Vocal Resonance: Using Internal Body Voice for Wearable Authentication

RUI LIU, Department of Computer Science, Dartmouth College
CORY CORNELIUS*, Intel Labs
REZA RAWASSIZADEH*, Department of Computer Science, University of Rochester
RONALD PETERSON, Department of Computer Science, Dartmouth College
DAVID KOTZ, Department of Computer Science, Dartmouth College

We observe the advent of body-area networks of pervasive wearable devices, whether for health monitoring, personal assistance, entertainment, or home automation. For many devices, it is critical to identify the wearer, allowing sensor data to be properly labeled or personalized behavior to be properly achieved. In this paper we propose the use of *vocal resonance*, that is, the sound of the person's voice as it travels through the person's body – a method we anticipate would be suitable for devices worn on the head, neck, or chest. In this regard, we go well beyond the simple challenge of speaker recognition: we want to know who is *wearing* the device. We explore two machine-learning approaches that analyze voice samples from a small throat-mounted microphone and allow the device to determine whether (a) the speaker is indeed the expected person, and (b) the microphone-enabled device is physically *on* the speaker's body. We collected data from 29 subjects, demonstrate the feasibility of a prototype, and show that our DNN method achieved balanced accuracy 0.914 for identification and 0.961 for verification by using an LSTM-based deep-learning model, while our efficient GMM method achieved balanced accuracy 0.875 for identification and 0.942 for verification.

CCS Concepts: • **Security and privacy → Authentication**; **Biometrics**; • **Human-centered computing → Ubiquitous and mobile devices**;

Additional Key Words and Phrases: Biometric, vocal resonance, authentication, wearable device, mobile system security

ACM Reference Format:
Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ronald Peterson, and David Kotz. 2018. Vocal Resonance: Using Internal Body Voice for Wearable Authentication. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 19 (March 2018), 23 pages. https://doi.org/10.1145/3191751

1 INTRODUCTION

With continuing advances in the development of low-power electronics, including sensors and actuators, we anticipate a rapid expansion of wearable and pervasive computing. Today, it is common for people to carry

*This work was done while the author was in Dartmouth College.

Authors' addresses: Rui Liu, Department of Computer Science, Dartmouth College, Hanover, NH, 03755, Rui.Liu.GR@dartmouth.edu; Cory Cornelius, Intel Labs, Hillsboro, Oregon, 97124, cory.cornelius@intel.com; Reza Rawassizadeh, Department of Computer Science, University of Rochester, Monroe County, NY, 14627, rrawassizadeh@acm.org; Ronald Peterson, Department of Computer Science, Dartmouth College, Hanover, NH, 03755, Ronald.A.Peterson@dartmouth.edu; David Kotz, Department of Computer Science, Dartmouth College, Hanover, NH, 03755, David.F.Kotz@dartmouth.edu.

2474-9567/2018/3-ART19 $15.00
https://doi.org/10.1145/3191751

multiple computing devices, such as smart phones, smart watches, and cameras; increasingly, they also carry, hold, or wear devices to measure physical activity (e.g., Fitbit [17] or Lumo Run [32]), to interact with entertainment devices (e.g., Virtual Reality (VR) headsets), or to monitor their physiology (e.g., a cardiac patient concerned about heart arrhythmia, a diabetic managing her blood glucose, or a woman tracking her fertility cycle).

Many more devices have been proposed or developed as research prototypes [16, 28]. These unobtrusive wearable devices make it possible to continuously or periodically track many health- and lifestyle-related conditions at an unprecedented level of detail. Wireless connectivity allows interaction with other devices nearby (e.g., entertainment systems, climate-control systems, or medical devices). Sensor data may be automatically shared with a social-networking service, or (in the case of health applications) uploaded to an Electronic Medical Record (EMR) system for review by a healthcare provider.

In this paper, we focus on a critical authentication problem involving wearable devices: *who is wearing the device?* For a group-shared device, such as a VR headset, it can recognize the user and load the right game profile or music playlist (identification) and confirm the user's identity when the user tries to purchase through a web store (verification). Most compellingly, we are concerned about health-monitoring devices. Such a device may monitor the user's health data, and transfer the data automatically to the user's medical record. For example, consider a headset that monitors the user's brain data through EEG sensors, transfers them to a smartphone, and detects stress and seizures [22]. A mix-up in use of such wearables may lead to incorrect treatment or diagnosis decisions, and cause serious harm to the patient. An attacker may tempt the user's health database by injecting incorrect data.

In our vision, a person should be able to simply attach the desired set of devices to their body – whether clipped on, strapped on, stuck on, slipped into a pocket, or even ingested, and have the devices *just work*. That is, without any other action on the part of the user, the devices discover each other's presence, recognize that they are on the same body (as opposed to devices in radio range but attached to a different body), develop shared secrets from which to derive encryption keys, and establish reliable and secure communications. Furthermore, for many of the interesting applications described above, the devices must also identify *who* is wearing them so that the device data can be properly labeled (for storage in a health record) or the devices may be used in the context of the user's preferences. Indeed, security and privacy are particularly important in both health-related pervasive applications [24] and wearables [37].

Devices that can automatically recognize their wearer can be smaller and simpler than devices that require user input, because they need no interface for user identification (or PIN or password for authentication). Cornelius et al. developed a method for a networked set of devices to recognize that they are located on the same body; their approach uses correlations in accelerometry signals for this purpose [9]. If even one device can identify *which* body, then transitively the set of devices know who is wearing them. Indeed, it is unlikely that every device will have the technology, or suitable placement, to biometrically identify the user; in our model, only one such device needs to have that capability.

One easy solution, common in many devices today, is for the device to be permanently associated with a given user. This smartphone is *my* phone, whereas that fitness sensor is *your* fitness sensor. The device is assumed to be used by only that user; any data generated by a sensor is associated with that user. There are many situations where this model fails, however. In some households, a given device might be shared by many users (e.g., a blood-pressure cuff). In other settings, two people might accidentally wear the wrong sensor (e.g., a couple who go out for a run and accidentally wear the other's fitness tracker). In some scenarios, a person may actively try to fool the system (e.g., a smoker who places his "smoking" sensor on a non-smoking friend in order to receive incentives for smoking cessation).

Thus, what we need is a simple, wearable device that uses biometric techniques to identify the user, then share that identity with a body-area network of other devices (earlier confirmed to be on the same body [9, 12]). This device should be trained once, for each user that might wear it, but thenceforth be completely automatic and

unobtrusive. Cornelius et al. developed a biometric wristband that used the physical property of bioimpedance to recognize its wearer [11, 12]. While that method worked well, it requires the use of a wristband; we seek an alternative biometric, notably, one that might work for devices mounted on the head, neck, or chest.

We propose to use *vocal resonance*, that is, the sound of the person's voice as it travels through the person's body. Note that *vocal resonance* is related to the approach used by *bone-conduction* headphones, in which sound travels through bones to the inner ear; in vocal resonance, however, the voice passes through bones and tissues from the voice box to a microphone mounted on the surface of the body. In our method, a microphone is placed into contact with the body. It records audio samples and compares them with a model built earlier during a training phase. If the samples fit the model, then we conclude that (a) the speaker is indeed the person for whom we trained the model, and (b) the microphone device is physically *on* the speaker's body. If we train the device for a set of users, e.g., the members of a household, then the device should be able to identify which of those people is wearing the device, or that none of them are wearing the device.

In this paper we explore two machine-learning approaches to the use of vocal resonance as a biometric. Our first *on-device* GMM approach runs *stand-alone*, processing all data within the microphone-enabled wearable device. Our second *off-device* DNN approach relies on a *remote host* to run a more expensive deep-learning algorithm; here, we aim to achieve higher accuracy in anticipation of the day when these complex algorithms can be migrated into wearable or portable devices. The idea of vocal resonance as a passive biometric was presented in a one-page abstract for a poster [29]. An unpublished technical report presented our preliminary work and part of experimental results [10]. However, more experiments and discussions were needed; for example, to explore the various microphone locations used, window sizes and window overlaps, the robustness against attacks, and to optimize accuracy. This paper is an extension of our previous work and presents the full idea and experimental results.

Challenges and contributions: In this paper we present *vocal resonance* as a novel, unobtrusive biometric measurement that can support user authentication (identification or verification) in wearable body-mounted devices. We evaluate the feasibility of this biometric through two distinct machine-learning algorithms, and evaluate their performance on data from 29 volunteer subjects. The goal of this work is to evaluate whether vocal resonance is a biometric that could support authentication rather than demonstrating an improved method for speaker identification, which is a well-studied field [39, 44]. Our results show that vocal resonance could be used as a biometric using concepts drawn from traditional speaker-identification approaches. There were two critical challenges.

Distinguish individuals: We sought to support both identification and verification. We propose two methods: a *Gaussian Mixture Model (GMM)* method and a *Deep Neural Network (DNN)* method to authenticate the individuals. We found that our GMM method achieved balanced accuracy of 0.875 for identification and 0.942 for verification, while the DNN method with Long Short Term Memory (LSTM), in combination with a fully-connected layer architecture, achieved balanced accuracy of 0.914 for identification and 0.961 for verification. With these experimental results, we demonstrate that it is possible to achieve reliable speaker authentication through a wearable, body-contact microphone that can reliably distinguish among multiple individuals. Furthermore, this system could successfully authenticate the wearers after a period of time (for example, two weeks).

Distinguish 'body voice' from 'air voice': In the context of vocal resonance, it is challenging to distinguish 'body voice' and 'air voice' from multiple users. To distinguish the two kinds of voices, we propose to use two sets of GMMs in the GMM method and a fully connected dense layer in the DNN method. For near-body 'air voice' that is several inches away from the microphone, the GMM method limits attackers to a 3.1% success rate, and the DNN method kept attackers 3.6% success rate. For other-body 'air voice' that is one meter away from the microphone, the GMM method restrains attackers to 0.3% success rate, while DNN method limits attackers to 0.1% success rate. Thus, this system can distinguish between the situation where the microphone is on the body of the enrolled speaker and where the microphone is simply nearby, even on another body. We also demonstrate

that vocal resonance is robust against replay attacks in which an enrolled user's air voice is replayed through another person's body.

Furthermore, we implemented two wearable prototypes and verified that the algorithms have acceptable latency and energy consumption when used for occasional or periodic identification or verification.

The remainder of this paper is organized as follows. In the next section, we provide more background on biometrics. Then in Sections 3 and 4 we detail our two models and describe our machine-learning methods, respectively. In Section 5 we describe our implementation of a wearable prototype on the Raspberry Pi Zero platform. In Section 6 we present an evaluation of our methods on measurements from human subjects. Finally, we compare our approach with related work in Section 7, discuss limitations and findings in Section 8, and conclude in Section 9.

2 BIOMETRICS

Biometrics seek to learn some tell-tale characteristic of the person, and use this characteristic to determine whether that same person is present at some later time. This problem, called biometric authentication, is well studied [6]. Biometrics leverage physiological or behavioral characteristics of a person to accomplish identification [5]. Physiological characteristics range from non-invasive characteristics like facial features and hand geometry to more invasive characteristics like the impression of a finger, the structure of the iris, or the makeup of DNA. Behavioral characteristics include things like the dynamics of using a keyboard, the acoustic patterns of the voice, the mechanics of locomotion, and how one signs a signature. To qualify as a biometric, the chosen characteristic must have at least three properties: universality, uniqueness, and permanence [23]. A *universal* characteristic is one that every person possesses. Although everyone may possess such a characteristic, the characteristic must also be individually *unique* within a given population. Lastly, the characteristic must have some *permanence* such that it does not vary over the relevant time scale. These properties, with their stated assumptions, are necessary but not sufficient for a biometric that we desire.

In the context of pervasive applications and particularly personal health sensors, a biometric must also be unobtrusively measured yet difficult to circumvent. The ability to *unobtrusively measure* a biometric stems from our desire to provide usable security for personal health-sensing systems. Apart from attaching the sensors to their body, a person should expect the system to automatically and unobtrusively identify whom the system is sensing. Likewise, a biometric needs to be *difficult to circumvent* because there are incentives for people to circumvent them. For example, a person might want to game their insurance provider or fool a physician into believing they have a certain ailment for prescription fraud. Thus, a sufficient biometric will be *universal*, *unique*, *permanent*, *unobtrusively measurable*, and *difficult to circumvent*.

Not all of the above-mentioned biometrics are suitable for our purposes. While the makeup of DNA, the structure of the iris, and the impression of a finger may be difficult, if not impossible, to forge, they are also difficult to unobtrusively measure. Each of the examples above requires the user to stop what they are doing to measure the biometric. The behavioral characteristics mentioned above are, however, more amenable to unobtrusive measurement since they can be collected as the person goes about their day. On the other hand, they might be easier to circumvent because they can be easily measured. A microphone can capture a person's voice, a malicious application could learn one's typing rhythm [2], or smartphone-usage data can create a user profile [36]. A biometric suited for our purposes would incorporate the difficulty of circumventing a physiological biometric with the measurability of a behavioral biometric.

3 MODELS

We propose using a person's *vocal resonance* as a biometric. Vocal resonance is measured by a microphone placed on a person's body. By virtue of being attached to the person's body, we can use speaker-authentication

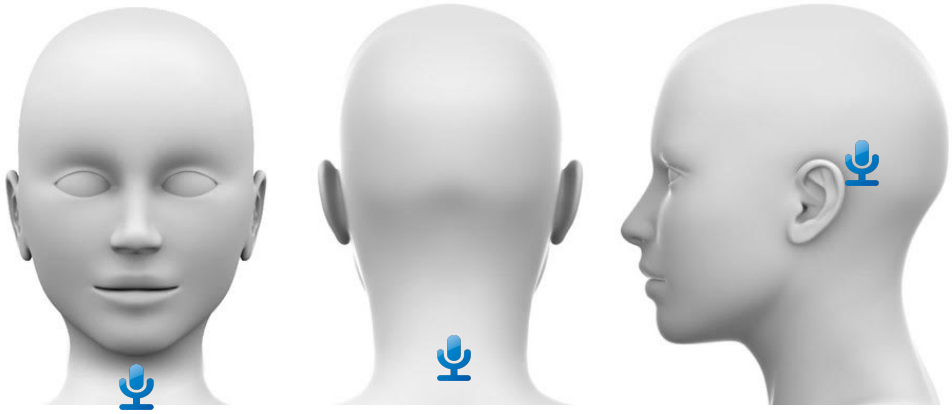

Fig. 1. Possible placements are marked by the icons.

techniques to determine the speaker while simultaneously guaranteeing that the microphone is attached to the speaker's body. Like a typical speaker-authentication system, the microphone hears the person's voice, but unlike a typical speaker-authentication system, the microphone is hearing the voice as it travels through the body itself, rather than through the air. By training the system using on-body voice recordings through the contact microphone, the system can later identify who is speaking *and* verify that the detected voice is coming through the body and not through the air.

3.1 Microphone Locations

A traditional speaker-authentication system makes no guarantees about the placement of the microphone; it may or may not be attached to the person's body. In fact, most traditional speaker-authentication systems make no guarantees that the person is even *present*, because they can be fooled by capturing the person's voice and playing it back through a suitable speaker. Some systems alleviate this concern by employing a challenge-response scheme, whereby the person is asked to speak a randomly selected phrase. However, this approach is obtrusive and thus unsuitable. Capturing the vocal resonance of a person is unobtrusive: all the user must do is talk as they go about their day. Unlike a traditional speaker-authentication system, however, it is difficult to circumvent because an adversary would need to physically attach a microphone to the target individual.

The microphone's location will be critical to the success of the system. A microphone placed near the chest would pick up a person's voice better than a microphone placed on their leg. Figure 1 shows three possible placements: throat, back of the neck, or back of the ear. We imagine a piece of jewelry, such as a necklace, earpiece, or glasses, that would contain a contact microphone to sample vocal resonance and another microphone to sample ambient sound. Such a form factor has several technical advantages. First, these items are worn the same way each time, more or less; issues with placement of the microphone are diminished because it can sense data from nearly the same location each time. Second, the device can be instrumented to detect when it has been placed on and taken off a person using, for example, a switch embedded in a clasp and sensors that detect skin contact. These mechanisms (with details outside the scope of this paper) allow the device to conserve energy by performing identification only when the microphone first comes into contact with a person.

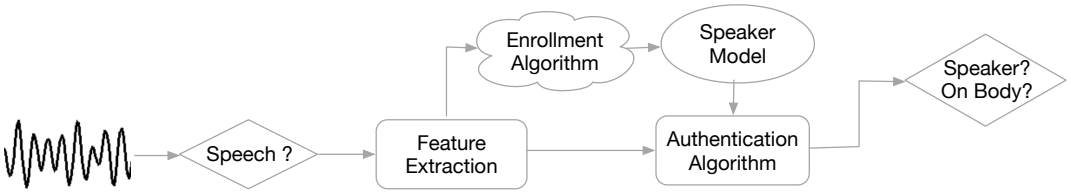

Fig. 2. The major components of our method. Most of the components would be computed on the device itself, except for the enrollment algorithm (which can be done in the cloud). The DNN implementation performs its authentication algorithm on the cloud.

3.2 Adversary & Threat Model

In any system there is a set of assumptions the designers make about the intended adversaries the system is designed to handle. We state these assumptions here.

The device cannot know *a priori* the universe of people, so we assume there is a known set of people who intend to wear the device. The device needs to determine whether it is on the body of an intended person and correctly identify that intended person using the data from its microphone. It should also correctly reject any situation when an unintended person wears the device in the presence of speech by an intended person, whether on a body (of an unintended person) or not.

Our prototypical attacker is passive. They are a person who mistakenly wears the device they believe they were intended to wear. A couple, for example, might have two of these devices and accidentally mix them up. The device should be able to handle this case properly.

We also consider attackers who are actively trying to fool the device. An active attacker might wear the device and play a recording of an intended person's air voice to fool the device into thinking it is on that intended person's body. They might also try to imitate an intended person's voice or introduce noise. They may even replay a recording of an intended person's voice through their own body, e.g., by holding a speaker up to the skin of their neck. However, we assume they will not physically alter the device or its firmware.

4 METHODS

We are inspired by the techniques from the speaker-authentication literature but account for the unique nature of the data collected via a contact microphone [40]. Figure 2 shows the major components of our methods (on-device or off-device), both of which follow a similar workflow. The system has two modes: *enrollment mode* and *authentication mode*. The *enrollment mode* trains the device to recognize the users' voice, and the *authentication mode* authenticates the user to the wearable device. To collect the users' voice, we use two microphones: a *contact microphone* to collect the user's vocal resonance and an *ambient microphone* to collect the user's air voice.

Enrollment mode: When users first receive the device, they use the enrollment mode to establish identity. In the enrollment mode, the device simultaneously collects audio data from both the contact and ambient microphones, and then uses these data to create two models of the person's voice using an *enrollment algorithm*. The *enrollment algorithm* should model the user's body voice and air voice separately. Thus, they could be used to distinguish the user's body voice from air voice later. The model computed from the contact microphone models the speaker's vocal resonance ("body voice"), while the model computed from the ambient microphone models the speaker's voice through the air ("air voice"). Typically, these models would be computed off-device because the computational requirements for learning such a model are high.

Authentication mode: Once enrolled, the users use the authentication mode to get access to the wearable devices. However, when should the wearable device collect audio samples and run the authentication algorithm?

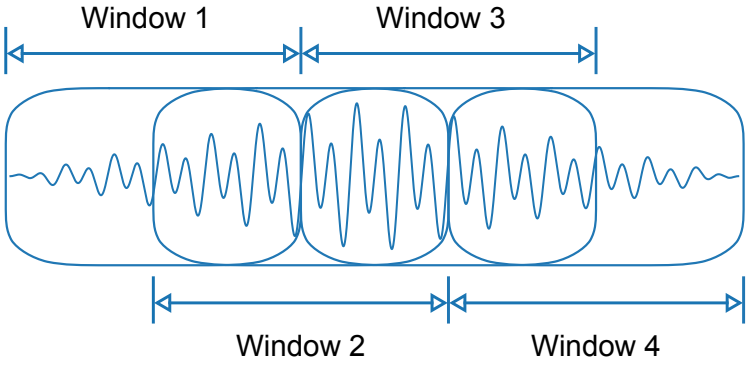

Fig. 3. Sliding windows and overlaps.

We assume the device can detect when it is placed on a body; it then wakes up periodically to collect a sample of audio from the contact microphone. If the device determines the audio sample to contain speech, the device then applies an *authentication algorithm* to this audio sample. At first, it uses an *identification algorithm* to determine which enrolled user, if any, is the speaker. Later, it uses a *verification algorithm* to determine whether the speaker is still the user identified earlier.

We consider two methods: a Gaussian Mixture Model (GMM) [38] and a Deep Neural Network (DNN), specifically, one based on Long Short Term Memory (LSTM) [27]. In both cases we anticipate running the computationally expensive enrollment (training) procedure on a remote host, such as a personal computer or a cloud service. The GMM method is compact and fast, enabling the authentication (classifier) to run on the local (wearable) device; the DNN is computationally expensive [3], so our implementation performs the authentication (classifier) on a remote host. The on-device GMM method provides greater privacy and the remote DNN method provides greater accuracy; we anticipate that future advances in deep-learning algorithms and wearable-computing hardware will enable the DNN method to eventually run the classifier on-device (or, at least, on a nearby personal computing device such as a smartphone).

Before we present the enrollment and authentication algorithms, we define the audio-segmentation and feature-extraction algorithms used by both methods.

4.1 Audio Segmentation

As Figure 3 shows, the microphone delivers a series of audio *samples*, which are then divided into small *windows* to ease processing. We experimented with several window sizes and overlap percentages (see Section 6.3) and chose the one that resulted in the best accuracy metrics in authentication algorithms. Because not all audio data contain speech, we first determine whether the window contains speech with a speech detection module; this step is accomplished efficiently using time-domain features combined with a decision tree as described by Lu et al. [31]. Specifically, the decision tree uses zero crossing rate (ZCR) and root mean square (RMS) to determine whether the window contains speech. Windows that do not contain speech are discarded. (During enrollment, corresponding windows from the contact and ambient microphones are discarded if either segment is determined not to contain speech.)

4.2 Feature Extraction

Given a window of audio, we first extract some features that capture characteristics of the person's voice. We use the set of Mel-Frequency Cepstral Coefficients (MFCCs) [20], which characterize the cepstrum of the audio

segment (the power spectrum of the log of the power spectrum). Because the power spectrum of the segment is first mapped onto the mel scale, which is empirically based on the frequency scale of the human auditory system, these coefficients model how we hear sound [46]. For this reason, MFCCs have been successful in many voice-processing tasks.

For both GMM and DNN methods, inspired by other speech-recognition approaches [25, 38], we selected the MFCC and their deltas (delta-spectral cepstral coefficients [25]) for a total of 26 features. This results in a *feature vector* for each window; a feature vector computed from audio data sampled from the contact microphone is called a *contact feature vector*, and a feature vector computed from the ambient microphone is called an *ambient feature vector*.

4.3 Enrollment Algorithms

During enrollment, we collect audio from both the contact microphone and the ambient microphone. The prospective user is asked to read a short passage of text; in our experiments (Section 6) the subjects took under two minutes (107 seconds on average) to read this passage. As above, these recordings are segmented into short 40 ms windows; windows without speech are discarded; features are extracted from each of the remaining windows; and the sequence of feature vectors is fed to the enrollment algorithm to train machine-learning models. The GMM method builds four Gaussian mixture models for each speaker, two for each microphone; the DNN method builds one model for all speakers and both microphones.

4.3.1 GMM method. A GMM models the distribution of observations using a weighted linear combination of Gaussian density functions, where each Gaussian is parameterized by a mean vector and covariance matrix. We use GMMs to model the distribution of feature vectors for a given speaker. It is important to distinguish the user's *body voice* and *air voice*. To distinguish the two voices, for each enrolled user, we construct two sets of GMMs: one modeling the characteristics of user's *body voice*, and the other modeling the characteristics of user's *air voice*. It is also important to separate the user-dependent characteristics; we use one GMM to model the characteristics of user's voice and the other GMM to model the user-independent characteristics of the voice. Thus, we construct four GMMs in total for each user: GMM_b using the 'contact' feature vectors of the target user; GMM_{bu} using the 'contact' feature vectors of all other users; GMM_a using the 'ambient' feature vectors of the target user; GMM_{au} using the 'ambient' feature vectors of all other users. We use GMM_{bu} and GMM_{au} as *universal background models* to represent the speaker-independent distribution of feature vectors [40]. To learn the underlying distribution of feature vectors, we use the Expectation-Maximization (EM) algorithm [13] to iteratively refine the mixture of Gaussian densities until the maximum likelihood remains stable (i.e., the difference between successive iterations is less than 10^{-4}) or after a maximum number of iterations (5000). We use the EM algorithm because 1) the EM algorithm is efficient to optimize the GMMs and 2) it achieves the maximum likelihoods [13].

We choose initial Gaussian densities by clustering the set of feature vectors using k-means clustering [18], where k is set to the desired number of Gaussian densities. We iteratively refine these initial Gaussian densities using the EM algorithm, until it achieves the maximum likelihood. In our experiments, all the EM were finished within 1000 iterations, which implies that all the GMMs achieved local maximum likelihood. We use diagonal covariance matrices for the following three reasons: 1) a GMM with full covariance matrix could be equally achieved by a larger-order diagonal covariance; 2) diagonal-covariance GMMs are computationally efficient; 3) it has been shown that using a larger-dimensional diagonal covariance matrix outperforms a smaller-dimensional full covariance matrix [4].

The *likelihood* of a feature vector, given a GMM, is the weighted linear combination of the probability density function of each Gaussian given the feature vector. From a speaker's GMM models we compute the log-likelihood of a given contact feature vector corresponding to each model. For the given contact feature vector, we compute LL_b from GMM_b, LL_{bu} from GMM_{bu}, LL_a from GMM_a, and LL_{au} from GMM_{au}. We compute a *contact likelihood*

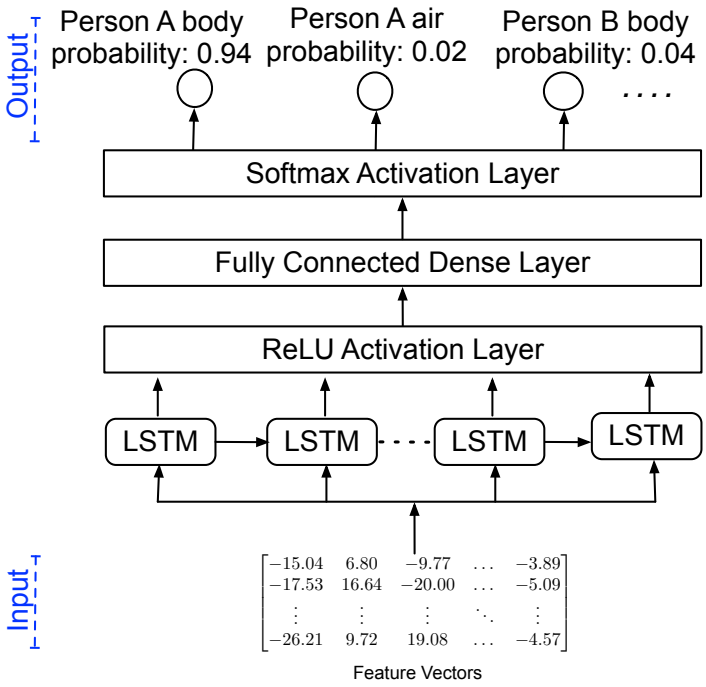

Fig. 4. Our proposed deep-neural-network model.

ratio LR_b by subtracting the contact background likelihood ($LR_b = LL_b - LL_{bu}$) and similarly an *ambient likelihood ratio* ($LR_a = LL_a - LL_{au}$). Then we compute the *difference likelihood* with $LD = LR_b - LR_a$.

Given a predetermined acceptance threshold τ, we say the audio segment corresponding to the contact feature vector matches the speaker's vocal resonance ("Accept") if $LD \geq \tau$; otherwise it does not fit the model and therefore does not match the speaker's vocal resonance ("Reject"). We choose τ to achieve the highest balanced accuracy on a small random partition (5%) of the training data.

4.3.2 DNN method. We also employ a Deep Neural Network (DNN) inspired from other work investigating speaker authentication algorithms [21]. A Recurrent Neural Network (RNN) is a Deep Neural Network architecture that is useful for processing sequential data [27], such as the sequence of feature vectors extracted from a sequence of audio windows observed by our hypothetical wearable device. We use a specific RNN layer known as Long Short Term Memory (LSTM) [27]. LSTM processes a new event by referring back to the previous event. Others have demonstrated that LSTM is more effective than traditional deep neural networks, or conventional RNN models, for acoustic modeling [42].

The DNN model is made up sequentially of a LSTM layer, a Rectified Linear Unit (ReLU) activation layer, a fully connected dense layer, and a softmax activation layer, as shown in Figure 4. The LSTM cells in the LSTM layer generate the representations that capture the temporal dependencies in the feature vectors. To accelerate the convergence of stochastic gradient decent, we employ a ReLU activation layer [34] after the LSTM layer. Then, the fully connected dense layer and the softmax activation serve as a classifier of the temporal dependencies captured by the LSTM and ReLU layers. Specifically, the fully connected dense layer aggregates the output from the ReLu activation. In the last layer, we use softmax activation to output the probabilities of the classes. To

distinguish among the users and the user's *body voice* and *air voice*, the model's output includes two classes for each of N known persons: the *body class* representing their voice captured by the body microphone, and the *air class* representing their voice represented by the ambient microphone. Each of these $2N$ classes indicates the probability that the input feature vector is from that person and corresponding microphone. We built one DNN model for each microphone location in Figure 1.

The DNN method then applies a prediction algorithm to the outputs of the neural network, given a predetermined acceptance threshold θ. If none of the $2N$ output probabilities are greater than θ, the predictor outputs "Reject". Otherwise, the predictor considers the highest-probability model output; if that output is an *air class*, the predictor outputs "Reject"; otherwise it outputs the identity of the user corresponding to that highest-probability output. Thus, the DNN method outputs one of $N + 1$ possibilities every time it receives a feature vector as input: either one of the N known users, or "Reject" (representing an unknown user or a non-body voice). As with the GMM method, we choose θ to achieve the highest balanced accuracy on a small random partition (5%) of the training data.

4.4 Authentication Algorithm

During the authentication phase we use the contact microphone only, sampling the microphone for 1 second. We segment the 1-second chunk into windows, discard windows that do not pass the speech-detection test, then extract a contact feature vector for each window, as described previously. We wish to determine whether these newly measured contact feature vectors match the vocal resonance of an enrolled speaker. The window sizes and overlap percentages used for GMM and DNN methods in this section are explained in Section 6.3.

GMM: For the GMM method we extract a feature vector every 10 ms. The feature vector is computed over a 40 ms window, and neighboring windows have 30 ms (75%) overlap so we have a new window and thus a new feature vector every 10 ms. Each Gaussian mixture outputs a likelihood for each feature vector, i.e., LL_b, LL_{bu}, LL_a, and LL_{au}. Then we compute the *difference likelihood* LD for each feature vector. To achieve stability, we average over a series of LD values for all the feature vectors; in our experiments there are 99 feature vectors in each second of audio. To identify the wearer, the GMM method finds the model with highest log-likelihood LR_b and outputs that model's identity as long as its averaged LD exceeds τ. To verify the wearer, the GMM method examines only the target user's model and reports 'Accept' if the averaged LD is greater than the threshold τ.

DNN: For the DNN method, we extract a feature vector every 30 ms for a 40 ms window. The feature vector is computed over a 40 ms window as with the GMM method, and neighboring windows have 10 ms (25%) overlap, so we have a new window and a new feature vector every 30 ms. For our DNN experiments, we ran $n = 33$ feature vectors (1 second of audio) simultaneously through the model. To identify the wearer, the DNN model outputs the user with highest 'body class' probability as long as it exceeds θ. To verify the wearer, the DNN model checks whether the probability of the 'body class' of the target user exceeds the threshold θ.

5 IMPLEMENTATION

In this section, we describe our implementation on a wearable device, and in the next section, we show how we collected voice recordings from 29 human subjects to explore our methods' ability to authenticate people.

For our prototype, we used the Raspberry Pi Zero Wireless platform [35]. The Pi has a 1GHz ARM Cortex-A8 processor, 512MB RAM, and a microSD slot, and it supports Wi-Fi (802.11b/g/n) and Bluetooth 4.1+BLE. For the DNN algorithm, we used an Ubuntu Linux server with 2.20GHz Intel 12 Core Xeon and 64GB RAM, with an NVidia Geforce GTX 1080 GPU card (8GB RAM, and 1607MHz base clock speed). We used the CUDA API for the NVIDIA GPU to implement the DNN algorithm.

Figure 5 shows the mobile system: a Raspberry Pi Zero Wireless board, an external USB soundcard connected to the USB On-The-Go port, and a piezo contact microphone connected to the soundcard. We used an external

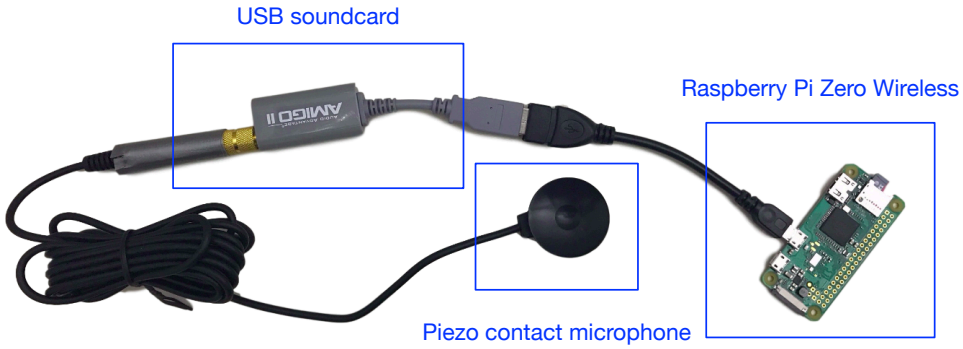

Fig. 5. Raspberry Pi Zero W, external USB soundcard, and piezo microphone. An amplifier was only required due to a software driver issue and hence was left out of the picture. The photo only shows one microphone but the our experimental apparatus included four similar microphones connected through a USB hub.

USB soundcard because the Pi is not equipped with any built-in audio ports. Specifically, we used the Amigo II USB sound card by Turtle Beach. We used Radio Shack Mini Audio Amplifiers to amplify the microphones. These amplifiers were connected to the USB external sound cards. We used the `arecord` utility for the ALSA soundcard driver on the Raspberry Pi to capture the audio data at 44,100 Hz with 16-bit resolution.

We implemented audio segmentation (Section 4.1), feature extraction (Section 4.2), and the GMM authentication algorithm (Section 4.4), on the Raspberry Pi. We implemented the DNN authentication algorithm on the server, as noted above, and the Pi uses Wi-Fi to transfer the feature vectors to the server and receive back the DNN predictor's output. We used the server because we could not run the authentication algorithm on the Raspberry Pi, due to the low computation capacity of Raspberry Pi and high computation capacity required. We implemented both the GMM and DNN enrollment (training) algorithms on the server; our purpose here is to evaluate the speed and energy cost of the authentication algorithm, because enrollment is a one-time operation that requires more computation than feasible on a wearable device like the Raspberry Pi. We recorded audio samples on the Pi using the `arecord` utility. We used the openSMILE tool to extract the MFCCs in the feature-extraction step; we cross-compiled it for the Raspberry Pi [14]. For the GMM implementation, we used Python version 3.5 and scikit-learn version 0.18.2. For the DNN implementation we used Tensorflow version 1.2.0 and CUDA Toolkit version 8.0.

Implementing the DNN identification step on a remote host may allow for faster processing and higher accuracy but yields two major disadvantages over the GMM method. First, it requires a network connection and access to a remote service that hosts the trained model and runs the identification algorithm. Second, it may raise privacy concerns because it may be possible to invert the feature vectors sent to the remote host and recover a meaningful chunk of the wearer's speech [7]. For now, we assume the remote host can be trusted with this task; in future work we will explore means for moving the DNN computations to a trustworthy companion device (such as a smartphone) or to the wearable itself. Or, explore a different set of features that cannot be inverted to recover meaningful speech.

6 EVALUATION

In this section we explore the viability of vocal resonance as a biometric. Recall that we require a biometric to be universal, unique, permanent, unobtrusively measurable, and difficult to circumvent. For the purposes

of this study, we assume that speech is universal. We explore the uniqueness, permanence, measurability, and circumventability properties.

6.1 Data Collection

Participants: We collected data from 29 human subjects using an IRB-approved protocol. We enrolled 18 males and 11 females (age mean = 26, standard deviation = 3.83). All the experiments were conducted in English.

Apparatus: We used AXL PG-800 External Piezo Transducer suction cup microphones (as seen in Figure 5); we anticipate that such microphones could be integrated into glasses, earpieces, VR headsets, and possibly necklace pendants. We recorded data from five locations, including three locations on the subject's body (*body* microphones), one in the air (*air* microphone) and another on another person's body (*other* microphone). The speaker attached one microphone to his or her throat, one microphone at the back of the neck, and one microphone at the back of the ear, as shown in Figure 1. The speaker held the *air* microphone six inches from his or her mouth. The *body* microphones simulated the case when an enrolled person is wearing the device; we call these microphones the *throat* microphone, the *neck* microphone, and the *ear* microphone respectively. The *air* microphone simulated the case when an enrolled person is not wearing the device, but the device could still hear them speaking. The *other* microphone was placed on the throat of a listener sitting about one meter away from the speaker; this microphone simulated the case when another person, enrolled or unenrolled, is wearing the device and an enrolled speaker is speaking nearby. The *body* and *other* microphones were secured by an elastic bandage and a surgical tape.

Procedure: We instructed the subjects to read two passages. We selected the *Rainbow Passage* and the *Wind in the Willows* because they encompass most of the phonemes of the English language [15, 19]. We used the first passage (*Rainbow Passage*) to train the GMM and DNN models. The second passage (*Wind in the Willows*) acted as a control passage. We chose the first 24 lines from the *Wind in the Willows* as a common test phrase for all subjects. Subjects took an average of 107 seconds to read the *Rainbow Passage*, and 92 seconds to read the *Wind in the Willows* lines. To explore permanence of vocal resonance, we repeated the experiment for the same subjects with the same procedure after a 2-week period.

Training: For training purposes, we used data from the *throat* microphone to represent the 'body class' and data from the *air* microphone for the 'air class'.

Testing: For evaluation purposes, we used data from the *throat* microphone as positive test cases and from the *air* microphone as negative test cases. In Section 6.6 (only) we explore the other two *body* microphone locations, *neck* and *back*, to explore whether accuracy varied with microphone location. To evaluate how well the vocal resonance distinguishes the 'body voice' from 'air voice', we use the voice collected from *air microphone* and *other* microphone in Section 6.7.

In both training and testing procedures, we used only speech data to demonstrate the vocal resonance as a passive biometric and non-speech data were eliminated by the speech-detection module described in Section 4.

6.2 Metrics

Verification: Verification algorithms output positive (the speaker is the expected subject and the audio collected is classified as a 'body class') or negative (otherwise); we thus have four cases: TP, where the algorithm outputs positive and is correct; TN, where the algorithm outputs negative and is correct; FN, where the algorithm outputs negative and is incorrect; and FP, where the algorithm outputs positive and is incorrect. We use *Balanced Accuracy* (BAC = $\frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$) to report the accuracy of verification, because the testing data is biased (only 3.45% of the testing cases are positive). We also report the *false accept rate* (FAR = $\frac{FP}{TN+FP}$) and *false reject rate* (FRR =

$\frac{FN}{TP+FN}$). FAR is the fraction of negative cases that were misclassified (i.e., they were classified as positive), while the FRR is the fraction of positive cases that were misclassified (i.e., they were classified as negative). We also present a Receiver Operating Characteristic (ROC) curve and report the Area Under Curve (AUC). The ROC curve plots the True Positive Rate ($TPR = \frac{TP}{TP+FN}$) against the False Positive Rate ($FPR = \frac{FP}{FP+TN}$) and the AUC is the area under the ROC curve.

Identification: Identification algorithms output the identity of a subject, or "Reject" if the audio collected by the device is classified as an 'air class' or if none of the subject's models output high probability. We have four cases: True Positive (TP), where the output identity is the same as the speaker's 'body voice'; True Negative (TN), where the algorithm outputs "Reject" and is correct; False Positive (FP), where the algorithm outputs an identity and is incorrect; False Negative (FN), where the algorithm outputs "Reject" and is incorrect. We compute *Accuracy* (ACC = $\frac{TP+TN}{TP+TN+FP+FN}$). When used for identification, *Balanced Accuracy* (BAC) is equal to Accuracy, because we have the same number of positive and negative samples. For simplicity, we refer to BAC for both verification and identification henceforth.

6.3 Window Size and Overlap Percentage

Both GMM and DNN are parameterized by the window size and the overlap percentage of the neighboring windows. We explored combinations of window sizes {20 *ms*, 40 *ms*, 80 *ms*} and overlaps {25%, 50%, 75%} for each model on our dataset. We used the *Rainbow Passage* to train models with different parameters and tested the models on the *Wind in the Willows* passage to compute the accuracy metrics. Then, we chose the parameters that led to the highest BAC for each model, from the following results.

GMM: We evaluated each combination with BAC for identification and verification. As Table 1 suggests, we chose windows of 40 ms with 75% overlap (30 ms).

Table 1. GMM BAC for identification and verification with window sizes and overlaps

Window Size	Overlap					
	25%		50%		75%	
	Identification	Verification	Identification	Verification	Identification	Verification
20 ms	0.869	0.940	0.874	0.940	0.874	0.945
40 ms	0.868	0.936	0.876	0.938	**0.880**	**0.949**
80 ms	0.847	0.925	0.862	0.932	0.873	0.938

DNN: Table 2 shows BAC for identification and verification for each combination. We chose windows of 40 ms with 25% overlap (10 ms).

Table 2. DNN BAC for identification and verification with window sizes and overlaps

Window Size	Overlap					
	25%		50%		75%	
	Identification	Verification	Identification	Verification	Identification	Verification
20 ms	0.891	0.946	0.913	0.960	0.893	0.951
40 ms	**0.914**	**0.961**	0.909	0.955	0.895	0.950
80 ms	0.898	0.958	0.903	0.955	0.900	0.953

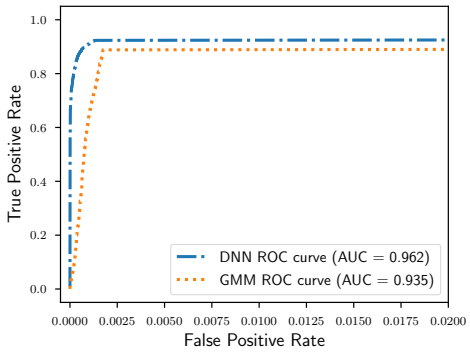

Fig. 6. ROC curves for GMM and DNN models in the uniqueness test. The x-axis represents the FPR; notice it is truncated to 0.02. The y-axis represents the TPR.

6.4 Uniqueness

We consider vocal resonance to be unique over a given population if every person in that population can be individually determined by their vocal resonance. To validate our method, we ran the described methods on our dataset to see how well the method could accurately classify individuals.

GMM: For each user in our dataset, we trained four GMM models (two for body voice and two for air voice) and learned a threshold of the difference likelihood using their training data.

DNN: We trained a single DNN model for all $N = 29$ users and two microphones (air and body).

Table 3 shows the accuracy metrics for both GMM and DNN models. Both methods achieved high BAC (> 0.87), low FAR (< 0.04), and low FRR (< 0.01) for both identification and verification. The table also shows that the DNN method outperformed the GMM method in all the metrics. Figure 6 shows the ROC curve for GMM and DNN models. The AUC is 0.962 for DNN and 0.935 for GMM, which demonstrates both models could distinguish the different individuals accurately using vocal resonance.

Table 3. Uniqueness

method	Identification			Verification			
	BAC	FAR	FRR	BAC	FAR	FRR	AUC
GMM	0.875	0.012	0.005	0.942	0.037	0.002	0.935
DNN	0.914	0.012	0.004	0.961	0.032	0.001	0.962

6.5 Permanence

The permanence property requires the biometric to remain stable over a period of time. To evaluate the permanence of vocal resonance, we used the data collected during the second visit to test the GMM and DNN models trained with the data from the first visit; Table 4 reports the results. When compared to Table 3, most of the accuracy metrics for both models dropped. Nevertheless, both methods maintained relatively high accuracy on the second visit. We note that the metrics for DNN did not drop as much as the metrics for GMM, which suggests that the DNN method was not as sensitive to the biometric changes over time. We conclude that vocal resonance may have sufficient permanence, though more extensive study (more subjects and longer periods) will be needed to

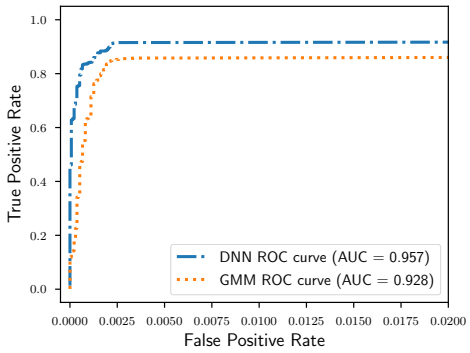

Fig. 7. ROC curves for GMM and DNN models in the permanence test. The x-axis represents the FPR; notice it is truncated to 0.02. The y-axis represents the TPR.

ensure permanence. Figure 7 shows the ROC curve for GMM and DNN models. The AUC is 0.957 for DNN and 0.928 for GMM, which demonstrates that both models are able to distinguish among the different individuals accurately over a period of time using vocal resonance.

Table 4. Permanence

method	Identification			Verification			
	BAC	FAR	FRR	BAC	FAR	FRR	AUC
GMM	0.814	0.020	0.012	0.926	0.098	0.011	0.928
DNN	0.875	0.016	0.005	0.957	0.075	0.006	0.957

6.6 Measurability

Measurability is the property of being easy and unobtrusive to measure. The unobtrusiveness of the device will highly depend upon its form factor. We argue for integration into existing devices, like a necklace or ear piece, that people already wear. However, the ease of measuring vocal resonance will also depend on the placement of the microphones. In Figure 8, we evaluate the accuracy metrics of vocal resonance on three locations: throat, back of the neck, and back of the ear. The figure shows that different placements of the device had limited impact on the accuracy metrics: for identification the differences among placements were generally small, although the GMM method had some trouble with the 'neck' location.

6.7 Near-body and Other-body Test

To evaluate how well the vocal resonance distinguishes the *body voice* from *air voice*, we used the data collected from the air microphone (near-body voice) and the other microphone (other-body voice) to test the GMM and DNN models. The GMM and DNN models should *reject* the near-body voice and the other-body voice, which represent accidental or malicious attack situations. We define the *Failure Rate* (FR) to be the fraction of attempts in which a non-body voice is verified as an enrolled user ($FR = \frac{FP}{TN+FP}$). We use the FR to evaluate how well vocal resonance could distinguish the on-body from near-body or other-body voices. Table 5 shows that the FR of near-body voice is higher than the FR of other-body voice, which but both are nonetheless very low. The

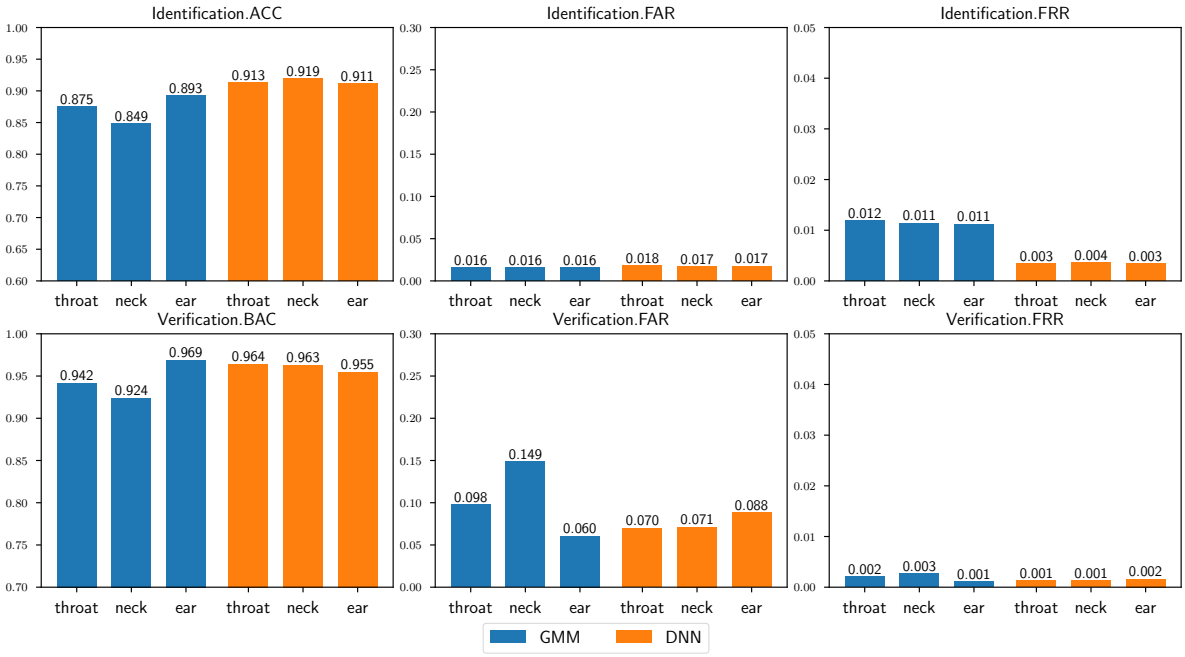

Fig. 8. Accuracy metrics (BAC, FAR, FRR) for identification and verification on three different locations with GMM and DNN models. The x-axes represent the different locations of the devices. The y-axes represent the values of the accuracy metrics. GMM and DNN models are grouped to ease the comparison among the three locations.

results demonstrate that vocal resonance is able to distinguish on-body voice from near-body or other-body voice ($FR < 0.04$). The FR is essentially the success rate for a nearby attacker, and was less than 4%.

Table 5. FR for near-body and other-body voice

method	Failure Rate (FR)	
	near-body voice	other-body voice
GMM	0.031	0.003
DNN	0.036	0.001

6.8 Circumventability

An active *attacker* may try to fool the method by introducing specially crafted audio into the environment. For example, he could capture an enrolled person's air voice and then replay that audio through his own body so as to fool a device on his body into believing the enrolled person is wearing the device. In this scenario, we call the enrolled person the *victim*. We define the *Attack Success Rate* (ASR) to be the fraction of attack attempts in which the attacker is verified as the victim ($ASR = \frac{FP}{TN+FP}$). ASR is a special case of FR, in which a non-body voice is replaced with an attack. To test vocal resonance against this attack, we placed a loudspeaker on the throat of the $(i+1)$th subject (*attacker*) and played back the ith subject's air voice (*victim*). The GMM achieved average ASR

0.017 across all the subjects, while the DNN achieved average ASR 0.010. Both models are demonstrated to be robust (ASR < 0.02) when faced with this attack.

6.9 Power Consumption and Response Time

For wearable devices, the energy capacity is always limited [37]. We measured the energy used by the Raspberry Pi Zero for audio recording, feature extraction and identification/verification algorithms over 100 runs with a Monsoon Power Monitor [33]. We excluded the amplifier from the power measurement since it would not have been necessary if our USB soundcard had had enough gain. Likewise, we measured average end-to-end latency incurred for feature extraction and speaker-authentication algorithms. Table 6 presents the results. If the devices wake up to authenticate 1 second of audio every 10 mins, with a rechargeable battery pack (2200 mAh and 5 V [1]), the GMM method could last 232 hours, while the DNN method could last 248 hours. Although these results are encouraging, they reflect only a simple prototype built on off-the-shelf hardware; a true wearable device would have power-optimized hardware that would likely process data much more efficiently.

Table 6. Power consumption and response time for a 1 second sample, averaged over 100 runs

method	Power consumption (J)		Response time (s)	
	Identification	Verification	Identification	Verification
GMM	10.46	8.76	4.53	2.37
DNN	8.62	8.50	2.21	2.17

6.10 Threshold Sensitivity

The thresholds τ and θ will affect the performance of the verification algorithms. As described in Section 4, we predetermine thresholds τ for GMM and θ for DNN models from a partition of the training data. Figure 9 depicts the accuracy metrics of vocal resonance with varying thresholds τ and θ for GMM and DNN models. The figure shows that lower thresholds yield lower FARs and higher FRRs for both τ and θ in GMM and DNN models, while higher thresholds yield higher FARs and lower FRRs. Note that the scales of the x-axes for GMM and DNN models are not identical, because we use log-likelihoods in GMM and likelihoods in DNN models, and thus we do not compare the threshold sensitivity across the two algorithms. For strong security, the thresholds chosen should yield a high BAC, a low FAR, and a low FRR. Figure 9 shows that the accuracy metrics were optimal when $\tau = -99.468$ for GMM algorithm and $\theta = 0.04$ for DNN algorithm. Thus, we chose $\tau = -99.468$ and $\theta = 0.04$ in our experiments.

7 RELATED WORK

Others have proposed some promising approaches to identify or verify users of pervasive devices, such as smartphones [8] or tablets [43]. Here, we focus on a different, unobtrusive method for identifying the user of a wearable device – one that does not require a display or touch screen or any particular user interface, only a contact microphone. Speaker-identification and -verification systems have been studied for some time [4, 38, 40, 41]. State-of-the-art speaker-authentication systems use features and methods similar to the one described in this paper. With the advent of Deep Learning, we anticipate a significant improvement over the quality of speech, speaker and language recognition algorithms [41].

The most similar research to our own is by Yegnanarayana et al. [48]. They study the feasibility of a speaker-recognition system for samples collected by a throat contact microphone in a simulated noisy environment, compared to a microphone placed close to the speaker. To compare both microphones, they collected data simultaneously from both microphones from 40 speakers. They note that the throat contact microphone is mostly

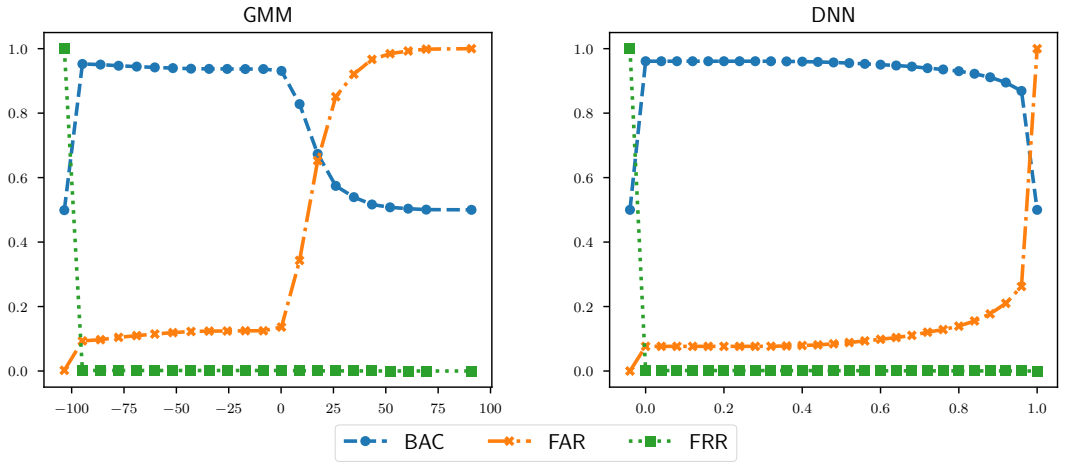

Fig. 9. Accuracy metrics (BAC, FAR, FRR) for thresholds with GMM and DNN models. The x-axes represent the values of thresholds. The y-axes represent the values of the accuracy metrics. Note that the scales of the x-axes for GMM and DNN models are not identical.

immune to noise and reverberation, unlike the close-speaking microphone, but it also suffers from the attenuation of higher formants in speech. To determine feasibility of speaker recognition, they extracted 19 linear predictive cepstral coefficients as features from the audio, and use an auto-associative neural network to model these features. They show that the performance of the system using throat and close-speaking microphones is the same in a noise-free environment. In a noisy environment, the close-speaker microphone system degrades in the presence of noise while the throat microphone system does not. Our work is complementary to theirs as we study the feasibility of vocal resonance as a biometric. Furthermore, we build a novel speaker-authentication approach using both DNN and GMM models, and implement those models on a wearable prototype device.

Tsuge et al. looked at bone-conductive microphones for speaker verification [45]. (A bone-conductive microphone picks up sounds conducted through the skul much as our contact microphone picks up sounds passing through bones and tissues.) They use this kind of microphone to study the feasibility of a speaker-verification system over a dataset with more than 600 speakers. They extract a 25-dimensional feature vector from 12 MFCCs and use 64 vector-quantization centroids to model a speaker. Their experiments show that the bone-conducting microphone performs poorer than an air-conducting microphone, due to placement and noise. However, when the two microphones are combined, the equal-error rate improves by 16% over just an air-conducting microphone. As with Yegnanarayana et al. [48], this paper is focused on speaker-recognition, whereas our work explores the potential for vocal resonance as a biometric that can also distinguish on-body from off-body cases.

Feng et al. proposed a continous authentication system *VAuth* for voice assistants, such as Alexa, Siri and Google Now [16]. The VAuth system collects the body-surface vibrations of the user from wearable devices and matches the vibrations with the speech signal received by the voice assistant's microphone. They achieved 97% detection accuracy and less than 0.1% false-positive rate. They also showed that VAuth is robust against replay attacks, mangled voice attacks and impersonation attacks. Their work focused on authenticating the wearer of a wearable device to voice assistants, while we focus on authenticating a person to a wearable device, a complementary and critical feature they did not address.

There are efforts to make deep learning possible on devices with limited hardware capacities. Zeng et al. proposed a system to recognize a pharmaceutical pill from images with a model compression framework that significantly reduces the size of the deep-learning model without deteriorating its recognition performance [49]. Lane et al. showed the feasibility of running deep neural networks with fully connected layers for audio-sensing applications on low-power mobile digital signal processors [26]. These results give hope that our DNN model may someday be possible to run on a wearable device.

In short, we are the first to explore the use of vocal resonance as a biometric, and furthermore, to use it to allow devices to distinguish between on-body and near-body situations.

8 DISCUSSION AND LIMITATIONS

Our experiments demonstrate the efficiency of vocal resonance as a biometric for personalizing wearable devices; however, several important issues need to be addressed to incorporate this new biometric into real devices. In the following we describe these challenges and our planned extensions to this work.

Length of audio samples: We conducted all the experiments in Section 6 with 1-second audio samples. For audio samples that are longer than 1 second, we could partition each audio sample into smaller audio chunks, perform authentication on each audio chunk, and use majority-voting to determine the authentication decision. An interesting future work would be to study the length of audio samples and utilization of majority voting.

Wake-up mechanism: The wake-up mechanism depends on the specifications of the wearable. If the wearable device is equipped with a sensing module that detects when the wearer puts on/off the device, then authentication could be triggered at the moment when the device is put on. Or, the wearable device should wake up periodically, for example, every 10 mins, to authenticate/re-authenticate the wearer. We did not analyze the wake-up period in detail. On the one hand, the period should be short so the wearable device can quickly detect context changes and take action. On the other hand, the period should be long, to reduce battery usage. The best choice will depend on application needs, and on patterns in user behavior, which is a study for future work.

Size of cohort: For identification, we tested accuracy metrics on the entire cohort of 29 subjects in Section 6. Smaller or larger cohorts may also influence the identification accuracy metrics; for example, intuitively it is easier to distinguish between two people rather than among a group of 100 people. To understand how the sizes of cohorts affect the identification accuracy metrics, we need to conduct experiments on cohorts of different sizes.

Other body parts: This work is focused on head-mounted wearable devices and we show in Section 6 that vocal resonance to be viable for throat, back of the neck, and back of the ear. Human voice also resonates through other body parts, and thus it may be possible to collect vocal resonance from those body parts, for example, the chest or lower body. To learn how other locations perform, we need to collect sound from those locations, although we anticipate that locations distant from the head and chest will have poorer results.

Features used: We employed 26 MFCC features in the evaluation of GMM and DNN algorithms, in Section 6. We need to conduct extensive experiments with different feature subsets to explore effects on accuracy metrics, energy consumption, and the response latency. Liu et al. analyzed different feature-selection algorithms and different feature subsets from time-domain and frequency-domain features for speaker identification and speaker verification on wearable devices [30]. We could employ the analysis suggested by this work.

Permanence: Our studies in Section 6.5 demonstrate the robustness of vocal resonance in authentication under changes of physiological conditions over a period of time (2 weeks). We need to consider longer periods. Moreover, all the subjects were comfortable and calm when we collected vocal resonance data. However, vocal resonance's physiological features may change due to illness, stress or intensive exercise. Physiological changes such as a hoarse voice, when the subject has a cold, may yield low accuracy metrics. Such sensitivity is common to other biometrics as well. For example, a gait-based authentication system will be heavily influenced by injury, footwear, or surface conditions. To eliminate these unwanted sensitivities, we need to learn how vocal resonance

changes under certain physiological conditions. To understand how the physiological conditions affect our proposed system, we need to perform more extensive studies, over longer periods and in more contexts.

Native languages: In Section 6, we conducted all the vocal resonance experiments in English. Our subjects were drawn from two populations: 10 native English speakers and 19 native Mandarin Chinese speakers. Curious about the potential effects of "accent" on speaker authentication, we tested our methods on the two populations separately, using models built from the universal training data. Table 7 shows the BAC for authentication of native English speakers and native Mandarin Chinese speakers. For both methods and both authentication problems, the balanced accuracy of native Mandarin Chinese speakers was statistically significantly higher ($p < 0.02$) than those of native English speakers. It appears that native Mandarin Chinese speakers were more distinguishable than native English speakers when speaking English. Similarly, Van Leeuwen et al. found that the 'air voices' from native Mandarin Chinese speakers are more distinguishable than 'air voices' from native English speakers when they are speaking English [47]. However, we need larger populations for both languages to demonstrate this argument is also true in terms of 'body voices'. It would be an interesting follow-up study to determine how accents and native languages affect our system, via experiments for different languages and including native speakers from other languages.

Table 7. BAC on identification and verification

method	native English speakers		native Mandarin Chinese speakers	
	Identification	Verification	Identification	Verification
GMM	0.870	0.929	0.888	0.960
DNN	0.898	0.935	0.920	0.970

Resource use: In this work we propose two approaches, GMM in a stand-alone mode on the wearable device, and DNN on a remote server or cloud service. The stand-alone approach has weaker accuracy and higher energy use, but does not require network access and maintains privacy by processing all audio on-board. The off-board approach has higher accuracy, but requires network access and places trust in the remote server to process speech features that may contain personal or sensitive information. We plan to seek optimizations in the algorithm to allow a personal smartphone to be used as a trusted server to run the DNN algorithms. There are promising works in progress to implement resource-efficient deep-learning algorithms for small devices [3].

Signal quality: We collected vocal resonance data from subjects in a quiet lab environment. The signal quality may change due to noisy environments, for example, busy streets. Moreover, the accuracy metrics may be affected by the signal quality, specifically, the signal-noise-ratio (SNR). In our dataset, the average SNR of vocal resonance from the subjects is 26.9 dB. To learn how signal quality affects our system, we need to collect vocal resonance data from other environments and perform extensive studies.

9 CONCLUSION

In this paper we present *vocal resonance* as a novel, unobtrusive biometric that can support user authentication (identification and verification) in small, wearable pervasive devices. Notably, it goes well beyond traditional speaker-recognition methods by specifically confirming whether the device is *on* the speaker's body, not simply nearby. In addition, we implemented a wearable prototype and tested it with two different machine-learning methods, an on-board GMM method and a remote-assisted DNN method. Using data from 29 subjects, we found that the DNN method achieved balanced accuracy 0.914 for identification and 0.961 for verification, while the GMM method achieved 0.875 and 0.942 respectively. Our results show that 1) it is possible to achieve reliable speaker authentication through a wearable, body-contact microphone, that can reliably distinguish among multiple individuals, 2) it can distinguish between the situation where the microphone is on the body of the

enrolled speaker and where the microphone is simply nearby, even on another body, 3) it can authenticate the wearers after a delay of two weeks, and 4) it is robust against replay attack when one's air voice is replayed through another body. Our prototype, based on a Raspberry Pi and a USB sound card, was able to collect and process the data in a reasonable amount of time and with a reasonable battery lifetime, given a suitable duty cycle. In future work we anticipate refining these methods, optimizing their performance, and using the smartphone as a platform to execute the DNN classifier.

ACKNOWLEDGMENTS

This research results from a research program at the Institute for Security, Technology, and Society at Dartmouth College, supported by the National Science Foundation under award number CNS-1329686, CNS-0910842. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

REFERENCES

[1] Adafruit. 2017. USB Battery Pack. (Aug. 2017). https://www.adafruit.com/product/1959

[2] Salil P. Banerjee and Damon L. Woodard. 2012. Biometric Authentication and Identification using Keystroke Dynamics: A Survey. *Journal of Pattern Recognition Research* 7, 1 (July 2012), 116–139. https://doi.org/10.13176/11.427

[3] Sourav Bhattacharya and Nicholas D. Lane. 2016. Sparsification and Separation of Deep Learning Layers for Constrained Resource Inference on Wearables. In *Proceedings of the ACM Conference on Embedded Network Sensor Systems (SenSys)*. ACM, 176–189. https://doi.org/10.1145/2994551.2994564

[4] Frédéric Bimbot, Jean-Franc Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Téva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A. Reynolds. 2004. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Advances in Signal Processing* 2004, 4 (Jan. 2004), 430–451. https://doi.org/10.1155/s1110865704310024

[5] Jorge Blasco, Thomas M. Chen, Juan Tapiador, and Pedro P. Lopez. 2016. A Survey of Wearable Biometric Recognition Systems. *Journal ACM Computing Surveys (CSUR)* 49, 3 (Sept. 2016), 43:1–43:35. https://doi.org/10.1145/2968215

[6] Ruud M. Bolle, Jonathan H. Connell, Sharanthchandra Pankanti, Nalini K. Ratha, and Andrew W. Senior. 2004. *Guide to biometrics* (1 ed.). Springer. https://www.springer.com/computer/image+processing/book/978-0-387-40089-1

[7] Laura E. Boucheron and Phillip L. De Leon. 2008. On the inversion of Mel-frequency cepstral coefficients for speech enhancement applications. In *Proceedings of the International Conference on Signals and Electronic Systems*. IEEE, 485–488. https://doi.org/10.1109/icses.2008.4673475

[8] Shaxun Chen, Amit Pande, and Prasant Mohapatra. 2014. Sensor-Assisted Facial Recognition: An Enhanced Biometric Authentication System for Smartphones. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 109–122. https://doi.org/10.1145/2594368.2594373

[9] Cory Cornelius and David Kotz. 2012. Recognizing whether sensors are on the same body. *Journal of Pervasive and Mobile Computing* 8, 6 (Dec. 2012), 822–836. https://doi.org/10.1016/j.pmcj.2012.06.005

[10] Cory Cornelius, Zachary Marois, Jacob Sorber, Ron Peterson, Shrirang Mare, and David Kotz. 2014. *Vocal resonance as a biometric for pervasive wearable devices*. Technical Report TR2014-747. Dartmouth Computer Science. http://www.cs.dartmouth.edu/reports/TR2014-747.pdf

[11] Cory Cornelius, Ronald Peterson, Joseph Skinner, Ryan Halter, and David Kotz. 2014. A wearable system that knows who wears it. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 55–67. https://doi.org/10.1145/2594368.2594369

[12] Cory T. Cornelius. 2013. *Usable Security for Wireless Body-Area Networks*. Ph.D. Dissertation. Dartmouth College Computer Science, Hanover, NH. http://www.cs.dartmouth.edu/reports/TR2013-741.pdf Available as Dartmouth Computer Science Technical Report TR2013-741.

[13] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* 39, 1 (April 1977), 1–38. https://doi.org/10.2307/2984875

[14] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 835–838. https://doi.org/10.1145/2502081.2502224

[15] Grant Fairbanks. 1960. *Voice and Articulation Drillbook* (2nd ed.). Harper & Row. 127 pages. https://doi.org/10.1288/00005537-194112000-00007

[16] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous Authentication for Voice Assistants. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom)*. ACM, 343–355. https://doi.org/10.1145/3117811.3117823
[17] Fitbit. 2016. Fitbit Alta HR. (Nov. 2016). https://www.fitbit.com
[18] Jean-Luc Gauvain and Chin-Hui Lee. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing* 2, 2 (April 1994), 291–298. https://doi.org/10.1109/89.279278
[19] Kenneth Grahame. 1908. *The wind in the willows*. Bantam Classics. 232 pages. https://books.google.com/books?isbn=0752548727
[20] Rashidul Hasan, Mustafa Jamil, and Golam Rabbani Saifur Rahman. 2004. Speaker Identification using Mel Frequency Cepstral Coefficients. *Variations* 1 (Dec. 2004), 4. https://doi.org/10.1109/CONIELECOMP.2012.6189918
[21] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. 2016. End-to-end text-dependent speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5115–5119. https://doi.org/10.1109/icassp.2016.7472652
[22] Emotiv Inc. 2017. Emotiv Insight. (Oct. 2017). https://www.emotiv.com/insight/
[23] Amit K. Jain, Arun Ross, and Sanjay Prabhakar. 2004. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 14, 1 (Jan. 2004), 4–20. https://doi.org/10.1109/tcsvt.2003.818349
[24] David Kotz, Carl A. Gunter, Santosh Kumar, and Jonathan P. Weiner. 2016. Privacy and Security in Mobile Health: A Research Agenda. *IEEE Computer* 49, 6 (June 2016), 22–30. https://doi.org/10.1109/MC.2016.185
[25] Kshitiz Kumar, Chanwoo Kim, and Richard M Stern. 2011. Delta-Spectral Cepstral Coefficients for Robust Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4784–4787. https://doi.org/10.1109/ICASSP.2011.5947425
[26] Nicholas D. Lane, Petko Georgiev, and Lorena Qendro. 2015. DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments Using Deep Learning. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp)*. ACM, 283–294. https://doi.org/10.1145/2750858.2804262
[27] Yann LeCun, Yoshua Bengio, and Geoffery Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444. https://doi.org/10.1038/nature14539
[28] Feng Lin, Chen Song, Yan Zhuang, Wenyao Xu, Changzhi Li, and Kui Ren. 2017. Cardiac Scan: A Non-contact and Continuous Heart-based User Authentication System. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom)*. ACM, 315–328. https://doi.org/10.1145/3117811.3117839
[29] Rui Liu, Cory Cornelius, Reza Rawassizadeh, Ron Peterson, and David Kotz. 2017. Poster: Vocal Resonance as a Passive Biometric. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 160. https://doi.org/10.1145/3081333.3089304
[30] Rui Liu, Reza Rawassizadeh, and David Kotz. 2017. Toward Accurate and Efficient Feature Selection for Speaker Recognition on Wearables. In *Proceedings of the Workshop on Wearable Systems and Applications (WearSys)*. ACM, 41–46. https://doi.org/10.1145/3089351.3089352
[31] Hong Lu, A. J. Bernheim Brush, Bodhi Priyantha, Amy K. Karlson, and Jie Liu. 2011. SpeakerSense: Energy Efficient Unobtrusive Speaker Identification on Mobile Phones. In *Proceedings of International Conference on Pervasive Computing*. Springer, 188–205. https://doi.org/10.1007/978-3-642-21726-5_12
[32] Lumo Bodytech. 2017. Lumo Run. (Oct. 2017). https://www.lumobodytech.com
[33] Monsoon Solutions, Inc. 2017. Monsoon Power Monitor. (Aug. 2017). https://www.msoon.com/LabEquipment/PowerMonitor
[34] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the International Conference on Machine Learning (ICML)*. Omnipress, 807–814. https://doi.org/10.1.1.165.6419
[35] Raspberry Pi Foundation. 2017. Raspberry Pi Zero Wireless. (March 2017). https://www.raspberrypi.org
[36] Reza Rawassizadeh, Elaheh Momeni, Chelsea Dobbins, Joobin Gharibshah, and Michael Pazzani. 2016. Scalable Daily Human Behavioral Pattern Mining from Multivariate Temporal Data. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (July 2016), 3098–3112. https://doi.org/10.1109/TKDE.2016.2592527
[37] Reza Rawassizadeh, Blaine A. Price, and Marian Petre. 2015. Wearables: Has the Age of Smartwatches Finally Arrived? *Communications of the ACM* 58, 1 (Jan. 2015), 45–47. https://doi.org/10.1145/2629633
[38] Douglas A. Reynolds. 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17, 1-2 (Aug. 1995), 91–108. https://doi.org/10.1016/0167-6393(95)00009-d
[39] Douglas A. Reynolds. 2002. An Overview of Automatic Speaker Recognition Technology. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, 4072–4075. https://doi.org/10.1109/ICASSP.2002.5745552
[40] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing* 10, 1-3 (Jan. 2000), 19–41. https://doi.org/10.1006/dspr.1999.0361
[41] Fred Richardson, Douglas Reynolds, and Najim Dehak. 2015. Deep Neural Network Approaches to Speaker and Language Recognition. *IEEE Signal Processing Letters* 22, 10 (April 2015), 1671–1675. https://doi.org/10.1109/LSP.2015.2420092
[42] Hasim Sak, Andrew W Senior, and Françoise Beaufays. 2014. Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. In *INTERSPEECH*. International Speech Communication Association (ISCA), 338–342. http://www.isca-speech.org/archive/interspeech_2014/i14_0338.html

[43] Michael Sherman, Gradeigh Clark, Yulong Yang, Shridatt Sugrim, Arttu Modig, Janne Lindqvist, Antti Oulasvirta, and Teemu Roos. 2014. User-generated Free-form Gestures for Authentication: Security and Memorability. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 176–189. https://doi.org/10.1145/2594368.2594375

[44] Sreenivas Sremath Tirumala and Seyed Reza Shahamiri. 2016. A Review on Deep Learning Approaches in Speaker Identification. In *Proceedings of the International Conference on Signal Processing Systems (ICSPS)*. ACM, 142–147. https://doi.org/10.1145/3015166.3015210

[45] Satoru Tsuge, Takashi Osanai, Hisanori Makinae, Toshiaki Kamada, Minoru Fukumi, and Shingo Kuroiwa. 2008. Combination Method of Bone-Conduction Speech and Air-Conduction Speech for Speaker Recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 1929–1932. https://doi.org/10.1109/ISPACS.2009.5383806

[46] Sharath Umesh, Lawrence Cohen, and David Nelson. 2002. Frequency Warping and the Mel Scale. *IEEE Signal Processing Letters* 9, 3 (Aug. 2002), 104–107. https://doi.org/10.1109/97.995829

[47] David A. van Leeuwen, Alvin F. Martin, Mark A. Przybocki, and Jos S. Bouten. 2006. NIST and NFI-TNO evaluations of automatic speaker recognition. *Computer Speech & Language* 20, 2-3 (April 2006), 128–158. https://doi.org/10.1016/j.csl.2005.07.001

[48] Bayya Yegnanarayana, A. Shahina, and M. R. Kesheorey. 2004. Throat Microphone Signal for Speaker Recognition. In *Proceedings of International Conference on Spoken Language Processing (INTERSPEECH)*. ISCA, 2341–2344. http://www.isca-speech.org/archive/interspeech_2004/i04_2341.html

[49] Xiao Zeng, Kai Cao, and Mi Zhang. 2017. MobileDeepPill: A Small-Footprint Mobile Deep Learning System for Recognizing Unconstrained Pill Images. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, 56–67. https://doi.org/10.1145/3081333.3081336

Received August 2017; revised November 2017; accepted January 2018.