# Vocal resonance as a passive biometric

Cory Cornelius, Zachary Marois, Jacob Sorber, Ron Peterson, Shrirang Mare, David Kotz

Dartmouth College — Institute for Security, Technology, and Society

Dartmouth Computer Science Technical Report TR2014-747

## ABSTRACT

We anticipate the advent of body-area networks of pervasive wearable devices, whether for health monitoring, personal assistance, entertainment, or home automation. In our vision, the user can simply wear the desired set of devices, and they "just work"; no configuration is needed, and yet they discover each other, recognize that they are on the same body, configure a secure communications channel, and identify the user to which they are attached. This paper addresses a method to achieve the latter, that is, for a wearable device to identify the wearer, allowing sensor data to be properly labeled or personalized behavior to be properly achieved. We use *vocal resonance*, that is, the sound of the person's voice as it travels through the person's body. By collecting voice samples from a small wearable microphone, our method allows the device to determine whether (a) the speaker is indeed the expected person, and (b) the microphone device is physically *on* the speaker's body. We collected data from 25 subjects, demonstrate the feasibility of a prototype, and show that our method works with 77% accuracy when a threshold is chosen *a priori*.

## 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 not uncommon for people to carry multiple computing devices, such as smart phones, music players, and cameras; increasingly, they also carry, hold, or wear devices to measure physical activity (e.g., Fitbit [8]), to interact with entertainment devices (e.g., the Wii), or to monitor their physiology (e.g., a cardiac patient concerned about heart arrhythmia or a diabetic managing her blood glucose). Many more have been proposed or developed as research prototypes. 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 fundamental problem involving wearable devices: who is wearing the device? This problem is key to nearly any application. For an entertainment device, it can recognize the user and load the right game profile or music playlist. For a home climate control, it can adjust the environment to the wearer's preference. Most compellingly, for a health-monitoring device, it can label the sensor data with the correct identity so that it can be stored in the correct health record. (A mixup of sensor data could lead to incorrect treatment or diagnosis decisions, with serious harm to the patient.)

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 implanted or 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.

We have earlier developed a method for a networked set of devices to recognize that they are located on the same body; our approach uses correlations in accelerometry signals for this purpose [4]. 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 statically 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 sensor). 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 [4]). This device should be trained once, for each user that might wear it, but thenceforth be completely automatic and unobtrusive.

Our approach is to use *vocal resonance*, that is, the sound of the person's voice as it travels through the person's 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.

Such solutions have many advantages. Not all wearable devices need have the ability to identify the user; only one device need do so, assuming it can communicate the identity to other devices proven to be on the same body. The devices may be smaller and simpler, needing no interface for user identification (or PIN or password for authentication). Use of a biometric provides important security and privacy properties, preventing unauthorized users from either accessing sensitive data (e.g., in which an adversary Alice tricks Bob's sensor into divulging his activity data to her smart phone), and preventing the mis-labeling of sensor data that might later be used for medically important decisions. Privacy is particularly important in health-related pervasive applications [1]. Furthermore, these methods can support personalization techniques so often envisioned in pervasive computing.

### *Contributions.*

In this paper we present a novel method for an unobtrusive biometric measurement that can support user identification in small, wearable pervasive devices. Drawing on known methods for speaker identification, we show that it is possible to achieve reliable speaker identification through a wearable, body-contact microphone, that can reliably distinguish among multiple individuals sharing a household, and indeed that it can distinguish between the situation where the microphone is on the body of the identified speaker and where the microphone is simply nearby, even on another body. We evaluate the feasibility of vocal resonance as a biometric using data collected from 25 subjects. In addition, we implemented a wearable prototype and tested it in stationary and mobile settings in both quiet and noisy environments. Our method achieves 77% accuracy when an *a priori* threshold is optimized for minimizing the false acceptance rate.

In the next section, we provide more background on biometrics. Then in Sections 3 and 4 we detail our model and describe our method, respectively. In Section 5 we describe our implementation of a wearable prototype based on the Gumstix platform. In Section 6 we present our evaluation of the method as a suitable biometric based on measurements from human subjects. Finally, we compare our work with related work in Section 7, discuss our findings in Section 8 and our conclusions in Section 9.

## 2. BIOMETRICS

To attach an identity to the sensor data, we first need some method of identifying whom the device is sensing. One approach is 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 [3]. Biometrics leverage physiological or behavioral characteristics of a person to accomplish identification. 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 the following properties: universality, uniqueness, and permanence. A *universal* characteristic is one that every person (or most people) possess. 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.

Furthermore, in the context of pervasive applications and particularly personal health sensors, a biometric needs to 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 not have to do anything more but 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 camera can capture one's gait, or a malicious application could learn one's typing rhythm [13]. A biometric suited for our purposes would incorporate the difficulty of circumventing a physiological biometric with the measurability of a behavioral biometric.

## 3. MODEL

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 a person's body, we can use speaker identification techniques to determine the speaker while simultaneously guaranteeing that the microphone is attached to the speaker's body. Like a typical speaker-identification system, the microphone hears the person's voice, but unlike a typical speaker-identification system, the microphone is hearing the voice as it travels through the body itself, rather than through the air. Thus, the system needs to identify who is speaking, and verify

Figure 1: Two types of microphones. On the left is a commercial throat microphone intended to be worn around the neck. On right are the contact microphones intended to be used for a guitar. These are not on the same scale.

that the detected voice is coming through the body and not through the air or some other medium.

A traditional speaker-identification 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-identification 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. Most systems alleviate this concern by employing a challenge-response scheme, whereby the person is asked to speak a randomly generated phrase. However, this is obtrusive and thus unsuitable. Capturing the vocal resonance of person is unobtrusive: all the user must do is talk as they go about their day. Unlike a traditional speaker-identification 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. One would imagine a microphone placed on the throat would be optimal for collecting a person's vocal resonance. However, it is difficult to imagine people opting to wear something like a traditional throat microphone (seen in Figure 1, at left). The mere presence of such a device on a person indicates to others that they are using some type of personal sensing system.

We imagine a piece of jewelry, not unlike a necklace or earpiece, that would contain a contact microphone to sample vocal resonance and another microphone to sample ambient sound. The form factor of a necklace or earpiece 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 necklace or earpiece can be instrumented to detect when it has been placed on and taken off a person. This can be detected, for example, by the ends of the necklace being clasped together or when the earpiece has sufficient contact with the ear by detecting properties of the skin such as temperature, moisture, or electrical impedance. Because we require the microphone to be in contact with the body and not all form factors will afford continuous contact, a mechanism to detect when the device is in contact with a body is necessary (but outside the scope of this paper). Such a simple detection mechanisms also allow us to conserve energy by only performing identification when the microphone is actually in contact with a person.

## 3.1 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, thus we assume there is a 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 detect and handle this case properly.

We consider some 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 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 into the environment. However, we do not assume they will physically alter the device or its firmware. We discuss how one might mitigate an active adversary in Section 8.

# 4. METHOD

We are inspired by the techniques from the speaker identification literature [19] but account for the unique nature of the data collected via a contact microphone. Figure 2 shows the major components of our model. When a person first receives the device, they must train the device to recognize their voice. In 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 model computed from the contact microphone models the speaker's vocal resonance, while the model computed from the ambient microphone models the speaker's voice through the air. Typically, these models would be computed off-device because the computational requirements for learning such a model are high. The enrollment algorithm produces two models of that particular speaker; the models are loaded into the device for later use. Because we anticipate the device being used by multiple people, it may be loaded with multiple speaker models.

Once its users are enrolled, the device performs periodic verification to determine whether it is on a person's body and to identify the speaker. The device first checks whether or not it has been placed on somebody, using the mechanisms described above; if so, it periodically collects a sample of audio from the contact microphone. If the device determines the audio sample to contain speech, the device then uses an *identification algorithm* to determine which enrolled person's models, if any, fit this audio sample.

Before we present the enrollment and identification algorithms, we define the audio-segmentation and feature-extraction methods common to both algorithms.

## 4.1 Audio Segmentation

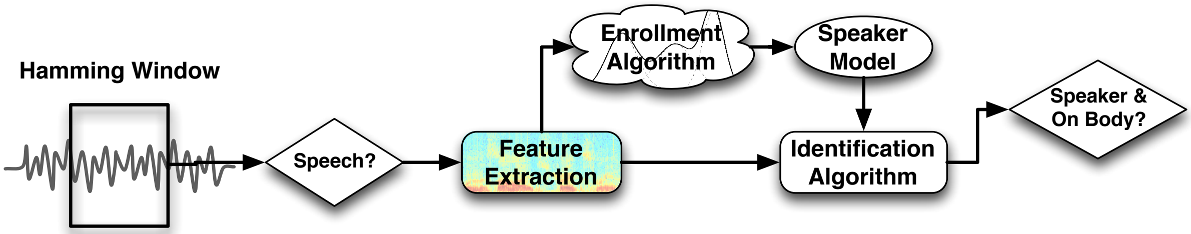


Figure 2: The major components of our method. Most of the components would be computed on the device itself, except for the enrollment algorithm (which can be done in the cloud).

Given audio data from a microphone, the first task is to divide it into smaller chunks to ease its processing. Our method examines a brief 20ms segment of audio data, a size at which speech is assumed to be quasi-stationary [16]. For each 20ms segment of audio data, we apply a Hamming window to deemphasize the sides of the audio segment and compensate for spectral leakage. Because the sides are deemphasized and to account for temporal effects, it is preferable to have the segments overlap. We use a standard 50% (10 ms) overlap [19].

Because not all audio data will contain speech, it is first necessary to determine whether the audio segment contains speech; this can be accomplished efficiently using simple time-domain features combined with a decision tree as described by Lu et al. [12]. If an audio segment is deemed not to contain speech, it is discarded. (During enrollment, corresponding segments from the contact and ambient microphones are discarded if either segment is determined not to contain speech.)

## 4.2 Feature Extraction

Given such an audio segment, we first extract some features that capture features of the person's voice. We use the set of Mel-Frequency Cepstral Coefficients (MFCCs) [14], 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 [20]. For this reason, MFCCs have been successful in many voice-processing tasks. Since the first coefficient models the mean value of the signal, we discard this coefficient. In addition to these coefficients, we include the first and second derivatives (velocity and acceleration) of each coefficient, to capture how the coefficients change over time. These derivatives account for the fact that the defining features of a person's voice, *formants*, vary over time. We approximate each derivative by computing the 5-point central difference:

$$v_t^{(1)} = \frac{c_{t-2}^{(1)} - 8c_{t-1}^{(1)} + 8c_{t+1}^{(1)} - c_{t+2}^{(1)}}{12}$$

where $v_t^{(1)}$ is the velocity of the first coefficient $c_t^{(1)}$ at time $t$. We compute acceleration in a similar matter except over velocities instead of coefficients. The result is a *feature vector* for each audio segment for some chosen number of coefficients and their respective derivatives. We call a feature vector computed from audio data sampled from the contact microphone a *contact feature vector*, and we call a feature vector computed from the ambient microphone an *ambient feature vector*. In our experiments, we vary the number of coefficients and fix the number of mel-scale filters equal to the number of desired coefficients.

## 4.3 Enrollment Algorithm

During enrollment, we collect audio from both the contact microphone and the ambient microphone, and build a model for each. To do so, we segment the audio, extract a feature vector for each segment, and statistically model the distribution of feature vectors. The most effective such model for speaker identification is a Gaussian Mixture Model (GMM) [17]. A GMM models the distribution of observations using a weighted linear combination of Gaussian densities where each Gaussian density is parameterized by a mean vector and covariance matrix. Thus, we use GMMs to model the distribution of feature vectors for a given speaker. We model the distribution of contact feature vectors and ambient feature vectors separately. To learn the underlying distribution of feature vectors, we use the Expectation-Maximization (EM) algorithm [5] 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 choose initial Gaussian densities by clustering the set of feature vectors using $k$-means clustering [11], where $k$ is set to the desired number of Gaussian densities. We iteratively refine these initial Gaussian densities using the EM algorithm. Modeling the covariance matrix in full is computationally expensive, so we use diagonal covariance matrices because it has been shown that using a larger-dimensional diagonal covariance matrix performs better than a smaller-dimensional full covariance matrix [2]. Similarly, because some values of the covariance matrix can become very small, as in the case of outliers, we enforce a variance floor of $10^{-5}$.

In essence, a GMM serves to summarize the set of feature vectors that capture the sound of a particular speaker's voice. The set of feature vectors is reduced to a diagonal covariance matrix, mean vector, and weight for each Gaussian density. We call a model trained on contact feature vectors a *contact model*, while a model trained on ambient feature vectors is called an *ambient model*. In our experiments, we vary the number of Gaussian densities for each model.

Once we have learned a GMM, we also need learn a threshold at which the likelihood of a contact feature vector should be accepted or rejected. The likelihood of a feature vector given a GMM is simply the weighted linear combination of the probability density function of each Gaussian given the feature vector. From a speaker's contact and ambient models we compute the likelihood of a given contact feature vector corresponding to each model. A *contact likelihood* is one computed from a contact feature vector given the contact model, and *ambient likelihood* is computed from a contact feature vector given the ambient model. Given these two likelihoods, we compute a *likelihood ratio* to decide whether the sample came from contact model or the ambient model. For numerical stability, we compute the log-likelihood which allows us to take the difference between the two log-likelihoods, which we call the *difference likelihood*. If the difference likelihood is greater than some threshold, then we say the audio segment correspond to the contact feature matches the speaker's vocal resonance; otherwise it does not fit the model and therefore does not match the speaker's vocal resonance. In this way the ambient model acts as a kind of background model [18], albeit not a universal one since it is only modeling the speaker's voice through the air.

Ideally a threshold should exactly classify each sample correctly, but there is no good way to choose a threshold *a priori*. Because we wish to use vocal resonance as a biometric, it makes sense to minimize the number of times the device will accept a contact feature vector that was computed from an audio segment collected from an unenrolled person. This is called a *false accept*, and we choose the threshold that minimizes the number of false acceptances in the training set. In Section 6 we evaluate how close this *a priori* threshold (as computed on the training set) is to the *a posteriori* threshold (as computed on the testing set).

## 4.4 Identification Algorithm

During the identification phase we use the contact microphone only. We segment the audio sample and extract a series of contact feature vectors; we wish to determine whether these newly measured contact feature vectors match the vocal resonance of an enrolled speaker. For a given speaker's pair of models, built during enrollment, we compute the difference likelihood for this new contact feature vector. Because we extract a feature vector every 10 ms, it is preferable to average over a series of difference likelihoods to achieve stability; we compute this *average likelihood* using a simple moving average over $n$ difference likelihoods. As above, if this average difference likelihood is greater than the pre-computed threshold, then we assume those segments of audio match the vocal resonance for that speaker.

Because we cannot know *a priori* who is wearing the device, we compute these average likelihoods for each model the device has trained. The device reports the speaker of the trained contact model as the current speaker if the average difference likelihood is greater than the model's threshold. If more than one model has a average likelihood greater than its respective threshold, then we consider this an unclassifiable segment of audio and the identification algorithm reports it as such. Over a window of $m$ predictions, we can further smooth these reports to reduce spurious misclassifications via a simply majority vote.

## 4.5 Parameter Settings

We explore several parameters that have a direct effect on the computational, storage, and energy resources required for enrollment and identification. First is the number of MFCCs in the feature vector, which will affect how much the higher frequencies contribute to the learning process. Second is the number of Gaussians, which will affect how well the GMM fits the underlying distribution. Third, we can vary the number of recent likelihoods we average, where an average taking into account more samples might be more accurate but will require longer audio captures and thus a delay in identification. Fourth, we can vary the number of predictions we smooth over, where a longer smoothing window will require more samples but might reduce spurious misclassifications. In Section 6 we explore settings of these parameters.

# 5. IMPLEMENTATION

To test our approach, we conducted two types of experiments. In this section, we describe how we implemented our approach on a wearable device, and the results of experiments with the prototype. In the next section, we show how we collected voice recordings from 25 users to explore our method's ability to identify and distinguish people.

For our prototype we used the Gumstix [9] platform, specifically, the Gumstix Overo Fire COM with the Tobi expansion board. The Overo Fire COM has a 700 MHz ARM Cortex-A8 processor, 512MB RAM, 512MB flash, a microSD slot, and it supports Wi-Fi (802.11b/g); using the Tobi expansion board, we can connect a microphone to the COM. Figure 3a shows the entire Gumstix setup: Overo Fire COM mounted on the Tobi expansion board, USB sound card, and wearable acoustic microphones; the Gumstix Overo COM itself is quite small (see Figure 3b).

Although the Gumstix expansion boards have built-in audio line-in ports, they have poor audio quality and we could not use them.[1] So we decided to use a USB sound card to get good-quality audio; we used the Amigo II USB sound card by Turtle Beach (as shown in the Figure 3a).

We implemented the audio segmentation (Section 4.1), feature extraction (Section 4.2), and the identification algorithm (Section 4.4) on the Gumstix processor. We also implemented the enrollment algorithm (Section 4.3), but it is a computationally heavy process, and needs to be done only once, so we ran it offline on a desktop. We used the FFTW [7] library for the feature-extraction step; we cross-compiled it for the Overo. Our implementation requires approximately 5,800 lines of C++ code. For the Overo, we built a console image (kernel version 3.0.0) using the OpenEmbedded build framework.

## 5.1 Experiments

To evaluate the performance of our prototype, we measured two metrics: latency, that is, how quickly the system identifies the user, and energy consumption. We recorded an audio sample on the Gumstix (using the `arecord` utility), and then ran the feature-extraction and identification algorithms on the Gumstix multiple times, each time with different parameters, and we measured the latency and energy consumption.

We ran the feature-extraction step for five different numbers of coefficients, shown in Table 1. In the feature-extraction step, we generate three features for each coefficient. Hence, as the number of coefficients increase, the MFCC step runs longer, increasing the latency and thus the energy consumed.

[1] We tried two expansion boards, the Tobi and Palo 35, and both the boards had poor quality audio from the line-in, the Tobi being slightly better.

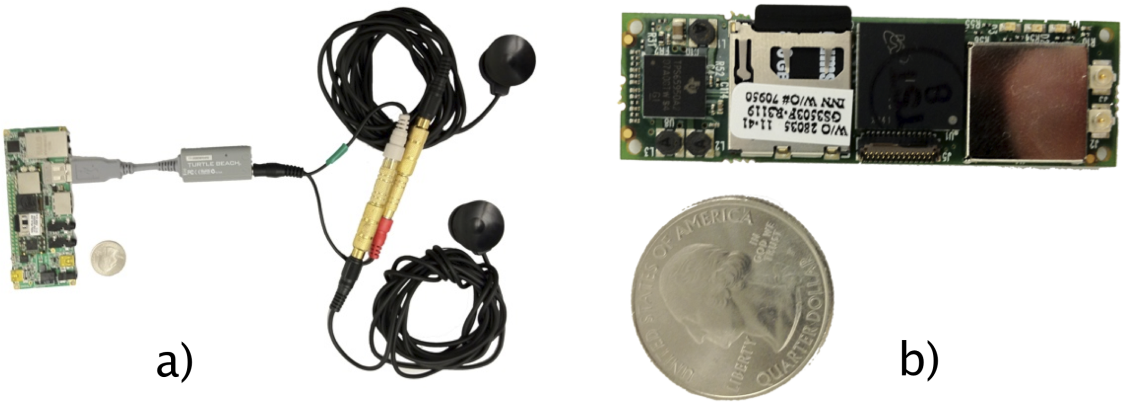


Figure 3: a) Gumstix setup: Overo mounted on Gumstix Tobi expansion board, USB sound card, and two microphones, b) Gumstix Overo Fire COM

| Coefficients | Energy (J) | Latency (sec) |
|---|---|---|
| 10 | 2.655 | 1.574 |
| 20 | 2.805 | 1.662 |
| 30 | 2.958 | 1.752 |
| 40 | 3.094 | 1.831 |
| 50 | 3.242 | 1.917 |

Table 1: Energy consumption and latency of the feature-extraction step for different number of coefficients.

We ran the identification step 35 times, with five different coefficient values (10–50), and using seven different GMMs based on the number of Gaussians. As shown in Table 2, increasing the number of Gaussians and the number of coefficients, increases the latency and energy of the identification step, as expected.

The total energy required by our system on the Gumstix for user identification is the sum of energy required to record and segment audio, extract feature vectors, and run the identification step on a single user's speaker model. The total time, however, is not the sum of these three steps, because extraction and identification overlap with the recording of the audio. For 20 coefficients, with 16 Gaussians, a window size of 20, and a step size of 10, the feature-extraction step takes about 1.66 seconds to run over a 20-second audio sample, and the identification step takes about 5.53 seconds. The required energy for recording and processing a 20-second audio sample is 30.4 J (required energy for recording 20 second audio) + 2.8 J (required energy for feature-extraction step) + 13.3 J (required energy for identification step) = 46.5 J; note that the energy consumption is dominated by the audio recording step, and the feature extraction and identification steps only consume about 34% of the total energy.

With a 645 mA battery, the Gumstix can do continuous recording and identification for an hour. However, continuous speaker-identification is not required, in most settings, and the system need only run periodically. If, for example, the system collects a 20-second sample every 10 minutes, this battery would last more than a day. If the device has the capability to detect when it is removed from a person, then it need only conduct verification when it is transferred to a new person, and can last for much longer.

# 6. EVALUATION

In this section we explore the viability of vocal resonance as a biometric. Recall that we require a biometric that is universal, unique, permanent, unobtrusively measurable, and difficult to circumvent. We explore the uniqueness, measurability, and circumventability properties and provide arguments for why the universality and permanence properties hold as well.

## 6.1 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 collected a dataset and ran the described method to see how well the method could accurately classify individuals in our dataset.

### *6.1.1 Dataset*

We collected data from 45 human subjects using an IRB-approved protocol. Three AXL PG-800 External Piezo Transducer suction cup microphones (as seen in Figure 1, at right) recorded data from three locations, as shown in Figure 4. The speaker attached one microphone to the side of his or her neck, and held the second microphone six inches from his or her mouth. The first microphone simulated the case when an enrolled person is wearing the device; we call this microphone the *body* microphone. The second microphone simulated the case when an enrolled person is not wearing the device, but the device could still hear them speaking; we call this microphone the *air* microphone. In addition, a listener sat two to three feet away from the speaker and had a microphone attached to his own neck. This microphone simulated the case when another person, enrolled or unenrolled, is wearing the device and an enrolled speaker is speaking; we call this microphone the *other* microphone. The microphones were secured to each subject's neck using 3M Micropore Tape. A pair of Radio Shack Mini Audio Amplifiers amplified the body and air microphones; these amplifiers were connected to the line-in port of a laptop. The

| | Coefficients | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Gaussians | 10 | | 20 | | 30 | | 40 | | 50 | |
| | J | sec | J | sec | J | sec | J | sec | J | sec |
| 16 | 5.75 | 3.03 | 9.06 | 5.53 | 13.33 | 8.15 | 17.64 | 10.79 | 20.36 | 10.79 |
| 32 | 9.11 | 5.56 | 20.00 | 12.92 | 26.38 | 16.14 | 36.17 | 22.16 | 40.44 | 21.43 |
| 64 | 17.95 | 10.97 | 36.30 | 23.25 | 52.56 | 33.02 | 70.04 | 42.80 | 80.84 | 43.74 |
| 128 | 35.70 | 21.81 | 70.65 | 43.07 | 105.30 | 64.38 | 146.91 | 89.74 | 165.88 | 87.25 |

Table 2: Energy consumption and latency of the identification step for different number of coefficients and different number of Gaussians.

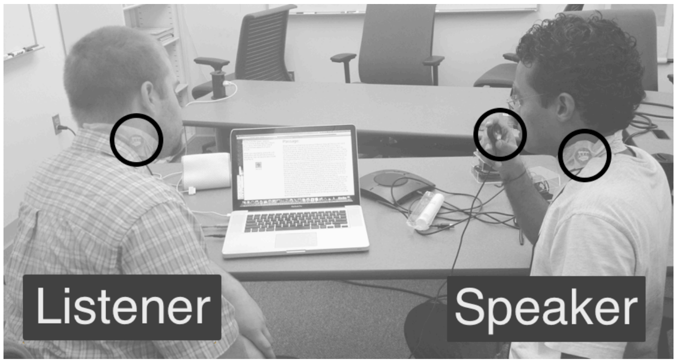


Figure 4: The sample collection setup. The microphone locations are indicated by circles.

other microphone connected to the laptop using a Turtle Beach Amigo II USB audio adapter, which has a built-in amplifier. We used the `PyAudio` and `wave` modules for Python to capture the audio data at 44100 Hz with 16-bit resolution. Figure 5 shows an example spectrogram for audio data collected at the described microphone locations.

We obtained usable data from 25 subjects (not 45, due to equipment malfunction). The average of these subjects, 17 males and 8 females, was 21 years. We instructed the subjects to read three passages. The first passage was used for training. We selected the *The Rainbow Passage* [6] because it encompasses most of the phonemes of the English language. The second passage 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. The third passage was selected randomly for each subject. We selected 20-26 consecutive lines from the first two chapters of *The Wind in the Willows* to act as a unique test phrase for each subject. Subjects took an average of 107 seconds to read The Rainbow Passage, 92 seconds to read the control test passage, and 91 seconds to read the randomized test passage.

### 6.1.2 Experiments

We ran several tests to experimentally evaluate the uniqueness of vocal resonance. As mentioned previously, our method can be parameterized by the number of extracted MFCCs, the number of Gaussian densities used to model them, and the number difference likelihoods used for smoothing. We explored these parameters to find an optimal setting that maximizes classification accuracy across all subjects.

#### Testing Procedure.

For each user in our dataset, we trained a GMM and learned a threshold using their training sample collected from their body microphone. Figure 6 shows a histogram of likelihoods at bottom and the ACC, FAR, and FRR of this training set for various thresholds. We computed these likelihoods from the same training samples that were used to compute the GMM. From this histogram of likelihoods, we learned the threshold, indicated by the vertical line in the figure, that minimizes the FAR.

We then tested the subject against all other subjects in our dataset using the collected testing data to determine how well the chosen threshold performed. A positive sample is any sample that came from the same subject and came from the body microphone. A negative sample is any other sample, including those containing speech from the same subject collected by either the air or other microphones. Thus, the number of negative samples outnumber the number of positive samples by a factor of 74 since we collected 3 separate sets of testing samples (one for each of the microphones) from each of the other 24 subjects.

#### Measures.

A sample is classified as positive if it is greater than the threshold, otherwise it is classified negative. Given a particular threshold, we report the *false accept rate* (FAR), *false reject rate* (FRR), *accuracy* (ACC), and *balanced accuracy* (BAC) for each test. For a set of classified samples, the FAR is defined as the fraction of negative samples that were misclassified (i.e., they were classified as positive), while the FRR is the fraction of positive samples that were misclassified (i.e., they were classified as negative). Since we can vary the threshold – which implies different FARs and FRRs for different thresholds – the EER is the rate at which the FAR equals the FRR. We cannot know the EER *a priori*, however it is useful to see how well an *a priori* threshold fares with an *a posteriori* selected threshold that is the EER. Balanced accuracy is the sum of half of the true accept rate (i.e., the fraction of positive samples that were correctly classified, or $1 - FRR$) and half of the true reject rate (the fraction of negative samples that were correctly classified, or $1 - FAR$). Balanced accuracy weights the negative and positive examples equally. Accuracy is ratio of correctly classified samples and the total number of samples. Because we perform each test for every subject, we report the average over all subjects.

#### Results.

We tested each subject with varying parameters. Fro each subject we varied the number of MFCCs from 10, 20, 30, 40, and 50; the number of Gaussian densities from 16, 32, 64, 128, 256, 512, and 1024; the length of the average window from 10, 100, and 1000; and the length of the smoothing window from

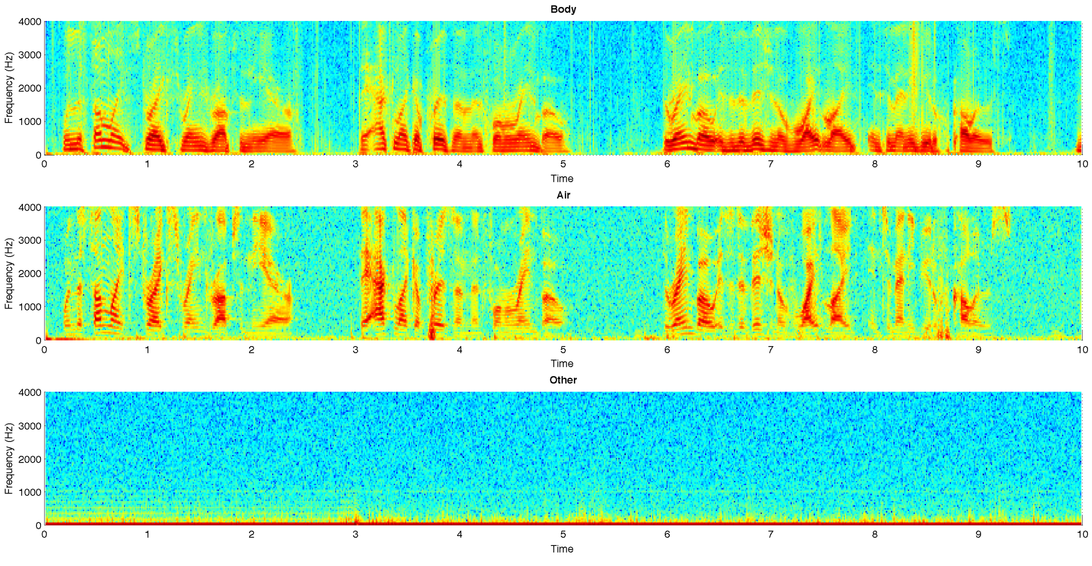


Figure 5: Spectrograms of a training sample collected from a subject. The top spectrogram was generated from audio collected via the microphone attached to the subject, the middle spectrogram was generated from audio collected via the microphone held in front of the subject, and the bottom spectrogram was generated by from audio collected via the microphone on another subject.

1, 17, 33, 65, 129, 257, 513, 1025. Each point in Figures 7a, 7b, 7c, and 7d represents a setting of these parameters where the value of that point is the BAC averaged over all users. The BAC tends to stay relatively the same, hovering around 65% for most parameter settings. However, the ACC goes as high as 100% when the number of Gaussians densities is low and the average window length is 1. This is a conservative setting because, although the FAR is very low, the FRR is also very high.

From Figures 7a, 7b, 7d, and 7c, we chose the parameter settings of 30 MFCCs, 128 Gaussian densities, length 10 averaging window, and length 129 smoothing window. On average, subjects *a priori* tested with 77% accuracy with an 8% standard deviation.

Finally, we determined the *a posteriori* threshold for each subject by searching for the threshold that maximizes EER for each subject. On average, subjects *a posteriori* tested with 75% accuracy with an 5% standard deviation. Thus our choice of threshold was not too too far off from the *a posteriori* threshold.

## 6.2 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 location of the device.

Background noise will also affect the measurability of vocal resonance, although we believe better contact microphones would alleviate most noise-related concerns. Table 3 shows how our method responds to different types of artificial noise. We trained GMMs for four users in a non-noisy environment, then we tested them in two simulated noisy environments – a pre-recorded restaurant, and white noise – along with a quiet environment for control. Table 3 also shows the effect on the mobility of the speaker on our methods. Moving samples were taken while the speaker paced across the room; non-moving samples were taken while the speaker sat calmly.

| Mobility & Noise Type | ACC | FAR | FRR |
|---|---|---|---|
| Moving & None | 86.71% | 5.15% | 35.56% |
| Moving & Restaurant | 86.90% | 3.23% | 41.31% |
| Moving & White | 87.50% | 0.26% | 48.70% |
| Non-Moving & None | 85.43% | 12.15% | 22.00% |
| Non-Moving & Restaurant | 84.70% | 9.53% | 31.53% |
| Non-Moving & White | 87.30% | 0.93% | 48.98% |

Table 3: Classification measures for different types of noise averaged across four users.

## 6.3 Circumvention

The number of negative misclassifications determines how easy it is to circumvent the biometric. An active attacker will try to fool the method by introducing specially crafted audio into the environment. For example, they could capture an enrolled person's voice and then replay that audio through their body so as to fool the device into believe the person is speaking. To simulate this attack, we placed a loudspeaker on the body of subject and played back an enrolled users voice. We then moved the loudspeaker up to 100cm away, with increments of 10cm, from the subject. Table 4 shows the effect on the classification measures when this scenario occurs for varying distances to the device.

An adversary could also swallow a loudspeaker and replay an enrolled user's voice through their own physiology. However, this is unpractical because identification occurs

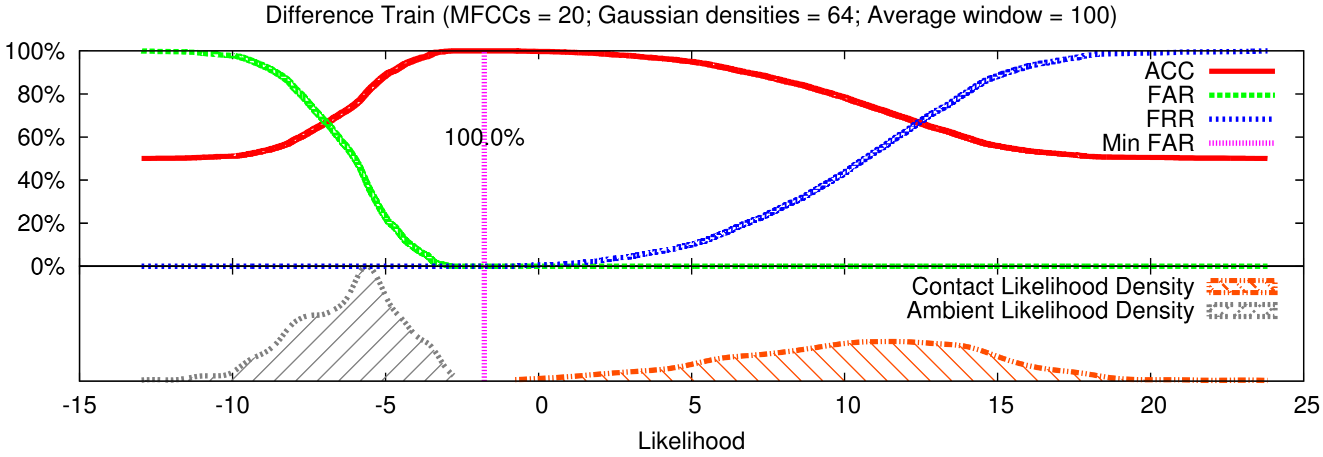


Figure 6: An example histogram of difference likelihoods computed from the training set of one subject, computed from the same samples used to learn the GMM. The number of MFCCs was fixed to 20, the Gaussian densities to 64, and the average window to 100. The vertical lines with percentages represent the learned thresholds and their corresponding ACC measures at that threshold.

| Distance from User | FAR |
|---|---|
| 0 cm | 7.9% |
| 10 cm | 0.58% |
| 20 cm | 0.33% |
| 30 cm | 0.95% |
| 40 cm | 1.23% |
| 50 cm | 3.23% |
| 60 cm | 0.75% |
| 70 cm | 1.93% |
| 80 cm | 0.65% |
| 90 cm | 1.43% |
| 100 cm | 1.23% |

Table 4: Classification rates when the distance of a hypothesized attacker varies.

repeatedly and it would be uncomfortable for an adversary to maintain a loudspeaker in their throat. A less invasive approach would be learning a filter that changes speech captured in the air to mimic like it was captured via the body. With such a filter, all an attacker would have to do is capture an enrolled person's voice, apply the filter, and play it back to the device. However, the feasibility of such an attack remains unclear, and we leave it for future work.

### 6.4 Universality

The universe of people that can use this device is limited to those that can speak, primarily because there must be some structure in the audio data collected by the microphone in the device. The major structures captured by our method are *formants*, which are the resonances of the human vocal tract. While our dataset consists of all English-speaking participants, our method could be used for other languages as well. We leave this testing of universality as future work.

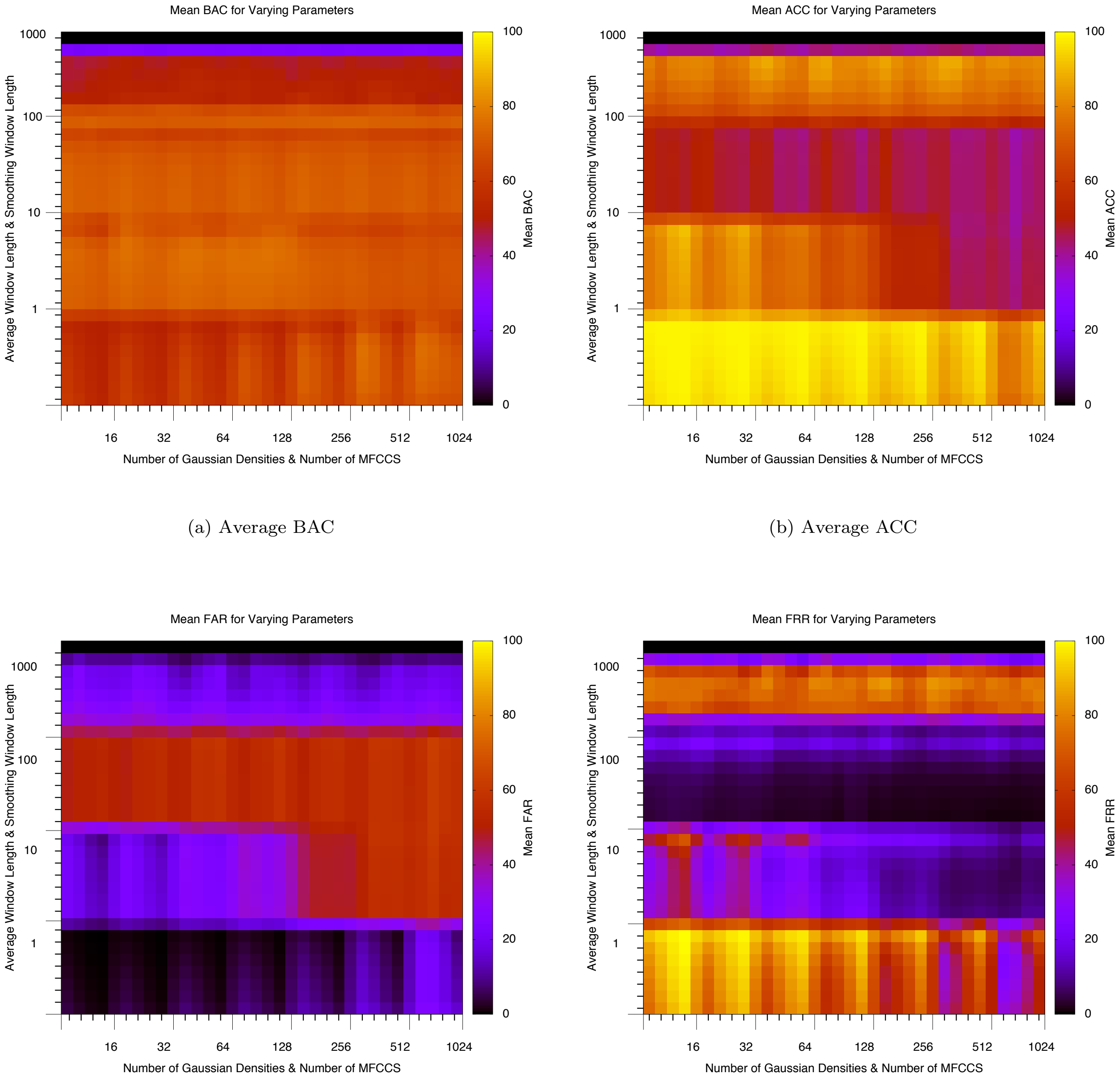
### 6.5 Permanence

Permanence is the property of remaining unchanged. It is well known, however, that each person's voice changes over their lifetime, both temporarily and permanently. A prominent physical changes occurs at puberty in the vocal cords, which are the primary source of speech production, causing them to thicken on into old age. Disease can also have a temporary or permanent effect on one's vocal cords; hoarse voice is a symptom of many diseases. Similarly, the physical dimensions of the person affect vocal resonance. The area which the device records vocal resonance can vary based upon the amount of fat, muscle, and bone density present. These could differ depending on caloric intake and exercise, or traumatic injury to the area.

Other than retraining, there is no good solution for abrupt changes to one's vocal resonance. Any kind of extraordinary change to the vocal tract would require retraining. For more subtle changes over time, we could employ a scheme similar to Darwin [15]. Darwin uses "classifier evolution," which is "an automated approach to updating models over time such that the classifiers are robust to the variability in sensing conditions common to mobile phones (e.g., phone in the pocket, in pocket bag, out of the pocket), and settings (e.g., noisy and loud environments)." Darwin's classifier evolution might be used to relearn a model used for speaker identification in a mobile setting.

## 7. RELATED WORK

Speaker-recognition and -identification systems have been studied for some time [2, 17, 19]. State-of-the-art speaker verification systems use features and methods similar to the one described in this paper. However, we differ from the state-of-the-art methods by not incorporating a background model. The purpose of a background model is to model all the possible audio segments that are not from the intended speaker; typically one learns a model of all other speakers in the dataset resulting in a "universal background model" [18]. Any kind of background model has the disadvantage that one must know beforehand all the representative speakers of the desired population. We sidestep this challenge by thresholding on the likelihood of the model trained on each enrolled user. While the accuracy of our method might be



(a) Average BAC

(b) Average ACC

(c) Average FAR

(d) Average FRR

Figure 7: The average BAC, ACC, FAR, and FRR computed over all subjects for all explored parameter settings.

lower than one that incorporates a background model, our method does not necessarily preclude the use of such a model. It is, *a priori*, difficult to collect representative background data and impossible to collect data from all other speakers the device might hear over its lifetime.

The most similar research to our own is by Yegnanarayana et al. [22]. They study the feasibility of a speaker-recognition system for samples collected by a throat 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 microphone is mostly 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 state-of-the-art speaker-recognition model using MFCCs and GMMs.

Tsuge et al. have looked at bone-conductive microphones for speaker verification [21]. (A bone-conductive microphone picks up sounds conducted through the skull.) 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.

# 8. DISCUSSION AND FUTURE WORK

Our experiments demonstrate the usefulness of vocal resonance as a biometric for personalizing wearable devices; however, a number of important issues still need to be addressed to incorporate this new biometric into real devices. The following paragraphs describe our goals and planned extensions to this work.

## 8.1 Improving Classification

Our work, to date, has focused primarily on the use of state-of-the-art speaker verification techniques to explore a novel biometric – vocal resonance. While this approach allows us to build on existing knowledge and works well in practice, these tools do not take full advantage of the unique characteristics of vocal resonance. For example, we assume that each user's vocal sounds are filtered differently, due to anatomical variation, as they pass through the body.

Moving forward, we plan to explore classification approaches that learn the parameters of a body's individual filter, to improve verification accuracy. Better understanding how the sound is filtered will also allow us to explore differential approaches using two microphones, that directly compare ambient sound to vocal resonance directly.

## 8.2 Optimizing Prototype Hardware

A major goal of this work is a working, wearable prototype device that can identify its wearer based on vocal resonance. We need to adapt our initial prototype so we can expand our experiments to include longitudinal studies that measure the effect of contact quality, time of day, exertion, stress, illness, and other potentially voice-altering phenomena on speaker identification.

Achieving this goal presents a range of challenges. Continuously analyzing audio samples requires significant processing, especially for a small resource-constrained device. While our initial feasibility experiments demonstrate that a software-only implementation achieves acceptable execution times on a low-power computer (Overo Gumstix), we are currently exploring techniques that improve both speed and energy efficiency.

A common solution for improving the efficiency is to employ specialized analog hardware to accomplish tasks that are much more computationally or energy intensive when performed digitally. Extracting spectral features (like MFCCs), from a signal using analog circuitry before converting before it is handed off to a DSP or other digital processor has been shown to dramatically improve both processing speed and energy efficiency [10]. In order to expand our ability to conduct longitudinal experiments on a much tighter energy constraints, we plan to implement the MFCC-extraction portion of our system in analog circuitry.

Finally, our prototype device is not very wearable. With custom hardware the device could easily be smaller and more amenable to the kind of wearable devices we see today. However, while miniaturization is possible, one also needs to account for the location of the microphone in the device itself and how the microphone is intended to hear the person's vocal resonance. We plan to explore these design considerations in future work.

## 8.3 Testing Stronger Attacks

Our experiments have evaluated the impact of simple threats—such as when the device is placed on the wrong person or when the owners voice is heard over the air but the devices is not being worn. Of course, other stronger attacks are possible. For example, an adversary might be able to fool the worn device by imitating the owner's voice, or by playing speech recorded from the device's owner through their own body using a contact speaker. In the future, we plan to evaluate how effective these attacks are and whether anatomical differences in bodies result in detectable differences in acoustic filtering.

We also plan to explore the use of better microphone isolation techniques as well as coordinated use of air and body microphones to correctly authenticate users even in the presence of these stronger adversaries.

# 9. SUMMARY

In this paper we present a novel method for an unobtrusive biometric measurement that can support user identification in small, wearable pervasive devices. We evaluate the feasibility of vocal resonance as a biometric using data collected from 25 subjects. In addition, we implemented a wearable prototype and tested it in stationary and mobile settings in both quiet and noisy environments. Our results show that it is possible to achieve speaker identification through a wearable, body-contact microphone, that can reliably distinguish among multiple individuals sharing a household, and indeed that it can distinguish between the situation where the microphone is on the body of the identified speaker and where the microphone is simply nearby, even on another body. Our prototype, based on a Gumstix processor 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. A purpose-built device – with hardware support for FFT and with extraneous hardware removed – would be substantially smaller and more efficient. In future work we anticipate refining the method, optimizing its performance, and testing it in realistic settings.

## 10. REFERENCES

[1] S. Avancha, A. Baxi, and D. Kotz. Privacy in mobile technology for personal healthcare. *ACM Computing Surveys*, 45(1), Nov. 2012. DOI 10.1145/2379776.2379779.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 2004(4):430–451, 2004. DOI 10.1155/S1110865704310024.

[3] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. *Guide to Biometrics*. Springer Professional Computing, 2004.

[4] C. Cornelius and D. Kotz. Recognizing whether sensors are on the same body. In *Proceedings of the 9th International Conference on Pervasive Computing (Pervasive)*, volume 6696 of *LNCS*, pages 332–349. Springer, June 2011. DOI 10.1007/978-3-642-21726-5_21.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977. DOI 10.2307/2984875.

[6] G. Fairbanks. *Voice and Articulation Drillbook*. Harper & Row, 2nd edition, 1960.

[7] FFTW. Online at http://fftw.org/, visited Dec. 2011.

[8] Fitbit. Online at http://www.fitbit.com/, visited Apr. 2012.

[9] Gumstix. Online at http://www.gumstix.com, visited Dec. 2011.

[10] P. Hasler and D. V. Anderson. Cooperative analog-digital signal processing. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–3972–IV–3975. IEEE, May 2002. DOI 10.1109/icassp.2002.5745527.

[11] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, July 2002. DOI 10.1109/tpami.2002.1017616.

[12] H. Lu, A. J. B. Brush, B. Priyantha, A. K. Karlson, and J. Liu. SpeakerSense: Energy efficient unobtrusive speaker identification on mobile phones. In *Proceedings of International Conference on Pervasive Computing*, volume 6696 of *Lecture Notes in Computer Science*, pages 188–205. Springer, 2011. DOI 10.1007/978-3-642-21726-5.

[13] P. Marquardt, A. Verma, H. Carter, and P. Traynor. (sp)iPhone: Decoding Vibrations From Nearby Keyboards Using Mobile Phone Accelerometers. In *Proceedings of 18th ACM Conference on Computer and Communications Security (CCS)*, pages 551–562. ACM, Oct. 2011. DOI 10.1145/2046707.2046771.

[14] P. Mermelstein. Distance Measures for Speech Recognition–Psychological and Instrumental. In *Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence*, pages 374–388. Academic, June 1976. Online at http://www.haskins.yale.edu/sr/sr047/SR047_07.pdf.

[15] E. Miluzzo, C. T. Cornelius, A. Ramaswamy, T. Choudhury, Z. Liu, and A. T. Campbell. Darwin phones: the evolution of sensing and inference on mobile phones. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys)*, pages 5–20. ACM, 2010. DOI 10.1145/1814433.1814437.

[16] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley Publishing Company, 1987.

[17] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, August 1995. DOI 10.1016/0167-6393(95)00009-D.

[18] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 963–966, September 1997. DOI 10.1006/dspr.1999.0361.

[19] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000. DOI 10.1006/dspr.1999.0361.

[20] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8(3):185–190, Jan. 1937. DOI 10.1121/1.1915893.

[21] S. Tsuge, D. Koizumi, M. Fukumi, and S. Kuroiwa. 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 (ISCA)*, 2009. DOI 10.1109/ISPACS.2009.5383806.

[22] B. Yegnanarayana, A. Shahina, and M. R. Kesheorey. Throat Microphone Signal for Speaker Recognition. In *Proceedings of 8th International Conference on Spoken Language Processing (INTERSPEECH)*, pages 2341–2344. ISCA, Oct. 2004. Online at http://www.isca-speech.org/archive/interspeech_2004/i04_2341.html.