10-2021

EarGate: Gait-based user identification with in-ear microphones

Andrea FERLINI

Dong MA
Singapore Management University, dongma@smu.edu.sg

Cecilia MASCOLO

Citation

FERLINI, Andrea; MA, Dong; and MASCOLO, Cecilia. EarGate: Gait-based user identification with in-ear microphones. (2021). *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, New Orleans, 2021 October 25-29,*. 337-349.
Available at: https://ink.library.smu.edu.sg/sis_research/6990

EarGate: Gait-based User Identification with In-ear Microphones

Andrea Ferlini†
University of Cambridge
af679@cl.cam.ac.uk

Dong Ma†
University of Cambridge
dm878@cl.cam.ac.uk

Robert Harle
University of Cambridge
rkh23@cl.cam.ac.uk

Cecilia Mascolo
University of Cambridge
cm542@cl.cam.ac.uk

† First authors with equal contribution in alphabetical order.

ABSTRACT

Human gait is a widely used biometric trait for user identification and recognition. Given the wide-spreading, steady diffusion of ear-worn wearables (*Earables)* as the new frontier of wearable devices, we investigate the feasibility of earable-based gait identification. Specifically, we look at gait-based identification from the sounds induced by walking and propagated through the musculoskeletal system in the body. Our system, EarGate, leverages an in-ear facing microphone which exploits the earable's *occlusion effect* to reliably detect the user's gait from inside the ear canal, without impairing the general usage of earphones. With data collected from 31 subjects, we show that EarGate achieves up to 97.26% Balanced Accuracy (BAC) with very low False Acceptance Rate (FAR) and False Rejection Rate (FRR) of 3.23% and 2.25%, respectively. Further, our measurement of power consumption and latency investigates how this gait identification model could live both as a stand-alone or cloud-coupled earable system.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

ACM Reference Format:
Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2022. EarGate: Gait-based User Identification with In-ear Microphones. In *The 27th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '21), January 31-February 4, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3447993.3483240

1 INTRODUCTION

Human gait has been shown to be unique across individuals and hard to mimic [26, 37]. As such, there have been a variety of attempts to use gait as a biometric for user recognition and identification. While computer vision based gait bio-metrics is widely spread, wearable-based gait tracking is particularly attractive for continuous identification and, potentially, authentication. Wearable gait based identification is an enabler for various applications, including: mobile devices to remain unlocked and ready for interaction when on the owner's person; health insurers to identify the device owner as the policy owner when rewarding healthy habits sensed via the wearable; automated entry systems for home, work or vehicles; automated ticket payment/validation for public transport [10, 21]; etc. There have even been examples of using wearable gait data to generate a secure key to pair devices worn on the same body [23].

Wearable-based gait tracking methods leverage sensor data collected from wearable devices worn by the user to capture their motion dynamics, typically through accelerometer analysis [17], or step sounds [8, 19]. To date, the focus has been on smartphones or smartwatches as the current mass-market wearables. In this work, instead, we look at the use of ear-based sensing (via so-called *earables*) for this task. The importance of the approach we present is further highlighted by the fact that, in the near-future, earables are likely to become stand-alone devices [16]. Hence, the need for earable-based identification schemes becomes more and more crucial [13]. Being able to seamlessly identify the earable wearer, it can act as an authentication accelerator for earables or mobile devices, bypassing traditional bio-metrics such as fingerprint (requires the integration of capacitive sensing pad on earables with limited size) and face recognition (impossible to capture front face image from an earbud). Occasionally, if the identity is mistakenly rejected, a request for a secondary authentication can be triggered. The secondary authentication method could be implemented on a companion device, such as smartphone/smartwatch, that has full access to fingerprint and human face. Additionally, once the user has been successfully identified by the earable, the earable itself can act as a hub to authenticate the user for access control (e.g. opening their office door, validating their ticket at the train station, etc.). Furthermore, with the increased sensing capabilities of earables, successfully identifying the earable's wearer becomes crucial in order to associate the sensitive bio-medical information collected by the earable to the right user.

Today's earables fall broadly into two categories: advanced playback devices (e.g., Apple AirPods) or assistive devices (e.g., advanced hearing aids). Notably, although being useful for all earables, such continuous identification systems suit hearing aids particularly well, as they are continuously worn throughout the day. Earables present potential advantages over other wearables. For instance, earables' location in-the-ear means that, unlike phones, they are in a consistent, stable location, firmly attached to the user's body; and unlike watches, they are not subject to confounding movements during walking (e.g., carrying objects, using tools or loose straps). Further, they also offer two independent signals (one per each ear).

Here we investigate acoustic-gait as a convenient alternative to inertial-based gait tracking. Specifically, we look at the possibility of gait-based identification from the *sounds* caused by the physical act of walking and transmitted internally via the musculoskeletal

ACM MobiCom '21, January 31-February 4, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8342-4/22/01.
https://doi.org/10.1145/3447993.3483240

system. Our novel earable-based *acoustic-gait* identification system, EarGate, is built around a cheap in-ear facing microphone that is already available on most earbuds and hearing aids (e.g., for noise cancellation purposes). We show that an earable equipped with a microphone *inside* the ear canal can not only detect motion signals, but those are also amplified by the combination of two biological phenomena: *bone conduction* and the *occlusion effect*. We conducted a user study, with 31 subjects of mixed gender, demonstrating how the acoustic-gait thus collected is a good candidate for a privacy-preserving identification system (benefit from the in-ear facing microphone). Further, we investigated the implications of running the identification-framework entirely on-device, as well as offloading the computation (either to the Cloud or to a companion device) to show the versatility of the approach.

In summary, *our approach leverages sensors (microphones) already inherently embedded in earables for main functions*. As mentioned, previous gait identification/recognition approaches have often relied on inertial sensors [17]. However, inertial sensors have made reasonable penetration into the high-end leisure devices market but not as much into cheaper earables, or the hearing aid market: the addition of inertial sensors to these devices would entail more complex system design and form factor [40], as a consequence, inertial-based earable solutions are likely to result in increased cost and delays in reaching the market. This paper offers an investigation into an alternative to inertial sensors through the use of microphones. Notably, while we acknowledge the merits of inertial-based gait recognition approaches, there is great value in showing the potential of a lesser explored modality, such as in-ear microphones. Particularly given what suggested by research and market trends, according to which, in the near future, miniaturized form factor will have a key role, especially as the distinction between hearing aids and earables is likely to become less marked. Further, in-ear based microphones offer complementary advantages as an overall lower cost, given its importance as a sensor to enable noise cancellation, a must-have feature for both high-end leisure earables and in hearing aids. While in this paper we focus on acoustic-gait recognition, we note this is only one of the possible use cases for in-ear bone-conducted sounds. For instance, these could indeed be used for activity recognition [24], as well as physiological sensing. In particular, the contributions of this work can be summarized as:

- We devise a novel earable-based gait identification system, EarGate, consisting of a hardware prototype and a software pipeline. EarGate, not only leverages a novel type of signal, in-ear bone-conducted sounds, to identify users based on their gait, but also is accurate, robust, and has improved usability (only a few steps are sufficient to identify the user, without the need for them to be continuously walking);
- To track gait cycles from earables, we leverage the occlusion effect, a natural enhancement of low-frequency components in an occluded ear canal [36]. In addition, we designed an end-to-end signal processing pipeline and some techniques to guarantee the reliable presence of the occlusion effect;
- We collected a one-of-its-kind dataset with 31 subjects under various conditions, which we released to the research community at Kaggle[1];

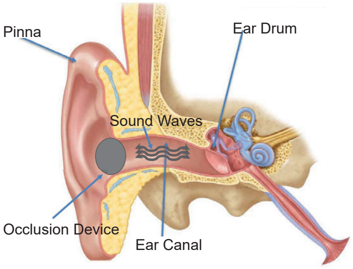


Fig. 1: Anatomy of human ear and occlusion effect. When the orifice is occluded, sounds are trapped in the ear canal, resulting in the amplification of low frequency components.

- We evaluate the identification performance of EarGate under various practical scenarios, showing we can achieve up to 97.26% Balanced Accuracy (BAC) with very low False Acceptance Rate (FAR) and False Rejection Rate (FRR) of 3.23% and 2.25%, respectively. Furthermore, we demonstrated that EarGate is robust to high-frequency internal (human speech) and external (music playback and phone calls) noises;
- Finally, we assessed the system performance of EarGate by measuring the power consumption and latency. We find EarGate can work in real-time (74.25 ms on-device identification latency) consuming acceptable energy (167.27mJ for one-time on-device identification). This confirms that EarGate could be deployed in potential new generation earables which will likely be standalone from the system perspective.

2 PRELIMINARIES

In this section, we first brief the reader on the rationale driving our work and then provide evidence of the feasibility of our approach.

2.1 Rationale: Occlusion Effect

In this work, we leverage the natural low-frequency boost (up to 40 dB depending on the frequency [11]) provided by the phenomenon known by the name of *occlusion effect*. This section provides the reader with a basic understanding of what the occlusion effect is in practice and how it can be exploited to facilitate in-ear human gait identification.

From a physiological point of view, the occlusion effect can be defined as a dominance of the low-frequency components of a bone-propagated sound due to the loss of relevance of the outer ear sound pathways whenever the ear canal orifice is sealed (i.e. occluded) [36]. For example, such phenomenon could be experienced in the form of echo-like/booming sounds of their own voice if a person is speaking and is obstructing her ear canals with a finger or an earplug. Essentially, sound is nothing but vibrations propagating as acoustic waves. Usually, it travels through bones and escapes the inner-ear via the ear canal orifice. However, if this opening is obstructed, the vibration waves are blocked inside the canal and are bounced back to the eardrum [34], as illustrated in Figure 1. As a result of that, the low-frequency-bone-conducted sounds are amplified [33]. A more precise definition of what the occlusion effect entails can be denoted by the ratio between the sound pressure inside the occluded ear canal and that in the open ear [35]. Specifically to our use case, the vibrations generated by a foot hitting the ground, as soon as a person takes a step (basic component

[1] https://www.kaggle.com/dongma878/eargate

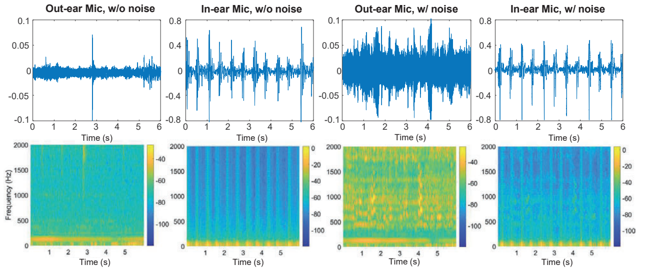


Fig. 2: Walking signals collected with out-ear and in-ear microphones under different conditions. The in-ear microphone data show higher gain at low frequencies (<50 Hz) due to the occlusion effect, and more resilience towards environmental noise.

of a gait cycle), propagate through the body via bone-conduction. Interestingly, these vibrations are amplified if the ear canal of the person is occluded by, for example, an earbud.

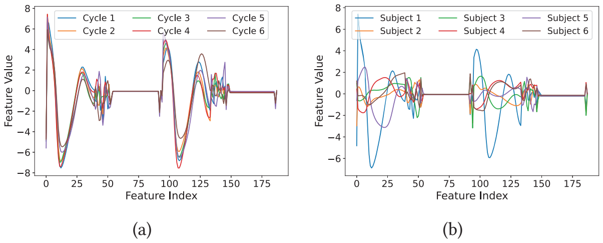
2.2 Out-ear Mic vs. In-ear Mic

Compared to the approaches using external microphones (out-ear mic) for gait recognition [8, 19], the use of occlusion effect and an inward-facing microphone brings the following advantages: **(1)** given an occluded ear canal, an inward-facing microphone, which mostly records bone-conducted sounds, results to be less susceptible to external sound and consequent environmental noise. This not only means our system is more robust to noise, but practically, it makes our approach more appealing from a privacy perspective: potentially sensitive external sounds, such as human speech, are hardly audible from our in-ear facing microphone. **(2)** another direct consequence of the occluded ear canal, is that the body-sounds we are interested in are relatively low-frequency and, therefore, we greatly benefit from the low-frequency amplification boost induced by the occlusion effect, resulting in an improved signal-to-noise-ratio (SNR) of the desired signal.

Figure 2 plots the raw signals and corresponding spectrograms collected with the out-ear microphone and in-ear microphone when a subject is walking, with and without environmental noise. From these graphs, we can clearly observe how the in-ear microphone overcomes the drawbacks of the out-ear microphone that captures air-conducted sounds. First, air conduction incurs large attenuation, while bone conduction and occlusion effect guarantee an excellent SNR. Second, the walking sound measured by the out-ear microphone resides in higher frequency (audible) and is completely mixed with other environmental noise (e.g., music and human speech). Consequently, they can not be separated even with a lowpass filter.

2.3 Human Gait Primer

The walking style of a person is commonly known as their gait. Medical and physiological studies [37] suggest that the human gait shows 24 different components. The differences between the gait of distinct subjects are caused by the uniqueness in their muscular-skeletal structure. The human gait is regulated by precise bio-physical rules [26]. These, in turn, are dictated by the tension generated by the muscle activation and the consequent movement of the joints. As a result of that, the forces and moments linked to the movement of the joints cause the movement of the skeletal links which, therefore, exert forces on the environment (e.g. the foot striking the ground). Hence, the human gait can be described as a generation of ground-reaction forces which are strongly correlated with the muscular-skeletal structure of each individual. In practice, differences in the body structure of individuals are among the factors that produce the interpersonal differences in walking patterns that enable gait-based identification.



(a)

(b)

Fig. 3: (a) Extracted feature vectors for four different gait cycles from the same subject, and (b) gait cycles from four different subjects. Figure 3(a) clearly shows how the gait cycles captured by EarGate are consistent for each individual (i.e. high intra-class similarity). Figure 3(b), on the other hand, shows how the gait cycles are distinguishable among subjects (i.e. high inter-class difference).

2.4 Feasibility Exploration

Although the in-ear microphone is capable of detecting human steps, whether we could extract the acoustic gait to differentiate people remains unclear. To demonstrate the feasibility of gait recognition, we need to prove that (1) gait cycles belonging to the same individual are consistent with each other (i.e., intra-class similarity) and (2) gait cycles belonging to different subjects show significantly different patterns (i.e., inter-class dissimilarity). Thus, we collected data from four subjects and extract features (see Section 3.3) to represent the user's gait. As shown in Figure 3, the extracted features exhibit high intra-class similarity and high inter-class difference. Therefore, it would be feasible to identify people with the acoustic signals measured with the in-ear microphone.

3 SYSTEM DESIGN

This section presents an overview of EarGate and its functionalities and a description of the proposed gait-based identification pipeline.

3.1 EarGate: System at a Glance

Initially, the user has to take part in an enrollment phase, a required stage during which the system acquires the data and process them (pre-processing and feature extraction) before training the model to recognize the acoustic-gait of the legitimate user. Notably, as we will discuss in Section 5.9, a small number of steps and, therefore, little time, is sufficient to achieve good identification performance. Once the enrollment phase is over, the system is ready to operate. As shown in Figure 4, EarGate silently collects, and pre-processes on-device, acoustic-gait data. The system is provisioned to either execute the (*limited*) computation required to identify the user

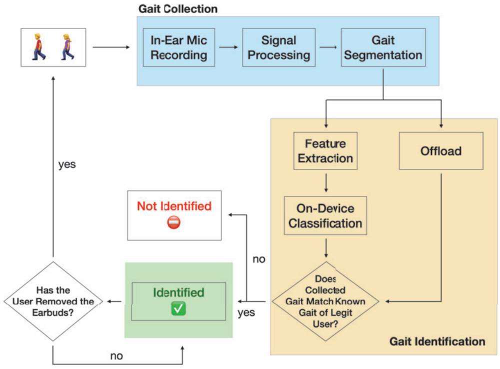


Fig. 4: EarGate Functioning.

(or reject them) on-device, or to offload it, e.g., to the cloud, a smartphone, or a remote server (Section 6).

Notably, there is NO need for the user to be *constantly* and *instantly* walking to be identified. Instead, EarGate can perform a one-time identification once the user wears the earbuds and walks a couple of steps (e.g. to grab a coffee, to go to the restroom, etc.); the identification result (either recognized as legit user, or not) is then considered valid until the user removes the earbuds (i.e., the only chance the wearer has of switching identity). Detecting such occurrence is, in practice, very simple, as the natural properties of the occlusion effect will suddenly disappear from the in-ear microphone recordings, whenever the user removes the earbuds. Besides, commercially available earbuds, like Apple AirPods already do that (i.e., to automatically stop the music whenever the user removes their earbuds). Such scheme significantly relieves the user burden of *"walk now to be identified"* and saves system energy as one identification could be valid for a longer time.

3.2 Signal Processing

Our prototype records the microphone outputs in 48 kHz, while the generated step sounds are at relatively low frequencies [2]. To minimize the computation overhead during processing, we first down-sample the recorded data to 4 KHz. Then, a low-pass filter with a cutoff frequency of 50 Hz is applied to eliminate the high-frequency noise. The choice of 50 Hz cutoff frequency is to guarantee a good signal to noise ratio of the walking signal, while retaining robustness to most environmental noise (typically higher than 50 Hz). Moreover, due to the large attenuation of sound in the air as well as the blockage of the ear canal opening, the majority of the noise is suppressed or canceled in the ear canal. Figure 5 shows the low-pass filtered signal from the left earbud when one participant is walking on tiles with sneakers. We can observe that an acoustic gait cycle is composed of two spikes (happening when the foot hit the floor in the *strike phase*) and two relatively flat (silent) periods (denoting the *swing phase*). This is different from sinusoidal-like patterns recorded by the IMU [32], therefore, existing gait segmentation approaches proposed for IMU data are not applicable. We propose a peak detection based algorithm to segment the signals and extract gait cycles.

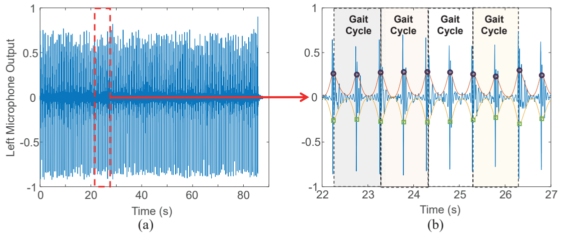


Fig. 5: (a) The low-pass filtered signal collected when one participant is walking, (b) a segment showing the performance of proposed gait segmentation algorithm.

Specifically, we first use the Hilbert transform to extract the envelopes of the filtered signal and apply a low-pass filter (with a cutoff frequency of 3 Hz) on the envelops to smooth them, as illustrated by the red (upper envelop) and yellow (lower envelop) curves in Figure 5 (b). Then, we perform peak detection on the filtered envelops and regard the peaks as the points when the human foot hits the ground. Whenever a pair of upper peak-lower peak is aligned, we treat it as a step. Next, we select the cycle start points by skipping one between every two peaks, as each gait cycle consists of two steps. Lastly, a gait cycle is extracted as the samples between every two cycle start points. Most of the extract cycles last for around one second, so we interpolate them into the same length of 4,000 samples using spline interpolation.

3.3 Feature Extraction

Then, EarGate extracts features that could represent the characteristics of user gait from each cycle. We do that using *librosa* [4], a Python package specific for audio processing. Specifically, we looked at:

- Mel-Frequency Cepstral Coefficients (MFCC): obtained from the short-term power spectrum, MFCC certainly is one of the most common and known features in audio processing [14];
- Chroma of Short-Time Fourier Transform (STFT);
- Mel Spectrogram: the signal spectrogram in the Mel-scale;
- Root-Mean-Squared Energy (RMSE): the Root-Mean-Squared (RMS) of the STFT of the signal, which provides information about the power of the signal;
- Onsets: the peaks from an onset strength envelope, result of a summation of positive first-order differences of every Mel-band.

Rather than only using the data recorded by either the left-earbud-microphone or those coming from the right-earbud-microphone, separately, we fuse them concatenating the features we extracted from each. Notably, unlike other wearables (e.g., smartwatches) earables are two and, therefore, it is possible to leverage, and fuse, their two independent measurements of the same phenomenon.

3.4 Identification Methodology

Like most identification systems (e.g., FaceID and TouchID), before online identification, EarGate requires an enrollment phase. It is during this phase that the legitimate users provide their gait data to system, thus training a model to classify the users as either legitimate users, or impostors. Notably, all the users that have not been

[2] Note that 'frequency' in this paper refers to the pitch of the step sound, instead of the cadence/speed of walking.

seen by the model in the enrollment phase will be regarded as impostors. In this work, we consider two enrollment schemes: **(i)** with and **(ii)** without impostor data. The former leverages a pre-trained model with benchmark impostor data, together with some data from the legitimate user. However, given it is not always possible to assume the availability of benchmark imposer data, especially little after the release of the system, we also look at a model solely trained on the legitimate user data. In either case, the system needs some walking data from the legitimate user and, therefore, will instruct the user to follow an enrollment protocol (basically walking for a few steps and training a model with such walking data). For the online-identification framework we adopt: **(i)** a two-class Support Vector Machine (SVM) classifier (if benchmark imposer data are available); or **(i)** a one-class SVM classifier (when we only have the data of the legitimate user), due to its high computational efficiency and low complexity [18].

Prior work unequivocally showed how gait is a unique user fingerprint, and how is very hard for an impostor to impersonate another person's gait [28, 29]. Hence, the aim of this paper is to use earables as a personal-identification device, we sought to assess whether our in-ear microphone-based approach is capable of recognizing the user in such way that, if others (i.e. *impostors*) were using the device, it would be able to spot it. To this extent, we consider both **(i)** *Replace Attacks* (a different user mistakenly tries to use the earables) and **(ii)** *Mimic Attacks* (a malicious attacker deliberately tries to use the earbuds by actually impersonating the user, i.e., simulating the user's gait). As highlighted by our evaluation, different users can be distinguished very clearly by our system, thus making mimic attacks even more unfeasible. To further clarify that, for example, let us assume there is a very well-trained imposer, who can accurately mimic the gait of the legitimate user, generating the very same vibrations whenever the feet hit the ground. However, even given all the most favorable conditions, when the vibrations propagate through the body and the bones, all the way to the ear canal, they will inevitably be different from those belonging to the legitimate user. This is because, as discussed in Section 2.1, the human body and skeleton act as a natural modulator. Hence, in the remainder of the paper we will focus on showing the performance of our system against the more common replace attacks.

4 IMPLEMENTATION

This section provides the design details our prototype and describes the data collection procedures we followed.

4.1 EarGate Prototype

To have full control on the data (sampling rate as well as unlimited access), we decided to build our own prototype. During the whole process, we were driven by the requirements we stated in section 2.4. In particular, the customization we carried out to manufacture our EarGate prototype is not altering the substantial functioning and design of the earable. In fact, in-ear-facing microphones are *already* embedded in commercial, off-the-shelf earbuds (e.g., Apple AirPods Pro) [1, 3]. However, they are mainly used for noise cancellation and, unfortunately, companies do not expose APIs to access in-ear audio recordings. Therefore, our hardware prototype simply consists of a cheap and easily available in-ear facing microphone placed in a commercially available pair of earbuds. More precisely, we chose a pair of MINISO Marvel earphones [6]. The reasoning behind our choice being: (1) their case is large enough to comfortably fit a Micro-Electro-Mechanical System (MEMS) microphone without the need for removing existing components; (2) the earphones have a removable silicone ear-tip, well suited to occlude people's ear canals. In addition, spare ear-tips of different sizes are also included. Regarding the microphone, we opted for an analog MEMS SPU1410LR5H-QB [5]. We selected this particular model given its wide frequency response, spanning flat from 20 Hz to 20 kHz.

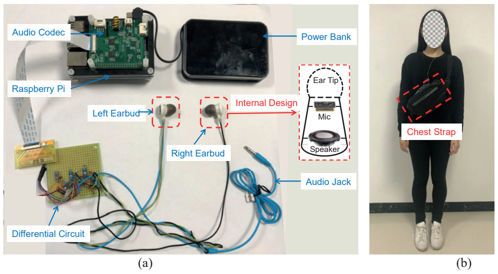


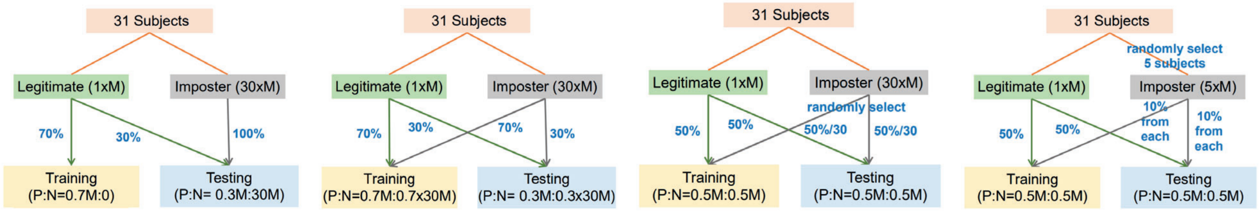
Fig. 6: (a) The designed earbuds prototype and accompanying data recording device, (b) illustration of a participant wearing the device.

Figure 6 (a) depicts our system in its entirety. In particular, it also reports the internal design of our customized bud, showing how we positioned the microphone immediately in the vicinity of the ear-tip. By doing so, we achieve a higher signal-to-noise ratio (SNR) than what we would have if we had placed the microphone behind the speaker. Notably, this modification does not lead to any deterioration in the audio playback quality. To assess that, we recruited 31 subjects and, after reproducing a music piece from both a regular and a modified pair of earbuds, we requested them to provide us with feedback concerning the audio quality. In this phase of the experiment, the subjects were agnostic of presence/absence of the microphone. Interestingly, 29 out of 31 people could not experience any difference in audio quality. Surprisingly, 2 even attributed a slightly better quality to the modified bud. We provide more details on our data collection campaign in Section 4.2.

In addition to the internal design of our customized earbuds, Figure 6 (a) presents the data-logging component we used during our data collection. Both the left and right buds are modified as described above. Notice the speakers in the buds are normally connected to the original 3.5 mm audio jack. To achieve the best possible SNR, we connected each microphone to a Differential Circuit right before sampling them with an Audio Codec. Specifically, the codec we opted for is an 8-channel audio recording with two ADC chips (AC 108) named ReSpeaker Voice Accessory HAT [7]. The audio codec is controlled by the Python script executed on a Raspberry Pi 4. The microphone recordings are sampled at 48 kHz. Power supply comes from an off-the-shelf power bank, thus allowing us to collect data with the subjects in motion. To reduced the burden of the volunteers when they were moving, we placed all the prototype components (but the earbuds) in a chest strap, as of in Figure 6 (b).

4.2 Data Collection

After having obtained clearance for carrying out the studies from the Ethics Board of our institution, we recruited 31 subjects. Out



(a) OC-SVM

(b) Bi-SVM (Imbalanced)

(c) Bi-SVM (Balanced-All)

(d) Bi-SVM (Balanced-5)

Fig. 7: Training and testing data splitting scheme for (a) one-class SVM, (b) imbalanced binary SVM, (c) balanced binary SVM with all subjects' data, and (d) balanced SVM with part of subjects' data. $P : N$ **represents the ratio between positive (legitimate) and negative (impostor) gait.**

of the 31 participants, 15 of them were females and the remainder 16 were males. The mean age of the participants was 26.6±5.8. All the participants received a voucher in exchange of their time. After having taken COVID-19 related precautions (i.e. face masks, hands sanitizing, etc.), we admitted the participants, one at a time, into the room of the experiment (a 12 × 6 square meters room). We instructed the participants to walk in circles following the perimeter walls of the room. The room was quiet, with a noise level of approximately 30 dB. We asked the user to keep their normal pace and walking style. As a gait cycle is composed of two consecutive steps, the subjects always started with their left foot and finished with the right foot when walking. We considered different ground materials, as tiles and carpet, and multiple conditions, with the participants walking barefoot, with slippers, with sneakers, and while speaking. Factoring in all these variables allows us to further assess the robustness, and generalizability, of EarGate . For each of these conditions, the participants walked continuously for 1.5 minutes (a session). During each walk, the subject counted the number of steps he/she took. This served as ground truth. Walking at normal speed, all the subjects made 156−176 steps per 1.5 minutes-long session. Eventually, each participant performed 8 (2 ground material × (3 footwear + 1 speaking)) different walking sessions, accounting for a total of 52,046 steps (i.e. 26,023 gait cycles).

5 PERFORMANCE EVALUATION

In this section, we present the data collection procedure, the training methodology, as well as different variables we consider while assessing the performance of our system.

5.1 Metrics

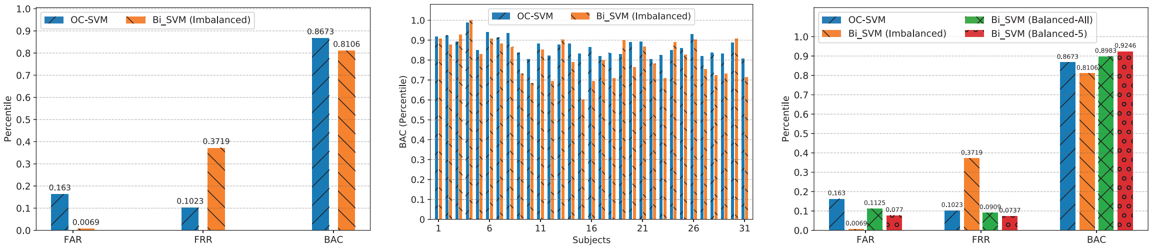
The metrics we consider to assess the goodness of the proposed identification system are:

- **False Acceptance Rate** (FAR): a common metric for an identification system that describes the system's likelihood of successfully identifying a non-legitimate-user;
- **False Rejecting Rate** (FRR): also known by the name of False Negative Rate (FNR), it indicates the identification system's likelihood of rejecting the legitimate-user;
- **Balanced Accuracy** (BAC): given we also assess the performance of our system in the case of unbalanced training and testing sets, we consider BAC to gauge the real accuracy of our system. Specifically, BAC is defined as $\frac{TPR+TNR}{2}$, where TPR and TNR are True Positive and True Negative Rate, respectively. The True Positive Rate of an identification system is its goodness in recognizing legitimate users, whilst True Negative Rate indicates the value of the system in protecting the user from attackers.

5.2 Training-testing protocol

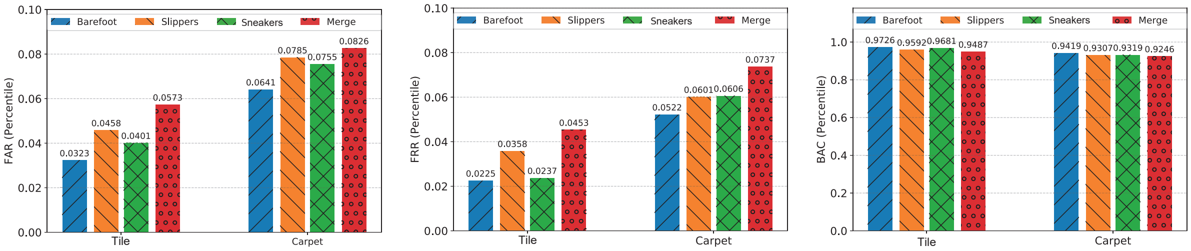
To evaluate our system, we propose four training-testing protocols, as shown in Figure 7. The number of gait cycles collected from each of the 31 subjects is denoted as M. The first two schemes ((a) and (b)) involve different amounts of positive and negative data during training and testing and such data imbalance might affect the performance. Thus, we further propose two protocols ((c) and (d)) that ensure balanced positive and negative samples during training and testing. All the four schemes consider adversarial attacks during testing. Notably, informed by the physiological explanation of gait (Section 2.3), we did not review mimic attacks: by being so strongly correlated with the muscular-skeletal structure of the individual, gait is extremely hard to mimic −if not impossible. Moreover, the fact that the in-ear (bone-conducted) audio we leverage is further modulated by the skeleton of the subject, constitutes an additional barrier against mimic attacks.

- **(a)** One-Class SVM (**OC-SVM**): one subject is iteratively selected as the legitimate user and the rest are regarded as impostors. Training dataset only consists of 70% (0.7 × M) data from the legitimate user, and the testing dataset is composed of 30% (0.3 × M) legitimate user data and all impostor data (30 × M).
- **(b)** Imbalanced Binary SVM (**Bi-SVM (Imbalanced)**): one subject is iteratively selected as the legitimate user and the rest are regarded as impostors. Training dataset consists of 70% legitimate user's data (0.7 × M) and 70% impostors' data (30 × 0.7 × M), and the testing dataset is composed of 30% (0.3 × M) legitimate user's data and 30% (30 × 0.3 × M) impostors' data.
- **(c)** Balanced Binary SVM with all subjects' data (**Bi-SVM (Balanced-All)**): one subject is iteratively selected as the legitimate user and the rest are regarded as impostors. Training dataset consists of 50% (0.5 × M) legitimate user's data and the same number (0.5 × M) of gait cycles that are randomly selected from the 30 impostors. The testing dataset is composed of the rest 50% (0.5 × M) legitimate user's data and another 0.5 × M randomly selected impostor data.



(a) (b) (c)

Fig. 8: (a) Overall performance of EarGate (averaged across 31 subjects), (b) BAC of each individual using OC-SVM and Bi-SVM (Imbalanced). (c) Comparison of the four proposed training-testing protocol, which also reflects the impact of data imbalance.



(a) (b) (c)

Fig. 9: Performance comparison (a) FAR, (b) FRR, and (c) BAC, of different footwear and ground material.

- **(d)** Balanced Binary SVM with part of subjects' data (**Bi-SVM (Balanced-5)**): one subject is iteratively selected as the legitimate user and five (5/30) subjects are randomly selected as impostors. Training dataset consists of 50% ($0.5 \times M$) legitimate user's data and 10% from each of the five impostors ($5 \times 0.1 \times M$). The testing dataset is composed of the rest 50% ($0.5 \times M$) legitimate user's data and another 10% ($5 \times 0.1 \times M$) data from each of the impostors.

In the remainder of this section, we present our experimental results under various conditions.

5.3 Overall Performance

We first evaluate the overall performance of EarGate by combining all the collected data together, i.e., different ground material and footwear. Figure 8(a) presents the results obtained with the first two training protocols (OC-SVM and Bi-SVM (Imbalanced)), which are averaged over the 31 subjects. We can observe that with both methods, EarGate can achieve over 80% balanced accuracy (BAC). The FAR and FRR are relatively high, which might be caused by the imbalanced dataset and the variability of different waking conditions. We will discuss this impact in the following sub-sections. Figure 8(b) plots the BAC of each individual with the two protocols. We can see that although the performance varies among subjects, most of the subjects achieve over 80% BAC. In addition, we found that the best training-testing protocol is subject-dependent as some users obtain higher BAC with OC-SVM, while others achieve better performance with Bi-SVM (Imbalanced). Thus, it is necessary to optimize the training protocol for each individual.

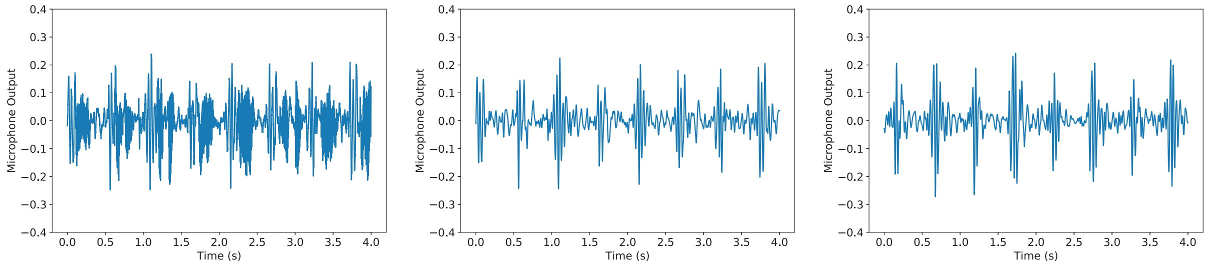
5.4 Impact of Data Imbalance

Intuitively, Bi-SVM is expected to perform better than OC-SVM. The reason is that OC-SVM is similar to a clustering problem and the model has to learn the correlation of the data without information on outliers. While for Bi-SVM, the inclusion of benchmark (impostor) data provides additional information about the negative samples, so that the model can learn a more accurate and converged representation of the legitimate user. However, Bi-SVM performs worse Figure 8(a). Specifically, FRR is dramatically higher than FAR. Such phenomenon is also observed in [42] and we suspect this originates from the imbalanced positive and negative samples during training and testing. Thus, we re-run the experiments using the proposed two training schemes with balanced samples.

As shown in Figure 8(c), the large difference between FAR and FRR disappears when the number of positive and negative classes is balanced. As expected, Bi-SVM (Balanced-All) and Bi-SVM (Balanced-5) obtains better performance (lower FAR/FRR and higher BAC), with BAC significantly enhanced, from 81% to 92.5%. Bi-SVM (Balanced-5) even performs slightly better as the impostor data is also balanced (10% from each). We therefore only report the results obtained with Bi-SVM (Balanced-5) in the rest of the evaluation.

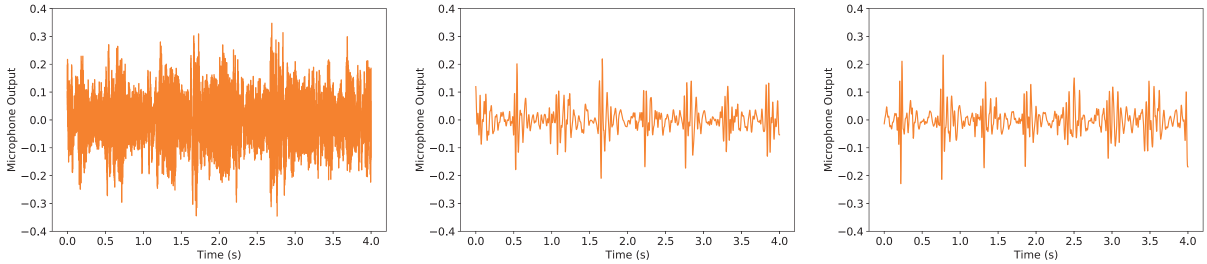
5.5 Impact of Walking Condition

The walking conditions, such as different footwear or ground material, might affect the gait identification performance. For instance, compared to tiles ground, walking on carpet will result in longer stance phase and weaker vibrations when hitting the ground. These conditions also introduce variations on human gait and therefore making the identification task tougher. Next, we investigate the impact of footwear and ground material.



(a) (b) (c)

Fig. 10: Impact of user speaking. We take one subject walking on tiles as an example, (a) original signal with speaking, (b) low-pass filtered signal with speaking, (c) filtered signal without speaking.



(a) (b) (c)

Fig. 11: Impact of music playback. We take one subject walking on tiles as an example, (a) original signal with music playing, (b) low-pass filtered signal with music playing, (c) filtered signal without music.

5.5.1 Footwear. Figure 9 shows the FAR, FRR, and BAC with data collected when the subjects were barefoot, wearing slippers and sneakers, as well as the combination of all these data, respectively. The results indicate that EarGate works well regardless of footwear, with BAC higher than 94%. Particularly, walking barefoot achieves the best performance, followed by wearing sneakers, whilst wearing slippers is the most challenging case, as slippers introduce more variations during walking. This observation is applicable to both the dataset collected on tiles and carpet. In addition, when evaluating the data collected in a single session, the BAC significantly increased from 92.46% (Figure 8(c)) to 97.26%.

5.5.2 Ground Material. Next, we explore the impact of ground material on the identification performance. As shown in Figure 9, both tiles and carpet achieve good performance, with BAC higher than 93%. In particular, for the same footwear (e.g. sneakers), tiles always obtains better performance (3% improvement on BAC) compared to carpet. This might because soft carpet counteracts part of the generated vibrations, and therefore the signal-to-noise ratio (SNR) is lower.

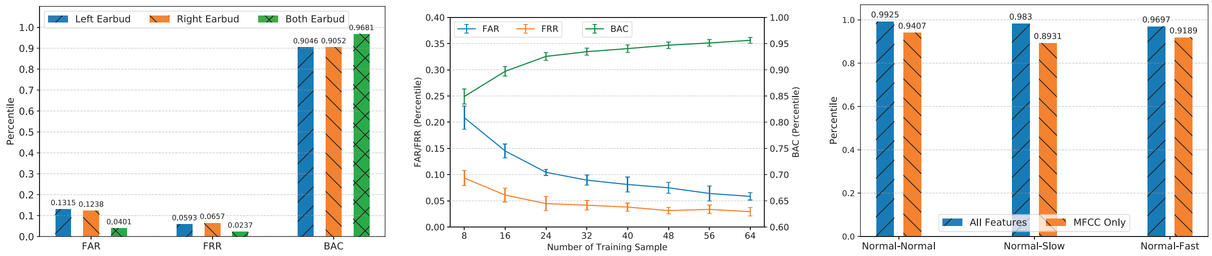
5.6 Impact of Human Speech

Human speech might have an impact on the proposed identification system. On one hand, the voice produced when people speak will be captured by the in-ear microphone, thereby polluting the recorded acoustic gait data. On the other hand, during speaking, the movement of mouth and jaw modifies the structure of human body. Consequently, the generated vibrations will experience a slightly different propagation path, resulting in different modulations on the gait signal. Thus, we asked the subjects to speak when walking on tiles with sneakers (the most common case for daily walking) and explore whether the performance would be affected.

Figure 10 plots the raw microphone data from the left earbud. It is clear that the gait signal is polluted by high-frequency human speech. Fortunately, the actual gait signal we are interested in can be distinguished in frequency domain as **(1)** the generated vibrations are in low frequencies and **(2)** the occlusion effect mainly emphasizes the low-frequency components of the bone-conducted sound. Figure 10(b) illustrates the low-pass filtered version of Figure 10(a). In addition, we plot the filtered signal of the same participant walking without speaking in Figure 10(c). Visually, the two filtered signals look quite similar and the impact of human speech has been completely removed. Using the Bi-SVM (Balanced-5) protocol, we run the experiments on the collected dataset with users speaking during walking. The results are satisfactory for both tiles (FAR=7.73%, FRR=4.37%, BAC=93.95%) and carpet (FAR=10.73%, FRR=6.98%, BAC=91.15%), denoting how human speech has little impact on EarGate .

5.7 Impact of Music Playback

The general purpose of earbuds is leisure/entertainment, particularly music playback. Due to the vicinity (less than 1 centimeter) of the speaker and in-ear microphone, we investigated whether music playback introduces interference in the gait identification system. To explore that, we asked one subject to walk when the earbud was playing music at an appropriate volume. Figure 11(a) plots the original signal from the left earbud, where the target gait signal is overwhelmed by the music. Similar to human speech, these audible sounds are in higher frequencies, while EarGate captures gait in low frequencies (<50 Hz). Thus, after applying the low-pass



(a) (b) (c)

Fig. 12: Impact of (a) fusing data from both the earbuds, (b) different training size, and (c) different pace and features, on FAR, FRR, and BAC.

filter, music can be eliminated and gait signal of interest is clearly observed in Figure 11(b). Also, we plot a trace (low-pass filtered) of the same subject walking without playing the music in Figure 11(c), which shows high similarity with signals in Figure 11(b).

To further demonstrate how EarGate is robust against different types of music, we conducted a spectrum analysis of the All-time Top 100 Songs launched by Billboard [2]. Specifically, we applied fast Fourier transform (FFT) analysis on each song and obtain the signal energy at each frequency band. Then, the energy of frequency band lower than 50 Hz was summed and we calculated the energy ratio by $E_{<50}/E_{total}$ ($E_{<50}$ refers to the total energy at frequencies below 50 Hz and E_{total} refers to the total energy of the song). Averaged among the 100 songs, only 1.5% energy is distributed under 50 Hz, suggesting that the impact of music playback is negligible for EarGate . We also considered the impact of phone calls. Evidence from the literature shows that the frequency range of human voice over telephony transmission is within 300-3400 Hz [15], which can be easily removed after low-pass filtering. Thus, EarGate is compatible with the general purpose of earbuds, without introducing any mutual interference. Besides, given many existing earbuds already features in-ear microphones for active noise cancellation, EarGate can be operated concurrently. Specifically, the data from in-ear microphones can be delivered to two independent pipelines: noise cancellation algorithm to actively generate the anti-noise and user identification framework to recognize the user.

5.8 Sensor Multiplexing

Unlike other wearable devices, earbuds possess a unique advantage of being able to sense with both the left and right earbud simultaneously, providing a sensor multiplexing gain. Taking the data collected when subjects walked on tiles with sneakers as an example, we study the performance of using solely the left or right earbud, as well as the achievable gain with both earbuds. As presented in Figure 12(a). when a single earbud is used, the left and right one achieves quite similar performance on the three metrics, suggesting both of them are effective in detecting user' gait. When combining the features extracted from both earbuds, the BAC lifts from 90.5% to 96.8%, indicating a great multiplexing gain when using two earbuds.

5.9 Impact of training size

To assess the overhead of EarGate during enrollment phase, we study the impact of training data size on the identification performance as the minimal number of gait cycles required for training is the actual burden on users. Specifically, we select 80 gait cycles for each subject from the tiles-sneakers dataset and run the experiment with Bi-SVM (Balanced-5) protocol. 20% (16) gait cycles are fixed as the testing dataset to ensure a fair comparison. Then, we continuously increase the number of training samples from 10% (8) to 80% (64). The experiment runs when one subject is iteratively selected as the legitimate user and Figure 12(b) shows the averaged performance over the 31 subjects. Obviously, the performance improves with the increase of training samples. The BAC reaches 93.5% when 32 gait cycles are used for training. Based on the walking speed of participants, the corresponding overhead for data collection is around 30 s continuous walk, which would be acceptable for most people.

In addition, EarGate can adapt its identification model by re-training it with new gait samples. For example, if a new gait cycle is recognized as positive/negative with high confidence, it can be added to the training set for model re-training. Thus, the performance of EarGate is expected to improve continuously after it has been put into practical use.

5.10 Transfer Learning

We sought to improve the performance of EarGate by looking at more complex, deep Learning based approaches. Although the performance of traditional deep learning techniques are hindered by the modest size of our dataset (the number of available data points is not sufficient for the model to properly learn the weights[3]), transfer learning may come handy [31]. In doing so, we leveraged VGGish, an audio-specific pre-trained model released by Google [20]. VGGish is a convolutional neural network (CNN) that is trained using a large-scale YouTube audio dataset. Concretely, we use VGGish to automatically extract audio features (*embeddings*) from the raw in-ear audio data. At this point, we combined the handcrafted features with the embeddings extracted by VGGish and trained two different classifiers: SVM (as in our previously reported results) and XGBoost, an advanced decision-tree based machine learning algorithm [12]. We report our findings in Table 1. As we can appreciate from Table 1, XGBoost consistently outperforms SVM when solely using the VGGish embeddings, without any handcrafted feature. Conversely, by feeding handcrafted features to SVM we achieve better results (both in terms of BAC as well as, in most cases, in terms of FAR and FRR). Notably, by combining the embeddings extracted with transfer learning and the manually engineered features, we can further boost the performance of EarGate, increasing the BAC

[3] Considering VGGish as an example, the number of trainable parameters is 4,499,712.

Table 1: Identification performance improvements achieved with transfer learning.

Tasks	Type of Features	SVM			XGBoost		
		FAR	FRR	BAC	FAR	FRR	BAC
Brick Barefeet	VGGish embeddings	0,730	0,231	0,520	0,093	0,092	**0,907**
	Handcrafted Features	0,081	0,023	**0,948**	0,077	0,070	0,927
	VGGish embeddings + hancrafted features	0,069	0,017	**0,957**	0,065	0,067	0,934
Brick Shoe	VGGish embeddings	0,645	0,333	0,511	0,133	0,088	**0,889**
	Handcrafted Features	0,048	0,022	**0,965**	0,060	0,059	0,940
	VGGish embeddings + hancrafted features	0,032	0,022	**0,973**	0,073	0,055	0,936
Brick Slipper	VGGish embeddings	0,640	0,316	0,522	0,052	0,113	**0,917**
	Handcrafted Features	0,085	0,026	**0,945**	0,071	0,064	0,932
	VGGish embeddings + hancrafted features	0,056	0,016	**0,964**	0,048	0,073	0,940
Carpet Barefeet	VGGish embeddings	0,716	0,202	0,541	0,132	0,147	**0,861**
	Handcrafted Features	0,096	0,034	**0,935**	0,089	0,101	0,905
	VGGish embeddings + hancrafted features	0,052	0,036	**0,956**	0,085	0,097	0,909
Carpet Shoe	VGGish embeddings	0,589	0,378	0,517	0,160	0,126	**0,857**
	Handcrafted Features	0,116	0,051	**0,917**	0,104	0,085	0,906
	VGGish embeddings + hancrafted features	0,056	0,048	**0,948**	0,093	0,086	0,910
Carpet Slipper	VGGish embeddings	0,548	0,404	0,524	0,129	0,141	**0,865**
	Handcrafted Features	0,069	0,043	**0,944**	0,115	0,107	0,889
	VGGish embeddings + hancrafted features	0,050	0,038	**0,956**	0,130	0,104	0,883

and lowering both FAR and FRR. This is due to the model learning from both the carefully selected features as well as from the more abstract representation of the data generated by VGGish. Interestingly, while in terms of accuracy the benefit of transfer learning is clear, it is also important to bear in mind the accuracy versus power consumption trade-off. The performance achievable with only the handcrafted features are indeed only marginally lower than those achieved by combining the transfer learning embeddings and the manually crafted features, in face of an inevitably higher power consumption (caused by the execution of the VGGish model). However, if a larger dataset is available, the performances associated with transfer learning could further improve.

5.11 Contribution of Specific Features

Our classifier was trained on a variety of features, including many related to the frequency spectrum. For an evaluation of our scale, it is likely that walking cadence will be distinct for all participants and, therefore, there is the risk that the model learned to strongly weight the frequency features closely associated with people's cadence. This would be problematic because walking cadence will not be distinct on the broader population level, and furthermore people move at different cadences (e.g., when walking alongside someone).

To confirm that our classifier was not simply a cadence classifier, we asked one subject to walk at three *speeds*: slow (1.59 step/s), normal (1.97 step/s), fast (2.13 step/s). After having trained the model with normal-speed walking data, we tested it on all the three instances. Figure 12(c) reports the BAC for normal, slow, and fast pace respectively: 99.25%, 98.30%, and 96.97% BAC. These results clearly show that the model is not biasing towards cadence as the identifier, and is instead learning something more fundamental to the user's movement.

Given this we also considered the value of the different features. After training our classifier individually on each of the features reported in Section 3.3, we plot in Figure 12(c) the BAC achieved training our system with only Mel-Frequency Cepstral Coefficients (MFCC) as well as with all the features. Notably, among all the features taken individually, MFCC achieved the best results. Interestingly, when analyzing the computation time of different features, we found that most of the feature extraction time (89%) is actually consumed on a feature called 'tonnetz' (with 6 values). So, we repeated the experiment after removing this feature and the identification accuracy almost remains the same. Therefore, it is possible to use a smaller set of features and reduce the overall end-to-end latency of the system.

6 SYSTEM CONSIDERATIONS

To gauge the system-level performance of EarGate, we conducted a power-consumption and latency investigation using the same prototype described in Section 4. We assume the model is pre-trained and only consider the real-time identification overhead. Given EarGate can either run the identification framework on-device, or offload it, we consider three different schemes which would impact differently power consumption and latency:

1. **On-device identification**: microphone data recording (MicRecd), as well as all the gait identification procedures, including low-pass filtering (LowPassFilt), feature extraction (FeatExtr), and inference, are performed on-device.
2. **Distant identification (raw data offloading)**: the earable records microphone data and directly transmit the raw data via WiFi or Bluetooth (BT), without any processing. The size of raw data is 16 KB [4]. The identification process would be performed on the cloud or companion smartphones and possibly the result will be communicated back to the device.
3. **Distant identification (features offloading)**: both low pass filtering and feature extraction are carried out on the earable, right after recording the microphone data. Only the extracted features (either all features or only MFCC features)

[4] With a sampling rate of 4000 Hz and gait cycle length of 1 second, the data from two microphones is $2 \times 4000 \times 2B = 16\ KB$.

Table 2: Power consumption and latency measurement of EarGate.

Scheme	Operation	Power (mW)	Latency (ms)	Energy (mJ)
On-device identification	MicRecd	120	1000	
	LowPassFilt	635	1.83	168.59 (All)
	FeatExtr (All/MFCC)	655/651	71.98/23.62	136.82 (MFCC)
	Inference	644	0.44	
Raw Data Offloading	MicRecd	120	1000	
	TX [OS+Air] (WiFi)	334	9.49+12.8	123.17 (WiFi)
	TX [OS+Air] (BT)	478	148.41+128	190.94 (BT)
Feature Offloading	MicRecd	120	1000	
	LowPassFilt	635	1.83	168.91 (WiFi, All)
	FeatExtr (All/MFCC)	655/651	71.98/23.62	172.27 (BT, All)
	TX [OS+Air] (WiFi)(All/MFCC)	332	1.81+0.59/0.39+0.26	136.67 (WiFi, MFCC)
	TX [OS+Air] (BT)(All/MFCC)	457	8.66+5.94/5.74+2.56	139.26 (BT, MFCC)

are transmitted via WiFi/BT and the result might be communicated back to the device. The size of features is 0.748 KB for all features and 0.16 KB for MFCC features [5].

Our aim here is to show the flexibility of our system to work, irrespective of the network architecture preferred or available.

For the offloading cases, we consider two types of radio frequency (RF) communications: Bluetooth (BT) and WiFi, as they are (or will soon be) the commonly available radio chips in earables. For WiFi, we consider a typical uplink throughput of 10 Mbps (similar to that of a domestic network) to compute the transmission latency over the air. Regarding Bluetooth, the version supported by the Raspberry Pi we use is BT 4.1. Compared to more recent BT standards (e.g., Bluetooth 5, available in Apple AirPods Pro), BT 4.1 offers less throughput (1 Mbps instead of 2 Mbps). As a consequence of that, the over-the-air transmission time reported is longer than what it would be if a more advanced version of BT were adopted. We measure the power consumption with a USB power meter. Latency measurements are obtained timing the execution of the software handling the operation of interest. The results are averaged over multiple measurements and presented in Table 2. The baseline power consumption of our Raspberry Pi (idle) is around 2,325 mW and the values reported in the table are additional power consumption. The energy column computes the total energy required to perform the operations for one gait cycle.

On-device identification performs the whole pipeline including microphone recording (MicRecd), filtering (LowPassFilt), feature extraction (FeatExtr), and inference on the device. The power column indicates that numerical computations (LowPassFilt, FeatExtr, and Inference) are intensive and more power-hungry than microphone recording. The latency column shows that most of the processing time is spent on feature extraction. Regardless of the time for data acquisition, the overall identification latency is within 100 ms. Concretely, the energy required for a one-time on-device identification is 168.58 mJ (using all features) and 136.82 mJ (using MFCC only, the most effective feature as described in Section 5.11).

When offloading the raw data to the cloud, only MicRecd is performed on-device. Here, we consider the latency as the sum of TX OS (the time that Pi requires to write the data in the buffer of the chosen interface, either Bluetooth or WiFi) and TX Air (the time it takes for the data to propagate over-the-air). We can observe that the latency is largely dependent on the throughput of the network. The energy required for WiFi offloading is 123.17 mJ, whilst for BT is 190.94 mJ. Conversely, when offloading the features, also LowPassFilt and FeatExtr are done on-device. Here, following Section 5.11, we compare the energy efficiency of sending all the features or just the MFCC features. Given there are 187 features in total and only 40 are MFCC features, the transmission latency for MFCC features only is shorter, overall below 100 ms.

In summary, we can conclude that (1) both on-device identification and features offloading guarantee a time delay (after data acquisition) of less than 100 ms, showing EarGate can work in real-time scenarios both on-device and while offloading features; (2) with respect to the energy consumption, all the schemes consume less than 200 mJ for one-time identification; (3) the latency and energy consumption for raw data offloading are strictly dependent on the quality of the communication link. Thus, offloading raw data would be the best option when the network is stable. Notably, in this section we focus on the SVM-based pipeline, only considering the case where handcrafted features are used to train the model. The reasons we do that are manifold. First, at the moment, deploying a complex model like VGGish on resource-constrained devices (like earbuds) is an extremely challenging task and there are no publicly available tools to supporting it. Second, given the performance improvements of the transfer learning approach over the handcrafted features one are marginal (Section 5.10). Third, running VGGish to extract the embeddings would entail far more operations than simply leveraging a limited set of handcrafted features to train SVM and, therefore, it is fair to assume that power and latency figures would be considerably higher than those for the traditional machine learning pipeline.

7 DISCUSSION

In this section, we talk through the limitations of our current work, the possible improvements, and the future directions we plan on pursuing. Despite the promising results, we are aware of some of the limitations of our approach. First and foremost, in order for our system to work, initial user data (in the enrollment phase) are required. While we acknowledge it would be ideal if such a phase did

[5] All features includes 187 feature from each microphone data and the size is $2 \times 187 \times 2B = 0.748\ KB$, MFCC features includes 40 features and the size is $2 \times 40 \times 2B = 0.16\ KB$.

not have to take place, most of the well-established biometric identification schemes, like facial recognition and finger-print, do require some initial user data. Besides, in our evaluation (Figure 12(b)) we show we only need a limited amount of user data to start offering acceptable performance.

Second, whilst Figure 9 shows the impact of different footwear is marginal, when the model is trained on them, to maintain high performance, the model should be partially re-trained to add new pairs of shoes to the legitimate user data. This could be done by the user walking and manually committing the new gait cycles to re-train the model; or it could happen in background if the user changes shoes while wearing the earable (provided the user has been successfully identified by the earable). We believe the latter is a reasonable assumption, especially for hearing aid users as such devices are continuously worn throughout the day. Hopefully, this will soon be the case for leisure earables, too, which, once the advancements in materials will guarantee better comfort, could also be worn for a longer amount of time. Further, although gait may slightly change over the years, continuously adapting the model (like we do for different shoes) could relieve this issue, too. Alternatively, the impact of different footwear could be obviated by mean of the combination of a general model and a personalized model. For instance, an auto-encoder [22] could be trained on all the users' data to obtain a general model.

Lastly, one other concern could be related to the obstruction of the ear canal orifice, and the consequent impact on audible sounds (which will result muffled due to the presence of an obstructing body). We are aware this could be a potential safety issue, therefore, similarly to the AirPods Pro *Transparency Mode*, it is possible to do the same with EarGate. By playing back the audio recorded by the external mic, the use will be able to hear as if there were no earbuds obstructing their ear canal. Notably, this does not affect our system as we can easily filter it out (leveraging the difference in frequency), like we did for music. In addition, we believe the power consumption and latency can be further reduced when specialized audio chips (e.g., Apple H1 Chip) are used and more advanced engineering work (e.g., dedicated PCB design) is implemented.

8 RELATED WORK

Gait is a well-studied human biometric, proven to be unique for each individual, and, therefore, often used as an identification biometric [26, 37]. Traditional approaches for gait recognition enumerate machine vision-based approaches [41], floor sensor-based techniques [27], wireless fingerprinting based method [38], and wearable sensor-based methods [17, 25]. These approaches own specific advantages (e.g., zero user effort and complete device-free) and disadvantages (high computation cost and privacy issue for vision-based method, and requirement of deploying the wireless transceivers for wireless fingerprinting based method), thereby complementing each other in different scenarios.

Different from the more traditional techniques, there are two works using acoustic as the modality for gait recognition. Geiger et al. [19] exploited a microphone attached to the foot to record the walking sounds when human feet hit the ground. The main drawback is that the step sounds change with different materials of the ground or shoe sole. In the extreme case, the microphone may not be able to observe noticeable sound when walking on the carpet with barefoot. Instead, our work leverages the occlusion effect to measure the bone-conducted sounds (essentially vibrations) in the ear canal, which are robust to footwear and ground material. Wang et al. [39] proposed a fingerprinting-based system (called Acoustic-ID) for human gait detection using acoustic signals. Specifically, by deploying a pair of acoustic transmitter (ultrasound) and receiver, gait pattern is extracted by measuring the reflected acoustic variations when human is walking within the sensing range. Unlike Acoustic-ID that requires to actively transmit ultrasound, our approach is completely passive and does not require the deployment of any additional hardware.

Instead of using gait, researchers have discovered other biometric acquired from human ear for person identification. Nakamura et al. [30] proposed to identify and authenticate users with in-ear electroencephalogram (EEG) measured by a customized earpiece. More recently, EarEcho [18] leverages the uniqueness of ear canal geometry to recognize the users. Our work differs from EarEcho in three aspects. First, the operation rationale is different. EarEcho is based on the uniqueness of the geometry of the ear canal, while EarGate leverages the uniqueness of the human gait to identify the user. Second, EarEcho requires the active transmission of a sound/ultrasound pulses so as to measure the echo sound, while EarGate is completely passive. Third, as demonstrated in [9], the geometry of the ear canal will change under different facial expressions, which might impact the effectiveness of EarEcho in daily life. Although gait can also change over time, it usually evolves in a large time span, so that the user can periodically adapt the trained model. Moreover, the uniqueness of ear canal geometry has not been tested with a large population (only 20 subjects involved in [18]). Ultimately, with this work, we propose the first earable-based acoustic-gait identification system, showing its potential applicability to various identification use cases.

9 CONCLUSION

We presented EarGate, an earable identification system based on user gait. Exploiting the occlusion effect, EarGate enables detection of human acoustic gait from an in-ear facing microphone. Experimenting with 31 subjects, we demonstrated that EarGate achieves robust and acceptable identification performance (up to 97.26% BAC, with low FAR and FRR of 3.23% and 2.25% respectively) under various practical conditions. Moreover, EarGate will not affect the general functionality of earbuds and is robust to high-frequency noises like music playback and human speech. We envision that EarGate can be an effective and robust way of identifying (standalone or companion with smartphone) earables. Particularly, its unobtrusiveness makes it an appealing replacement of FaceID for users wearing face masks during the COVID-19 pandemic.

10 ACKNOWLEDGMENTS

This work is supported by ERC through Project 833296 (EAR) and by Nokia Bell Labs through a donation. We thank D. Spathis and L. Qendro for the insightful discussions, the anonymous shepherd and reviewers for the valuable comments, and the volunteers for the data collection.

REFERENCES

[1] Online. AirPods Pro. https://www.apple.com/uk/airpods-pro/. (Accessed on November 12, 2020).

[2] Online. Billboard All-Time Top 100 Songs. https://www.billboard.com/articles/news/hot-100-turns-60/8468142/hot-100-all-time-biggest-hits-songs-list. (Accessed on November 12, 2020).

[3] Online. Honor Magic Earbuds. https://www.hihonor.com/global/products/accessories/honor-magic-earbuds/. (Accessed on November 12, 2020).

[4] Online. Librosa. https://librosa.org/. (Accessed on November 12, 2020).

[5] Online. Microphone SPU1410LR5H-QB. https://www.mouser.com/datasheet/2/218/SPU1410LR5H-QB-215269.pdf. (Accessed on November 12, 2020).

[6] Online. MINISO Marvel Earphones. https://www.miniso-au.com/en-au/product/145169/marvel-earphones/. (Accessed on November 12, 2020).

[7] Online. ReSpeaker Voice Accessory HAT. https://wiki.seeedstudio.com/ReSpeaker_4-Mic_Linear_Array_Kit_for_Raspberry_Pi/. (Accessed on November 12, 2020).

[8] M Umair Bin Altaf, Taras Butko, and Biing-Hwang Fred Juang. 2015. Acoustic gaits: Gait analysis with footstep sounds. *IEEE Transactions on Biomedical Engineering* 62, 8 (2015), 2001–2011.

[9] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial expression recognition using ear canal transfer function. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 1–9.

[10] Matei-Sorin Axente, Ciprian Dobre, Radu-Ioan Ciobanu, and Raluca Purnichescu-Purtan. 2020. Gait Recognition as an Authentication Method for Mobile Devices. *Sensors* 20, 15 (2020), 4110.

[11] Kévin Carillo, Olivier Doutres, and Franck Sgard. 2020. Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation. *The Journal of the Acoustical Society of America* 147, 5 (2020), 3476–3489.

[12] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[13] Romit Roy Choudhury. 2021. Earable Computing: A New Area to Think About. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 147–153.

[14] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* 28, 4 (1980), 357–366.

[15] D Esteban, C Galand, Daniel Mauduit, and J Menez. 1978. 9.6/7.2 kbps voice excited predictive coder (VEPC). In *ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3. IEEE, 307–311.

[16] Andrea Ferlini, Alessandro Montanari, Andreas Grammenos, Robert Harle, and Cecilia Mascolo. 2021. Enabling In-Ear Magnetic Sensing: Automatic and User Transparent Magnetometer Calibration. *The 19th International Conference on Pervasive Computing and Communications (PerCom 2021)* (2021).

[17] Davrondzhon Gafurov, Kirsi Helkala, and Torkjel Søndrol. 2006. Biometric Gait Authentication Using Accelerometer Sensor. *JCP* 1, 7 (2006), 51–59.

[18] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[19] Jürgen T Geiger, Maximilian Kneißl, Björn W Schuller, and Gerhard Rigoll. 2014. Acoustic gait-based person identification using hidden Markov models. In *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop*. 25–30.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[21] Andrew H Johnston and Gary M Weiss. 2015. Smartwatch-based biometric gait recognition. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–6.

[22] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A Hierarchical Neural Autoencoder for Paragraphs and Documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1106–1115.

[23] Qi Lin, Weitao Xu, Guohao Lan, Yesheng Cui, Hong Jia, Wen Hu, Mahbub Hassan, and Aruna Seneviratne. 2020. KEHKey: Kinetic Energy Harvester-based Authentication and Key Generation for Body Area Network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.

[24] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. OESense: employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 175–187.

[25] Dong Ma, Guohao Lan, Weitao Xu, Mahbub Hassan, and Wen Hu. 2020. Simultaneous Energy Harvesting and Gait Recognition using Piezoelectric Energy Harvester. *IEEE Transactions on Mobile Computing* (2020).

[26] Maria De Marsico and Alessio Mecca. 2019. A survey on gait recognition via wearable sensors. *ACM Computing Surveys (CSUR)* 52, 4 (2019), 1–39.

[27] Lee Middleton, Alex A Buss, Alex Bazin, and Mark S Nixon. 2005. A floor sensor system for gait recognition. In *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*. IEEE, 171–176.

[28] Bendik B Mjaaland, Patrick Bours, and Danilo Gligoroski. 2010. Walk the walk: Attacking gait biometrics by imitation. In *International Conference on Information Security*. Springer, 361–380.

[29] Muhammad Muaaz and Rene Mayrhofer. 2017. Smartphone-based gait recognition: From authentication to imitation. *IEEE Transactions on Mobile Computing* 16, 11 (2017), 3209–3221.

[30] Takashi Nakamura, Valentin Goverdovsky, and Danilo P Mandic. 2017. In-ear EEG biometrics for feasible and readily collectable real-world person authentication. *IEEE Transactions on Information Forensics and Security* 13, 3 (2017), 648–661.

[31] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[32] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. 2019. STEAR: Robust Step Counting from Earables. In *Proceedings of the 1st International Workshop on Earable Computing*. 36–41.

[33] Roman Schlieper, Song Li, Stephan Preihs, and Jürgen Peissig. 2019. The Relationship between the Acoustic Impedance of Headphones and the Occlusion Effect. In *Audio Engineering Society Conference: 2019 AES INTERNATIONAL CONFERENCE ON HEADPHONE TECHNOLOGY*. Audio Engineering Society.

[34] Stefan Stenfelt. 2011. Acoustic and physiologic aspects of bone conduction hearing. In *Implantable bone conduction hearing aids*. Vol. 71. Karger Publishers, 10–21.

[35] Stefan Stenfelt and Sabine Reinfeldt. 2007. A model of the occlusion effect with bone-conducted stimulation. *International journal of audiology* 46, 10 (2007), 595–608.

[36] Michael A Stone, Anna M Paul, Patrick Axon, and Brian CJ Moore. 2014. A technique for estimating the occlusion effect for frequencies below 125 Hz. *Ear and hearing* 35, 1 (2014), 49.

[37] Changsheng Wan, Li Wang, and Vir V Phoha. 2018. A survey on gait recognition. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–35.

[38] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 363–373.

[39] Yingxue Wang, Yanan Chen, Md Zakirul Alam Bhuiyan, Yu Han, Shenghui Zhao, and Jianxin Li. 2018. Gait-based human identification using acoustic sensor and deep neural network. *Future Generation Computer Systems* 86 (2018), 1228–1237.

[40] Yifan Zhang, Shuang Song, Rik Vullings, Dwaipayan Biswas, Neide Simões-Capela, Nick Van Helleputte, Chris Van Hoof, and Willemijn Groenendaal. 2019. Motion artifact reduction for wrist-worn photoplethysmograph sensors based on different wavelengths. *Sensors* 19, 3 (2019), 673.

[41] Guoying Zhao, Guoyi Liu, Hua Li, and Matti Pietikainen. 2006. 3D gait recognition using multiple cameras. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. IEEE, 529–534.

[42] Yongpan Zou, Meng Zhao, Zimu Zhou, Jiawei Lin, Mo Li, and Kaishun Wu. 2018. BiLock: User authentication via dental occlusion biometrics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–20.