6-2017

BreathPrint: Breathing acoustics-based user authentication

Jagmohan CHAUHAN
University of New South Wales

Yining HU
University of New South Wales

Suranga SEREVIRATNE
University of New South Wales

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Aruna SEREVIRATNE
University of New South Wales

See next page for additional authors

BreathPrint: Breathing Acoustics-based User Authentication

Jagmohan Chauhan [†⋆], Yining Hu [†⋆], Suranga Seneviratne [†⊥]
Archan Misra [‡], Aruna Seneviratne [†⋆], Youngki Lee [‡]

[†]Data61, CSIRO; [⋆]School of EET, UNSW, Australia; [⊥] School of CSE, UNSW, Australia;
[‡] School of Information Systems, SMU, Singapore

ABSTRACT

We propose *BreathPrint*, a new behavioural biometric signature based on audio features derived from an individual's commonplace breathing gestures. Specifically, *BreathPrint* uses the audio signatures associated with the three individual gestures: sniff, normal, and deep breathing, which are sufficiently different across individuals. Using these three breathing gestures, we develop the processing pipeline that identifies users via the microphone sensor on smartphones and wearable devices. In *BreathPrint*, a user performs breathing gestures while holding the device very close to their nose. Using off-the-shelf hardware, we experimentally evaluate the *BreathPrint* prototype with 10 users, observed over seven days. We show that users can be authenticated reliably with an accuracy of over 94% for all the three breathing gestures in intra-sessions and deep breathing gesture provides the best overall balance between true positives (successful authentication) and false positives (resiliency to directed impersonation and replay attacks). Moreover, we show that this breathing sound based biometric is also robust to some typical changes in both physiological and environmental context, and that it can be applied on multiple smartphone platforms. Early results suggest that breathing based biometrics show promise as either to be used as a secondary authentication modality in a multimodal biometric authentication system or as a user disambiguation technique for some daily lifestyle scenarios.

Keywords

Security, Usability, Authentication, Breathing Gestures

1. INTRODUCTION

Robust user authentication on small form-factor personal devices, such as smartphones and wearables, remains an important problem, especially as such devices increasingly become stores of sensitive personal data, such as daily digital payment traces, health/wellness records, or contact e-mails. Ensuring that the authentication mechanisms remain secure and user-friendly will be even more important with the upcoming surge in IoT-rich environments, where everyday objects such as coffee makers, cabinet doors, and pantries are embedded with sensing and actuation capabilities.

User authentication approaches are broadly organized according to three principles: (i) "what you know" approaches, such as PINs, passwords or memorable questions that leverage secret information that only you should be privy to, and have well known usability challenges (ii) "what you have" approaches, such as the use of smart tokens, that rely on hardware-generated temporary credentials that only the possessor of the device has access to, and are susceptible to device theft, and (iii) "what you are" approaches that utilize personal traits (biometrics) of an individual that are hard to reproduce or mimic. Within the "what you are" category, there are two classes of methods: *physiological biometrics* (such as fingerprints or iris scans) and *behavioural biometrics* (such as gait or heart rate measurements [75]).

In this paper, we introduce a new modality of behavioural biometrics based authentication called *BreathPrint*: it uses the sounds generated by a user's *breathing behavior*, as captured by a microphone sensor in close proximity to the user's nose. If the biometrics of "breathing sounds" are indeed *unique* (i.e., it can distinguish among multiple users) and *persistent* (i.e., an individual's signature remains relatively unchanged over time), then *BreathPrint* can provide a compelling mode of authentication. Among other advantages, (i) it is universal: all living users need to breathe!, whereas touch or gait-based biometrics pose challenges for people with hand or foot disabilities; (ii) it is highly usable: breathing is a naturally-occurring reflex activity that requires no additional cognitive or physical effort; (iii) it has low observability: the breathing sounds are essentially inaudible even to an observer that is 10–20 cms away, and (iv) it avoids specialized hardware: unlike past work which needed chest straps [42] or flow sensors [64] to monitor breathing, a commodity microphone sensor is almost universally available on mobile devices.

Key Research Questions and Contributions: Using a set of real world studies (10 users observed over a period of seven days), we first demonstrate that breathing gestures can indeed provide an individual specific fingerprint. We then develop the analytics pipeline on a microphone sensor based system (8 kHz sampling, placed 1–2 cm below the nose) that can extract and use such a fingerprint for practical authentication.

Our key challenges and contributions are:

- **What Natural Breathing Gestures Constitute a Fingerprint?** We develop an authentication approach that uses three distinct, natural breathing gestures: *sniff*, *normal* breathing and *deep* breathing. We show that users have distinctive differences in the amplitude and frequency components of the sound signals for these three breathing gestures, and that each individual breathing gesture achieves over 94% accuracy for user authentication (among 10 users).
- **What is the Processing Pipeline for Breathing-based Authentication?** We develop a classification pipeline that uses Gammatone Frequency Cepstral Coefficients (GFCC) as features as part of a GMM (Gaussian Mixture Model) based classifier. We describe how *BreathPrint* uses a segmentation technique to isolate individual breathing gestures, and how the selection of an appropriate number of GMM components helps the classifier achieve both high authentication accuracy and low false positive rates.
- **Is the Authentication Accurate & Resilient?** We show that *BreathPrint* is able to achieve between 90%–100% true positive rates (*TPR*) for different breathing gestures, especially by changing the number of GMM components. More importantly, we show that this approach is resilient to both impersonation (where an attacker hears the breathing audio records and tries to manually mimic the user) and replay attacks (where an attacker eavesdrops on the sounds from a short distance). The false positive rates (*FPR*) due to *impersonation* and *audio replay* attacks are less than 2%, and 0% respectively. This is lower than the FPR of comparable audio based authentication systems.
- **Are Breathing Biometrics Persistent?** Using studies that trained the classifier models on 50 samples of breathing data, we show that, when users attempt to authenticate 2–3 days later, the classification accuracy degrades somewhat for deep (8%) and sniff (6%) breathing gestures, with the drop being noticeably higher (23%) for normal breaths. To counteract these medium time-scale changes, we develop a simple online unsupervised learning technique to *continually* update the GMM-based classifier. This approach partially solves the persistency challenge – the breathing gesture classification accuracy increases by 3%. Note that the TPR for deep and sniff breathing gestures between the initial sessions and the session done after 2–3 days did not drop using the evolutionary technique, when the number of GMM components is small. *BreathPrint's* robustness to longer-term variation remains an open question.
- **Is the Authentication Robust to Contextual Changes?** *BreathPrint* appears to be robust to changes in the user's breathing pattern caused by selected natural daily activities: TPR remains high (over 80%) even after the episodes of walking for all the breathing gestures except the normal breathing gesture. Similarly, the microphone-based sensor is robust to background noise. TPR drops by only 4% when the noise level measured by a smartphone is 50 dB.

Overall, we believe that our work is the first to demonstrate the promise of a robust and non-intrusive breathing-based authentication mechanism. Besides enabling explicit user authentication (e.g., where a user breathes into a smartwatch or smartphone), our work opens up the possibility of continuous user authentication, perhaps by incorporating other breathing related features, e.g., the heatmap of breaths captured by an infra-red sensor embedded in a smartglass.

2. BACKGROUND

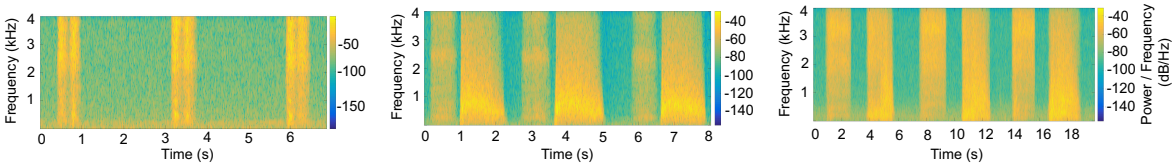
We start our exploration by providing a short background on the physiology of breathing, how it varies across different demographic segments and why we believe that breathing can act as a distinguishing behavioural biometric. Lastly, we describe three breathing gestures: *sniff*, *normal*, and *deep*.

2.1 Breathing Mechanics

Breathing occurs naturally, and a healthy adult takes around 12–20 breaths per minute. Normal breathing has two phases: inhalation (taking air in) and exhalation (pushing air out), and involves three major parts of the body: trachea, lung, and diaphragm. Air enters and leaves the body through the nose and reaches the lungs through the trachea. A large dome-shaped muscle under the lungs, called the diaphragm, aids the process. During inhalation, the diaphragm contracts, creating a vacuum that lowers the pressure in the lungs; consequently, air is sucked into the lungs. The reverse happens during exhalation: the diaphragm relaxes upwards, compressing the lungs and increasing the air pressure there, forcing air out through the trachea. Normally breathing occurs through the nose. We exclude the scenarios where breathing occurs from the mouth - e.g., when people are suffering from some medical condition.

Extensive medical and physiological research exists on respiratory sounds [45] and the effect of demographic attributes, such as age [46], weight [60], height [55], and gender [25] on respiration. These papers studied sounds coming from the lung and trachea, as measured either over the chest wall, at the suprasternal notch (hollow between the neck and the collarbone) or the lateral neck. Lung sound's amplitude differs between persons and different locations on the chest surface, and primarily varies with airflow in the lungs [23]. Moreover, as shown by Pasterkamp et al. [47], the spectral shape of tracheal sounds is highly variable across subjects.

Pasterkamp et al. [46] studied the properties of normal respiratory lung sounds and how they differ between children and adults. Authors showed that infants had higher median frequencies present, compared to older children and adults, and concluded that this was due to the presence of less power in low-frequency regions (small lungs and thin chest walls) among infants. However, the roll-off in the power of higher frequency components was similar at all ages. Sanchez et al. [55] found that tracheal sounds of children have higher cut-off frequency (the frequency after which the audio signal's power drops sharply) than adults due to shorter tracheal length. Sharp et al. [60] reported that the pressure, volume and air flow is very different between a normal and an obese person. The difference has mainly been attributed to the different efforts needed to move the ponderous thoracic wall (lungs, diaphragm) and abdomen during breathing. Gross et al. [25] found that the ratio of power between two frequency bands (60–330 Hz and 330–600 Hz) is very distinct between males and females for lung sounds.



(a) Sniff (b) Normal breathing (c) Deep breathing

Figure 1: Spectrograms for the three breathing gestures

The role of the diaphragm is studied by Mead et al. [43], where they demonstrated that the volume change in lungs during breathing is influenced more by the diaphragm than the rib cage. More recently, breathing-related parameters have been used to detect different medical conditions [22, 6, 28, 61]. Sinues et al. [61] used mass spectrometry to analyze the composition of the exhaled breath. They found that individual signatures of breath composition exist, implying that analysis of breathing can be used for metabolic studies. We believe that we are the first to utilize the audio signal of breathing to authenticate an individual.

2.2 Breathing Gestures

In our proposed approach, we focus on audio generated by three distinct breathing gestures: *Sniff*, *Normal*, and *Deep*. We have deliberately restricted ourselves to these three gestures, because they occur "naturally" and can thus be easily performed by a healthy person. In fact, we initially also considered a few additional more complex breathing gestures, such as "two quick inhales followed by an exhale" and vice versa, and "one quick exhale followed by two quick inhales" and vice versa. However, during initial exploratory studies, our participants indicated that they found such breathing gestures too cumbersome to perform and too difficult to remember. Such complex breathing gestures were thus excluded from deeper analysis.

Figure 1 shows the spectrograms for three breathing gestures: Sniff, Normal, and Deep. Each spectrogram contains three samples of each breathing gesture. The figure is obtained using data collected during our experiments with 10 participants (see Section 3). Sniff breathing gesture has the shortest duration of less than a second and consists of two successive quick inhalations. Normal breathing gesture lasts around 2–3 seconds and comprises an inhalation, followed by an exhalation. Deep breathing gesture has longer inhalation and exhalation phases than normal breathing gesture, and lasts around 4–5 seconds.

Motivated by the aforementioned research findings over respiratory sounds, we hypothesized that acoustic features derived from different breathing patterns, measured by a microphone sensor, should provide an effective approach for authenticating people. However, we needed to discover the key features (e.g., amount of air, forcefulness, time duration of inhalation and exhalation, etc.) that can both be detected through audio signal analysis and provide high discriminatory power across individuals.

3. DATASET

In this section, we describe the data collection process, including the experimental setup and user demographics. We also present results to show what makes user's breathing behaviour different across the three breathing gestures.

3.1 Data Collection

We recruited 10 volunteers (two females and eight males, aged between 20 and 35) to perform three types of breathing gestures: *sniff (two quick consecutive inhalations)*, *normal breathing*, and *deep breathing*. The participants were employees and postgraduate students affiliated with Data61, CSIRO. The user study was approved by the CSIRO Human Research Ethics Committee under Ref. No. 085/16 and was performed in a closed office room of size 1.7m (w) x 3.8m (l) x 2.7m (h). The hardware used for the data collection was a wired iPhone 6 earphone attached to a Dell XPS 13-9343 laptop running Windows 10 (see Figure 2a). We used the free version of Gold Wave software[1] to capture the acoustic signal generated by different breathing gestures.

Each user held the microphone of the earphone 1–2 *cm* below their nose and repeated each breathing gesture multiple times, while remaining seated. We used an 8000 Hz sampling rate to capture the generated acoustic signals and store them as *.wav* files. We repeated the experiment on three days; the first day (Session 1), 3–4 days after the first day (Session 2), and 6–7 days after the first day (Session 3). In the first two experiment sessions, the participants repeated 30 instances of each breathing gesture. During the third session, participants repeated 10 instances of each breathing gesture. They also performed impersonation attacks by trying to mimic a target user breathing gestures after listening to an audio recording. Additionally, to perform replay attacks, we also recorded the user's breathing sounds via a smartphone placed close to the user.

The acoustics of the breathing signal depends on various physical conditions. Also, our data acquisition process may be affected by ambient noise. To understand the effect of such factors, we carried out additional "contextual" experiments with three users (two of whom are part of the original pool of 10 participants). For these users, we captured the acoustic signal for the three breathing gestures in multiple scenarios: *i) normal experiment conditions (while sitting without background noise), ii) while standing, iii) after 2-3 min of walking, iv) after 2–3 min of going up and down the staircases, v) while sitting under three levels of ambient noise, and vi) while keeping the microphone at different distances from the user's nose.* The contextual experiments are only carried out once (one session) for each user. More details about the attacks performed and the different contextual experiments are presented later in Section 5.

[1]https://www.goldwave.com

We also evaluated the proposed biometrics on two smartphones: iPhone 6 and Nexus 6P, by asking the same three users to perform breathing gestures on the two devices. We used the smartphone's microphone to record the breathing gestures. We did not use the headphone mic, as it has a built-in AGC (Automatic Gain Control) function that we could not disable. Experimental setups: (a) of a laptop with an earphone and (b) smartphones only are shown in Figure 2.

(a) Laptop - Earphone (b) Smartphone

Figure 2: Experimental Setups

3.2 Breathing Gesture Characteristics

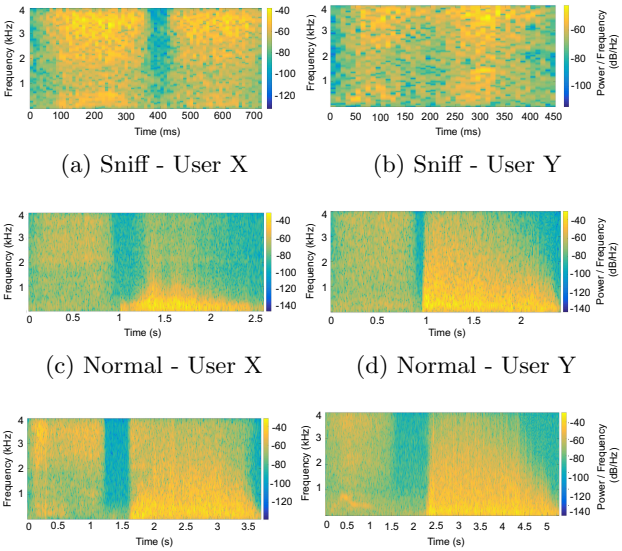
Here we discuss the characteristics of different breathing gestures and why they are different across individuals. Table 1 shows the statistics about the average, minimum, and maximum length of different breathing gestures obtained from the dataset. Sniff breathing gesture is the shortest in length with an average duration of less than a second. Normal breathing gesture length varies between 1.7–4.3 seconds, while deep breathing gesture is the longest gesture with an average duration of approximately 5 seconds.

Table 1: Breathing Gesture Lengths (Seconds)

Gesture Type	Average	Minimum	Maximum
Snifff	0.55	0.42	0.71
Normal	2.83	1.79	4.28
Deep	4.8	3.74	6.53

Figure 3 shows the spectrograms of the same breathing gesture from two users. We can see that the spectrograms are different across the two users. The energy for a sniff breathing gesture lies mostly in the higher frequency range; conversely, for normal breathing gesture, most of the energy lies in the low frequency range. Also, more power exists in the exhalation phase than the inhalation phase. The same is true for deep breathing gesture, except that the energy distribution across the inhalation and exhalation phases is more evenly distributed.

To get a clearer picture of why different people might have different breathing signatures, we plot the root mean square of the amplitude of the breathing signal and the power present in different frequency bands for all the three breathing gestures for all the participants. The frequency bands are chosen according to the Gammatone filterbank frequency ranges, described later in Section 4. The results are shown in Figure 4. The Root Mean Square (RMS) of the breathing signal for all the users is distinctive. Also, the frequencies, which separate the users in case of a sniff breathing gesture, lie in the high frequency range (2000–4000 Hz). On the other hand, the separating frequency range for normal and deep breathing gestures lies in lower frequency ranges, specifically normal breathing gesture between 100–1300 Hz and deep breathing gesture between 100–1600 Hz.



(a) Sniff - User X (b) Sniff - User Y

(c) Normal - User X (d) Normal - User Y

(e) Deep - User X (f) Deep - User Y

Figure 3: Spectrograms for Two Users

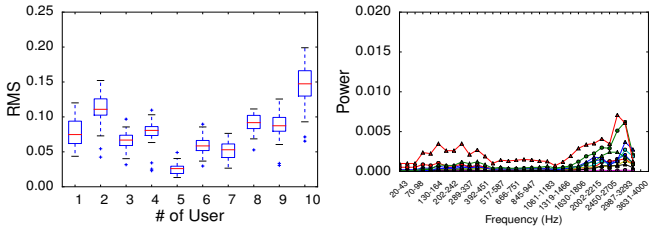
Inter-gesture correlation: We also investigated the correlation, in the average RMS value of the breathing signal across different users, for each breathing gesture pair. We found that this correlation was negligible (in the range of [-0.27, 0.17]) between sniff and normal breathing gesture or between sniff and deep breathing gesture. The correlation was highly significant (0.75) between normal and deep breathing gestures. These results on correlation, along with the robustness studies performed in Section 5.4, suggest that "sniff" and "deep" are the two distinct, independent breathing gestures that provide *BreathPrint* the highest discriminatory power.

4. THE BREATHPRINT PIPELINE

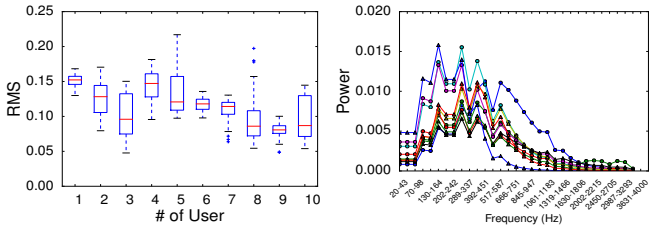
We first describe the overall design of *BreathPrint*, showing its different components. Next, we explain the feature extraction process we used to represent the acoustic signals in a form that can be fed into an appropriate classifier, which determines whether the presented acoustic signal was from a specific individual or not. The classifier input is a discrete breathing gesture. Finally, we present a segmentation algorithm that automatically extracts episodes of individual breathing gestures from an underlying continuous sequence of audio sensor data.

4.1 Overall System

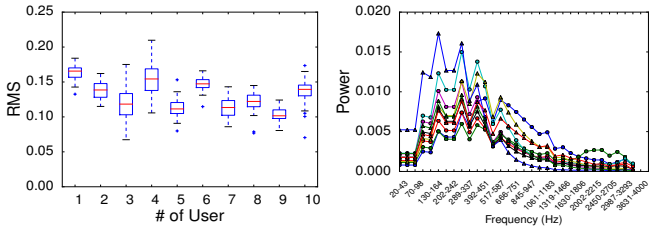
In this subsection, we describe the general working mechanism of *BreathPrint*. Figure 5 provides a schematic overview of the system setup used to generate our empirical results. The user breathes into a microphone, mounted near his/her nose. The acoustic signal from the microphone first goes through the segmentation algorithm that separates out each breathing gesture and also removes leading and trailing silence periods. After this step, the audio signal of the breath-



(a) Sniff - RMS (b) Sniff - Power



(c) Normal - RMS (d) Normal - Power



(e) Deep - RMS (f) Deep - Power

Figure 4: RMS and Power-Frequency Distribution

ing gesture is divided into non-overlapping frames of 10 ms. Acoustic feature vectors (described shortly in Section 4.2) are extracted from each frame. The feature vectors are then fed into both GMM_{user} model and GMM_{others} model to obtain two different log likelihood values: LL_{user} and LL_{other}. The resulting log likelihood ratio ($\frac{LL_{user}}{LL_{other}}$) is derived for each frame. Subsequently, we compute the average of this ratio across all the frames constituting the gesture, and accept (authenticate) the user only if this average ratio is greater than one. In case the value of average ratio is less than one, the user is not allowed to access the system.

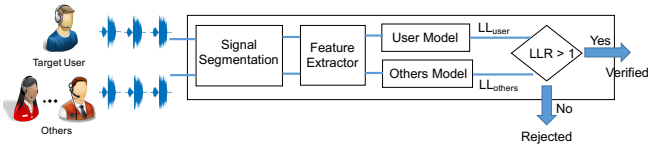


Figure 5: *BreathPrint*: System Architecture

4.2 Feature Extraction

We utilize the Gammatone Frequency Cepstral Coefficients (GFCC) as features, which are based on signal power in a set of frequency bands. While Mel Frequency Cepstral Coefficients (MFCC) is more widely used as features in audio signal processing, especially for human speech analysis, recent research [56, 66, 31] has shown that Gammatone filterbanks [2] can offer higher accuracy than Mel filterbanks. The major reason why GFCC is more accurate than MFCC is because GFCC provides finer resolution at low frequencies than MFCC, making GFCC not only more robust in the presence of noise [77], but also better suited for detecting low frequency sounds. Given that a single breath is barely audible at larger distances, we anticipate that breathing has its dominant energy at lower frequencies.

Generally, a filterbank divides the frequency spectrum of the signal into multiple overlapping bands (adjustable) called filters and then calculates the log-energy (weighted sum of the FFT magnitudes) for each filter. As, the final step in calculating the cepstral coefficients, the Discrete Cosine Transformation (DCT) is applied to each of the log-energy. This yields number of cepstral coefficients equal to the number of filters. DCT is a standard orthogonal transformation technique; the most important information about the spectrum is typically embedded in the lower order DCT coefficients. Note that the DCT coefficients capture the energy variation across the entire spectrum; for example, the first DCT coefficient is the sum of all the log-energies. For our purpose, we used the GFCC implementation by Zhao et al. [76]. We extracted 32 GFCC as features using a 32-Gammatone filterbank, in which each filter is defined by a center frequency, f and a bandwidth, b. We also used GFCC *delta* and GFCC *double delta* as features, which capture the rate of change between two consecutive GFCC and GFCC delta coefficients respectively. In total, we used 96 features. For GFCC, the filterbank centre frequencies are distributed across frequency in proportion to their bandwidth, known as the ERB (Equivalent Rectangular Bandwidth) scale [24]. ERB scale provides an approximation as how the bandwidths of the filters should be divided. The center frequencies we used are in the range of 20–4000 Hz. The bandwidth B of each Gammatone filter is given by 1.019 ERB, where ERB is calculated as: $24.7*(4.37*0.001*f+1)$, where f being the filter's center frequency. For comparison purposes, we also implemented our classifier using MFCC. Similar to GFCC, we used MFCC, its delta and double delta as features.

4.3 Classifier

We used GMM (Gaussian Mixtures Model) as a classifier as it has been shown to perform better than DTW (Dynamic Time Warping) or supervised learning methods such as SVM for speech recognition classification tasks [33, 51]. GMM is a probabilistic model, which assumes that all the data points in a dataset are generated from a mixture of a finite number of Gaussian distributions. Each Gaussian distribution has its own mean and covariance and is represented as a component in the GMM based model.

By adjusting the number of components and the type of covariance matrix, one can attempt to represent a set of data points. We tried with 5, 10, 15, 20, 25, and 30 as the number of components in the classifier. We stopped at 30, as the classifier's accuracy drops below 50% beyond 30 components (as we shall see in Section 5). Moreover, larger volumes of training data are also needed to train a model with a higher number of GMM components. For example, speech recognition systems typically include 256–1024 GMM components, as they usually utilize hours of speech data [51]. In our case, we seek to train classifiers with only limited training data (at most a few minutes of human effort). Given

such limited training data, the use of a greater number of GMM components is likely to lead to overfitting.

The other parameter of the GMM, called covariance matrix type, can be chosen from four options: *full, tied, diagonal,* and *spherical. Full* covariance matrix means that each component has its own general covariance matrix, whereas *diagonal* implies that each component has its own diagonal covariance matrix. In this work, we used the *diagonal* option, as estimating the parameters of a full covariance GMM requires more training data and is more computationally expensive. Moreover, past empirical evidence suggests that diagonal matrix GMMs outperform full matrix GMMs [52].

To train the classifier, we first created two GMM models for each user. The first model GMM_{user} is constructed only using the breathing audio samples from the target user. The second model GMM_{others} is constructed using the samples from all other users (i.e. excluding the target user samples). Each audio sample was divided into $10\,ms$ frames with no overlap. 10–30 ms is the preferred choice of frames in audio processing tasks [70]. For each frame, a feature vector of GFCC is extracted and fed to the training module, which uses the well known Expectation Maximization (EM) technique to compute the GMM components. During the test phase, the incoming audio sample is divided into 10 ms frames, features are extracted and a log likelihood ratio, $LLR = LL_{user}/LL_{other}$ is calculated by classifying the incoming feature vectors with both GMM_{user} and GMM_{others} models.

4.4 Segmentation Algorithm

To automate the feature extraction process, we designed a waveform segmentation algorithm. While inspired by existing work, our segmentation algorithm is specially tuned to isolate the (start, stop) duration of each individual breath. In contrast, the majority of the respiration monitoring works focus only on breathing rate estimation [14, 7, 71] - e.g., reliably isolating the peak of each breath. In addition, segmentation in speech recognition is threshold based and is used to separate speech vs. non-speech frames [32, 38] – such approaches do not directly work in our case as the thresholds need to be different for different frames, over the full breathing gesture signal. When a raw recording is given to the algorithm, it calculates the breathing rate (number of breathing cycles per unit time), a suitable window size (for each breathing gesture), and the starting points of each window. The steps of segmentation algorithm are illustrated in Figure 6.

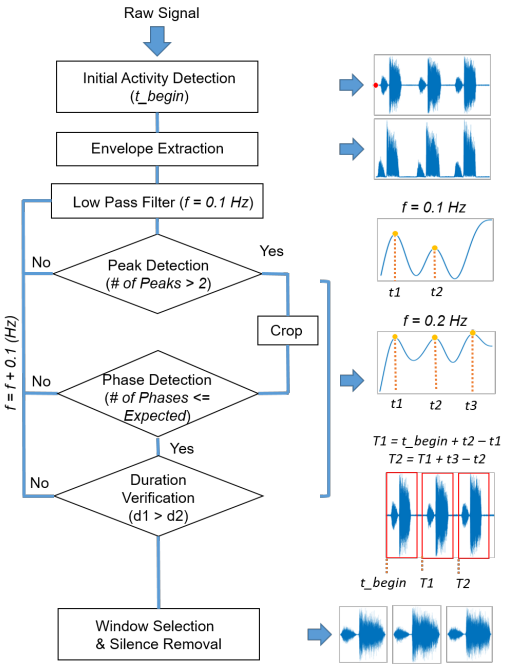


Figure 6: Segmentation Algorithm

i) Initial activity detection - To identify the start of the activity (i.e. where a user starts performing repeated breathing gestures), we use a simple threshold method. The raw signal is separated into frames of $10\,ms$. If five consecutive frames (i.e., $50\,ms$) each have signal power higher than an empirically predetermined threshold, the starting point of the first frame t_{begin} is considered as the starting point.

ii) Envelope extraction - We then extract the envelope of the acoustic signal by applying Hilbert transformation. For acoustic signals, applying Hilbert transformation smooths the signal and eliminates negative values.

iii) Low pass filter - The signal envelope is then passed through a low-pass Butterworth filter as suggested in [14] to get the fundamental frequency of the signal; i.e. the breathing rate. The starting frequency (f) is set to be $0.1Hz$ (as the corresponding window size of $10s$ is sufficient for any breathing gesture) and is iteratively incremented by $0.1Hz$ until three criterion are met: (a) *Number of peaks*, (b) *Number of phases*, and (c) *Duration*.

a) Peak detection - When f is close enough to the fundamental frequency, peaks on the filtered envelope are uniformly distributed; otherwise, the filtered signal is irregular in shape without clear peaks. Thus, when the *number of peaks* criterion is met, the algorithm determines that it has correctly estimated the breathing rate. Once this criterion is met, the window length is calculated as the average distance between adjacent peaks and the starting points of the windows are estimated by adding t_{begin} to the distance between the target peak to the first peak. The raw signal is then cropped into distinct windows using the average window length and window locations.

b) Phase detection - A segment detection method is used to check the *Number of phases*. A phase in a breathing cycle corresponds to either inhalation or exhalation. Therefore, the number of expected phases for normal breathing, deep breathing, and sniffing breathing gestures are two, two, and one respectively. Note that since sniffing breathing gesture is two quick inhalations it does not have two phases, as there is no significant time gap between the two inhalations. The algorithm compares the number of phases inside each window and if most windows contain more phases than expected, the window size is shortened by increasing f (which indirectly decreases the distance between adjacent peaks).

c) Duration verification - For some users, we observe that the inhalation phase for normal or deep breathing gesture is weak and hence buried in the noise. For such users, for breathing gestures with two expected phases, two exhalations belonging to two adjacent but separate breathing cycles may be cropped into one window. To avoid such fusion of multiple breathing gestures, the algorithm performs

an additional *duration verification* step. The *duration verification* step compares the duration $d1$ of each detected phase and the distance $d2$ between the two phases inside a window. If $d1 < d2$, then it implies that the two breathing cycles are incorrectly grouped together and the window size is shortened by increasing f.

iv) Window selection and silence removal - To ensure that there is only one complete breathing gesture inside each window, the algorithm selects windows with silence periods at both the beginning and at the end. This is based on the assumption that the user brings the microphone close to his/her nose only during the authentication step and otherwise usually the microphone is sufficiently away from the nose at other times. A frame level, threshold based method is used to eliminate windows with consecutive high power frames on the sides and the silence periods (that occur due to the pauses between the breathing gestures) of the selected windows are later removed.

Segmentation Performance: To study the performance of the segmentation algorithm, we fed it five waveforms containing a total of 100 individual breathing gesture instances (counted manually). For sniff, normal breathing, and deep breathing gestures the segmentation algorithm was able to correctly segment 95%, 93%, and 84% respectively, of the manually annotated individual gesture instances. A segment generated by the algorithm was considered correct if it contains a one complete (individual) breathing gesture when manually inspected. The segmentation process applied to continuous breathing activity does not relate to the accuracy of authentication of individual breathing gestures. Even though our segmentation method cannot capture all breathing cycles, a user can simply repeat the breathing gesture until a correct segment is extracted.

5. RESULTS

As discussed in Section 3.1, we collected the data from 10 volunteers for the three breathing gestures in three sessions: first day (Session 1), 3–4 days after the first day (Session 2), and 6–7 days after the first day (Session 3). Each user held the microphone of the earphone 1–2 *cm* below their nose. We collected 30 instances of each breathing gesture in the first two sessions and 10 instances in the last session. In the last session, the volunteers were also asked to do impersonation attacks for a target user 10 times for each breathing gesture. Three volunteers were picked to perform data collection (out of which two were from the initial pool of 10 participants) for an additional set of "contextual" experiments, and to test robustness of BreathPrint on two additional smartphones.

5.1 Baseline: Intra-Session User Verification

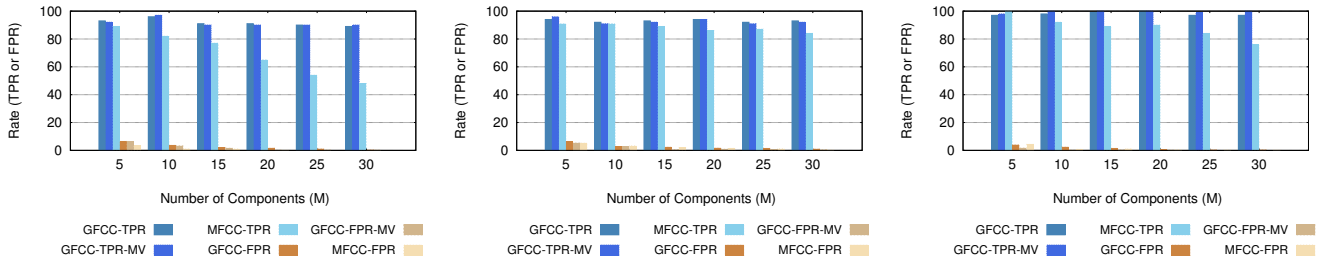
We first used all the 30 samples from Session 1 and first 20 samples from Session 2 to train the classifier for each user and each breathing gesture. To create the "others" model, 50 samples from each other user were taken (total, 9*50=450 samples per breathing gesture). The rest of the 10 samples from Session 2 were used as the testing set (total, 10*10=100 samples per breathing gesture). We report the values of performances metrics – *TPR* and *FPR* – achieved by the *BreathPrint* classifier. Note that TPR in this context is the percentage of instances where a user is correctly verified (i.e. calculated based on 100 predictions where each user's 10 samples are fed into their corresponding GMM). FPR is the percentage of instances where a classifier misclassifies other user's samples as the legitimate (actual) user's sample (i.e. calculated based on 900 predictions where all other users' samples are fed into one user's GMM). In the context of user verification, by True Positive we mean how many times a genuine user was correctly verified to be the claimed user on the basis of some threshold to get legitimate access to the system. While, False Positive means how many times an illegitimate user gets access to the genuine user's account because the system fails to detect the fraudulent access. We compare the performance of *BreathPrint's* classifier, built with GFCC, against an alternative built using the MFCC.

Figure 7 shows the results, as the number of GMM components is varied between 5 to 30, in increments of 5. For all the breathing gestures, TPR varies between 90% and 99% when GFCC are used. MFCC perform worse than GFCC as the TPR drops to 50%. For GFCC, FPR is between 0.4% and 6.4% and for MFCC, it is between 0% and 5.2%. Although, MFCC provides a slightly better FPR than GFCC, it comes at the cost of dropped TPR. We attribute this notable performance difference to the fact that GFCC provides finer resolution at low frequencies where most of the energy of the breathing signal is located. The performance of the classifier is promising and comparable to the other forms of behavioural biometrics as discussed in Section 7.

Figure 7 also shows that number of GMM components (M) plays a significant role in classification accuracy. TPR and FPR both decrease with the number of GMM components. TPR drops as M increases; 2%–7% for GFCC and 7%–40% for MFCC. This might be due to overfitting. As the number of components increases, the ability of the classifier to correctly classify the actual user diminishes. However, the affect is more drastic for MFCC than GFCC. Overall, the experimental results suggests that GFCC provide a better balance, compared to MFCC, between the usability and security of the system.

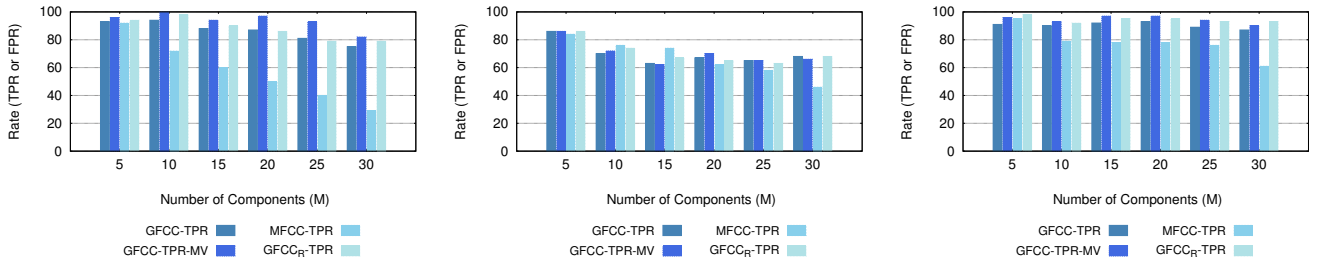
Until now, we have presented results to verify a user based on the prediction of a single test sample. We also increase the number of testing samples at the time of user verification from one to three and used a majority voting based decision to check how it affects the accuracy of the system, in terms of TPR and FPR for GFCC based classifier only as GFCC provided better results than MFCC. We pick three random test samples out of 10 at each iteration and successfully authenticate a user if the classifier verifies at least two of the three test samples correctly. We run 10 such iterations for each user. Referring back to Figure 7, we can see that majority voting (GFCC-TPR-MV) increases the TPR for deep breathing by 1.5% (average) across all the components while for the other two breathing gestures it has a minimal impact. For FPR (GFCC-FPR-MV), we a see a drop of 0.6%, 0.9% and 1.2% on average against GFCC-FPR across all components for sniff, normal and deep breathing gestures.

The running time for building a classifier for all the 10 users using GFCC varied between 32-47 seconds, 110-275 seconds, and 165–435 seconds for sniff, normal, and deep breathing gestures respectively across different GMM components we tested. The time taken for testing a sample was 2 seconds for any of the three breathing gestures irrespective of the number of GMM components. The time was calculated on a laptop running Mac OS X (Sierra version) with 2.6 GHz Intel Core i5 processor and 8 GB of RAM.



(a) Sniff (b) Normal (c) Deep

Figure 7: TPR and FPR for the three breathing gestures



(a) Sniff (b) Normal (c) Deep

Figure 8: TPR and FPR for the three breathing gestures - Evolution

5.2 Evolution: Inter-Session User Verification

We tested the classifier performance (built using Session 1 and Session 2 data) on 10 samples collected on Session 3 (i.e. 6-7 days after Session 1). The results are shown in Figure 8 in the form of TPR and FPR similar to Section 5.1. As Figure 8 shows the TPR varies between 75%–93%, 68%–86%, and 87%–91% for sniff, normal, and deep breathing gestures respectively. Compared to the intra-session TPR, there is an average 8%, 23%, and 6% drop in inter-session TPR for sniff, normal, and deep breathing gestures respectively. This can be ascribed to temporal factors such as user's physiological condition and activity levels. More notably, if the number of components M=5, the TPR for sniffing does not drop while it drops only by 8% and 6% for the normal and deep breathing gestures. Lower TPR for the normal breathing gesture was due to the presence of a single *outlier* user, for whom none of the 10 breathing gestures were correctly authenticated.

To counteract these medium scale temporal changes, we develop a simple online unsupervised learning technique to continually update the GMM-based classifier. We replaced the first 10 samples belonging to Session 1 from the user's model with the 10 testing samples from Session 2. The results are labeled as GFCC$_R$. This approach partially solves the persistency issues as it improves the authentication accuracy by 3%. The TPR for deep and sniff breathing gestures did not drop using the technique, when the number of GMM components is small.

We also performed majority voting for inter-session user verification. As we can see from Figure 8, majority voting increases the TPR for all three breathing gestures: 7% for sniff, 0.3% for normal, and 4% for deep on an average across all the components. Overall, majority voting improves the accuracy of the system by decreasing the FPR while having slight increase in the TPR.

5.3 Security Evaluation: Attacks

We evaluated the security of the breathing based biometrics system under three different types of attacks; i) *random attacks*, ii) *impersonation attacks*, and iii) *audio replay attacks*. In random attacks, an attacker does not know how a target user performs a certain breathing gesture. To simulate random attacks, we used last 10 samples from Session 2 of each user against a target user's classifier and obtained the FPR at different GMM components. This is exactly the process that is used to calculate the FPR in Section 5.1. FPR is in the range of 0.4%–2% when the number of GMM components is between 15–30 indicating that the proposed biometrics is safe against random attacks.

Recall that in the final session (c.f. Section 3.1), all the users were asked to act as an attacker by trying to *impersonate* a target user. The attackers had to listen to the audio recording of the target user breathing gesture and then attempt to mimic the breathing gesture 10 times. The attacker could play the audio recordings any number of times. We collected 10 impersonation attack samples from each user and in total 100 samples for each breathing gesture.

To simulate *audio replay attacks*, we placed a smartphone near the laptop at a distance of 40-50 cm. The smartphone was recording the breathing gestures. In total, 100 samples for each breathing gesture were collected for audio replay attacks. These samples were then replayed on the microphone of the earphone so that the system thinks that an actual user is performing the breathing gesture. One can argue that recording breathing sounds from a smaller distance (4–5 cm) will increase the likelihood of a successful attack against a target user. However, launching attacks from smaller dis-

tances will be easily observable to the target user and hence would be difficult to be launched successfully.

Table 2 shows the results. For impersonation attacks, the number of successful attacks is: 7% for sniff breathing gesture and more than 10% for normal and deep breathing gestures. However, as the number of components increases the successful impersonation attacks drop to less than 4% across all the breathing gestures. The same is true for audio replay attacks. The successful attacks go to 0% across all the breathing gestures when using GFCC. More replay attacks on sniff and deep breathing gestures are successful because sniff and deep breathing gestures contain higher energy than a normal breathing gesture, which makes them audible at a larger distance compared to a normal breathing gesture.

Table 2: Results of Attack Scenarios

		GFCC		MFCC	
Gesture Type	**M***	**Imper.**	**Replay**	**Imper.**	**Replay**
Sniff	5	7	7	7	12
	10	3	0	0	10
	15	1	0	0	7
	20	2	0	0	5
	25	1	0	0	4
	30	0	0	0	0
Normal	5	16	10	19	0
	10	3	1	9	0
	15	3	1	9	0
	20	2	1	8	0
	25	2	1	6	0
	30	2	0	3	0
Deep	5	18	19	26	10
	10	6	0	7	10
	15	9	0	6	10
	20	3	0	4	10
	25	2	0	6	10
	30	2	0	1	10

(M* - Number of GMM components)

To summarize, our results show that highest TPR is achieved when the number of GMM components (M) is 5. However, the system also becomes more prone to impersonation and audio replay attacks at lower values of M. When the number of GMM components is 30, the exact opposite is true. The TPR is too low for a usable system but very high security (low FPR) can be achieved by the system. Nonetheless, *BreathPrint* can achieve a good balance between security and usability by using GFCC and a GMM classifier with 15–20 components. The number of components can be increased at the expense of collecting more training data. Our results suggest that the *deep* breathing gesture provides the best balance between TPR and FPR while preventing targeted attacks. The deep breathing gesture also has the highest TPR rate during evolution. For the rest of this section, we use GFCC for classification purposes as they perform better than MFCC.

5.4 Contextual Results

We check the accuracy of the classifier in following scenarios: (a) having the user perform some physical activity, (b) having the user stand, (c) increasing distance between the user (3–5 cm and 6–8 cm) and the microphone of the earphone, and (d) increasing the level of ambient noise. These results are obtained using data from three users (c.f. Section 3). 30 samples are collected for training from each user for each breathing gesture under normal experiment conditions (sitting without background noise). Then 10 samples are collected for each breathing gesture under eight different scenarios. The physical activity scenarios include: (a) walking for 2-3 minutes and (b) going up and down the staircase for 2-3 minutes. The ambient noise is generated by playing music from a smartphone at three levels. We chose music because playing a recorded audio can be done repeatedly easily for all the participants. The noise levels are achieved by dividing the maximum volume of the smartphone speaker into three equal parts. The average decibel values measured from the user device side are 50 dB, 54 dB, 66 dB for Level 1, Level 2, and Level 3 respectively. 50–55 dB is equivalent to general activities in a business office. 65 dB corresponds to average road traffic heard at a distance of 25 m.

Table 3 shows the results. We can clearly see that the accuracy of the classifier only lies between 40%–60% in most of the cases across breathing gestures for 3–5 cm. For 6–8 cm, the accuracy hovers around 33%. This implies that the microphone needs to be really close (0–2 cm) to the user's nose for *BreathPrint* to perform well. However, on the other hand, this very disadvantage makes breathing gestures extremely safe against audio replay and similar attacks. With respect to ambient noise, the breathing gestures are quite resilient at Level 1 and Level 2 except the normal breathing gestures for which accuracy ranges between 40%–60%. This is due to the fact that more energy is present in sniff and deep breathing gestures, thereby making them highly unsusceptible to the ambient noise. The standing posture also has relatively low impact on the accuracy of the classifier.

For activity scenarios, we find that the *after stairs climbing* results are worse than after walking scenario. Naturally, walking puts less strain on the user body and hence the chances of breathing gestures being changed are less compared to the stairs scenario. We also observe better results for deep and sniff breathing gestures compared to normal breathing gesture, which suggests that normal breathing is affected more compared to the other two gestures after a physical activity is performed by a person. We expected the accuracy of the classifier to go down with physical activity scenarios. We leave increasing the accuracy of our system in different physiological contexts as future work.

5.5 Multi Platform Results

We tested the feasibility of using breathing gestures on two other devices: an iPhone 6 and a Google Nexus 6P Android smartphone. The three users who participated in contextual experimental scenarios performed each breathing gesture 30 times on both the smartphones. The first 20 samples are used for training and the remaining 10 are used for testing. The TPR of the classifier is 100% on both the devices for all the breathing gestures across all the GMM components. However, we used a different threshold to remove silence from the segmented breathing gestures in case of iPhone 6 compared to Android smartphone breathing samples. For samples collected using iPhone, we used 0.1 times the threshold value being used for samples collected using Android phone. The samples collected using Android phone used the same threshold as used for the earphones.

6. RELATED WORK

We divide related work into i) detecting non-speech body sounds and their applications, ii) breath control for interactive gaming, iii) breath monitoring for respiratory disorder detection, and iv) behavioural biometrics for authentication.

Table 3: Contextual Results (TPR)

Gesture Type	M*	Distance		Ambient Noise			User Activity		
		3-5cm	6-8cm	Level 1	Level 2	Level 3	After Stairs	After Walking	Standing Idle
Sniff	5	70	66	100	73	76	73	83	96
	10	50	33	96	56	60	46	83	86
	15	43	33	93	60	46	46	83	90
	20	43	33	96	90	46	43	83	83
	25	36	33	96	83	40	43	83	76
	30	40	33	90	76	40	50	80	73
Normal	5	63	23	100	63	36	63	53	76
	10	63	20	100	43	36	43	33	56
	15	60	30	100	53	36	50	33	66
	20	56	16	100	50	40	46	26	76
	25	56	36	100	43	40	53	36	73
	30	66	20	96	43	40	50	33	76
Deep	5	43	33	100	96	36	66	86	100
	10	40	50	100	96	36	63	83	100
	15	43	33	100	96	33	66	90	100
	20	43	36	100	100	36	66	96	100
	25	43	43	100	100	33	66	93	100
	30	43	43	100	100	33	66	90	100

(M* - Number of GMM components)

i) Non-speech body sounds: Bodybeat [49], SymDetector [62], BodyScope [74], CoughSense [35], iSleep [26], Woosh [50], UbiBreathe [1], Zephyr [4], and ApneaApp [44], are some of the earlier works on detecting various types of body sounds for applications under different contexts. BodyBeat [49] is a mobile sensing platform that captures and recognizes four different types of non-speech body sounds: eating, drinking, breathing, and coughing using a custom-built piezoelectric microphone placed near the user's throat. Such a system can be used for food journaling and detecting illness conditions like coughing. SymDetector [62] detects multiple respiratory sounds such as coughing, sneezing, sniffling, and throat clearing with a smartphone microphone. It handles users' privacy and reduces power consumption to record users' respiratory symptoms for a longer duration of time.

BodyScope [74] aims at detecting sounds produced in a user's throat area, such as eating, drinking, speaking, laughing, and coughing using a special sensor working at 22 kHz. A similar mobile system named CoughSense [35] aims at detecting cough sounds with high accuracy while maintaining the users' privacy using a smartphone worn around the user's neck or placed in a shirt pocket. Woosh [50] relies on air blown from the mouth as a way to interact with smartwatches. UbiBreathe [1] uses WiFi RSS (Received Signal Strength) patterns to enable ubiquitous and non-invasive respiratory rate estimation. ApneaApp [44] monitors the minute movements in chest and abdomen caused by breathing on smartphones to detect sleep apnea. Our research is different compared to the earlier works because we are interested in using different type of breathing sounds for user authentication on commodity hardware (an earphone or a smartphone) using a microphone placed near the user's nose.

iii) Breath control for interactive gaming: Breath can be used to control a wide range of gaming applications [42, 64]. Marshall et al. [42] demonstrated the feasibility and benefits of breath control over thrill rides. Breathing phases including inhaling, exhaling, and holding the breath were detected by a chest strap and directly mapped to the horizontal rotation of a thrill chair. Tennent et al. [64] examined breath control via a gas mask embedded with flow sensors. Five customized games were chosen to investigate breath as a mechanism to control them. Besides the aforementioned breathing phases, breathing frequency was used as an additional control mechanism. Instead our primary focus is to use breathing gestures as a means for user authentication.

iii) Breath monitoring for respiratory disorder detection: Auscultation of body sounds has long been an important part in both sleep and breathing related studies. Many systems were proposed to continuously monitor and detect abnormal breathing activities such as wheezes [63, 29], crackles [40, 58], and snores [5, 73]. Azarbarzin et al. [5] devloped a snoring detection algorithm for breathing sounds captured by two microphones, one over the tracheal and the other one in the air. A vertical box method was applied to detect activity episodes and frame level features were then extracted in each episode. Using an unsupervised fuzzy C-means clustering algorithm, episodes with snoring activities were labeled. A similar segmentation method was also used by Yadollahi et al. [73] to study sleep apnea detection. Jain et al. [29] developed a wheeze detection method using frame-level FFT on 50 ms Hamming windows. A frame was selected if a frequency peak is detected and subsequent frames have frequency peaks within the 30 Hz range. A continuity check was performed to find episodes with more than 250 ms of duration.

In our case, one recording only contains one periodically repetitive pattern and no classification is needed for the initial cropping stage. Therefore instead of using moving windows and comparing signal powers across the whole original recording, our cropping algorithm simply cuts the recording into pieces according to the detected frequency.

iv) Behavioural biometrics for user authentication: Acoustic features from user speech have been used to design systems for speaker recognition [19, 57, 52, 10, 17, 8]. However, voice can be easily spoofed, especially using replay [39] (27.3% and 70.0% FPR), text to speech [16], and voice conversion attacks [48, 34]. Various works [20, 37, 59, 54, 72, 11] have used touch gestures such as taps, swipes, and multiple fingers for user authentication on touchscreen devices. These works report a very low EER (Equal Error Rate) in the range of 1%-5%. However, all these works fail to test the security of touch gestures in the presence of sophisticated attacks such as shoulder surfing or video based observation

Table 4: Comparison Between Different Systems
(x – Not Applicable)

Work	Modality	Hardware	Application	FPR (Random)	FPR (Targeted)	TPR
Woosh [50]	Blow air (mouth)	Commodity	Gesture Recognition	x	x	x
Tennet et al. [64]	Breath Flow	Specialized	Game Control	x	x	x
Zhao et al. [75]	ECG	Specialized	Authentication	8.7	x	91.3
Venugopalan et al. [67]	EMG	Specialized	Authentication	10.0	x	97.3
Eberz et al. [18]	Eye Movements	Specialized	Authentication	4.0	x	96.0
Shahzad et al. [59]	Swipes	Commodity	Authentication	4.0	2.1	94.6
Gafurov et al. [21]	Gait	Commodity	Authentication	13.0	15.0	87.0
Hautamaki et al. [27]	Audio	Commodity	Authentication	9.0	11.6	91.0
Ours	Breathing Gestures	Commodity	Authentication	0.5	2.0	98.0

attacks. Recently, a study from Khan et al. [30] pointed out that touch gestures are highly susceptible to such attacks. The way a person walks (gait) has also been proposed in the literature [41, 65, 78] for user authentication. The EER achieved with gait in these works varied between 4%-9%. However, gait based biometrics are vulnerable to sophisticated attackers [21]. Additionally, a person has to walk in order to be authenticated.

Behavioural biometrics based on heart rate (ECG) [75], electrical activity of skeleton muscles (EMG) [9], eye movements [18], and electrical activity of the brain (EEG) [13] have been proposed in the literature. All of these methods require special apparatus to be attached to the body of the person. On the contrary, the major advantage of our proposed modality is its pervasive nature: a simple smartphone-embedded microphone is more widely available and ubiquitously usable than specialized hardware (fingerprint sensor, or heart rate sensor), which can achieve higher accuracy but often needs to be mounted on specific body locations (e.g., on the chest). Table 4 shows the comparison between our work and some of the existing works. For our work we show the results of deep breathing gesture as it provides the best balance between TPR and FPR while preventing targeted attacks.

Complementing the earlier works, we have introduced a new behavioural biometric, which is usable, easily deployable, accessible, highly resilient to several kind of attacks and can be combined with other modalities to provide highly secure multimodal biometric authentication systems.

7. DISCUSSION AND LIMITATIONS

Additional Attack Modalities: We have demonstrated that our proposed biometric is resilient to basic impersonation and audio replay attacks. However, we have not considered more sophisticated attacks, such as the voice conversion attack (well-known in speaker recognition systems). In this attack, an expert attacker can perform fine-grained manipulation of the frequency components and amplitude of audio samples. In another example, a sophisticated attacker equipped with a large directional array of sensitive microphones and good signal processing skills can pick up the signal from a considerable distance (a microphone array could be built using tightly time synchronized discrete units rather than a large visible physical array of antenna). We believe that although such attacks are possible, they require highly motivated attackers with significant time and resources at their disposal. Also, as breathing-based authentication involves both frequencies, as well as temporal, components (the duration of breathing gestures), it may prove more resilient to such attacks. Nonetheless, they need to be investigated. Techniques from speech verification systems to counter replay attacks [68, 69] and voice conversion attacks [3] can be borrowed to prevent such attacks in any acoustics based authentication system.

Physiological Context: Our studies in Section 5.4 demonstrate the robustness of the system under certain short-term *extrinsic* contextual factors, such as physical activity. However, breathing-related features are also likely to exhibit longer timescale changes, due to *intrinsic* physiological changes, such as aging and decrease in the lung capacity. Longitudinal studies are needed to establish how sensitive *BreathPrint* is to such changes. We also did not study *BreathPrint's* sensitivity to short-term intrinsic context, such as when someone is sick or has breathing difficulties. Such sensitivity is common to other biometrics-based authentication approaches as well. For example, fingerprint-based authentication is known to fail if there is a severe cut on the user′s finger. Additionally, to understand other possible unexpected context related deficiencies, we need to perform an extensive longitudinal study, which involves collecting data from a large number of users. We are developing a smartphone app to enable such large-scale, longitudinal data collection. As collecting data for all possible contexts is challenging, we plan to also use data augmentation techniques [53, 36] to synthesize new data.

Intermittent vs. Continuous Authentication: Our canonical usage model has involved the use of *deliberate* breathing gestures, performed solely for authentication purposes–e.g., the user breathing into the microphone of a smartphone. Given the non-stop nature of breathing, future possibilities include the use of one or more microphones that provide continuous, breathing-based authentication of a user in a non-intrusive way. *A particularly exciting case might be that of a smartglass*: a microphone embedded in the lower rim of the glass should be able to pick up the natural breathing sounds. Such breathing-based authentication and interaction might be especially appealing in assisted-living environments, where breathing gestures may be used to control lights, turn on/off appliances etc.

Realizing this vision will require additional advances, such as (i) Improved segmentation: our current pipeline segments breathing gestures based on the knowledge that the user is deliberately performing one of the three predefined breathing gestures. This does not apply to the case of continuous, natural breathing; and (ii) Energy efficiency: Microphones on wearable platforms typically incur high energy overhead and cannot be operated continuously. However, it may be possible to design special-purpose microphones, operating

at low duty cycles and low sampling frequency for detecting the low-frequency breathing-based features.

Multimodal Sensing and Authentication: *BreathPrint* currently uses only audio sensing of selected breathing gestures. There is a body of work that uses additional sensor data as an attestation mechanism–e.g., the use of motion estimates, based on inertial sensing to provide greater resiliency to 2-D face authentication systems against photo/video attacks [12]. Similar approaches can be applied to breathing gestures as well, especially if additional sensors are placed at locations that can sense the breathing-related motion artifacts (e.g., inertial sensing on a smartglass).

Device-independence: In this study, we primarily evaluated the model using the data collected from same hardware (either a laptop or a smartphone). To explore the device-dependence of such models, we need to verify: i) Is it possible to train the model on one device and test on another device?, and ii) what parameters need to be adjusted to make cross device transferability viable? Robust cross-device models may prove to be widely deployable in emerging IoT-rich smart environments. To this end we are considering to use transfer learning methods [15] to create cross device adaptive models.

8. CONCLUSION

We have proposed a new authentication mechanism, called *BreathPrint*, for mobile and wearable devices. *BreathPrint* requires the user to perform three distinct, but commonplace, breathing gestures: sniff, normal, and deep, and uses audio captured by a microphone sensor to authenticate individuals. *BreathPrint* offers several advantages, such as accessibility, usability, deployability, and security. We show that *BreathPrint* has TPR rates of over 94% across all three gestures and also incurs very low FPR ($< 2\%$), even under directed *replay* and *impersonation* attacks. The authentication pipeline also proves to be quite robust to a selected set of artifacts, such as user activity and background noise. While larger, longer-lasting user studies are necessary, *BreathPrint* shows promise as a pervasive and novel method for identifying users in a variety of settings – e.g., in disambiguating between two sleeping individuals. We believe that *BreathPrint*, in tandem with other sensor modalities, can also prove to be an easy to use and robust mechanism for continuous authentication in future wearable and IoT platforms.

9. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers and the shepherd Mary Baker (HP Labs) for constructive feedback on preparation of the final version of this paper.

This research is supported partially by Singapore Ministry of Education Academic Research Fund Tier 2 under research grant MOE2014-T2-1063, and by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative. All findings and recommendations are those of the authors and do not necessarily reflect the views of the granting agency, or SMU.

10. REFERENCES

[1] H. Abdelnasser, K. A. Harras, and M. Youssef. Ubibreathe: A ubiquitous non-invasive wifi-based breathing estimator. In *MobiHoc*, pages 277–286. ACM, 2015.

[2] A. Aertsen, P. Johannesma, and D. Hermes. Spectro-temporal receptive fields of auditory neurons in the grassfrog. *Biological Cybernetics*, 38(4):235–248, 1980.

[3] F. Alegre, A. Amehraye, and N. Evans. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *ICASSP*, pages 3068–3072. IEEE, 2013.

[4] H. Aly and M. Youssef. Zephyr: Ubiquitous accurate multi-sensor fusion-based respiratory rate estimation using smartphones. In *INFOCOM*, pages 1–9. IEEE, 2016.

[5] A. Azarbarzin and Z. M. Moussavi. Automatic and unsupervised snore sound extraction from respiratory sound signals. *IEEE Transactions on Biomedical Engineering*, 58(5):1156–1162, 2011.

[6] A. Bajtarevic, C. Ager, M. Pienz, M. Klieber, K. Schwarz, M. Ligor, T. Ligor, W. Filipiak, H. Denz, M. Fiegl, et al. Noninvasive detection of lung cancer by analysis of exhaled breath. *BMC cancer*, 9(1):1, 2009.

[7] L. Boccanfuso and J. M. O'Kane. Remote measurement of breathing rate in real time using a high precision, single-point infrared temperature sensor. In *Biomedical Robotics and Biomechatronics*, pages 1704–1709. IEEE, 2012.

[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE signal processing letters*, 13(5):308–311, 2006.

[9] J. Cannan and H. Hu. Automatic user identification by using forearm biometrics. In *Advanced Intelligent Mechatronics*, pages 710–715. IEEE, 2013.

[10] M. J. Carey, E. S. Parris, and J. S. Bridle. A speaker verification system using alpha-nets. In *ICASSP*, pages 397–400. IEEE, 1991.

[11] J. Chauhan, H. J. Asghar, M. A. Kaafar, and A. Mahanti. Gesture-based Continuous Authentication for Wearable Devices: The Smart Glasses Use Case. In *ACNS*, pages 648–665. Springer, 2016.

[12] S. Chen, A. Pande, and P. Mohapatra. Sensor-assisted facial recognition: an enhanced biometric authentication system for smartphones. In *MobiSys*, pages 109–122. ACM, 2014.

[13] J. Chuang, H. Nguyen, C. Wang, and B. Johnson. I think, therefore i am: Usability and security of authentication using brainwaves. In *FC*, pages 1–16. Springer, 2013.

[14] P. Corbishley and E. Rodríguez-Villegas. Breathing detection: towards a miniaturized, wearable, battery-operated monitoring system. *IEEE Transactions on Biomedical Engineering*, 55(1):196–204, 2008.

[15] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *NIPS*, pages 353–360, 2008.

[16] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2280–2290, 2012.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker

verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[18] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic. Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics. In *NDSS*. Internet Society, 2015.

[19] K. R. Farrell, R. J. Mammone, and K. T. Assaleh. Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on speech and audio processing*, 2(1):194–205, 1994.

[20] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. *IEEE TIFS*, 8(1):136–148, 2013.

[21] D. Gafurov, E. Snekkenes, and P. Bours. Spoof attacks on gait authentication system. *IEEE TIFS*, 2(3):491–502, 2007.

[22] G. Gamez, L. Zhu, A. Disko, H. Chen, V. Azov, K. Chingin, G. Krämer, and R. Zenobi. Real-time, in vivo monitoring and pharmacokinetics of valproic acid via a novel biomarker in exhaled breath. *Chemical Communications*, 47(17):4884–4886, 2011.

[23] N. Gavriely, M. Nissan, A. Rubin, and D. W. Cugell. Spectral characteristics of chest wall breath sounds in normal subjects. *Thorax*, 50(12):1292–1300, 1995.

[24] B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

[25] V. Gross, A. Dittmar, T. Penzel, F. Schuttler, and P. Von Wichert. The relationship between normal lung sounds, age, and gender. *American journal of respiratory and critical care medicine*, 162(3):905–909, 2000.

[26] T. Hao, G. Xing, and G. Zhou. isleep: unobtrusive sleep quality monitoring using smartphones. In *SenSys*, page 4. ACM, 2013.

[27] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *INTERSPEECH*, pages 930–934. Citeseer, 2013.

[28] J. He, P. M.-L. Sinues, M. Hollmén, X. Li, M. Detmar, and R. Zenobi. Fingerprinting breast cancer vs. normal mammary cells by mass spectrometric analysis of volatiles. *Scientific reports*, 4, 2014.

[29] A. Jain and J. Vepa. Lung sound analysis for wheeze episode detection. In *Engineering in Medicine and Biology Society*, pages 2582–2585. IEEE, 2008.

[30] H. Khan, U. Hengartner, and D. Vogel. Targeted Mimicry Attacks on Touch Input Based Implicit Authentication Schemes. In *MobiSys '16*, pages 387–398. ACM, 2016.

[31] C. Kim and R. M. Stern. Power-normalized cepstral coefficients (pncc) for robust speech recognition. In *ICASSP*, pages 4101–4104. IEEE, 2012.

[32] T. Kinnunen, E. Karpov, and P. Franti. Real-time speaker identification and verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):277–288, 2006.

[33] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010.

[34] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *ICASSP*, pages 4401–4404. IEEE, 2012.

[35] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel. Accurate and privacy preserving cough sensing using a low-cost microphone. In *UbiComp*, pages 375–384. ACM, 2011.

[36] H. Lee, P. Pham, Y. Largman, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *NIPS*, pages 1096–1104, 2009.

[37] L. Li, X. Zhao, and G. Xue. Unobservable Re-authentication for Smartphones. In *NDSS*. Internet Society, 2013.

[38] Q. Li, J. Zheng, A. Tsai, and Q. Zhou. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions on Speech and Audio Processing*, 10(3):146–157, 2002.

[39] J. Lindberg, M. Blomberg, et al. Vulnerability in speaker verification-a study of technical impostor techniques. In *Eurospeech*, volume 99, pages 1211–1214, 1999.

[40] X. Lu and M. Bahoura. An automatic system for crackles detection and classification. In *Canadian Conference on Electrical and Computer Engineering*, pages 725–729. IEEE, 2006.

[41] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *ICASSP.*, volume 2, pages 973–976. IEEE, 2005.

[42] J. Marshall, D. Rowland, S. Rennick Egglestone, S. Benford, B. Walker, and D. McAuley. Breath control of amusement rides. In *CHI*, pages 73–82. ACM, 2011.

[43] J. Mead and S. H. Loring. Analysis of volume displacement and length changes of the diaphragm during breathing. *Journal of Applied Physiology*, 53(3):750–755, 1982.

[44] R. Nandakumar, S. Gollakota, and N. Watson. Contactless sleep apnea detection on smartphones. In *MobiSys*, pages 45–57. ACM, 2015.

[45] H. Pasterkamp, S. S. Kraman, and G. R. Wodicka. Respiratory sounds: advances beyond the stethoscope. *American journal of respiratory and critical care medicine*, 156(3):974–987, 1997.

[46] H. Pasterkamp, R. E. Powell, and I. Sanchez. Lung sound spectra at standardized air flow in normal infants, children, and adults. *American journal of respiratory and critical care medicine*, 154(2):424–430, 1996.

[47] H. Pasterkamp, J. Schafer, and G. R. Wodicka. Posture-dependent change of tracheal sounds at standardized flows in patients with obstructive sleep apnea. *CHEST Journal*, 110(6):1493–1498, 1996.

[48] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet. Voice forgery using alisp: Indexation in a client memory. In *ICASSP*, pages 17–20. IEEE, 2005.

[49] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury. Bodybeat: A mobile system for sensing non-speech body sounds. In *MobiSys*, pages 2–13. ACM, 2014.

[50] G. Reyes, D. Zhang, S. Ghosh, P. Shah, J. Wu, A. Parnami, B. Bercik, T. Starner, G. D. Abowd, and W. K. Edwards. Whoosh: non-voice acoustics for low-cost, hands-free, and rapid input on smartwatches. In *ISWC*, pages 120–127. ACM, 2016.

[51] D. Reynolds. An overview of automatic speaker recognition. In *ICASSP*, pages 4072–407. IEEE, 2002.

[52] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1):72–83, 1995.

[53] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *NIPS*, pages 3108–3116, 2016.

[54] N. Sae-Bae, K. Ahmed, K. Isbister, and N. Memon. Biometric-rich gestures: A novel approach to authentication on multi-touch devices. In *CHI*, pages 977–986. ACM, 2012.

[55] I. Sanchez and H. Pasterkamp. Tracheal sound spectra depend on body height. *American Review of Respiratory Disease*, 148:1083–1083, 1993.

[56] R. Schluter, I. Bezrukov, H. Wagner, and H. Ney. Gammatone features and feature combination for large vocabulary speech recognition. In *ICASSP*, volume 4, pages 649–652. IEEE, 2007.

[57] M. Schmidt and H. Gish. Speaker identification via support vector classifiers. In *ICASSP*, volume 1, pages 105–108. IEEE, 1996.

[58] G. Serbes, C. O. Sakar, Y. P. Kahya, and N. Aydin. Effect of different window and wavelet types on the performance of a novel crackle detection algorithm. In *International Conference on Hybrid Information Technology*, pages 575–581. Springer, 2011.

[59] M. Shahzad, A. X. Liu, and A. Samuel. Secure Unlocking of Mobile Touch Screen Devices by Simple Gestures: You Can See It but You Can Not Do It. In *MobiCom*, pages 39–50. ACM, 2013.

[60] J. Sharp, J. Henry, S. Sweany, W. Meadows, and R. Pietras. The total work of breathing in normal and obese men. *Journal of Clinical Investigation*, 43(4):728, 1964.

[61] P. M.-L. Sinues, M. Kohler, and R. Zenobi. Human breath analysis may support the existence of individual metabolic phenotypes. *PloS one*, 8(4):e59909, 2013.

[62] X. Sun, Z. Lu, W. Hu, and G. Cao. Symdetector: Detecting sound-related respiratory symptoms using smartphones. In *UbiComp*, pages 97–108. ACM, 2015.

[63] S. A. Taplidou and L. J. Hadjileontiadis. Wheeze detection based on time-frequency analysis of breath sounds. *Computers in biology and medicine*, 37(8):1073–1083, 2007.

[64] P. Tennent, D. Rowland, J. Marshall, S. R. Egglestone, A. Harrison, Z. Jaime, B. Walker, and S. Benford. Breathalising games: understanding the potential of breath control in game interfaces. In *International Conference on Advances in Computer Entertainment Technology*, page 58. ACM, 2011.

[65] H. M. Thang, V. Q. Viet, N. D. Thuc, and D. Choi. Gait identification using accelerometer on mobile phone. In *ICCAIS*, pages 344–348. IEEE, 2012.

[66] X. Valero and F. Alias. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Transactions on Multimedia*, 14(6):1684–1689, 2012.

[67] S. Venugopalan, F. Juefei-Xu, B. Cowley, and M. Savvides. Electromyograph and keystroke dynamics for spoof-resistant biometric authentication. In *CVPR Workshops*, pages 109–118. IEEE, 2015.

[68] J. Villalba and E. Lleida. Detecting replay attacks from far-field recordings on speaker verification systems. In *European Workshop on Biometrics and Identity Management*, pages 274–285. Springer, 2011.

[69] J. Villalba and E. Lleida. Preventing replay attacks on speaker verification systems. In *ICCST*, pages 1–8. IEEE, 2011.

[70] L. Weisi, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo, and H. Wang. *Multimedia analysis, processing and communications*, volume 346. Springer Science & Business Media, 2011.

[71] D. Wertheim, C. Olden, E. Savage, and P. Seddon. Extracting respiratory data from pulse oximeter plethysmogram traces in newborn infants. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 94(4):F301–F303, 2009.

[72] H. Xu, Y. Zhou, and M. R. Lyu. Towards Continuous and Passive Authentication via Touch Biometrics: An Experimental Study on Smartphones. In *SOUPS*, pages 187–198. ACM, 2014.

[73] A. Yadollahi, E. Giannouli, and Z. Moussavi. Sleep apnea monitoring and diagnosis based on pulse oximetery and tracheal sound signals. *Medical & biological engineering & computing*, 48(11):1087–1097, 2010.

[74] K. Yatani and K. N. Truong. Bodyscope: A wearable acoustic sensor for activity recognition. In *UbiComp*, pages 341–350. ACM, 2012.

[75] C. X. Zhao, T. Wysocki, F. Agrafioti, and D. Hatzinakos. Securing handheld devices and fingerprint readers with ecg biometrics. In *BTAS*, pages 150–155. IEEE, 2012.

[76] X. Zhao, Y. Shao, and D. Wang. Casa-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(5):1608–1616, 2012.

[77] X. Zhao and D. Wang. Analyzing noise robustness of MFCC and GFCC features in speaker identification. In *ICASSP*, pages 7204–7208. IEEE, 2013.

[78] Y. Zhong, Y. Deng, and G. Meltzner. Pace independent mobile gait biometrics. In *BTAS*, pages 1–8. IEEE, 2015.