Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Feature Fusion and Voiceprint Based Access Control for Wireless Insulin Pump Systems

YUAN PING^{1,2}, (Member, IEEE-CS), BIN HAO^{2}, XIALI HEI^{2}, (Senior Member, IEEE), YAZHOU TU^{2}, (Member, IEEE), XIAOJIANG DU^{3}, (Senior Member, IEEE), JIE WU^{3}, (Fellow, IEEE)

[1]School of Information Engineering, Xuchang University, Xuchang 461000, China (e-mail: pyuan.lhn@xcu.edu.cn)
[2]School of Computing and Informatics, University of Louisiana at Lafayette, LA 70503, USA (e-mail: bin.hao@louisiana.edu; xiali.hei@louisiana.edu; yazhou.tu1@louisiana.edu)
[3]Department of Computer and Information Science, College of Science and Technology, Temple University, Philadelphia, PA 19912, USA (e-mail: dxj@ieee.org; jiewu@temple.edu)

Corresponding author: Bin Hao and Xiali Hei (e-mail: {bin.hao, xiali.hei}@louisiana.edu).

Y. Ping and B. Hao are co-first authors. This work was supported in part by US NSF under grant CNS-1812553, and the Program for Science & Technology Innovation Talents in Universities of He'nan Province under Grant No. 18HASTIT022.

ABSTRACT Without effective cryptographic mechanisms, the wireless channel between the USB/uploader and insulin pump frequently suffers from vulnerabilities. Either eavesdropping or therapy manipulation attacks would put the patients in a life-threatening situation. Towards tackling this problem, we propose an access control scheme by introducing feature fusion and voiceprint. Featured by the anti-replay speaker verification and voiceprint-based key agreement, it secures communications over the wireless channel. Through a cascaded fusion of speaker verification and anti-replay countermeasure, the anti-replay speaker verification guarantees that the pump can only be accessed after the verification. When defending against zero-effort and replay impostors with our scheme, the equal error rate can be reduced to 2.22%. Furthermore, to generate a common key for the wireless channel, in the voiceprint-based key agreement, we present a non-interactive energy-difference-based voiceprint extraction and adaptive Reed-Solomon coding based fuzzy extractor. Thus, it enhances the communication encryption which protects the pump from eavesdropping and therapy manipulation attacks. Also, with an appropriate constraint on voiceprints similarity, the key agreement lowers the risk of channel establishment from device locating outside the pump's close proximity.

INDEX TERMS Wireless insulin pump, feature fusion, voiceprint, access control, acoustic channel.

I. INTRODUCTION

IN the U.S., 9.4% of the population are suffering from diabetes. Among them, about 5% of patients with type 1 diabetes use insulin pumps [1]. Similar to other digital medical devices [2], [3], insulin pump systems prefer wireless channels to form a closed-loop system. But few of them are equipped with effective cryptographic mechanisms. Lacking enough security mechanisms, insulin pump systems frequently suffer from vulnerabilities. Patients who use such devices face security threats. For instance, wireless signals between the glucose sensor and the management device can be intercepted [4] while data transmission may be captured [5]. The prior usually causes inaccurate display while the latter leads to malicious control of the insulin pump, e.g., delivering fatal doses. Besides, Li et al. [6] and Marin et al. [7] conducted reverse-engineering towards the communication protocols. They confirmed that attacks such as eavesdropping, impersonation replay attacks, message injection attacks, and privacy attacks on the insulin pump system compromise both the privacy and safety of the patient. Consequently, appropriate security mechanisms are urgently required to secure wireless insulin pump systems.

Towards this requirement, we focus on protecting the wireless channel between the USB/uploader (e.g., CareLink USB [8], hereinafter called USB) and insulin pump. Some early solutions, e.g., certificate-based or token-based schemes, adopted either complicated key management [7] or additional devices [9], [10]. In [7], an AES-MAC based cryptographic solution with an optimization of the message format was presented. However, the requirement of sharing two independent symmetric keys increases the system complexity of key management. By introducing additional external devices, Denning et al. [9] and Inchingolo et al. [10] investigated fail-open defensive techniques to strike a balance between safety in the common case and security under adversarial conditions. Furthermore, Hei et al. [11] designed a patient infusion pattern based access control scheme to defend against single acute overdose and chronic overdose attacks, but it needs glucose level data measured by finger-stick. These attacks and solutions above bring up a question: *Is there an effective way of establishing a secure channel between unacquainted devices without pre-shared common keys in a closed-loop insulin pump system?*

To answer this question, in this paper, a novel feature fusion and voiceprint based access control scheme is proposed. To establish a secure channel between the USB and insulin pump (also called "the two devices"), it adopts two audio sensors separately embedded in each device to avoid introducing permanent key sharing or additional devices. Before data request or dosage adjustment, the pump grants permission to the USB through random passphrases speaking. After successful speaker verification, a secure channel can be established by the proposed key agreement.

Intuitively, the core ideas of our proposal comprise: 1) *automatic speaker verification (ASV)* utilizing cascaded fusion of speaker verification and *anti-replay countermeasure (CM)*. It makes the pump be accessible to the USB of the legitimate user (not a replay impostor); 2) *secure key agreement* introducing non-interactive (independent) energy-difference-based voiceprint extraction and *adaptive* Reed-Solomon (RS) coding [12] based fuzzy extractor [13]. Base on the protocol, the required common cryptography (temporary) key can only be generated when the user/speaker and the two devices are in close proximity. The *adaptive* means that the RS coding setting is not fixed but determined by real-time voiceprints.

Our voiceprint extraction scheme is non-interactive and real-time and only needs to compute one voiceprint in each device while the algorithm in [13] needs to align the voices and compute hundreds of voiceprints to find two matched voiceprints. The RS-coding-based fuzzy extractor utilizes the voiceprint f_1 as a secret to hide a confidential random number r (i.e., key seed) in such a way that only a similar voiceprint f_2 can decrypt/decode the original number r. We define the error tolerance threshold as τ, which is the ratio of the number of max tolerable bit differences (i.e., Hamming Distance) to the length of the voiceprints (f_1 and f_2 should be aligned to the same length). Only if the two voiceprints (generated in the insulin pump and USB, respectively) have a similarity $\eta \geq 1-\tau$, the common secret/key can be generated and be employed to be (or further generate) a session key for message encryption or appending MAC. Then attacks such as message eavesdropping and remote dosage setting can be resisted well. Besides, Gaussian mixture model (GMM) is adopted for anti-replay speaker verification, which only stores the target user's speaker models while reaches a low equal error rate (EER). EER is the threshold when the false acceptance rate (FAR) equals the false rejection rate (FRR).

The main contributions are as follows:

- The proposal of a feasible voiceprint based access control scheme which employs both anti-replay speaker verification and voiceprint-based key agreement protocol to secure the communication between the USB and the pump.
- The design of anti-replay speaker verification method which introduces a cascaded fusion of speaker verification and anti-replay strategy. For speaker verification, we extend the major implementations of feature extractions, including Rectangular Filter Cepstral Coefficients (RFCC) [14], Subband Spectral Centroid Frequency Coefficients (SCFC) [15], Subband Spectral Centroid Magnitude Coefficients (SCMC) [15], Subband Spectral Flux Coefficients (SSFC) [16], and Relative phase shift (RPS) [17]. And then we apply linear scale (LIN), mel scale (MEL) and inverted mel scale (IMEL) filter banks to these features and form new features, i.e., RFCC-MEL, RFCC-IMEL, SCFC-MEL, SCFC-IMEL, SCMC-MEL, SCMC-IMEL, RPS-LIN, and RPS-IMEL. Besides, we propose the using of Root Mean Square (RMS) energy for replay detection and form three features: RMSCC-LIN, RMSCC-MEL, and RMSCC-IMEL. Evaluation results show that the fusion of Linear Frequency Cepstral Coefficients (LFCC) [18] and RMSCC-MEL achieves the best performance.
- The non-interactive voiceprint extraction algorithm and an RS-coding-based fuzzy extractor. Taking the voiceprint extraction and fuzzy extractor as basic units, we present a secure key agreement for two unacquainted devices whose feasibility is evaluated by experiments using voice samples recorded by two smartphones in 27 distance settings. In the key agreement, the voiceprint similarity threshold ensures that the secure channel can only be established between devices in close proximity.
- On ASVspoof 2017 datasets, our anti-replay speaker verification scheme reduces EER to 2.22% with the existence of zero-effort and replay impostors.

The remaining parts of this paper are organized as follows. In Section II, we briefly review the related works. The general models for the system and attacker are described in Section III. Then, Section IV and V respectively presents our feature fusion and voiceprint based access control scheme for wireless insulin pump systems and the essential security analysis. Experimental results are presented in Section VI. We also make overhead analysis and emergency handling for the proposed scheme in Section VII. Finally, conclusions are drawn in the last section.

II. RELATED WORK

A. MEDICAL DEVICE AUTHENTICATION

By introducing reverse-engineering technology, Li et al. [6] showed that both passive attacks and active attacks threaten the insulin pump, and then compromise both the privacy and safety of patients . To ease these threats, many researchers tried to add authentication schemes to medical devices. For instance, Li et al. also proposed two possible CMs based on cryptographic protocols (rolling code) and body-coupled communication to protect the wireless links [6]. Fully reverse-engineering the wireless communication protocol in the insulin pump system, Marin et al. [7] further extended their attacks, and then provided an AES-MAC based cryptographic solution which needs sharing two symmetric keys.

Kune et al. measured the susceptibility of analog sensors in implantable medical devices by using specially crafted electromagnetic interference (EMI) [19]. They showed that EMI could inhibit pacing and induce defibrillation shocks on implantable cardiac electronic devices and then developed a defense mechanism using analog shielding components. Inchingolo et al. [10] and Denning et al. [9] proposed authentication schemes using additional external devices, which may be forgotten, lost or stolen, and could potentially disclose a patient's status. Hei et al. [20]–[23] proposed various access control schemes for wireless medical devices. However, most of the solutions above suffer from remote attacks if the attacker has sufficient knowledge of the radio propagation patterns. Different from these works, we prefer an acoustic channel as a source of proximity declaration which successfully utilizes the change of sound quality with respect to distance. Thus our scheme can construct secure channel between unacquainted devices in proximity.

B. ANTI-REPLAY VOICE AUTHENTICATION

Based on physiological features (face, fingerprint, iris, etc.) or behavioral features (voice, gait, keystroke dynamics, etc.), biometric identification systems are widely adopted by healthcare providers. However, these systems frequently suffer from spoofing attacks using methods such as artifact, mutilations, and replay to achieve impersonation or concealment. Focusing on voice or speaker authentication systems, the spoofing attacks consist of impersonation, voice conversion, speech synthesis, and replay [24]. In this study, the major concern is the anti-replay voice authentication. To mitigate replay attacks, some insightful CMs can be found in speech recognition or speaker verification systems. Among them, Nuance is a commercial voice authentication system which adopts a challenge-response strategy and requires the users' explicit cooperation (i.e., repeating sentences in a closed set). Refs. [25]–[27] agree that discriminative features are critical for building spoofing CMs. Thus, they provided acoustic feature-based methods. Besides, Zhang et al. found that liveness detection is also helpful to design anti-replay methods [28]. They proposed VoiceGesture which uses a smartphone to leverage the human's articulatory gesture while avoiding additional cumbersome operations.

C. SECURE CHANNEL ESTABLISHMENT

Some schemes have been proposed to establish secure communication between two devices with no prior trust [13], [29], [30]. Roeschlin et al. [30] proposed a device pairing protocol to ensure that the secure channel can only be constructed if both of the two devices are held by one person. The protocol uses an unauthenticated wireless channel to achieve Diffie-Hellman key agreement and the read-only human body channel to implement key confirmation. Rostami et al. [31] introduced a physiologically-based IMD device pairing protocol, called *Heart-to-Heart* (H2H), which uses the extracted time-varying randomness (Physiological Value, PV) from ECG signals as an authenticated mechanism. Then the programmer can access a patient's IMD if and only if physically contacting the patient's body. Zheng et al. [32] proposed the Finger-to-Heart (F2H), which is another biometric-based IMD authentication scheme using fingerprints to secure the IMD. The F2H scheme does not require the IMD to capture and preprocess the fingerprint image and extract the minutiae. It exhibits a lower resource consumption but assumes an authenticated, encrypted channel between the IMD and programmer, which is the goal of our paper. To establish a secure channel between unacquainted devices, the solution from [13] is conditioned on similar ambient audio patterns. The ambient audio fingerprints are introduced to generate a common key for two devices in proximity. In the solution, error-correcting codes are explored to account for noise in the fingerprints. However, pricey computations are consumed by voice alignment and voiceprints extraction before finding out two matched voiceprints. This problem motivates us to build a non-interactive scheme by which the voiceprints can be extracted *independently* in each of the two devices. In addition, Karapanos et al. [33] proposed Sound-Proof, a two-factor authentication (2FA) based on ambient sound, which also utilizes a previously established secure channel (symmetric encryption) as in [32]. Based on ultrasonic distance bounding, Rasmussen et al. [29] proposed a proximity-based access control scheme, which enables an implanted medical device to be accessed only by devices in close proximity if the devices have effective RF shielding.

III. SYSTEM AND ATTACKER MODEL

A. SYSTEM MODEL

1) Background

Fig. 1 shows a typical insulin pump system (e.g., Medtronic Paradigm). It is a real-time closed-loop control system which comprises the insulin pump and its accessories, e. g., blood glucose meter, remote control, glucose sensor & transmitter, and the upload device (USB). From finger-stick tests, the blood glucose meter gets blood glucose readings which can be automatically transmitted to the programmed insulin pump via wireless link 2. Meanwhile, the glucose level is read by the glucose sensor and sent to the insulin pump via wireless link 4. Then, the pump delivers insulin to the patient. Besides, the insulin pump works following the instructions from the remote control operated by the patient.

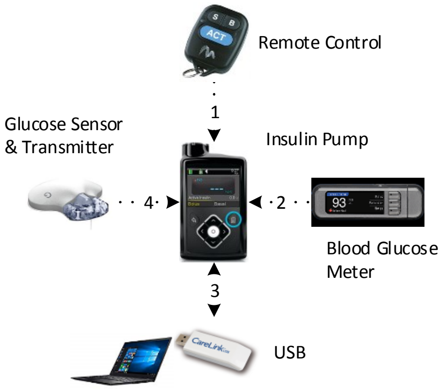

FIGURE 1: A real-time closed-loop control system of insulin pump system.

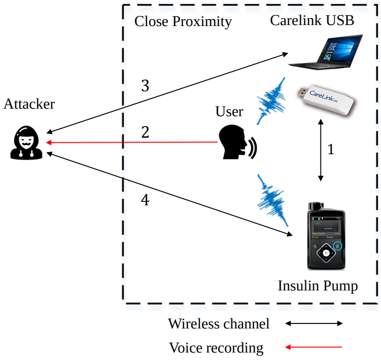

FIGURE 2: Communication channels among legitimate user, USB, insulin pump, and attacker.

Instructions such as suspending and resuming basal dosage are transmitted via wireless link 1 from a distant location. The last wireless channel is link 3 via which the USB requests report from the insulin pump and uploads data to a diabetes management system.

Via link 3, the remote attackers can passively eavesdrop data from the insulin pump and use the USB to change settings (e.g., dosage level) in the pump by forging radio signals. Therefore, in this study, our works investigate an innovative defense strategy over link 3 towards easing the potential threats on the privacy and safety of the user.

2) General system model

To mitigate the attacks above, a human-aware and acoustic channel based access control scheme is proposed in this work. Intuitively, we introduce access control to the phase of the USB accessing the insulin pump for data or modifying the therapy settings. In this phase, the USB should acquire the access permission first by sending a request. This behavior activates the speaker verification of the pump to avoid illegal attacker being granted. Generally, each pump should have its specific user(s). With the proposed speaker verification protocol and embedded audio sensor, the pump will bootstrap the key agreement protocol with the specific USB in close proximity. To match the pump, the USB device should be equipped with an audio sensor or use the microphone in the connected computer. Given the two separately collected voiceprints, the proposed system effectively generates a shared temporary secret/key for communication encryption of the secure channel.

B. ATTACKER MODEL

Generally, an attacker launches attacks to steal sensitive data or manipulating the therapy settings over channel 1 of Fig. 2. These attacks bring privacy leakage or put the patient in critical danger. In this proposal, we consider two work modes for the pump, i.e., *normal mode* and *emergency mode* [29]. The *normal mode* means the legitimate user has to pass the speaker verification, and the USB should be staying in close proximity, while only the proximity constraint for the USB is employed by the pump in the *emergency mode*. Meanwhile, the *emergency mode* tolerates the two devices sharing a common key generated from the voiceprints extracted from anyone.

Therefore, the proposed scheme are supposed to defend against three different adversaries:

- Remote impersonation. As shown in Fig. 2, the attacker wants to pass the speaker verification and launches the key agreement with the pump over channels 3 and 4. Furthermore, the attacker is not in the required proximity. But he can participate in the authentication process by remotely recording the voice in real-time or replaying the previously recorded voice over channel 2. This attack frequently appears when the pump is accidentally activated, and the user is speaking.
- Passive eavesdropping. Attacker eavesdrops on the messages over channels 1, 3 and 4, and then records the voice of the legitimate user over channel 2. Using the spied messages, the attacker may infer sensitive information related to the shared secret/key. The recorded voice can also be replayed by the attacker to impersonate the legitimate user/speaker.
- Man-in-the-middle (MITM). According to participating in the authentication process, the attacker makes the two devices believe that the shared secret/key is successful working. Unfortunately, the established so-called secure channels are connecting with the attacker.

IV. VOICEPRINT-BASED ACCESS CONTROL SCHEME

The proposed voiceprint-based access control scheme comprises two critical components. Speaker verification system is the first component featured with speaker-dependent and text-independent properties. The prior means it only accepts the legitimate user, and the latter allows the user to use random (and long enough) passphrases. The key agreement protocol is the second component which prefers a strategy of the combination of non-interactive energy-difference-based voiceprint extraction and adaptive RS-coding-based fuzzy extractor to complete the authentication between the pump and the USB. So the scheme executes in 4 phases:

1) **Speaker & Anti-replay Verification** accepts the legitimate user and rejects the zero-effort (impersonation) and replay impostor;
2) **Energy-difference-based Voiceprint Extraction** computes voiceprints using recorded passphrases in USB and insulin pump, respectively;
3) **Secret Agreement using Fuzzy Extractor (SAFE)** aborts the authentication procedure if the voiceprint similarity check fails;
4) **Key Agreement** establishes a secure wireless channel between the pump and USB.

We will explain phase 1 in Section IV-A and Section IV-B and explain phases 2, 3, and 4 in Section IV-C. Before the detailed explanation of the access control scheme, we list the following notations in Table 1, which are used in the subsequent sections to facilitate the understanding of the proposal.

TABLE 1: Notations

Symbol	Explanation
P	Passphrase or speech recorded by microphone and spoken by the user or the imposter
f_s	Sampling frequency of a passphrase
l	Sampling points number in each frame of a passphrase
L	Sampling points number of a passphrase
M	filter banks number when partitioning the whole frequency interval
sca	Type of filter-banking scale (linear, MEL, or IMEL)
N	The number of frames when framing a speech
X_n	Each of frames after framing a speech
F_n	Fourier transformation of X_n
SF	Vector containing $M+2$ frequency points spaced in $[{f_s}_{min}, {f_s}_{max}]$ according to sca, ${f_s}_{min}=0$, ${f_s}_{max}=f_s$ in our scheme
FB_m	Frequency bands partitioning SF, $m \in \{1,\ldots,M\}$
f	Voiceprint extracted from P using energy-difference-based voiceprint extraction

A. ACOUSTIC CHANNEL VERIFICATION

To introduce an acoustic channel verification to the wireless link 3, several challenges can be found in the embedded systems. Firstly, the insulin pump is usually a resource-constraint system with very limited computation and memory resources. Secondly, for a specific user, the speaker models should be pre-stored in the pump. Thirdly, the speaker verification system should achieve a high enough accuracy to make the scheme's applicability and security.

Motivated by the first requirement, we adopt features which can be extracted with low-cost computation, such as LFCC and RMSCC. To reduce memory occupation, we select a moderate number of filter bands who contribute to high verification performance, e.g., high accuracy and low ERR. Both of the corresponding computation complexities and storage overhead are analyzed in Section VII-A. Towards the second problem, we train the classifier based on our specific requirements. In practical use, for instance, there is no essential for the proposed system to support many speakers. Therefore, we can build a lightweight system to meet a specific user's requirement. The verification performance can be optimized for a particular scenario with only one user.

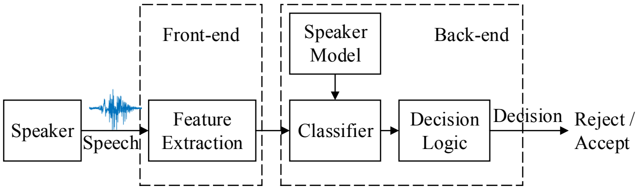

FIGURE 3: Automatic speaker verification system.

Speaker Verification Process. Typically, the speaker verification contains *enrollment* and *prediction* phases. In the enrollment phase, the user's utterances are carefully collected to train the speaker models and classifiers; whereas the prediction phase is to evaluate each test utterance based on the speaker models. Fig. 3 depicts a typical ASV system with *front-end* and *back-end* subsystems. The major tasks of the front-end are voice acquirement, feature extraction, and feature matrix generation. Each column (feature vector) of the feature matrix corresponds to a voice frame. Employing the trained classifiers, the back-end classifies the feature vectors and gives a verification result for decision.

1) Feature Extraction

In the proposed scheme, short-term power spectrum and short-term phase features (except CQCC) as in [25], [27] are considered. They include mel frequency cepstral coefficients (MFCC) [34], inverted MFCC (IMFCC) [35], linear frequency cepstral coefficients (LFCC) [18], linear prediction cepstral coefficients (LPCC) [36], constant-Q cepstral coefficients (CQCC) [37], subband spectral centroid frequency coefficients (SCFC) [15], subband spectral centroid magnitude coefficients (SCMC) [15], subband spectral flux coefficients (SSFC) [16], rectangular filter cepstral coefficients (RFCC) [14], all-pole group delay function (APGDF) [38], and relative phase shift (RPS) [17]. The extraction flow for each feature is shown in Fig. 4. Different from [27], which compared these features for synthetic speech detection, we evaluated these features for replay attack detection. Different from [25] making features analysis (except for APGDF and RPS) for replay attack detection, we focus on the fusion of these features and concentrate on the specific scenario with one speaker. In addition, we also extend the implementations of some of these features and introduce a new feature namely root mean square energy cepstral coefficients (RMSCC) which will be described in the following.

- **Linear, Mel-Scale, and Inverse Mel-Scale Filter Banking.** To extract the cepstral coefficients, the filter banking step is based on the power spectrum computed by signal framing and short-time Fourier transform (STFT) and followed by logarithmic compression and DCT to get the cepstral coefficients. There are two

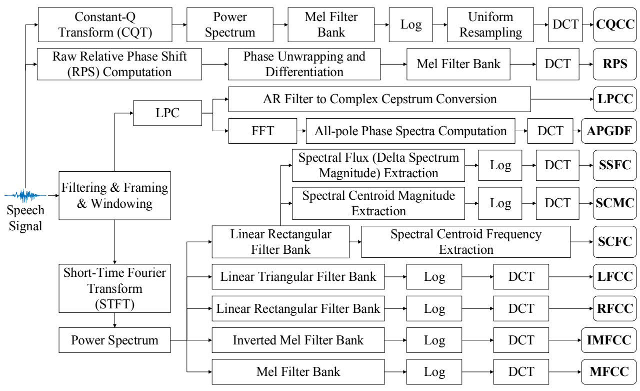
Constant-Q Transform (CQT)
Power Spectrum
Mel Filter Bank
Log
Uniform Resampling
DCT
CQCC
Raw Relative Phase Shift (RPS) Computation
Phase Unwrapping and Differentiation
Mel Filter Bank
DCT
RPS
LPC
AR Filter to Complex Cepstrum Conversion
LPCC
FFT
All-pole Phase Spectra Computation
DCT
APGDF
Spectral Flux (Delta Spectrum Magnitude) Extraction
Log
DCT
SSFC
Speech Signal
Filtering & Framing & Windowing
Spectral Centroid Magnitude Extraction
Log
DCT
SCMC
Linear Rectangular Filter Bank
Spectral Centroid Frequency Extraction
SCFC
Short-Time Fourier Transform (STFT)
Linear Triangular Filter Bank
Log
DCT
LFCC
Linear Rectangular Filter Bank
Log
DCT
RFCC
Power Spectrum
Inverted Mel Filter Bank
Log
DCT
IMFCC
Mel Filter Bank
Log
DCT
MFCC

FIGURE 4: Block diagram of the extraction flows of the evaluated features.

major types of filters, i.e., the rectangular window used in RFCC and the triangular window used MFCC, IMFCC, and LFCC, etc. There are three methods to place the filters: linear-scale, mel-scale, and inverted mel-scale. RFCC and LFCC use linear scale filters, which are evenly positioned in the whole frequency band. MFCC uses MEL filters, which places more filters in low-frequency band and less in high-frequency band whereas IMFCC uses inverted mel-scale filters, which places less filters in low-frequency band and more in high-frequency band. The original implementations of RFCC, SCFC, SCMC, and SCMC only use linear scale [27] and the RPS (only) uses MEL, we extend the implementations in [27], [39] and apply linear scale, MEL and inverted IMEL filter banks to these features and form new features: RFCC-MEL, RFCC-IMEL, SCFC-MEL, SCFC-IMEL, SCMC-MEL, SCMC-IMEL, RPS-LIN, and RPS-IMEL. Besides, we use RFCC-LIN, SCFC-LIN, SCMC-LIN, SCMC-LIN, and RPS-MEL to denote their original versions, respectively.

- **RMSCC.** The signal energy can be computed by taking the root average of the square of the amplitude, i.e., RMS [40]. RMS was used by [41] to fingerprint smart devices through embedded acoustic components. Different from [41], we compute the RMS in each of frequency subbands, get the RMSCC according to Algorithm 1, and then explore the possibility of applying the RMSCC feature to detect the attacks of replay impostors. Similar to the extension of RFCC, SCFC, SCMC, SCMC and RPS, we apply (rectangular) linear, mel-scale, inverted mel-scale filters to RMSCC (line 5 of Algorithm 1), and correspondingly form three features: RMSCC-LIN, RMSCC-MEL and RMSCC-IMEL.

Based on the results in Section VI, LPCC [36]) and LFCC [18] are chosen in the ASV model training for accuracy. Meanwhile, MFCC [34], SCMC-MEL, APGDF [38], and RMSCC-MEL are employed in the CM model training to resist replay attacks. Finally, they form 8 fusion systems whose performance are evaluated in Section VI.

Algorithm 1 RMSCC Extraction

```
Require: Passphrase $P$, sampling frequency $f_s$, frame length $l$, number of filter banks $M$, filter bank scale $sca$
Ensure: Feature $RMSCC$
1. Silence Removing [Optional]. Remove the silence segments in $P$ using voice activity detector (VAD) [42].
2. Prepossessing. Prepossess $P$ using digital filter.
3. Speech Framing. Partition $P$ of length $L$ into $N$ half-overlapping frames $X_1$, ..., $X_N$ of identical length $l$.
4. Short-Time Fourier Transformation (STFT). Transform each frame using fast fourier transformation (FFT) weighted by a hanning window (HW):
   $F_n = \mathrm{FFT}(\mathrm{HW}(X_n)), \quad n \in \{1, \ldots, N\}$.
5. Frequency Partitioning. Create a vector $SF$ of $M + 2$ points linearly, MEL, or IMEL spaced in the interval $V = [f_{smin}, f_{smax}]$ according to $sca$. Partition $V$ into $M$ frequency banks $FB_m$ using a triangular-shaped membership function (trimf) according to the points in $SF$.
6. Filter Banking and RMS Computing. Compute the RMS energy $RMS_{n,m}$ of each frequency band $FB_m$ per frame $F_n$ as follows:
   $$RMS_{n,m} = \sqrt{\frac{|F_n|^2 \cdot FB_m}{\mathrm{Len}(FB_m)}} \quad n \in \{1, \ldots, N\}, \quad m \in \{1, \ldots, M\} \tag{1}$$
   where $\mathrm{Len}(FB_m)$ gets the number of non-zero points in $FB_m$ and the symbol $\cdot$ denotes scalar dot product of vector $|F_n|^2$ and $FB_m$.
7. Feature Computation. Compute the static feature through log and discrete cosine transform (DCT): $RMSCC = \mathrm{DCT}(\mathrm{LOG}(RMS))$, where $RMS$ is feature matrix with elements $RMS_{n,m}$ ($n \in \{1, \ldots, N\}$, $m \in \{1, \ldots, M\}$).
8. Feature Velocity ($\Delta$) [Optional]. Compute the deltas (derivatives) of the static feature:
   $RMSCC_{\Delta} = \mathrm{Derivative}(RMSCC)$.
9. Feature Acceleration ($\Delta\Delta$) [Optional]. Compute the deltas (derivatives) of the feature velocity:
   $RMSCC_{\Delta\Delta} = \mathrm{Derivative}(RMSCC_{\Delta})$.
```

2) Classifier Training

- **GMM-ML Model.** In the scenario, attack detection is a two-class problem (genuine or replay). We adopt GMM with maximum likelihood estimation (GMM-ML) [43] as the anti-replay CM classifier in which GMM is the weighted combination of multivariate Gaussian distributions.
- **GMM-UBM Model.** Taking GMM as the likelihood function, GMM with universal background model (GMM-UBM) [44] uses the UBM to represent the alternative speakers and introduces Bayesian adaptation (e.g., maximum a posteriori, MAP) to generate speaker-specific models from the UBM. In this study, GMM-

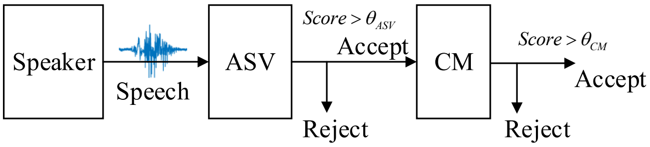

FIGURE 5: Fusion of ASV and CM.

UBM is employed in ASV training.

B. FUSION OF ASV AND ANTI-REPLAY CM

In practical use, both ASV and CM should accept genuine utterances, while either ASV or anti-replay CM should reject the utterances from (zero-effort or replay) impostors. Thus in [45], [46], a straightforward solution was formed in terms of a cascaded combination of ASV and CM. For further improvement in the particular scenario with one speaker, we propose a cascaded fusion system of ASV and CM whose framework is given by Fig. 5. If an utterance passes the ASV's verification ($score > \theta_{ASV}$), it continues to be checked by CM. If passing the verification of the CM ($score > \theta_{CM}$), a legitimate target speaker is declared.

C. VOICEPRINT-BASED KEY AGREEMENT

Even though the target speaker has passed the verification, the attacker still has a chance to tamper the dosage level before insulin delivery. Therefore, we propose a voiceprint-based key agreement protocol as a defense. By using the protocol, the chance can only be got by the USB who is in close proximity of the pump and the target speaker.

1) Non-interactive Extraction for Energy-difference-based Voiceprint

Following the speaker verification, the key agreement protocol will be started by the pump and the USB synchronously. At first, a voiceprint is separately extracted by each device from the sample audio. Towards reducing the complexity, we improve the voiceprint extraction scheme of [13], [47] to form a non-interactive scheme. In Algorithm 2, we use voice active detector (VAD) [42] to find the beginning and end of each voice. Then, the voiceprint of each frame is extracted independently according to the energy difference of adjacent filter banks. This strategy avoids the alignment between the two voices recorded in the insulin pump and USB. Therefore, this voiceprint extraction algorithm is non-interactive and real-time (only once computation required by each device) whereas the algorithm in [13] has to align the voices recorded and compute hundreds of voiceprints before finding out the two matched voiceprints.

Algorithm 2 Non-interactive Energy-difference-based Voiceprint Extraction

```
Require: Passphrase $P$, frame length $l$, preset sampling frequency $f_s$, number of filter banks $M$
Ensure: Voiceprint $f$
  1. Resampling. Resample $P$ to $f_s$ if its sampling frequency is not $f_s$.
  2. Silence Removing. Remove the silence before the beginning and after the ending of the active voice $P$ using voice active detector (VAD) [42] and get the active segments $X = \mathrm{VAD}(P, f_s)$.
  3. Framing. Partition $X$ of length $L$ into $N$ non-overlapping frames $X_1, ..., X_N$ of identical length $l$, such that $N = \mathrm{ceil}(^L/_l)$ (zero-padding for the last frame if $L$ is not a multiple of $l$).
  4. Short-Time Fourier Transformation (STFT). Transform each frame using Fast Fourier Transformation (FFT) weighted by a hanning window (HW):
     $F_n = \mathrm{FFT}(\mathrm{HW}(X_n)), \quad n \in \{1, \ldots, N\}$.
  5. Frequency Partitioning. Create a vector $SF$ of $M + 2$ points linearly spaced in the interval $V = [f_{s_{min}}, f_{s_{max}}]$. Partition $V$ into $M$ frequency banks $FB_m$ using a triangular-shaped membership function (trimf) according to the points in $SF$.
  6. Filter Banking and Energy Computation. Compute the energy $E_{n,m}$ of each frequency band $FB_m$ per frame $F_n$ by:
     $$E_{n,m} = |F_n|^2 \cdot FB_m, n \in \{1, \ldots, N\}, \quad m \in \{1, \ldots, M\} \qquad (2)$$
  7. Voiceprint Computation. Compute each bit $f_{n,m}$ of the voiceprint $f$ with length $N(M-1)$ bits by:
     $$f_{n,m} = \begin{cases} 1, & (E_{n,m} - E_{n,m+1}) > 0, \\ & n \in \{1, \ldots, N\}, \\ & m \in \{1, \ldots, M-1\} \\ 0, & \text{otherwise.} \end{cases} \qquad (3)$$
```

2) Key Agreement using Voiceprint

The pump and the USB may potentially use different audio sensors, and different hardware characteristics also exist in the same type of sensors. There is a high probability that two voiceprints respectively extracted by the two devices are similar yet not the same. The direct use of them as the common key is infeasible. Consider this problem, we propose a key agreement protocol which will abort if the two voiceprints' similarity η is less than a preset threshold, i.e., $\eta < 1 - \tau$. The protocol comprises an energy-difference-based voiceprint extraction algorithm (Algorithm 2) and an adaptive RS-coding-based fuzzy extractor (Algorithm 3). The employed fuzzy extractor is derived from the RS Codes [12], [13]. And we implement the RS coding by introducing the construction method of [48], [49]. Different from [13] using RS Codes with fixed parameters, we prefer an adaptive RS coding algorithm, which means that the length of the real-time voiceprints determines the parameter setting of RS.

When Algorithm 3 (SAFE) is finished, Alice and Bob share a common secret r, which can be used as the seed for

```
Algorithm 3 Secret Agreement using Fuzzy Extractor (SAFE)
Require: Alice's voiceprint $f_1$, Bob's voiceprint $f_2$, Error tolerance threshold $\tau$.
Ensure: Alice and Bob share a common secret $r$, or abort with an error.
1. Alice generates a random number (bit stream) $r$ with length $l_r = l_f(1-2\tau)$, $l_f$ is the length of $f_1$.
2. To tolerate or correct $t = l_f\tau$ bit errors, Alice encodes $r$ to get a codeword $c$ using $RS(n,k)$ coding algorithm satisfying: $k = n - 2t, k \geq l_r, n = (2^s - 1)$, $s$ is the chosen symbol size of RS coding and the lowest number such that $n \geq l_r + 2t$. If $k > l_r$, $r$ will be extended by adding $k - l_r$ zeros to its end before encoding. Note that, we use systematic code (native data bits are left unchanged and the parity bits are appended) and the zero-padding needs not to be transmitted. The receiver will reinsert the zero-padding before decoding.
3. Alice computes $h_1 = \mathrm{HASH}(\mathrm{m}_1), m_1 = c||f_1||n||k||l_r$ and sends $m = \langle c_1 = c + f_1, n, k, l_r\rangle$ and $h_1$ to Bob.
4. Bob adds $f_2$ to $c_1$ and gets $c_2 = c_1 + f_2 = c + f_1 + f_2$. Then Bob decodes $c_2$ using $RS(n,k)$ decoding algorithm to get $c$ and $r$ (first $l_r$ bits of $c$) if the Hamming Distance between $f_1$ and $f_2$ $\mathrm{HammingDistance}(f_1, f_2) \leq t$ and the voiceprint similarity $\eta \geq 1 - \tau$. Bob then gets Alice's voiceprint $f_1$ by computing $f_1 = c_1 + c$. At this time, Bob can check the integrity of $m$ ( and $m_1$) by comparing $\mathrm{HASH(c||f_1||n||k||l_r)}$ and the received hash value $h_1$ to confirm that $m$ is not modified. If the integrity check fails, Bob will abort the protocol and report an error. Otherwise, Bob sends $\mathrm{MAC}_r(\mathrm{f_1})$ to Alice.
5. Alice checks $\mathrm{MAC}_r(\mathrm{f_1})$ using $r$ and $f_1$ and reports an error if the check fails.
```

common key generation. Note that the two voiceprints f_1 and f_2 in SAFE should have the same length or f_2 should be truncated or padded with zero to the end to have the same length with f_1. Using SAFE as a basic unit, Algorithm 4 shows the whole key agreement protocol.

V. SECURITY ANALYSIS

A. REMOTE IMPERSONATION

If an attacker locating outside the close proximity of the pump/USB/speaker wants to participate the key agreement, he must obtain an effective voiceprint to pass the similarity validation (step 4 in Algorithm 3). To achieve this objective, two methods may be possible, i.e., generating a random bit sequence, or extracting from previously (in close proximity) or currently (in a different context, e.g., outside the door) recorded voice. For the prior, the voiceprints frequently have high entropy [13]. If the voiceprint length is long enough (e.g., $\geq$ 512 bits), the probability of guessing a voiceprint which has the similarity η $\geq$ the preset threshold (e.g., $\eta \geq 85.00\%$, $\tau = 15\%$) is usually negligible. For the latter, as the evaluation results shown in Section VI-G, there is a significant gap between the two voiceprints (the maximal $\eta \leq 65.00\%$) even though the adversary eavesdrops the same audio source in another context. However, through remotely getting a user's voice or using the record, the attacker can hardly get a voiceprint with high similarity with that of in the pump due to different contexts.

```
Algorithm 4 Secure Key Agreement (SKA)
Require: Insulin pump (Alice), USB (Bob), Speaker or user, Error tolerance threshold $\tau$
Ensure: Common key $K$, or abort with an error
1. The speaker says a random passphrase to Alice and Bob after the access control functionality in the two devices is activated by the speaker, e.g. using a button in each device.
2. Alice and Bob extract voiceprints $f_1$ and $f_2$, respectively, from the voice of the speaker according to Algorithm 2.
3. Alice and Bob share a common secret $r$ using SAFE (Algorithm 3) and Bob gets the voiceprint of Alice $f_1$ if $^{\mathrm{HammingDistance}(f_1,f_2)}/_{\mathrm{Length}(f_1)} \leq \tau$ holds.
4. Alice and Bob compute the common key: $K = \mathrm{HASH}(r||\,\mathrm{HASH}(f_1))$ (The symbol $||$ denotes bit string concatenation), respectively.
```

B. PASSIVE EAVESDROPPING

During the procedure of key agreement, all the exchanged messages could be collected by an attacker through channel eavesdropping. Meanwhile, the attacker can record the voice from the legitimate speaker. These problems generally lead to replay attacks. By using the cascaded fusion of speaker verification and anti-replay CM, experiments in Section VI confirm the mitigation of the replay attacks with high accuracy brought by the proposed speaker verification scheme. Actually, whether the attacker can learn partial or all information with respect to the exchanged voiceprints or not is critical for the security of the key agreement. The attacker may use brute-force and message eavesdropping to reach his objective. However, for brute-force attack, the key agreement requires that the extracted voiceprints have a long length (e.g., $\geq$ 512 bits) and the user speaks random passphrases with a long duration (e.g., $\geq$ 2 s). Thus a high entropy and no bias in bit distribution of the voiceprints can be guaranteed. For message eavesdropping, our protocol adopts an RS-based fuzzy extractor (SAFE in Algorithm 3) to exchange a common secret. No information about the voiceprints is leaked given the following three conditions.

- The random number used in Algorithm 3 is independently and uniformly distributed.
- The setting $\langle 2^s, n, k\rangle$ of RS coding is strong enough and the space of data words and codewords is large enough
- The entropy of the voiceprints is high. This can be ensured by randomly spoken passphrases and long voice duration (long voiceprint).

In the implementation, we set the symbol size $s = 10$, the number of symbols (including data and parity) of each codeword $n = 2^s - 1 = 1023$, and the number of data symbols $k \geq 512$. The data symbols (i.e., random numbers) can be chosen from a space $\geq (2^{10})^{512}$, which is large enough to ensure the codewords have sufficient entropy and do not leak the voiceprint's information (f_1 in Algorithm 3).

C. MAN-IN-THE-MIDDLE

If an eavesdropper Eve performs the attacks as a middle person in between Alice and Bob, she has to tamper at least one message during the key agreement. Otherwise, this is a passive eavesdropping. According to the remote impersonation analysis, Eve can hardly complete the agreement protocol with either Alice or Bob. Any modification happens to any message in the protocol (i.e., $m, h_1, \text{MAC}_r(\text{f}_1)$ in Algorithm 3), Alice and Bob will fail to finish the key agreement.

VI. EXPERIMENTS EVALUATION

To verify the performance of the proposed scheme, a series of experiments are conducted in this section. To make a full description of the conducted experiments, this section consists of 6 parts. 1) Section VI-A gives the data description and metrics used for evaluations; 2) Section VI-B presents evaluations for different features and the candidates selection method for ASV; 3) Section VI-C and Section VI-D respectively conduct both training and testing for the standalone ASV using features chosen in case of zero-effort and replay imposters; 4) Section VI-E considers the training and testing for the standalone CM in case of replay imposters; 5) Section VI-F checks the performance of our cascaded fusion scheme in case of zero-effort and replay impostors; 6) Section VI-G demonstrates the feasibility of the voiceprint based key agreement. All the experiments were executed in a virtual machine server with 8 Cores (Intel Broadwell 2.29 GHz) and 64 GB RAM.

A. DATASETS, PROTOCOL, AND METRICS

Datasets. To conduct experimetns, we use the ASVspoof 2017 challenge corpus which is primarily employed for spoofing (replay) attack detection in [26]. It consists of genuine and replay/spoof recordings with the former coming from recent text-dependent RedDots corpus and the latter from the replay version of the former [50], [51]. The data set is separated into three subsets for training, development and evaluation which are described in Table 2. We use version 2.0 [52] and purge the evaluation protocol file by removing the file items which do not exist in the Evaluation subset.

TABLE 2: 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

Protocol. In the ASVspoof 2017, the Training and Development subsets are suggested for finding the replay CMs while the Evaluation subset is provided to demonstrate the accuracy and generalization capacity of the replay detectors. Since the proposed scheme concentrates on a particular scenario with only one speaker, we use the data set in different ways, which are explained in following subsections.

Metrics. In the context of the ASV, there are two types of inputs: genuine and zero-effort impostor speech. For the CM, there is another type namely spoof/replay impostor speech. Both the ASV and CM should accept the genuine speech; the ASV should reject the zero-effort impostor while the CM should reject the spoof/replay impostor. Following [53], the decisions of ASV and CM corresponding to genuine trial and zero-effort/replay trial are illustrated in Table 3. To evaluate the performance of the ASV and CM, the preferred metrics are as follows:

- False Acceptance Rate (FAR): Percentage of impostor trials (incorrectly) accepted by the ASV and CM system. It corresponds to a threshold θ computed by:

$$\text{FAR}(\theta) = \frac{\text{Number of impostor trials with score} > \theta}{\text{Number of all impostor trials}} \quad (4)$$

- False Rejection Rate (FRR): Percentage of genuine trials (incorrectly) rejected by the ASV and CM, which corresponds to a threshold θ. It is computed by:

$$\text{FRR}(\theta) = \frac{\text{Number of genuine trials with score} \leq \theta}{\text{Number of all genuine trials}} \quad (5)$$

- Equal Error Rate (EER): It corresponds to a threshold θ_{EER} at which FAR and FRR are equal or approximately equal. The threshold is determined in development phase utilizing a reference database $\mathcal{D}_{base}$ by:

$$\theta_{\text{EER}} = arg\min_{\theta} |\text{FAR}(\theta, \mathcal{D}_{base}) - \text{FRR}(\theta, \mathcal{D}_{base})| \quad (6)$$

TABLE 3: Trial Decisions of ASV and CM

Trials	**Decision**	
	Acceptance	*Rejection*
Genuine trial	Correct acceptance	False rejection
Zero-effort trial	False acceptance (ASV)	Correct rejection (ASV)
Spoof/Replay trial	False acceptance (CM)	Correct rejection (CM)

B. MODEL FEATURE SELECTION

Feature Selection. To choose the appropriate features for ASV, we have evaluated 12 different features (shown in Section IV, the same kind of features with different filter bank implementations are considered as one feature) using the ASVspoof 2017 Training subset.

Classifier and Parameters. To train ASV, we use the GMM-UBM model with 256 GMM mixtures in 20 iterations. It is implemented using the MSR Identity Toolbox V1.0 [54]. For each feature, we have evaluated three combinations of static and dynamic coefficients: static, static + deltas (Δ), static + deltas (Δ) + double-deltas ($\Delta\Delta$). The deltas (Δ)

TABLE 4: Samples Composition of Training Subset of ASVspoof 2017 Corpus

Training Subset of ASVspoof 2017 Corpus			
Speakers	*Genuine*	*Spoof*	*Total*
M0001	280	280	560
M0002	258	258	516
M0003	30	30	60
M0004	90	90	180
M0005	90	90	180
M0006	150	150	300
M0007	120	120	240
M0008	200	200	400
M0009	189	189	378
M0010	100	100	200
Total	1507	1507	3014

is the derivatives of the static coefficients and the double-deltas (ΔΔ) is the derivatives of the deltas (Δ). For CQCC (no using of STFT), the number of cepstral coefficients is 30 (with an additional 0th coefficient). For the other features (with STFT), we adopt 20ms frame length and 40 filter banks.

Training and Testing Datasets. To train and evaluate the GMM-UBM classifier, we utilize the Training set of ASVspoof 2017, which contains utterances of a total of 10 speakers. Table 4 shows the numbers of samples for each speaker. The least number of the utterances of the user (i.e., M0003) used in the experiments is 30 and the total duration of these utterances is about 90 s. 70% of all the genuine utterances of the Training set are used to train the UBM model, and 70% of genuine utterances per speaker are used to train the speaker-specific model. Then the remaining 30% utterances are used for verification. All possible model-test combinations are taken into the verification trials (30% of one target speaker vs. 30% of all the other impostors' utterances).

Performance. Table 5 lists the performance in terms of ERR. We find that voice activity detector (VAD) is critical for most features. Except CQCC and RPS, which achieve better performance without VAD, other features achieve better performance when using VAD. Of all the evaluated features, LPCC (static + Δ + ΔΔ), SCMC-MEl (static + Δ), LFCC (static + Δ + ΔΔ), CQCC (static + Δ), and RMSCC-LIN (static + Δ + ΔΔ) achieve better performance than others, the EERs of which are 0.15%, 0.22%, 0.25%, 0.25%, and 0.25%, respectively. Base on the above results, we chose 5 features: LPCC (static + Δ + ΔΔ), SCMC-MEl (static + Δ), LFCC (static + Δ + ΔΔ), CQCC (static + Δ), and RMSCC-LIN (static + Δ + ΔΔ) as candidates to train the ASV. Although LPCC (static + Δ) also achieves good performance (0.25%), we only chose LPCC (static + Δ + ΔΔ), which has the best performance and dropped LPCC (static + Δ) to evaluate more others features.

TABLE 5: Standalone ASV Feature Performance (% EER) Based on Traing Subset of ASVspoof 2017 Corpus

Features	*Selected Coefficients*		
	(Stat)	*(Stat+Δ)*	*(Stat+Δ + ΔΔ)*
CQCC (no VAD)	0.34	**0.25**	0.44
MFCC	0.54	0.39	0.89
IMFCC	0.89	0.89	0.74
LPCC	0.44	0.25	**0.15**
LFCC	0.66	0.49	**0.25**
RFCC-LIN	0.57	0.54	0.44
RFCC-MEL	0.69	0.44	0.66
RFCC-IMEL	1.33	1.11	1.33
SCFC-LIN	1.63	1.06	1.11
SCFC-MEL	1.16	0.89	1.11
SCFC-IMEL	3.10	1.77	3.32
SCMC-LIN	0.91	0.44	0.44
SCMC-MEL	0.49	**0.22**	0.66
SCMC-IMEL	1.11	0.66	1.11
SSFC-LIN	1.20	0.81	1.99
SSFC-MEL	2.24	1.55	1.79
SSFC-IMEL	2.68	2.43	3.39
APGDF	1.11	1.89	2.70
RMSCC-LIN	0.66	0.49	**0.25**
RMSCC-MEL	0.54	0.39	0.89
RMSCC-IMEL	0.89	0.89	0.74
RPS-LIN (no VAD)	5.75	5.31	5.73
RPS-MEL (no VAD)	6.24	7.30	8.14
RPS-IMEL (no VAD)	5.68	5.97	5.53

C. STANDALONE ASV PERFORMANCE AGAINST ZERO-EFFORT IMPOSTORS

Selected Features. Following Section VI-B, we evaluate the performance of LPCC (static + Δ + ΔΔ), SCMC-MEl (static + Δ), LFCC (static + Δ + ΔΔ), CQCC (static + Δ), and RMSCC-LIN (static + Δ + ΔΔ) in case zero-effort impostors use their own sounds to impersonate the genuine target speaker.

Training and Testing Datasets. We use the Training subset shown in Table 4 and Development subset shown in Table 6 to train and test the two-class GMM-UBM classifier. In each evaluation of each feature for each speaker, we use *except one cross validation*, i.e., one of the 10 speakers is chosen as the enrollment (target) speaker, and the remaining 9 speakers are treated as zero-effort impostors. Their utterances are employed in training the speaker-specific model and the impostor model, respectively. To train the two-class GMM-UBM model, 70% of genuine utterances from the target speaker are combined with all of the other 9 impostors' genuine utterances. Meanwhile, 30% genuine utterances of the target speaker combined with all genuine utterances of the Development subset containing utterances from 8 speakers are used to verify the performance.

Performance. From Table 7 (columns 2-6), we can apparently find that the largest EERs achieved by all the 5 features on all 10 speaker models are 2.50%, 2.50%, 3.81%, 6.67%, and 17.37%, respectively. The first three features which achieve better performance are LFCC (static + Δ + ΔΔ), RMS-LIN (static + Δ + ΔΔ), and LPCC (static + Δ + ΔΔ).

TABLE 6: Samples Composition of Development Subset of ASVspoof 2017 Corpus

Development subset of ASVspoof 2017 Corpus			
Speakers	*Genuine*	*Spoof*	*Total*
M0011	140	200	340
M0012	50	40	90
M0013	90	90	180
M0014	60	20	80
M0015	30	50	80
M0016	190	240	430
M0017	90	140	230
M0018	110	170	280
Total	760	950	1710

D. STANDALONE ASV PERFORMANCE AGAINST REPLAY IMPOSTORS

Selected Features. Based on the trials in Section VI-B and Section VI-C, we have evaluated the performance of the same features used in Section VI-C. Here, recordings of the target speaker (or someone else) are employed by the replay impostors.

Training and Testing Datasets. The Training subset shown in Table 4 was used to train a two-class GMM-UBM classifier. In each iteration, we choose 1 of the 10 speakers as the target speaker to train the speaker-specific model, and the others are considered as the replay impostors to build impostor model. To build the GMM-UBM model, 70% of the genuine utterances of the target speaker and all of the genuine utterances of the other 9 impostors are chosen. Meanwhile, we take 30% genuine along with all the spoof utterances of the target speaker and the whole Development subset shown in Table 6 to predict and check the performance. All the spoof utterances of the target speaker are used to evaluate the performance against replay attacks while the whole Development subset is used to evaluate zero-effort attacks.

Performance. As shown in Table 7 (columns 7-11), the performance against the replay impostors is worse with larger EERs than that against the zero-effort impostors. The largest EERs achieved by all the 5 features on all 10 speaker models are 8.28%, 9.03%, 9.03%, 13.33%, and 15.56%, respectively. The first three features, which achieve better performance are LPCC (static + Δ + $\Delta\Delta$), LFCC (static + Δ + $\Delta\Delta$), and RMSCC-LIN (static + Δ + $\Delta\Delta$). They also achieve better performance in the 5 evaluated features when against zero-effort imposters.

E. STANDALONE CM PERFORMANCE AGAINST REPLAY IMPOSTORS

Selected Features. To evaluate the performance of standalone CM against replay impostors, all the features used in Section VI-B are selected. We find the performance will deteriorate when adding VAD before feature extraction. So in all the trials, we do not use VAD.

Classifier and Parameters. The employed GMM-ML model for training CM is with 512 GMM mixtures and 100 iterations. The GMM-ML implementation is based on the baseline implementation of ASVspoof 2017 [26]. Same with the evaluation in Section VI-B, for each feature we evaluate three combinations of static and dynamic coefficients: static, static + deltas (Δ), static + deltas (Δ) + double-deltas ($\Delta\Delta$). For all features except LPCC and CQCC, we adopt 20 ms frame length and 40 filter banks. For LPCC, the configuration with the 200 ms frame length and 40 filter banks achieves better performance. For CQCC (not using STFT as other features), the number of cepstral coefficients is 30.

Training and Testing Datasets. We use the Development subset shown in Table 6 and Evaluation subset shown in Table 8 to build the two-class GMM-ML. Specifically, all the genuine utterances of the Evaluation subset are employed to train the genuine model, while its spoof utterances are used to train the spoof model. Meanwhile, to evaluate the trained models, we use the whole Development subset.

Performance. Table 9 shows the performance of the standalone CM against replay impostors. We find that, of all the evaluated features, APGDF (static + Δ), SCMC-MEL (static), MFCC (static), and RMSCC-MEL (static) outperform the others, the EERs of which are 4.48%, 4.86%, 4.94%, and 5.05%, respectively. Base on the results, we chose these 4 best features as candidates to train the CM in the fusion system, which will be shown in Section VI-F.

F. ASV & CM FUSION PERFORMANCE AGAINST ZERO-EFFORT AND REPLAY IMPOSTORS

Selected Features. Combining the results of standalone ASV (in Section VI-B, Section VI-C, and Section VI-D) and the results of standalone CM (in Section VI-E), we adopt LPCC (static + Δ + $\Delta\Delta$) and LFCC (static + Δ + $\Delta\Delta$) to train the standalone ASV models, and then use MFCC (static), SCMC-MEL (static), APGDF (static+Δ) and RMSCC-MEL (static) to train the standalone CM models. For the chosen features' diversity, we do not choose RMSCC-LIN (static + Δ + $\Delta\Delta$) to train the ASV models even though it archives an equivalent performance to LPCC (static + Δ + $\Delta\Delta$). Because we have chosen RMSCC-MEL (static) to train the CM. The serial concatenation of the standalone ASV and CM forms the fusion system to mitigate the attacks from both the zero-effort and the replay impostors.

Datasets and Thresholds Determination. The Training subset (Table 4) is employed in both of the ASV GMM-UBM model training and testing, as well as determining the threshold (θ_{ASV}). If an utterance gets a score $\leq \theta_{ASV}$, it might be a zero-effort or replay impostor rejected by the ASV; otherwise, it needs one more check in CM. The other two data sets, Development subset (Table 6) and Evaluation subset (Table 8), are used to train the GMM-ML model for CM. In the Evaluation subset, all the utterances of each speaker, both the genuine and spoof, are employed in training a GMM-ML model. The whole Development subset is used to test the model and compute a CM threshold (θ_{CM}). Similarly, each utterance with a score $\leq \theta_{CM}$ is considered as a spoof/replay attack. The second check in the CM is to find out the utterances falsely accepted by the ASV(i.e.,

TABLE 7: Standalone ASV Performance against Zero-effort and Replay Impostors (% EER) based on Training and Development Subsets of ASVspoof 2017 Corpus

Speakers	**Zero-effort Impostors**					**Replay Impostors**				
	CQCC (no VAD)	*LPCC*	*LFCC*	*SCMC-MEL*	*RMSCC-LIN*	*CQCC (no VAD)*	*LPCC*	*LFCC*	*SCMC-MEL*	*RMSCC-LIN*
M0001	1.05	2.38	1.19	0.13	1.19	0.55	1.11	1.19	0.15	1.19
M0002	0.00	0.13	0.13	0.13	0.13	0.91	1.02	0.91	0.30	0.91
M0003	0.00	0.92	1.45	2.76	1.45	3.28	2.13	7.36	11.11	7.36
M0004	1.58	0.00	0.00	0.39	0.00	3.06	3.67	3.11	3.70	3.11
M0005	2.37	1.71	0.26	2.37	0.26	1.06	0.61	0.56	1.78	0.56
M0006	9.47	**3.81**	**2.50**	2.22	**2.50**	**15.56**	**8.28**	**9.03**	**13.33**	**9.03**
M0007	0.13	0.00	1.58	0.00	1.58	0.05	0.87	0.77	0.05	0.77
M0008	**17.37**	1.05	0.53	**6.67**	0.53	10.00	3.35	1.67	6.67	1.67
M0009	1.75	0.92	0.00	1.05	0.00	0.05	2.42	0.05	1.26	0.05
M0010	1.18	0.26	0.26	0.13	0.26	0.50	1.65	2.15	0.11	2.15

TABLE 8: Samples Composition of Evaluation Subset of ASVspoof 2017 Corpus

Evaluation Subset of ASVspoof 2017 Corpus					
Speakers	*Genuine*	*Spoof*	**Speakers**	*Genuine*	*Spoof*
M0019	40	524	M0032	100	1240
M0020	119	1066	M0033	70	630
M0021	69	562	M0034	29	284
M0022	110	798	M0035	59	572
M0023	60	584	M0036	10	-
M0024	50	531	M0037	10	-
M0025	139	1501	M0038	10	-
M0026	50	516	M0039	10	-
M0027	89	756	M0040	10	-
M0028	60	674	M0041	10	-
M0029	90	801	M0042	10	-
M0030	40	437	Total	1294	11988
M0031	50	512	-	-	-

TABLE 9: Standalone CMs Replay Detection Performance (% EER) on the Development and Evaluation Subsets of ASVspoof 2017 Corpus

Features	**Selected Coefficients**		
	(Stat)	*(Stat+Δ)*	*(Stat+$\Delta + \Delta$)*
CQCC	9.33	9.22	8.33
MFCC	**4.94**	5.93	6.77
IMFCC	6.62	6.94	7.84
LPCC	6.97	6.53	6.97
LFCC	7.31	6.65	6.49
RFCC-LIN	7.95	6.60	6.76
RFCC-MEL	7.51	7.89	7.87
RFCC-IMEL	7.51	6.82	9.63
SCFC-LIN	29.06	29.06	29.06
SCFC-MEL	30.98	30.98	30.98
SCFC-IMEL	45.09	45.09	45.09
SCMC-LIN	7.14	6.03	6.45
SCMC-MEL	**4.86**	5.70	6.21
SCMC-IMEL	7.26	6.67	8.63
SSFC-LIN	9.34	8.18	9.95
SSFC-MEL	11.75	12.06	14.31
SSFC-IMEL	11.16	10.21	11.24
APGDF	5.66	**4.48**	5.21
RMSCC-LIN	7.33	6.39	6.78
RMSCC-MEL	**5.05**	6.07	5.79
RMSCC-IMEL	6.41	6.88	7.94
RPS-LIN	28.68	30.45	36.75
RPS-MEL	40.65	40.85	41.64
RPS-IMEL	31.64	32.87	35.42

false positive). Notice that the false negative rate (a genuine utterance is considered as a spoof one) potentially increases due to the fusion policy. In that case, one more trial is recommended to guarantee safety.

Performance of Fusion Systems. We have evaluated LPCC (static + Δ + $\Delta\Delta$) and LFCC (static + Δ + $\Delta\Delta$) in ASV and MFCC (static), SCMC-MEL (static), APGDF (static + Δ), and RMSCC-MEL (static) in CM. Table 10 shows the corresponding performance of the 8 different fusions. Apparently, the max EERs of all speaker models in each of 8 evaluated fusion systems are 4.44%, 6.67%, 3.33%, 3.33%, 3.33%, 6.67%, 6.67%, and 2.22%, respectively.

Fusion System with Best Performance. Fusion8 achieves the best performance with the EER of 2.22%. It is the fusion of ASV2 (LFCC) and CM4 (RMSCC-MEL). To defense against zero-effort and replay impostors, we fix the FRR of 10% and then evaluate the performance of standalone ASV2 and fusion8 in terms of EER (%) and FAR (%) for all the speakers' models in Training subset of AVSspoof 2017. Column 2 of Table 11 gives the number of samples chosen from Training subset and used to test each speaker model. Comparing with the standalone ASV, the fusion system reduces the max EER from 3.17% to 2.22% and reduces the max FAR from 1.90% to 0.92%. Therefore, we believe that the fusion of ASV and CM improves performance.

G. FEASIBILITY OF THE NON-INTERACTIVE VOICEPRINT EXTRACTION BASED KEY AGREEMENT

Once the speaker passes the verification, the voiceprint-based key agreement will be launched to bootstrap a secure communication channel. By taking the non-interactive voiceprint extraction (Algorithm 2) as a basic unit, we demonstrate the feasibility of the key agreement (Algorithm 4).

Voice Samples Collection. To collect voice samples, we use iPhone 5S and Honor 10 (H10) to record 135 passphrases, respectively (total 270). In each case, we take a person speaking the passphrase as a voice source, and the two devices are separately positioned at one of 27 different distance settings

TABLE 10: ASV and CM Fusion Replay Detection Performance (% EER) based on ASVspoof 2017 Corpus

System	Speakers									
	M0001	*M0002*	*M0003*	*M0004*	*M0005*	*M0006*	*M0007*	*M0008*	*M0009*	*M0010*
ASV1 (LPCC)	2.16	0.20	0.80	1.22	3.17	0.43	0.05	1.94	0.32	0.33
ASV2 (LFCC)	1.19	0.10	1.90	1.67	0.28	3.12	0.38	1.67	1.75	0.33
CM1 (MFCC)	4.08	4.04	4.62	4.48	4.48	5.45	4.37	4.27	4.37	5.18
CM2 (SCMC-MEL)	3.51	3.55	4.06	3.91	3.91	4.91	3.84	3.67	3.95	4.28
CM3 (APGDF)	4.54	3.36	3.86	3.81	3.75	3.60	4.22	4.72	3.43	4.40
CM4 (RMSCC-MEL)	4.47	4.53	5.17	4.99	4.99	6.04	4.91	4.74	4.96	5.34
Fusion1 (LPCC+MFCC)	2.11	0.15	0.52	0.06	2.67	**4.44**	0.05	1.31	0.26	0.11
Fusion2 (LPCC+SCMC-MEL)	2.11	0.15	0.52	0.06	2.72	2.22	0.05	1.20	0.32	**6.67**
Fusion3 (LPCC+APGDF)	2.11	0.10	0.40	0.06	2.61	2.22	0.05	1.67	0.11	**3.33**
Fusion4 (LPCC+RMSCC-MEL)	2.11	0.15	0.52	0.06	2.72	2.22	0.05	1.20	0.32	**3.33**
Fusion5 (LFCC+MFCC)	1.19	0.05	0.80	0.11	0.11	2.22	0.38	1.20	1.75	**3.33**
Fusion6 (LFCC+SCMC-MEL)	1.19	0.05	0.98	0.06	0.17	2.31	0.38	1.15	1.75	**6.67**
Fusion7 (LFCC+APGDF)	1.19	0.05	0.52	0.06	0.11	2.22	0.33	**6.67**	1.75	3.33
Fusion8 (LFCC+RMSCC-MEL)	1.19	0.05	0.92	0.06	0.17	**2.22**	0.38	1.15	1.75	0.11

TABLE 11: Standalone ASV and Fusion8 (ASV LFCC + CM RMSCC-MEL) Performance in Terms of EER (%) and FAR (%) for a Fixed FRR of 10% for All the Speakers Against Zero-effort and Replay Impostors

Speakers	# Samples	ASV1		Fusion8 (ASV1+CM4)	
		EER	*FAR*	*EER*	*FAR*
M0001	2074	2.16	0.00	1.19	0.00
M0002	2045	0.20	0.05	0.05	0.05
M0003	1749	0.80	**1.90**	0.92	**0.92**
M0004	1827	1.22	1.11	0.06	0.06
M0005	1827	**3.17**	0.11	0.17	0.06
M0006	1905	0.43	0.59	**2.22**	0.59
M0007	1866	0.05	0.00	0.38	0.00
M0008	1970	1.94	0.26	1.15	0.10
M0009	1956	0.32	0.00	1.75	0.00
M0010	1840	0.33	0.17	0.11	0.11

TABLE 12: Average Similarity of Voiceprints Generated by Two Devices at Different Distances to the Same Voice Source

Distance (cm)	Average voiceprints similarity (%)					
	5S 20	*5S 30*	*5S 50*	*5S 150*	*5S 300*	*5S outside*
H10 20	88.20	87.73	85.84	78.48	78.96	59.43
H10 30	85.55	85.64	84.48	79.07	79.59	61.78
H10 50	85.83	84.82	84.59	78.69	76.74	58.45
H10 150	84.15	84.42	82.89	-	-	-
H10 300	79.25	82.47	83.13	-	-	-
H10 outside	64.16	59.96	62.67	-	-	-

from the person. The employed distances are shown in the first column and the second row of Table 12. In each distance setting, the speaker speaks 5 sentences. Each sentence should contain either 4 or 5 English words, or 5 numbers (in closed interval [0, 9]), and persist 1 $\sim$ 3 seconds. We use different distances among the speaker and the mobile phones to demonstrate the relative positions among the pump, the USB, and the user (or the attacker). Table 12 illustrates all the adopted distance settings and the similarity evaluations for voiceprints generated by the two devices.

Voiceprint Extraction Setting. For voiceprint extraction shown in Algorithm 2, we resample all the passphrases to 16 kHz sampling frequency and use 63 ms frame length and 17 frequency filter banks. The sampling frequencies of the native voice recording application in iPhone 5S and Honor 10 are 44.1 kHz and 48 kHz, respectively.

Performance. Table 12 gives all the experimental results. Some conclusions are revealed. (1) When the distance between the two devices and the voice source is smaller than 30 cm, the average voiceprint similarity (AVS) is larger than 85.00%. (2) If the distance between one device and the speaker is about 300 cm, the AVS drops down to 76.74%. Especially, if one device locates outside the closed door of a room (about 320 cm, mean ambient loudness in the room: 38 dB, outside: 47 dB), the AVS is only 58.45%. Therefore, an attacker cannot get a voiceprint which has a high similarity with the one in the pump or USB in a different context.

VII. DISCUSSION

A. OVERHEAD ANALYSIS

In this section, we analyze the storage consumption, computation complexity, and communication complexity of the whole authentication process, including feature fusion (of ASV and CM) and voiceprint-based key agreement. Considering that the pump can be powered by a battery and the USB by PC, we do not make power consumption analysis in this paper, which can be one of our future work.

Storage consumption. For the proposed scheme, the pump only store classifier models for one patient. Intuitively, this consumes much less storage than the conventional speaker recognition systems. Table 13 (columns 2-7) shows the storage comparisons for the 8 fusion systems. We choose fusion8 as the candidate to be used in the insulin pump system because it outperforms the others. By using fusion8, the storage consumption for the ASV includes one GMM UBM model (482 KB with 256 GMM components), one GMM user model (482 KB), and one GMM background users model (482 KB). For CM, the requirements are one GMM-ML Genuine model (324 KB with 512 GMM components) and one GMM-ML Spoof model (324 KB). So, the total permanent storage consumption is only about 2 MB. Actually, it is not a pricey consumption for the next generation insulin pump which may adopt higher hardware configuration (e.g., with ARMv7/v8 CPU and Flash Memory $\geq$32 MB).

Computation complexity. First, we evaluate the time-costs of feature extraction and score evaluation for all the 8 candidate fusion systems in a laptop with Intel Core (TM) i7-5500 CPU (2.40 GHz) and 8 GB RAM. The results are shown in Table 13 (columns 8-9). The average feature extraction time and score computation time, corresponding to two features (ASV2 LFCC and CM4 RMSCC-MEL) extracted from a voice sample, are 0.0256 s and 0.0145 s, respectively. Then, we get the executing time for the key modules of the fusion8 system in the Raspberry Pi 3 Model B+ with 1.4 GHz Broadcom BCM2837B0 CPU. We select the recorded voices with the duration of around 2s. For speaker verification, the ASV part comprises one audio read ($\approx$ 0.01 s), one feature extraction (LFCC, $\approx$ 0.05 s, including VAD), and one log likelihood computation ($\approx$ 0.04 s); the CM part consumes one feature extraction (RMSCC-MEL, $\approx$ 0.02 s), and one log likelihood computation ($\approx$ 0.04 s). To extract the voiceprint, the scheme generates 16 bits voiceprint for each frame of the voice. The lengths of all the evaluated passhprases range from 17 to 42 frames, i.e., from 272 to 672 bits. For security concern, we prefer the voiceprint with the length $\geq$ 512 bits and adaptive $RS(n = 2^{10} - 1, k = n - 2t = n - 2l_f\tau)$. Thus the number of data symbols k is determined by the length of real-time voiceprint l_f and the error tolerance threshold τ. For the voiceprints with the maximal length $l_f = 672$ bits, we set $\tau = 0.1$, get $k = 1023 - \lceil 2 \cdot 672 \cdot 0.1 \rceil = 888$ and chose $RS(1023, 888)$ in SAFE (Algorithm 3). In the key agreement protocol, the voiceprint extraction algorithm spends $\approx$ 0.05 s (672 bits, including silence removing in Algorithm 2); the creating of the RS sequential-root-generator-polynomial [49] takes $\approx$ 0.002 s, the RS encoding $\approx$ 0.01 s, and the RS decoding $\approx$ 0.02 s. The running time for each of other operations (6 HASHs and 2 MAC in Algorithm 3 and 4) is $\leq$ 0.0001 s, totally $\leq$ 0.001 s. Actually, the computational time for the whole access control is about 1 s.

Communication complexity. For the proposed scheme, message exchange only happens in the voiceprint-based key agreement (SAFE) between the pump and USB. The pump sends 1 HASH (h_1, 256 bits using SHA256), 1 codeword ($m1$, the max length $\approx$ 1024 bits) and receives 1 MAC (160 bits using EVP SHA1) during the SAFE protocol (Algorithm 3). In total, the size of the transmitted data is $\leq$ 2 Kbits. They can be exchanged within 1 s using the RF channel (frequency of pump to USB: 961.5 MHz, bandwidth: 185 kHz). Taking the computational time into account, the whole access control consumes about 2 s if the voice has been recorded. Generally, 3 s is enough for voice recording. Therefore, we can finish this voiceprint based access control within 5 s.

B. HANDLING EMERGENCY SITUATION

In the literature, how to achieve easy access to medical devices under emergencies is an orthogonal problem. Many researches [9], [10], [29] recommend granting open access to clinical staff during emergencies. [21] and [55] presented schemes for emergency cases. In this study, we can deactivate the speaker verification by a button in each device and execute only the key agreement by using the voiceprints extracted from the voice of anyone in case the target user has a throat sickness and cannot say a passphrase accepted by the access control. Although the patient cannot participate in the procedure and grant permission specifically, the insulin pump and USB can still establish a secure channel if the two devices are in close proximity.

VIII. CONCLUSION

Security of the wireless channel between the insulin pump and the USB/uploader is closely related to the patient's safety. In this paper, we have proposed a novel feature fusion and voiceprint based access control scheme. The scheme comprises an anti-replay speaker verification and a voiceprint-based key agreement. Based on the scheme, the insulin pump can only be accessed by the USB after the legitimate user passes the identity verification, and the pump establishes a secure channel only with the device in its close proximity. To generate a common key for communication encryption, we adopt non-interactive and real-time energy-difference-based voiceprint extraction and adaptive RS-coding-based fuzzy extractor to the key agreement protocol. It protects the insulin pump from message eavesdropping and parameters manipulation attacks while avoiding complex computations and data exchanges over the wireless channel. Furthermore, certificates, permanent shared keys, and additional devices are no more required. Finally, how to make the proposed scheme be suitable for various infusion systems or lightweight access control scenarios is to be investigated in the future.

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] S. B. Schwartz, "The medical device ecosystem and cybersecurity — building capabilities and advancing contributions," https://www.fda.gov/NewsEvents/Newsroom/FDAVoices/ucm624749.htm,Nov.1,2018.

[3] J. Butts, and B. Rios, "Understanding and Exploiting Implanted Medical Devices," https://infocondb.org/con/black-hat/black-hat-usa-2018/understanding-and-exploiting-implanted-medical-devices,Aug.9,2018.

[4] 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

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

[6] 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.

[7] E. Marin, D. Singelée, B. Yang, I. Verbauwhede, and B. Preneel, "On the feasibility of cryptography for a wireless insulin pump system," in Proc. of ACM CODASPY, pp.113-120, 2016.

[8] Medtronic, "CareLink USB," https://eshop.medtronic-diabetes.com.au/en/pumpsupplies/otherpumpsupplies/USB#, 2019.

[9] 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.

[10] 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.

TABLE 13: Storage and Computation Overhead of the Candidate Fusion Systems

Systems	**Storage Overhead (KB)**						**Computation Overhead (s)**	
	ASV GMM-UBM Models			*CM GMM-ML Models*		*Total Storage*	*Feature Extraction*	*Score Computation*
	UBM	*Speaker*	*Background Speakers*	*Genuine*	*Spoof*			
Fusion1 (LPCC+MFCC)	482	482	482	324	324	2094	0.0264	0.0142
Fusion2 (LPCC+SCMC-MEL)	482	482	482	324	324	2094	0.0284	0.0146
Fusion3 (LPCC+APGDF)	482	482	482	644	644	2734	0.0420	0.0168
Fusion4 (LPCC+RMSCC-MEL)	482	482	482	324	324	2094	0.0276	0.0143
Fusion5 (LFCC+MFCC)	482	482	482	324	324	2094	0.0244	0.0144
Fusion6 (LFCC+SCMC-MEL)	482	482	482	324	324	2094	0.0264	0.0148
Fusion7 (LFCC+APGDF)	482	482	482	644	644	2734	0.0400	0.0170
Fusion8 (LFCC+RMSCC-MEL)	482	482	482	324	324	**2094**	**0.0256**	**0.0145**

[11] X. Hei, X. Du, S. Lin, and I. Lee, "PIPAC: Patient infusion pattern based access control scheme for Wireless insulin pump system," in Proc. of IEEE INFOCOM, pp.3030-3038, 2013.

[12] I. S. Reed and G. Solomon, "Polynomial codes over certain finite field," Journal of the Society for Industrial and Applied Mathematics, 8(10), pp. 300-304, 1960.

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

[14] T. Hasan, S. Sadjadi, and G. Liu, et al., "UTD-CRSS systems for 2012 NIST speaker recognition evaluation," in Proc. of IEEE ICASSP, pp. 6783-6787, 2013.

[15] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition," in Proc. of ODYSSEY, 2010.

[16] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in Proc. of IEEE ICASSP, 1997.

[17] I. Saratxaga, I. Hernáez, D. Erro, E. Navas, and J. Sánchez, "Simple representation of signal phase for harmonic speech models," Electronics Letters, 45(7), pp. 381-383, 2009.

[18] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in Proc. of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1-8, 2013.

[19] D. F. Kune, J. Backes, and S. S. Clark, et al., "Ghost Talk: Mitigating EMI signal injection attacks against analog sensors," in Proc. of the 34th IEEE Symp. on SP, pp. 1-15, May 2013.

[20] X. Hei, X. Du, J. Wu, and F. Hu, "Defending resource depletion attacks on implantable medical devices," in Proc. of IEEE GLOBECOM, pp. 1-5, 2010.

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

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

[23] X. Hei, X. Du, and S. Lin, "Poster: Near field communication based access control for wireless medical devices," in Proc. of ACM MobiHoc, 2014.

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

[25] 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.

[26] T. Kinnunen, M. Sahidullah, and H. Delgado, et al., "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in Proc. of INTERSPEECH, 2017.

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

[28] 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, pp. 57-71, 2017.

[29] K. B. Rasmussen, C. Castelluccia, T. Heydt-benjamin, and S. Capkun, "Proximity-based access control for implantable medical devices," in Proc. of ACM CCS, pp. 410-419, 2009.

[30] M. Roeschlin, I. Martinovic, and K. B. Rasmussen, "Device pairing at the touch of an electrode," NDSS'18, Feb. 18-21, 2018, San Diego, CA, USA.

[31] M. Rostami, A. Juels, F. Koushanfar, "Heart-to-Heart (H2H): Authentication for implanted medical devices," in Proc. of ACM CCS, pp. 1099-1111, 2013.

[32] G. Zheng, W. Yang, and C. Valli, et al., "Finger-to-heart(F2H): authentication for wireless implantable medical devices,", IEEE Journal of Biomedical and Health Informatics, 23(4), 1546-1557, 2019.

[33] N. Karapanos, C. Marforio, C. Soriente, and Srdjan Čapkun, "Sound-Proof: Usable Two-Factor Authentication Based on Ambient Sound," in Proc. of the 24th USENIX Security Symposium, pp. 483-498, 2015.

[34] 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.

[35] 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.

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

[37] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q Cepstral Coefficients," in Proc. of ODYSSEY, 2016.

[38] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition," in Proc. of INTERSPEECH, pp. 2489-2493, 2013.

[39] B. Hao, X. Hei, Y. Tu, X. Du, and J. Wu, "Voiceprint-based Access Control for Wireless Insulin Pump Systems," in Proc. of IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), pp.245-253, 2018.

[40] L. O. Lartillot and T. Petri, "A Matlab toolbox for musical feature extraction from audio," International Conference on Digital Audio Effect, Bordeaux, 2007.

[41] A. Das, N. Borisov, M. Caesar, "Do you hear what I hear? Fingerprinting smart devices through embedded acoustic components," in Proc. of ACM CCS, pp. 441-452, 2014.

[42] T. Giannakopoulos, "A method for silence removal and segmentation of speech signals, implemented in Matlab," University of Athens, Athens 2 (2009).

[43] 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.

[44] 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.

[45] H. Delgado, M. Todisco, and H. Yu, et al., "Integrated spoofing countermeasures and automatic speaker verification: An evaluation on ASVspoof 2015," in Proc. of INTERSPEECH, 2017.

[46] M. Todisco, H. Delgado, and K. A. Lee, et al., "Integrated presentation attack detection and automatic speaker verification: Common features and Gaussian back-end fusion," in Proc. of INTERSPEECH, 2018.

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

[48] M. Riley and I. Richardson, "An introduction to Reed-Solomon codes: principles, architecture and implementation," https://www.cs.cmu.edu/~guyb/realworld/reedsolomon/reed_solomon_codes.html, 2019.

[49] Schifra, "Reed Solomon Error Correcting Library, Release Version 0.0.1," https://github.com/ArashPartow/schifra, 2019.

[50] T. Kinnunen, M. Sahidullah, and M. Falcone, et al., "RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research," in Proc. of ICASSP, 2017.

[51] K. A. Lee, A. Larcher, and G. Wang, et al., "The reddots data collection for speaker recognition," in Proc. of INTERSPEECH, 2015.

[52] H. Delgado, M. Todisco, and M. Sahidullah, et al., "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," in Proc. of ODYSSEY, pp.296-303, 2018.

[53] B. Hao and X. Hei, "Book chapter: Voice Liveness Detection for Medical Devices," In D. R. Kisku, P. Gupta, and J. K. Sing (Ed.), Design and Implementation of Healthcare Biometric Systems, 2019.

[54] 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.

[55] J. Sun, X. Zhu, C. Zhang, and Y. Fang, "HCPP: Cryptography based secure EHR system for patient privacy and emergency healthcare," in Proc. of ICDCS, pp. 373-382, 2011.

YUAN PING received the B.S. degree in electronics and information engineering from Southwest Normal University in 2003, the M.S. degree in mathematics from He'nan University in 2008, and the Ph.D. degree in information security from Beijing University of Posts and Telecommunications in 2012. He is an associate professor with Xuchang University and a visiting scholar with the School of Computing and Informatics, University of Louisiana at Lafayette. He was a visiting scholar with the Department of Computing Science, University of Alberta. His research interests include machine learning, public key cryptography, data privacy and security, cloud and edge computing.

BIN HAO received the B.S. degree in electronic and information engineering from China Agricultural University, Beijing, China, in 2004, the M.S. degree in signal and information processing from North China University of Technology, Beijing, China, in 2007, and the Ph.D. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China, in 2012. He is currently a Postdoctoral Fellow at School of Computing and Informatics, University of Louisiana at Lafayette, Louisiana, USA. His research interests focus on acoustical channel based access control, wireless device security, key agreement protocol, and trusted computing.

XIALI HEI received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2002, the M.S. degree in software engineering from Tsinghua University, Beijing, China, in 2005, and the Ph.D. degree in computer science from Temple University in 2014. She is an assistant professor in the School of Computing and Informatics at the University of Louisiana at Lafayette. Prior to joining the University of Louisiana at Lafayette, she was an assistant professor at Delaware State University from 2015-2017 and Frostburg State University 2014-2015. Her research interests are secure real-time wireless medical devices, vulnerability assessment and malware detection on Android, and efficient encryption schemes design. She was awarded NSF CRII grant and Delaware DEDO grant.She got several awards such as: ACM 2014 MobiHoc Best Poster Runner-up Award, Dissertation Completion Fellowship, The Bronze Award Best Graduate Project in Future of Computing Competition, IEEE INFOCOM and IEEE GLOBECOM student travel grant, etc. She is the TPC member of USENIX Security, IEEE GLOBECOM, IEEE ICC, WASA, etc.

YAZHOU TU received the B.S. degree in software engineering from Wuhan University, Wuhan, China, in 2013, and the M.S. degree in software engineering from Tsinghua University, Beijing, China, in 2016. He is currently a M.S. student in the School of Computing and Informatics at the University of Louisiana at Lafayette. His research interests include security and privacy of embedded devices.

XIAOJIANG DU (SM'09) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the M.S. and Ph.D. degrees from the University of Maryland at College Park, in 2002 and 2003, respectively, all in electrical engineering. He is currently a tenured Professor with the Department of Computer and Information Sciences, Temple University, Philadelphia, USA. His research interests are wireless communications, wireless networks, security, and systems. He has authored over 200 journal and conference papers in these areas, as well as a book published by Springer. He is a Life Member of the ACM. He received over $5 million U.S. dollars research grants from the U.S. National Science Foundation, Army Research Office, Air Force, NASA, the State of Pennsylvania, and Amazon. He was a recipient of the Best Paper Award at IEEE GLOBECOM 2014 and the Best Poster Runner-Up Award at ACM MobiHoc 2014. He serves on the editorial boards of three international journals.

JIE WU is the Associate Vice Provost for International Affairs at Temple University. He also serves as Director of the Center for Networked Computing and Laura H. Carnell professor in the Department of Computer and Information Sciences. Prior to joining Temple University, he was a program director at the National Science Foundation and was a distinguished professor at Florida Atlantic University. His current research interests include mobile computing and wireless networks, routing protocols, cloud and green computing, network trust and security, and social network applications. Dr. Wu regularly publishes in scholarly journals, conference proceedings, and books. He serves on several editorial boards, including IEEE Transactions on Service Computing and the Journal of Parallel and Distributed Computing. Dr. Wu was general cochair/chair for IEEE MASS 2006, IEEE IPDPS 2008, IEEE ICDCS 2013, and ACM MobiHoc 2014, as well as program co-chair for IEEE INFOCOM 2011 and CCF CNCC 2013. He was an IEEE Computer Society Distinguished Visitor, ACM Distinguished Speaker, and chair for the IEEE Technical Committee on Distributed Processing (TCDP). Dr. Wu is a CCF Distinguished Speaker and a Fellow of the IEEE. He is the recipient of the 2011 China Computer Federation (CCF) Overseas Outstanding Achievement Award.