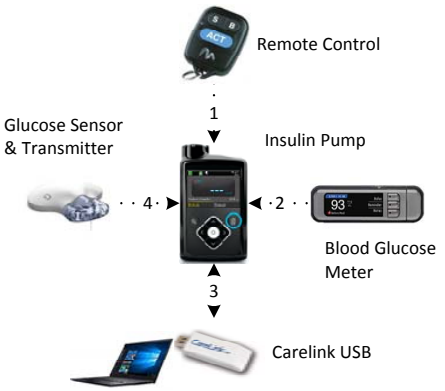

Fig. 1: A real time insulin pump system

proximity-based access control scheme enabling an implanted medical device to be accessed only by devices in its close proximity. The devices supporting this scheme need effective RF shielding; otherwise, a strong attacker can send data over the sound channel at a speed faster than that of sound, which breaks the assumption of the scheme.

III. System and Attacker Model

A. System Model

1) Background and the problem: The components of a Medtronic Paradigm real-time insulin pump system are shown in Fig. 1, which demonstrates a closed-loop control system. A typical insulin pump system consists of the insulin pump and its accessories (blood glucose meter, remote control, transmitter connecting to glucose sensor, and Carelink USB as an upload device). The blood glucose meter obtains blood glucose readings from the patients' finger-stick tests. The insulin pump can be programmed to automatically receive blood glucose readings from the blood glucose meter via wireless link 2 (each wireless link is numbered in Fig. 1). The glucose sensor tests the glucose level in the fatty layer under the skin. The transmitter connected with the glucose sensor then sends the readings to the pump via wireless link 4. The insulin pump delivers insulin to the patient. The remote control is operated by the patient to send instructions (such as suspend and resume basal dosage) to the insulin pump via wireless link 1 from a distant location. Via wireless link 3, the Carelink USB requests reports on blood glucose readings and patterns, then uploads data to a web-based diabetes management system using a USB port on a Laptop or PC.

Compromising wireless links 1, 3, 4 was demonstrated in [3] and [2]. The authors of [7] [8] have proposed security mechanisms for wireless links 1 and 4. In this paper, we will investigate an innovative approach to defend against passive and active attacks through link 3, which can be easily launched by remote adversaries. Besides passively eavesdropping the data from an insulin pump, the attackers can remotely change settings (such as dosage level) on the pump. These attacks bring potential threats to patient's privacy and safety, which need to be mitigated.

2) System model: We propose a human-aware, acoustic channel based access control scheme to mitigate above-mentioned attacks. In our considered scenarios, access control means the CareLink USB (or an attacker) wants to acquire access to an insulin pump to request data or remotely modify the therapy settings. When the Carelink USB wants to access the pump, first it sends request to the pump to activate the access privilege. The pump then starts the speaker verification to ensure that the access can be granted to the Carelink USB only when the legitimate user (e.g., the patient) passes the speaker verification. This verification needs the pump to embed speaker verification protocol and an audio sensor. After successful verification, the pump then bootstraps a key agreement with the Carelink USB, which also needs to embed an audio sensor or directly uses the microphone of the connected PC or laptop . This process takes as inputs two voiceprints and generates a shared temporary secret/key used to establish a secure channel between the two devices. The authentication process can be achieved in the case that the speaker passes the speaker verification and the Carelink USB is in close proximity to the pump, e.g., in the same diabetes consulting room.

B. Attacker Model

We consider two attack scenarios: the attacker wants to steal data (such as dosage history and patient personal information) transmitted from the pump or manipulate the therapy settings of the pump. The first attack can incur leakage of patient's privacy and the second may launch a maximum dosage injection, which would put the patient in critical danger. As in [17], we suppose the pump can work in two modes. In the *normal mode*, the pump needs the legitimate user to pass speaker verification while in the *emergency mode* the pump deactivates speaker verification and only needs the Carelink USB to be in close proximity. In the second mode, the pump and the Carelink USB can share a common key using the voiceprint extracted from ambient audio [16].

During the access control process, the proposal is supposed to defend against three different adversaries as follows:

- Remote impersonation. The attacker tries to pass speaker verification and perform key agreement with the pump. The attacker is not in close proximity (same context, e.g., clinic room) to the pump, but can participate in the authentication process by remotely receiving the user's voice or just using the voice previously recorded. This kind of attack can be launched when the pump is accidentally activated and the patient is speaking.
- Passive eavesdropping. The attacker eavesdrops on the messages transmitted over the wireless channel and records the voice of the legitimate user. The spied messages can be used to extract information about the shared secret/key. The recorded voice can be used to launch voice replay attacks to impersonate the legitimate speaker.
- Man-in-the-middle. The attacker tries to actively participate in the authentication process, making the pump and Carelink USB believe that they have successfully computed a shared secret/key but they haven't. They have

actually established a secure connection with the attacker, respectively, not each other.

IV. Voiceprint-based Access Control Scheme

The proposed scheme comprises a speaker-dependent (only legitimate user can pass the verification) and text-independent (the user can use any passphrases) speaker verification system using the acoustic channel to ensure user permission, and a key agreement protocol using energy-difference-based voiceprint extraction and secure multi-party computing (SMC) to add authentication between the pump and USB.

A. Acoustic Channel Verification

As shown above, we focus on adding acoustic channel verification for the wireless link 3 to make it safer. In embedded systems, we face several issues and challenges in implementing the above verification: First, the insulin pump system has limited computing power and memory; Second, we need to store speaker models in the insulin pump, which has limited storage capacity; Third, the accuracy of speaker verification must be high enough to guarantee the safety.

To solve the first problem, we choose the feature (i.e., a type of voiceprint) whose extraction has lower computation complexity. For lower memory occupation, we reduce the number of filter bands (feature dimension) while keeping appropriately high verification performance (e.g., low EER and high verification accuracy). For the second problem, we train the classifier based on our particular requirements. We need not to build a large-scale system used to verify many speakers. The model is lightweight because there is only one patient/speaker in each system. For the same reason, the accuracy of our system is higher than large-scale systems since we concentrate on only one speaker and optimize the verification for this particular scenario.

Speaker Verification Process. The speaker verification process typically has two phases: enrollment and prediction. In the enrollment phase, we collect the utterances of a user and then train them based on the algorithms stated below. While in the prediction phase, based on the trained model, the system makes prediction and computes a score for each test utterance. A typical automatic speaker verification (ASV) system shown in Fig. 2 comprises two subsystems: front-end and back-end. The front-end subsystem acquires voice from the speaker, implements feature extraction, and generates a feature matrix from the voice. Each column (feature vector) of the feature matrix corresponds to one frame of the voice. The back-end classifies the feature vectors using the trained classifiers based on speaker models, then outputs a verification result (accept or reject) using decision logic. Further processing of the utterances contains the following steps:

Step 1: Feature Extraction. We focused on short-term power spectrum features (except CQCC) and evaluated 9 different features [20], [21]: Mel Frequency Cepstral Coefficients (MFCC), Inverted MFCC (IMFCC), Linear Frequency Cepstral Coefficients (LFCC), Linear Prediction Cepstral Coefficients (LPCC), Constant-Q Cepstral Coefficients (CQCC),

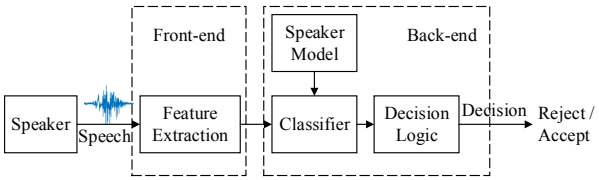

Fig. 2: Automatic speaker verification (ASV) system

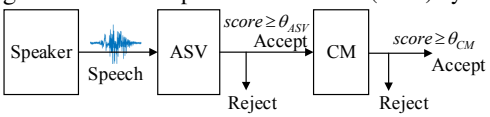

Fig. 3: Fusion of automatic speaker verification (ASV) and countermeasure (CM)

Subband Spectral Centroid Frequency Coefficients (SCFC), Subband Spectral Centroid Magnitude Coefficients (SCMC), Subband Spectral Flux Coefficients (SSFC), and Rectangular Filter Cepstral Coefficients (RFCC). Based on the evaluation results, We selected MFCC [22], CQCC [24], and LPCC [25]) to train the ASV model for high accuracy and IMFCC [26] to train the countermeasure model for resisting replay attacks.

Step 2: Classifier Training. Gaussian mixture model using maximum likelihood estimation (GMM-ML) [27] is used as a two-class classifier (genuine or replay) in our work. GMM is a weighted combination of multivariate Gaussian distributions. We use GMM-ML to train the classifier for detecting replay attacks. Gaussian mixture model with universal background model (GMM-UBM) [28] adopts GMM for likelihood functions, a universal background model (UBM) to represent alternative speakers, a kind of Bayesian adaptation such as maximum a posteriori (MAP) to generate speaker-specific models from the UBM. We use GMM-UBM to train the ASV.

B. Fusion of ASV and Anti-replay Countermeasure

Our speaker verification scheme can be considered as a kind of ASV system known to be vulnerable to replay attacks (a kind of spoofing). Some dedicated countermeasures (CMs) aiming to detect replay attacks have been proposed. Since genuine utterances should be accepted by both ASV and CM and either ASV or anti-replay CM should reject the utterances from impostors, a cascaded combination of ASV and CM forms a straightforward solution [29]. To our best knowledge, we are the first group to propose a cascaded fusion of ASV and CM using evaluation on the ASVSpoof 2017 datasets. The fusion framework is shown in Fig. 3. If an utterance passes the verification of ASV ($score \geq \theta_{ASV}$), it would continue to be verified by CM. If passing CM verification($score \geq \theta_{CM}$), a live target (legitimate) speaker is declared.

C. Voiceprint-based Key Agreement

After the target speaker passes verification, the attackers still have an opportunity to modify the dosage before delivery. To avoid this threat, we propose a voiceprint-based key agreement protocol, which ensures that access to the insulin pump can be granted to the Carelink USB only when the latter is in close proximity to both the insulin pump and the target speaker.

1) Energy-difference-based Voiceprint Extraction: After speaker verification, if the speaker is the claimed one, the pump and Carelink USB would begin the key agreement process. Firstly, each device needs to extract a binary characteristic sequence (called voiceprint) from the sampled audio. We adopt an energy-difference-based voiceprint extraction scheme, the principle of which comes from [16], [30]. Our scheme aims to extract voiceprint from a speaker's voice while [30] and [16] achieve that through music and ambient audio, respectively. The extraction algorithm proceeds as follows:

- Partition audio sample X of length L into N non-overlapping frames $X_1, ..., X_N$ of identical length L/N.
- Transform each frame using Fast Fourier Transformation (FFT) weighted by a hanning window (HW):

$$F_n = FFT(HW(X_n)), \quad n \in \{1, \ldots, N\} \tag{1}$$

- Partition the frequency into M non-overlapping frequency bands (filters) FB_m at linear space, compute the energy $E_{n,m}$ of each frequency band FB_m per frame F_n:

$$E_{n,m} = |F_n|^2 * FB_m, \ n \in \{1, \ldots, N\}, \ m \in \{1, \ldots, M\} \tag{2}$$

- Compute $(N-1)*(M-1)$ bits of the voiceprint by

$$f(n,m) = \begin{cases} 1, & (E_{n,m} - E_{n,m+1}) - \\ & (E_{n-1,m} - E_{n-1,m+1}) > 0, \\ & n \in \{2, \ldots, N\}, \ m \in \{1, \ldots, M-1\} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

2) Key Agreement using Voiceprint: Because the pump and CareLink USB may potentially adopt different types of audio sensors, and even the same type of sensors have different hardware characteristics, the voiceprints computed in these two devices are similar but there is a high probability they are not identical; they cannot be directly used as a key. We propose a key agreement protocol based on secure multi-party computing and energy-difference-based voiceprint extraction, which is shown in Fig. 4.

In the key agreement, we use Secure Three-party Sum Protocol (Sec3Sum) [31] as a voiceprint transmission method. Different from the standard protocol in [31], we consider one device (e.g., pump) as two participants, P_1 and P_2, and the other (e.g., USB) as the third participant P_3, and vice versa. The message exchange between P_1 and P_2 does not happen in practice. Suppose P_1 (Alice) computes voiceprint f_1, P_2 (Alice) generates a random number c_1, P_3 (Bob) computes voiceprint f_2, we want to securely compute the three sum $s = f_1 + c_1 + f_2$, but no one leaks private information (f_1, c_1, f_2) to others or public. Sec3Sum executes as follows:

- P_i $(i = 1, 2, 3)$ generates random shares $v_{i,j}$ $(j = 1, 2, 3)$, such that $f_1 = \sum_{j=1}^{3} v_{1,j}$, $c_1 = \sum_{j=1}^{3} v_{2,j}$, $f_2 = \sum_{j=1}^{3} v_{3,j}$.
- P_i $(i = 1, 2, 3)$ transmits $v_{i,j}$ to P_j $(j = 1, 2, 3$ and $j \neq i)$.
- P_i $(i = 1, 2, 3)$ gets all $v_{j,i}$ $(j = 1, 2, 3)$ and computes $v_i' = \sum_{j=1}^{3} v_{j,i}$, then broadcasts v_i'.
- P_i $(i = 1, 2, 3)$ computes the sum $s = \sum_{i=1}^{3} v_i' = f_1 + c_1 + f_2$.

After executing Sec3Sum, Alice (P_1, P_2) gets $s = f_1 + c_1 + f_2$ and computes the voiceprint f_2 of Bob by subtracting

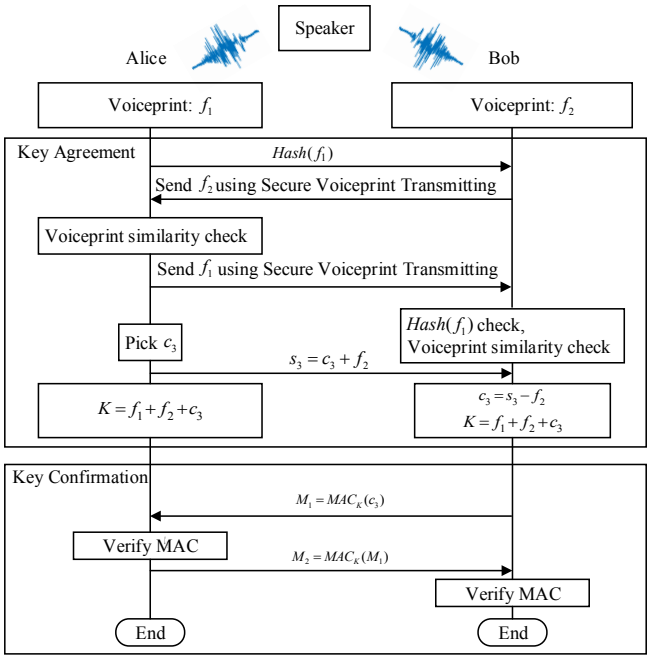

Fig. 4: Voiceprint-based key agreement

$f_1 + c_1$ from s. We also call this process Secure Voiceprint Transmitting (SVT) by which Bob securely sends voiceprint to Alice, and vice versa. Using Sec3Sum as a basic unit, the key agreement shown in Fig. 4 processes as follows:

- Alice and Bob extract voiceprints from the voice of the speaker, f_1 and f_2 (same length), respectively.
- Alice sends $Hash(f_1)$ to Bob as a commitment for f_1.
- Bob sends f_2 to Alice using SVT (Sec3Sum).
- Alice computes the voiceprint similarity: $HanmingDistance(f_1,f_2)/Length(f_1)$, aborts if the similarity is less than the preset threshold (e.g., 80%).
- Alice sends f_1 to Bob using SVT (Sec3Sum). Bob compares the hash of f_1 with $Hash(f_1)$ received previously from Alice, checks voiceprint similarity: $HanmingDistance(f_1,f_2)/Length(f_2)$. If any check (hash or voiceprint) fails, Bob aborts the process.
- Alice then picks a random number c_3, transmits $s_3 = f_2 + c_3$ to Bob.
- Upon receiving s_3, Bob gets c_3 by subtracting f_2 from s_3.
- Alice and Bob compute a key: $K = f_1 + f_2 + c_3$, respectively. The key agreement finishes.
- Key confirmation begins immediately after the key agreement. Bob generates a message authentication code (MAC) M_1 of c_3 using K and sends M_1 to Alice. Alice verifies M_1, generates a MAC M_2 of M_1 using K and sends M_2 to Bob. If all MACs pass the verification, Alice and Bob share a common key; otherwise, the scheme will terminate and report an error.

V. Security Analysis

A. Remote Impersonation

An attacker who is not in close proximity (e.g., same clinic room) to the pump or the speaker tries to perform

the key agreement with the pump or with the USB. The attacker must get a voiceprint which can pass the similarity validation. There are three possible methods to achieve this: generating a random bit sequence; extracting from previously or currently recorded voice; impersonating the pump to get the voiceprint of the USB, then making an agreement with the pump. For the first, the voiceprints have high entropy [16]; the probability of guessing a voiceprint with similarity greater than or equal to the preset threshold (e.g., 80%) is negligible if the voiceprint length is long enough (e.g., $\geq$ 512 *bits*). For the second, according to the results of [16], the gap between voiceprints is significant even if the adversary can listen to same audio source in a different room, and the differences between voiceprints extracted at different times are of high significance. The attacker cannot get a voiceprint having high similarity to the one in the pump in another context by remotely receiving the user's voice or using a previously recorded one. For the third, our protocol first requests the pump (or an impersonation attacker) sends the hash of voiceprint($Hash(f_1)$) to make commitment of having a voiceprint. Even if getting the voiceprint of the USB, the attacker cannot pass the hash check given that $Hash$ function is preimage resistant.

B. Passive Eavesdropping

By channel eavesdropping, the attacker can collect all the messages exchanged during the key agreement and record the voice of the legitimate speaker. Our speaker verification scheme uses the fusion of speaker verification and anti-replay countermeasure, which can mitigate the replay attacks with high accuracy as shown in Section VI. The security of the key agreement relies on whether or not the attacker can learn partial or all information about the exchanged voiceprints, which can be achieved by two methods: brute-force and message eavesdropping. For brute-force attack, according to the results of [16], the voiceprints have high entropy and no bias in bit distribution. We recommend the usage of long length of voiceprint (e.g., 512bits) and random passphrases (a sequence of words spoken by the user) to strengthen the security. For message eavesdropping, our protocol adopts a secure three-party sum protocol (Sec3Sum) to exchange voiceprints; no information about voiceprints is leaked given that the random numbers used in the protocol are uniformly distributed. Our scheme uses hash ($Hash(f_1)$) for voiceprint commitment and uses two MACs (M_1, M_2) for key confirmation. The protocol leaks no information about the voiceprint and key if the Hash function can resist preimage attacks and the MAC can resist second-preimage attacks.

C. Man-in-the-middle

Suppose an eavesdropper Eve to be located in the middle of Alice and Bob, she must modify or replace at least one message during the key agreement; otherwise, this would be passive eavesdropping. As shown in remote impersonation analysis, Eve cannot successfully finish the protocol with Alice and Bob, respectively. If Eve modifies any message in the protocol ($Hash(f_1)$, random numbers exchanged in Sec3Sum, c_3, M_1, M_2), Alice and Bob would compute different sum and/or different key, which finally makes the key confirmation fail.

VI. Experiments Evaluation

In this section, we present experiments to show how our cascaded fusion of speaker verification and anti-replay countermeasure is achieved and demonstrate the feasibility of the energy-difference-based voiceprint extraction scheme. We break down the experiments into the following: 1) describe the datasets used for evaluating our scheme; 2) evaluate different features and select candidates for speaker verification; 3) train and test the stand-alone ASV using selected features in case of zero-effort and replay imposters, respectively; 4) train and test the standalone CM in case of replay imposters; 5) evaluate the performance of our cascaded fusion scheme; 6) demonstrate the feasibility of energy-difference-based voiceprint extraction.

A. Dataset

We used the datasets of ASVspoof 2017 challenge, the primary technical goal of which is to evaluate spoofing (replay) attack detection accuracy of the countermeasures [32]. The datasets consisting of genuine and replay/spoof recordings are separated into three subsets comprising Training, Development and Evaluation set, details of which are listed in Table I. In ASVspoof 2017 challenge, the Training and Development subsets are provided for the design of replay countermeasures and the Evaluation subset is used to test the accuracy and generalization capacity of submitted replay detectors.

B. Model Feature Selection

To select the appropriate features used to implement automatic speaker verification (ASV), we evaluated 9 different features (shown in Section IV) using the ASVspoof 2017 Training subset. We used GMM-UBM model with 256 GMM mixtures and 20 iterations to train ASV. The GMM-UBM implementation is based on the MSR Identity Toolbox V1.0 [23]. For each feature except CQCC, we adopted 20ms frame length and 40 filter banks. For CQCC (not using short-time Fourier analysis as other features), the number of cepstral coefficients is 29 with the appended 0th coefficient. We used 70% of all the genuine utterances to train UBM model, and used 70% genuine utterances per speaker to train speaker-specific model and keep the remaining 30% utterances for test. Verification trials consist of all possible model-test combinations (30% of one target speaker vs. 30% of all the other impostors' utterances). The performance is shown in Table II. We find that voice activity detector (VAD) is critical. Without VAD, there is no successfully trained classifier except CQCC, which is not sensitive to VAD. Finally, we chose MFCC, LPCC, and CQCC

TABLE I: Statistics of the ASVspoof 2017 corpus.

Subset	**# Speakers**	**# Utterances**	
		Genuine	***Spoof***
Training	10	1507	1507
Development	8	760	950
Evaluation	24	1294	11988
Total	42	3561	14445

TABLE II: Standalone ASV feature performance (%EER) based on ASVspoof 2017 Traing subset

Features	*Training set (VAD)*	*Training set (No VAD)*
CQCC	0.66	**0.44**
MFCC	**0.54**	50.89
IMFCC	0.88	50.89
LPCC	**0.44**	55.56
LFCC	0.66	50.89
RFCC	0.57	50.89
SCFC	1.62	50.89
SCMC	0.88	50.89
SSFC	1.20	55.56

as candidates to train ASV because MFCC and LPCC achieve better performance than other features and CQCC achieves the best performance without VAD and outperforms all features in evaluation of the ASVspoof 2015 datasets [24].

C. Standalone ASV Performance of Zero-effort Impostors

Based on the trials above, we selected MFCC, CQCC, and LPCC as candidates and evaluated the performance of these three features in case zero-effort impostors try to impersonate the genuine target speaker just using their own sounds. We used the Training subset containing utterances of the 10 speakers to train a two-class GMM-UBM classifier. One of the 10 speakers was chosen as the enrollment (target) speaker whose utterances were used to train the speaker-specific model, and the 9 other speakers as zero-effort impostors whose utterances were used to train the impostor model. 70% of genuine utterances of the target speaker combined with the other 9 impostors' genuine utterances were used to train the two-class GMM-UBM model. 30% genuine utterances of the target speaker combined with all genuine utterances of the Development subset were used to predict and evaluate the performance, which is shown in Table III (columns 2-4). We can see that for most speakers MFCC achieves better performance, and LPCC achieves significantly better performance for speaker M0003 and M0008.

TABLE III: Standalone ASV performance of zero-effort and replay impostors (%EER) based on ASVspoof 2017 Datasets (Training and Development subsets)

Speakers	Zero-effort Impostors			Replay Impostors		
	MFCC	*CQCC*	*LPCC*	*MFCC*	*CQCC*	*LPCC*
M0001	0.00	1.45	4.20	0.05	1.19	2.56
M0002	0.00	0.13	1.30	0.20	0.25	2.03
M0003	3.55	2.37	**0.66**	14.54	11.11	**10.40**
M0004	1.32	0.00	3.70	4.78	2.94	3.70
M0005	0.39	2.63	0.13	0.56	1.56	0.44
M0006	1.97	4.21	3.68	16.16	**8.89**	12.58
M0007	0.00	0.26	0.26	0.22	0.22	0.11
M0008	10.39	11.67	**1.67**	8.69	6.67	**2.46**
M0009	1.75	0.26	1.18	1.75	0.42	1.75
M0010	0.39	0.53	0.92	0.22	0.22	1.88

D. Standalone ASV Performance of Replay Impostors

We evaluated the performance of the above 3 features in case replay impostors try to impersonate target speaker using recordings of target speaker (or someone else). We still used the Training subset to train a two-class GMM-UBM classifier. For each iteration, we chose 1 of the 10 speakers as the target speaker whose utterances were used to train speaker-specific model, and others as replay impostors whose utterances were used to train the impostor model. 70% of the target speaker's genuine utterances and all of the other 9 impostors' genuine utterances were chosen to train the two-class GMM-UBM model. We used 30% genuine along with all the spoof utterances of the target speaker and all genuine and spoof utterances of the Development subset to predict and evaluate the performance. The comparisons among MFCC, CQCC, and LPCC are shown in Table III (columns 5-7). We can see that the performance of the replay impostors is significantly lower than that of the zero-effort impostors, and that CQCC has comparable performance with LPCC while CQCC has significantly better performance for speaker M0006 and LPCC still achieves significantly better performance for speaker M0003 and M0008.

E. Standalone CM Performance of Replay Impostors

We trained a two-class GMM-ML model using the Development (Dev) and Evaluation (Eval) subsets of ASVspoof 2017. Specifically, we used all genuine utterances of Eval or Dev to train the genuine model while using its spoof utterances to train the spoof model, and then made Dev-Eval cross-validation. Table IV shows the performance. The second column shows the result with Dev as enrollment set and Eval as prediction set. The third column shows the result with Eval as enrollment set and Dev as prediction set. We chose the latter (showing IMFCC feature achieves the best performance) as the reference for subsequent fusion evaluation.

F. ASV & CM Fusion Performance of Replay Impostors

In this section, we demonstrate the performance of the fusion of ASV and CM. We used the Training subset to train and test the ASV GMM-UBM model and computed a threshold (θ_{ASV}). Each utterance getting a score below θ_{ASV} would be considered as zero-effort or replay impostor resulting in a rejection. The utterance getting a score $\geq \theta_{ASV}$ would continue to be double-checked in CM. Dev and Eval subsets were used to train GMM-ML model for CM. All the genuine and spoof utterances per speaker in Eval set were used to train a two-class GMM-ML model. All utterances, both the genuine and spoof of Dev, were used to test the model and compute a CM threshold (θ_{CM}). Each utterance getting a score below θ_{CM} would be considered as a spoof. The CM double-check is used to detect the utterance which is false positive. The fusion

TABLE IV: Standalone countermeasures (CMs) replay detection performance (%EER) based on ASVspoof 2017 Development (Dev) and Evaluation (Eval) subsets

Features	Enrollment/Prediction dataset	
	Dev/Eval set	*Eval/Dev set*
CQCC	27.58	8.94
MFCC	38.78	8.00
IMFCC	34.67	**6.57**
LPCC	30.90	8.42
LFCC	37.06	7.23
RFCC	36.14	8.04
SCFC	**25.11**	29.05
SCMC	34.97	8.14
SSFC	33.94	8.03

policy potentially increases the false negative rate (a genuine utterance is considered as a spoof one). When this happens, the user can try again. For the safety of the user, we think this inconvenience is acceptable. We evaluated MFCC, CQCC, and LPCC in ASV and IMFCC in CM, which results in 3 different fusions, the performance of which is shown in Table V. We find that in most cases the fusion of MFCC/LPCC ASV and IMFCC CM gets a lower EER than a standalone ASV or CM. The fusion of LPCC ASV and IMFCC CM achieves the best performance: the maximal EER value for all evaluated speakers is 4.02%.

G. Feasibility of Energy-difference-based Voiceprint Extraction

After the speaker has passed the speaker verification, the insulin pump and Carelink USB would start the voiceprint-based key agreement to bootstrap a secure communication channel. We made trials to demonstrate the feasibility of the voiceprint-based key agreement. Using the microphones in iPhone 5S and Honor 10, we recorded 270 passphrases, i.e., 135 for each device. In each test case, a person spoke the passphrase as a voice source, and these two devices were positioned in the way of one of total 27 different distance settings relative to the voice source (speaker). In each distance setting, the speaker spoke 5 sentences, each of which contains either 4 or 5 English words or 5 numbers between 0 and 9. The duration of each sentence is about between 1 and 3 seconds. The distance setting among the speaker and the mobile phones simulates the relative positions among the insulin pump, Carelink USB, and the patient (or the attacker). All the distance settings and results of similarities of voiceprints generated by these two devices are shown in Table VI. For voiceprint extraction, we used 16 kHz sampling frequency, 63 ms frame length, and 17 frequency filter banks. From all the experimental results, we get: (1) the average voiceprint similarity (AVS) is larger than 80% when the two devices are positioned within distance less than 30cm to the voice source; (2) the AVS drops down to 75.21% when one device is 300cm away from the speaker, and 61.71% when one device is at outside of the closed door of room (320 cm, mean ambient loudness in room: 38 dB, outside: 47 dB), which shows that the attacker cannot get a voiceprint having high similarity to the one in the pump or USB in another context by remotely receiving the user's voice.

VII. Discussion

A. Overhead Analysis

Storage overhead. We only need to store classifier models for one patient in the pump, which reduces the storage comparing with other speaker recognition systems. In the fusion system, the feature dimension adopted is 40. For ASV, we need to store one GMM UMB model (162 KB with 256 GMM components), one GMM user model (162 KB), and one GMM background users model (162 KB). For countermeasure, we need to store one GMM Genuine model (324 KB with 512 GMM components) and one GMM Spoof model (324 KB). The total permanent storage needed is only about 1 MB.

Computation complexity. We evaluated the executing time of the key modules in the environment of Raspberry Pi 1 Model B+ with 700 MHz Broadcom BCM2835 CPU. We suppose the duration of each recorded voice is around 2 s. For the speaker verification system, the ASV part needs one audio read (0.02 s), one feature extraction (0.15 s), and one log likelihood computation (0.23 s); the CM part needs one feature extraction (0.15 s), and one log likelihood computation (0.23 s). In the voiceprint extraction process, the scheme generates 16 bits voiceprint for each voice frame. All the tested passphrases have the length between 19 and 42 frames, so the length of voiceprints is between 304 and 672 bits. For high security, we adopt the voiceprint with the length $\geq$ 512 bits. During the key agreement, the voiceprint extraction spends 0.04 s; the running time of each of other operations (1 Hash, 8 Random number generations, 2 MACs, and 18 additions) is $\leq$ 0.001 s, totally $\leq$ 0.03 s. So the total computation time of the whole access control is around 1 s.

Communication complexity. Only the voiceprint-based key agreement needs message exchanges between the insulin pump and Carelink USB. The pump needs to transmit 1 Hash (256 bits using Sha256), 2 MACs (2x160 bits using EVP Sha1), 7 random numbers (7x512 bits) during two Sec3Sum protocols, and one random number s_3 (512 bits), and receive 7 random numbers during two Sec3Sum protocols. The total received and transmitted data of the pump (almost equivalent to Carelink USB) is $\leq$ 10 Kbits, which can be exchanged within 1 s using the RF channel (Pump to Carelink USB Frequency: 961.5 MHz, Bandwidth: 185 kHz). Combined with the computation analysis, the whole access control can be finished within around 2 s after the voice recording is finished.

B. Emergency Situation Handling

Allowing easy access to medical devices under emergencies is an orthogonal problem. Many proposals (e.g., in [9], [10], and [11]) suggested to grant open access to clinical staff during emergencies. Some literatures (e.g., in [6] and [12]) proposed schemes for emergency cases. To handle the emergency case, we can deactivate the speaker verification and execute the key agreement using the voiceprints extracted from ambient audio, the same case handled in [16]. In this situation, although without the participation and permission of the patient (unable to participate under emergency situation such as coma), the insulin pump and Carelink USB can still establish a secure channel so long as they are in close proximity to each other.

VIII. Conclusion

In this paper, we propose a novel voiceprint-based access control scheme comprising anti-replay speaker verification and voiceprint-based key agreement to secure the channel between the insulin pump and Carelink USB. We present a scheme that makes sure the insulin pump can be accessed by Carelink USB only after the legitimate user passes the identity verification, and the pump establishes a secure channel only with the device in its close proximity. Our scheme uses energy-difference-based voiceprint extraction and secure multi-party computing

TABLE V: ASV and CM fusion replay detection performance (%EER) based on ASVspoof 2017 Datasets

System	Speakers									
	M0001	***M0002***	***M0003***	***M0004***	***M0005***	***M0006***	***M0007***	***M0008***	***M0009***	***M0010***
ASV1 (MFCC)	0.20	0.00	11.11	3.70	0.78	4.78	0.16	6.67	1.75	0.17
CM1 (IMFCC)	7.11	6.06	6.96	6.71	6.67	6.31	7.59	6.61	6.11	6.81
Fusion1 (MFCC+IMFCC)	1.19	0.00	4.60	0.83	0.72	2.22	0.16	8.33	1.75	0.06
ASV2 (LPCC)	3.17	1.30	**0.80**	1.44	0.50	2.22	0.22	1.67	1.63	0.50
CM2 (IMFCC)	7.11	6.06	6.96	6.71	6.67	6.31	7.59	6.61	6.11	6.81
Fusion2 (LPCC+IMFCC)	**4.02**	2.60	0.52	0.50	0.22	1.77	0.11	**3.33**	1.63	0.33
ASV3 (CQCC)	2.31	0.20	4.94	0.56	0.89	4.44	0.05	6.67	0.00	0.22
CM3 (IMFCC)	7.11	6.06	6.96	6.71	6.67	6.31	7.59	6.61	6.11	6.81
Fusion3 (CQCC+IMFCC)	7.14	11.69	1.38	0.00	3.70	6.67	0.05	9.84	1.75	0.17

TABLE VI: Average similarity of voiceprints generated by two devices at different distances to the same voice source

Distance (cm)	Average voiceprints similarity (%)					
	S5 20	***S5 30***	***S5 50***	***S5 150***	***S5 300***	***S5 outside***
H10 20	81.55	80.22	80.78	78.66	74.97	64.36
H10 30	81.35	80.37	77.69	78.34	75.89	62.27
H10 50	80.73	80.50	78.45	78.32	77.00	61.11
H10 150	75.29	76.17	74.17	-	-	-
H10 300	75.06	75.79	72.55	-	-	-
H10 outside	60.39	60.90	61.22	-	-	-

to generate a common cryptography (temporary) key between the Carelink USB and insulin pump, which can be used to encrypt the subsequent communication while protecting the insulin pump from message eavesdropping and parameters manipulation attacks, such as remote dosage setting. Our proposal does not need certificates, permanent shared key or additional devices, which we believe will be an attractive solution to access control problem for insulin pump systems. Finally, our scheme may be generalized to other infusion systems as well, which can be our future work.

ACKNOWLEDGMENT

This work was supported in part by US NSF under grant CNS-1812553.

REFERENCES

[1] Centers for Disease Control and Prevention, "National diabetes statistics report, 2017," Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services, 2017.

[2] B. Jack, "Insulin pump hack delivers fatal dosage over the air," http://www.theregister.co.uk/2011/10/27/fatal_insulin_pump_attack.

[3] J. Radcliffe, "Hacking medical devices for fun and insulin: Breaking the human SCADA system," https://media.blackhat.com/bh-us-11/Radcliffe/BH_US_11_Radcliffe_Hacking_Medical_Devices_WP.pdf

[4] X. Hei et al., "PIPAC: Patient infusion pattern based access control scheme for wireless insulin pump system," in *Proc. of IEEE INFOCOM'13*, Turin, Italy, Apr. 2013.

[5] X. Hei et al., "Defending resource depletion attacks on implantable medical devices," in *Proc. of IEEE Globecom'10*, pp. 1-5, 2010.

[6] X. Hei and X. Du, "Biometric-based two-level secure access control for implantable medical devices during emergencies," in *Proc. of IEEE INFOCOM'11*, pp. 346-350, 2011.

[7] C. Li, A. Raghunathan, and N. K. Jha, "Hijacking an insulin pump: Security attacks and defenses for a diabetes therapy system," in *Proc. of the 13th IEEE Intl. Conf. on e-Health NAS*, pp. 150-156, 2011.

[8] E. Marin et al., "On the feasibility of cryptography for a wireless insulin pump system," in *Proc. of CODASPY'16*, pp.113-120, 2016.

[9] P. Inchingolo, S. Bergamasco, and M. Bon, "Medical data protection with a new generation of hardware authentication tokens," in *Proc. of Mediterranean Conf. on Medical and Biological Engineering and Computing*, 2001.

[10] T. Denning, K. Fu, and T. Kohno, "Absence makes the heart grow fonder: New directions for implantable medical device security," in *Proc. of the 3rd Conf. on Hot topics on security*, pp. 1-7, 2008.

[11] K. B. Rasmussen et al., "Proximity-based access control for implantable medical devices," in *Proc. of ACM CCS '09*, pp. 410-419, 2009.

[12] J. Sun et al., "HCPP: Cryptography based secure EHR system for patient privacy and emergency healthcare," in *Proc. of ICDCS'11*.

[13] X. Hei and X. Du, "Emerging security issues in wireless implantable medical devices," *Springer*, 2013.

[14] X. Hei, X. Du, and S. Lin, "Poster: Near field communication based access control for wireless medical devices," in *Proc. of ACM MobiHoc'14*.

[15] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proc. of ACM CCS'17*, pp. 57-71, 2017.

[16] D. Schürmann, and S. Sigg, "Secure communication based on ambient audio," *IEEE Trans. on Mobile Computing*, 12(2), pp. 358-370, 2013.

[17] K. B. Rasmussen et al., "Proximity-based access control for implantable medical devices," in *Proc. of ACM CCS'09*, pp. 410-419, 2009.

[18] M. Roeschlin, I. Martinovic, and K. B. Rasmussen, "Device pairing at the touch of an electrode," *NDSS'18*, Feb. 18-21, San Diego, CA, USA.

[19] Zh. Wu et al., "Spoofing and countermeasures for speaker verification: A survey", *Speech Communication*, 66, pp. 130-153, 2015.

[20] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," *INTERSPEECH*, 2015.

[21] R. Font, J. M. Espín, and M. J. Cano, "Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 challenge", *INTERSPEECH*, 2017.

[22] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357-366. 1980.

[23] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1.0: A MATLAB toolbox for speaker recognition research", *Microsoft Research Technical Report*, 2013.

[24] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q Cepstral Coefficients," in *Proc. of ODYSSEY'16*.

[25] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 29(2), pp. 254-272. 1981.

[26] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *International Journal of Signal Processing*, 4(2), pp. 114-121, 2007.

[27] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. On Speech And Audio Processing*, 3(1), pp. 72-83, 1995.

[28] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 10, pp. 19-41, 2000.

[29] H. Delgado et al., "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015," *INTERSPEECH*, 2017.

[30] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Pro. of the 3rd International Conference on Music Information Retrieval*, October 2002.

[31] Y. Yao and F. Yu, "Privacy-preserving similarity sorting in multi-party model," *International Journal of Network Security*, vol.19, no.5, pp. 851-857, Sep. 2017.

[32] T. Kinnunen et al., "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," *INTERSPEECH*, 2017.