

PA2BLO: Low-Power, Personalized Audio Badge

Hemanth Sabbella*, Dulaj Sanjaya Weerakoon*, Manoj Gulati†, Archan Misra*

* School of Computing and Information Systems, Singapore Management University

† School of Computing, National University of Singapore

Email: *hemanthrs@smu.edu.sg, *sanjayawm@smu.edu.sg, †manojg@nus.edu.sg, *archanm@smu.edu.sg

Abstract—We present the hardware design and software pipeline for an ultra-low power device, in the form factor of a wearable badge, that supports energy efficient sensing, processing and wireless transfer of human voice commands and interactions. The proposed system, called *PA2BLO*, is envisioned to support both: (a) real-time, scalable, authorized voice based interaction and control of devices and appliances, and (b) longitudinal, low-power logging of natural voice interactions. *PA2BLO* introduces two key novel capabilities. First, it includes a low power, low-complexity voice authentication module that is able to reliably authenticate an authorized user only using low sampling rate (500 Hz) audio data. Second, to reduce concerns around inadvertent leakage of voice biometrics to less secure voice-driven services, *PA2BLO* uses a power-efficient, randomized pitch shifting technique that dramatically lowers the ability to perform speaker recognition while preserving instruction/speech comprehensibility. We describe *PA2BLO*'s Cortex M4F-based micro-controller based hardware implementation, which is carefully designed to eliminate redundant processing and consumes less than 50J of energy per hour of active voice capture and processing. Through both controlled and naturalistic studies, we show that the *PA2BLO* prototype is capable of authenticating user voice segments reliably (accuracy > 89.8%) and can operate for well over a day (using a supercapacitor charged within just one minute) while capturing 2+ hours of active speaker data.

Index Terms—Sub-kHz Speaker Authentication, Speech Anonymization, Ultra-low Power Wearable Badge

I. INTRODUCTION

Wearable devices, placed on various body locations including the wrist [1], [2], ear [3], [4], fingers [5] and around the neck [6], have been used to capture a variety of human activity context, such as exercising and eating. Smart badges offer an interesting, ergonomic wearable form factor, as they can be easily attached to various parts of a human's clothing, including the chest, waist and collar. Smart-badge based wearable sensors have been used, for example, to (a) monitor various kitchen activities [7], (b) track physical and sleep activity [8] and (c) sense face-to-face interactions [9]. However, the need to periodically charge their batteries, for a duration lasting an hour or longer, creates adoption challenges, especially among the elderly [10].

This paper presents the design and evaluation of *PA2BLO*¹, a wearable, ultra-low power badge with an embedded microphone sensor capable of capturing, authenticating and transferring an authorized user's voice commands or speech snippets. In addition, to mitigate growing concerns about generative AI-based cloning of voiceprints [11],

PA2BLO supports privacy-compliant operation, where the wearable device performs speaker authentication but subsequently transmits only *de-personalized*, but comprehensible, voice commands. To our knowledge, ours is the *first ultra-low power (~14 mW) wearable "badge" that is capable of performing voice-activated processing and transfer of audio streams, enabling several hours of operation using a rapidly-charged 15F supercapacitor with a capacity of ~156J*. (In fact, *PA2BLO*'s low-power consumption theoretically translates into ~15+ days of continuous operation, assuming 4 hours/day of active voice commands, using a ~900 mWh smartwatch rechargeable battery.) While past work, such as WiWear [12] or Solar [13], has demonstrated the possibility of ultra-low power sensing on wearable devices, our achievement is notable as it involves the processing and transmission of audio data, which involves a significantly higher sampling frequency (16 KHz and higher) with a greater computational complexity.

PA2BLO adopts an intermittent, event-driven sensing paradigm [14], where the badge micro-controller operates in an ultra-low power mode (<40 μ W) until it detects a signal *matching* the voice of a single, authorized user. In addition, the badge includes a BLE radio to support energy-efficient transfer of the voice signals. To support ultra-low energy consumption (an average of ~48J for capturing a full one hour of human speech, as detailed in Section VI-B), *PA2BLO* utilizes and integrates two novel concepts: (a) First, it explicitly separates the step of user voice authentication from that of processing an authorized individual's voice, and performs authentication using a personalized model using only low sampling frequency (500 Hz) audio. This helps to avoid the unnecessary execution of the full audio processing pipeline, over 16 KHz data, for extraneous audio signals (e.g., background noise, speech by another person). (b) Second, to support de-personalized voice-based interaction, *PA2BLO* employs a highly-optimized, *pitch shifting* technique that preserves intelligibility while masking key features of an individual's voiceprint. Using an ARM Cortex M4-based hardware implementation, we shall show how *PA2BLO* successfully addresses the combined challenges of ultra-low power operation, voiceprint obfuscation and comprehension-preserving wireless communication with voice-enabled services (such as Amazon's Alexa).

PA2BLO's design helps support two classes of voice-based applications (further detailed in Section III-A), while overcoming key concerns on scalability and privacy. The first case involves *authenticated interaction* with stationary devices such as a voice-enabled home controller or a voice-

¹Personalized Authenticated Audio Badge with Low-Energy Operation

controlled semi-public IoT device. For example, an elderly person currently in the bedroom may wish to remotely instruct the Amazon Echo Dot located in her living room to “dim down the kitchen lights” or an authorized employee can instruct a coffee maker located in a public lounge to “prepare an espresso”. *PA2BLO* enables these interactions by either transferring the voice command beyond the audible range or by absolving the IoT device from performing audio authentication over a potentially large user pool (100s or 1000s of users). The second case involves pervasive logging and *offline profiling* of natural voice-based interactions—for example, using daily conversation data to profile an individual’s emotional states.

Key Contributions: We make the following key contributions:

- *Low-Power, Low Sampling Rate Voice Authentication:* We develop a low-power speaker authentication module that utilizes an audio stream sampled only at 500 Hz (far lower than the 16 KHz sampling rate for audio applications). The module utilizes a low complexity (4-stage), individualized neural network model that learns each individual’s *distinct* low-frequency voice features, across both male and female speakers. Experimental studies with our prototype show that this authentication module consumes only 13.59 mW of sampling + processing power, achieving an authentication F1-score of ~ 0.97 over 3 sec audio segments.
- *Low-Power, Randomized Pitch Shifting:* To support reliable but privacy-preserving interaction by a user with a potentially untrusted voice-enabled service, we utilize the concept of randomized pitch shifting, where the frequency components of individual 16 msec segments of audio stream are shifted by a random value between 0-200 Hz. We show how such shifting is accomplished through cheaper time domain multiplication operations (instead of expensive FFT/IFFT computations), resulting in $\sim 30x$ savings in processing latency and energy.
- *Prototype PA2BLO Design:* We embed the concepts of low frequency authentication and pitch shifting into a fully-functional *PA2BLO* prototype based on the XIAO nRF52840 micro-controller. Our prototype is powered by (a) a 15F super-capacitor that can be fully charged by near-field charging within 1 min and yet provides enough energy for 24-hour standby operation, (b) an ultra-low-power, wake-on microphone that allows the Cortex M4F processor to be asleep during periods of ambient silence, and (c) a BLE 5.0 wireless transceiver that transfers voice commands/segments over 10m distances. Using controlled experimental studies, we show that this prototype is capable of capturing and transmitting between 1.1-2.4 hours of *active*, authenticated voice data using a single charge, with an overall average power consumption of 13.3 mW.
- *User-Study based validation of PA2BLO:* We demonstrate *PA2BLO*’s overall reliability and usability via a realistic user study with 6 participants grouped into 3 pairs, with each pair of individuals wearing their own audio-badge while performing a set of collaborative, interactive tasks. We show that the *PA2BLO* prototype is (a) *accurate*, isolating and

capturing each individual’s voice samples accurately (accuracy= 89.8% and 99.8%, over 1 sec and 3 secs voice snippets respectively), (b) *energy-efficient*, consuming an average of $\sim 60J$ over a 30-minute session, and (c) *highly usable*, with participants providing an average System Usability Scale (SUS) rating= 86.4.

II. RELATED WORK

We discuss prior work that encompasses smart wearables (especially badges), lightweight audio classification models, and ultra-low power wearable systems.

Wearables & Badges: Prior works on smart wearables, such as [7], [15]–[17] explore the use of touch and 6-DoF motion as input modalities for human activity recognition and for interacting with smart objects. [7] utilizes sensors embedded in a smart badge for recognizing kitchen-related activities, utilizing Convolutional Neural Networks (1D-CNN and 2D-CNN) that concatenate input from multiple sensors. The badge consumes a current of 154mA. [18] and [16] focus on earable systems, with [18] performing voice activity recognition using bone conduction microphones, while [16] utilizes brain biosignals and face/eye movement markers to perform microsleep detection. While the power consumption reported in [18] is fairly low (5.11 mW) when actively executing an RNN model on the Apollo MCU, it excludes the significant additional energy needed for wireless data transmission.

Optimized Audio Processing: A relatively large body of work has addressed the challenge of deep learning based speaker/audio processing on resource-constrained, embedded devices. [19]–[21] discuss speaker identification in resource-constrained hardware such as wearables, smartphones, and smartwatches. [21] utilizes multi-processor hardware, on a smartphone device, for speaker identification and reports a power consumption of 771mW. [20] explores the time-domain and frequency-domain feature selection for wearables having limited energy budget and computational capabilities, performing speaker verification with a power of 263.05 mW. To support higher-layer speech based tasks, such as question answering, the DeQA system [22] utilizes a set of memory optimizations to achieve a 13x decrease in the latency of executing question answering on mobile devices. However, these systems are not amenable to execution using a capacitor-powered, wearable device.

Low-Power Audio & Battery-less Devices: Approaches for low-power audio sensing systems include SPIDR [23], which uses an acoustic sensor+ speaker for depth sensing with $\sim 10mW$ power, and [24], which distinguishes between speech and non-speech segments with a power of $0.4\mu W$. A variety of work has also investigated the possibility of battery-less pervasive sensing via energy harvesting. The Flicker platform [25] utilized multiple energy harvesting modes, such as solar and kinetic, to operate a variety of sensors (e.g., barometer, temperature). The WiWear wrist-worn wearable [12] demonstrated the use of directional WiFi transmissions to charge a capacitor, which was subsequently used to intermittently capture (but not transmit) data from a low-power accelerometer ($6\mu A$ current

draw at 50 Hz). More recently, the Solar platform [13] utilized a wrist-worn solar cell both for energy harvesting and for activity recognition. However, unlike our proposed *PA2BLO* smart badge, none of these devices perform authentication or transmission of *audio* data.

III. MOTIVATING SCENARIOS & *PA2BLO* SYSTEM LEVEL ARCHITECTURE

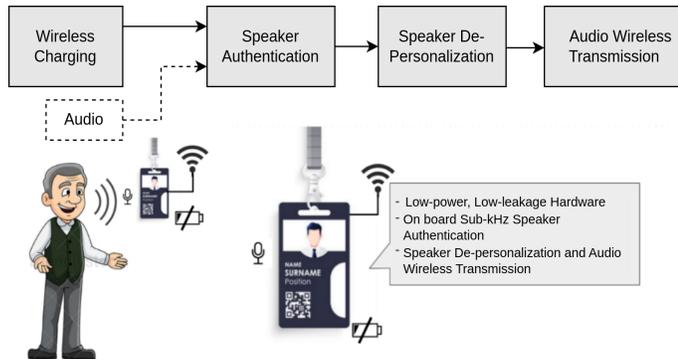


Fig. 1: *PA2BLO* Functional Architecture & Components

Before detailing the algorithmic components and hardware details of *PA2BLO*, we first outline some intended uses of *PA2BLO*, as well as its high-level architectural components.

A. Motivating Scenarios

PA2BLO's ability to capture and transmit an authenticated individual's speech/audio stream can support a variety of innovative applications.

Range-Extension to Home Voice Assistant Devices: By capturing and transmitting a user's voice segments over a distance of $\sim 10\text{m}$, *PA2BLO* enables an individual to securely issue voice commands to stationary voice assistant-based devices (such as Amazon Echo, or Google Nest) over longer distances. Using *PA2BLO*, an elderly person can issue instructions to a home device controller (e.g., to turn down the lights in a different room) even while being out of earshot. Being located on the speaker's body, *PA2BLO* is likely to capture the voice instructions more accurately, even in a noisy ambient environment.

Personal ID-Badge: In offices and other semi-public spaces, *PA2BLO* can be integrated with existing employee ID badges to support personalized and *authenticated* voice-based control to *access-restricted* IoT objects—e.g., to a shared “employees-only” coffeemaker device in a public lounge. Note that voice-activated devices can authenticate only modest number of user profiles—e.g., both Siri-based Homepods and Google Home can support only 6 users, implying that is infeasible to directly support reliable voice-based user identification, for say 250+ users, on such a coffeemaker. Instead, by decoupling the authentication steps into (a) audio-based user authentication on the badge and (b) device authentication of the badge to the IoT device (using standard wireless security protocols), *PA2BLO* permits authenticated voice-based appliance control to seamlessly scale up to hundreds of

users while also being resistant to physical loss of the badge. In addition, because the voice signal is sufficiently ‘scrambled’, it also ensures that an individual's voiceprint is sufficiently protected from untrusted, third-party voice services.

Pervasive Logger: *PA2BLO*'s ability to capture an individual's voice interactions with low power allows it to be used as a pervasive voice logging device, say for health and wellness applications. For example, the captured voice snippets can be stored on a remote server and subsequently analysed to infer emotional and cognitive states, such as frustration, stress or confusion. Such extracted mental markers can be useful for studying longitudinal behavioral trends.

While sharing the common functionality of speech-driven voice capture and transfer, these applications generate slightly different workloads and constraints on *PA2BLO*'s design. The Voice-Assistant and Personal ID-Badge applications typically involve interactive, short duration voice segments ($\sim 5\text{-}10$ secs) requiring appliance control to be activated with low latency. In contrast, the non-interactive Logger application may need to capture longer-lasting (several minutes) conversation snippets and buffer data locally prior to bulk wireless transfer.

B. System Components

We envision the badge to include a low-power, resource-constrained microcontroller, equipped with an energy-efficient processor, such as the Cortex M4F. Figure 1 illustrates the key components of our envisioned an ultra-low power badge, which support the envisaged *PA2BLO* functionality as follows:

- 1) *Supercapacitor & Wireless Charging:* *PA2BLO* utilizes resonant wireless coils to charge a super-capacitor that serves as a slowly-decaying energy storage component. The Charging Subsystem utilizes resonant coils (compatible with Qi™ wireless chargers) to perform rapid, non-contact charging of this supercapacitor. We emphasize that this is a pure implementation choice for an initial prototype, based on convenience and speed of recharge; alternative options, such as rechargeable batteries with two-contact chargers, are certainly feasible and will be discussed in Section VIII.
- 2) *Low Leakage and Current Supply:* A high-side switch-based architecture is used to ensure that the core processor remains disconnected until the supercapacitor voltage surpasses the desired operating value. Hardware components are carefully chosen for their low leakage, which ensures that the badge hardware current drain is ultra-low when it is in ‘idle’ mode.
- 3) *Wake-on Microphone and Speaker Authentication:* The badge uses an ultra-low power, wake-on microphone sensor that triggers the micro-processor only when the ambient noise level exceeds a minimum threshold, thereby avoiding superfluous processing during periods of ambient silence. While the actual acoustic sensing is performed by a *separate*, higher-quality microphone, we shall see how the speaker authentication mechanism

is optimized to use low-frequency (500 Hz) sampling, thereby reducing its power drain.

- 4) *Speaker De-Personalization*: This sub-system is invoked only when the current audio stream has been successfully authenticated. This system performs randomized frequency shifting of successive segments of the incoming audio stream (sampled post-authentication at 16 KHz) to generate the de-personalized audio data.
- 5) *Audio Wireless Transmission*: The de-personalized audio data is then buffered and subsequently transmitted using a wireless transceiver to the backend server/device for subsequent application-dependent processing. As most smart speakers and home voice-assistant devices utilize BLE 5.0 as the communication protocol, our prototype implementation of *PA2BLO* shall use a BLE 5.0 radio.

IV. NOVEL FUNCTIONAL COMPONENTS

PA2BLO's ability to support ultra-low power, long-lived processing of human audio data is driven by its novel systems-level implementation of two key capabilities: (a) Speaker Authentication with sub-kHz audio input, and (b) Speaker De-Personalization via Frequency Shifting. We now describe each of these two core components, and their algorithmic implementation designed to support low-power execution on micro-controller hardware.

A. Speaker Authentication with Sub-kHz Audio

For power-efficient operation, given *PA2BLO*'s goal of only capturing and transmitting the speech data of a single authorized user, the processing power of high-frequency audio data should ideally be confined to only audio segments where this user is speaking. However, because the start of a user's speech segment is not known a-priori, the authentication component must run continually in the background and thus needs to have very low energy drain. We empirically observed that the energy overhead of audio processing on a micro-controller is lower when the microphone sampling rate decreases, as this also reduces the amount of computational processing performed per second. We thus hypothesize that it is *possible to perform accurate speaker authentication while dramatically reducing the sensor sampling frequency*. Our hypothesis is driven by the observation that while a typical human's vocal range extends over a wide frequency range of 100 Hz-17 KHz, each individual has a unique *voiceprint*—a set of fundamental pitch values, driven by the shape of their vocal chords—even in the lower 100-250 Hz frequency band. Accordingly, based on Nyquist's theorem, *it should be possible to train a person-specific model to perform 1-class classification, using features from a 500 Hz audio signal*.

To validate this hypothesis, we first conducted a comprehensive study on Speaker Authentication using the benchmark VoxCeleb 1 dataset [26]. Very specifically, we downsampled the dataset from its original 16 kHz sampling rate to the following frequencies: {8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz, 250 Hz}. We segmented each voice sample into 1 second chunks and extracted features from the voice samples

using three different widely used audio feature extractors: Spectrograms, Mel Frequency Cepstral Coefficients (MFCC), and Mel-Frequency Energy banks (MFE). We then trained a lightweight 1-D CNN model (Figure 2), which was empirically validated to be computationally compatible with our eventual Cortex M4F resource-constrained hardware platform (detailed in Section V). The 1-D CNN model uses a single reshape layer, four 1-D convolution layers (COV layers) with Max pooling and dropout, followed by a flatten and output layer.

For our initial investigations on low-frequency authentication, we trained a multi-class person classifier (80-20 train-vs.-test split) involving 15 different individuals, selected at random from the VoxCeleb dataset. Note that such a multi-class classification is a pessimistic approximation of the authentication problem, which eventually (Section VI-A) involves training a *single-class* (user vs. all) model for each user. Nonetheless, such a model helps us establish an initial understanding of whether such low-frequency audio based user identification is even possible.

Figure 3 plots the classification accuracy of the 1-D CNN based classifier, for varying sampling rates and feature extractors. We observe that MFE and Spectrogram-based features have higher accuracy than MFCC, achieving accuracy values of 89.04% and 92.24%, respectively on 16 KHz voice samples. As anticipated, the classification accuracy decreases as the sampling rate reduces. However, the reduction in accuracy is fairly modest even at 500 Hz (with Spectrogram-based features achieving the highest accuracy of 72.61%), but then drops dramatically as the sampling rate is further reduced to 250 Hz. This result is not surprising as human voice pitch lies roughly between 80-250 Hz, with female voice pitch (170-250 Hz) typically being an octave higher than that of a male (80-155 Hz). Consequently, the sampling rate must be at least 500 Hz to accurately capture the key person-specific pitch value and other vocal frequencies.

B. Audio De-Personalization via Frequency Shifting

Pitch manipulation has been suggested [27] as a means of effectively masking an individual's unique voiceprints. Traditionally, such pitch manipulation involves computation of a speech segment's frequency components (e.g., using an FFT), followed by manipulation of the spectral coefficients and eventual regeneration (using the IFFT) of a modified time-domain signal. However, on micro-controller platforms, FFT computation is viewed as an expensive, floating point operation that can consume both high latency and power.

We thus utilize an alternative pitch manipulation approach that performs computations purely in the time domain. Our method capitalizes on the core concepts of sinusoidal basis functions to multiply the speech signal $x(t)$ by a cosine function, as demonstrated in equation 1. This is equivalent to a convolution operation in the frequency domain, leading the audio signal's spectrum to *shift* by $f + f_{\text{shift}}$ and $f - f_{\text{shift}}$. By considering only the positive shift, we can use this time-domain operation to effectively frequency shift the user's

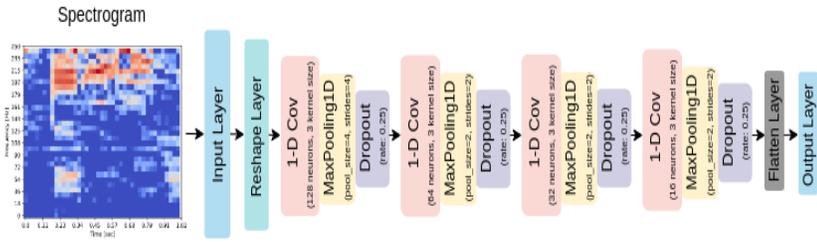


Fig. 2: Authentication Model: 1-D CNN Architecture

speech signal, thereby modifying both the voice pitch and associated nearby frequency components.

$$x_{\text{pitch-shifted}}(t) = x(t) \cdot \cos(2\pi f_{\text{shift}} \cdot t) \quad (1)$$

To prevent an adversary from reconstructing the personalized voice signal by performing an appropriate reverse shift, we do not shift the pitch by a constant value f_{shift} , but utilize a *randomized pitch shifting* technique, illustrated in Algorithm 1. For this, the speech signal is broken up into 16 msec segments (256 samples@16 KHz), and each segment is then shifted by a randomly chosen value of f_{shift} . As a consequence, the modified voice signal appears to be a sequence of differentially pitched segments. In practice, this is achieved by storing several in-memory tables of pre-computed coefficients, each table corresponding to a specific f_{shift} value and storing the values of $\cos(2\pi * f_{\text{shift}} * n)$ for different discretized values of $n = \{1, 2, \dots\}$. To demonstrate the feasibility and effectiveness of this idea, we study whether this approach is (a) energy-efficient, (b) effective (i.e., prevents speaker re-identification) and (c) comprehensible (i.e., allows for comprehension of the underlying speech by a regular speech recognition engine).

Algorithm 1 Real-time Time-Domain Pitch Shifting

```

1: procedure PITCHSHIFT( $x(t)$ ,  $N$ ,  $f_{\text{shift}1}$ ,  $f_{\text{shift}2}$ ,  $\dots$ ,  $f_{\text{shift}k}$ )
2:   for  $k = 1$  to  $K$  do ▷  $K$  signifies frequency shifts
3:     for  $i = (k - 1)N$  to  $kN$  do
4:        $sample \leftarrow x(i)$  ▷ Obtain current sample
5:        $cosine\_factor \leftarrow \cos(2\pi f_{\text{shift}k} \cdot i)$ 
6:        $processed\_sample \leftarrow sample \cdot cosine\_factor$ 
7:       Output  $processed\_sample$ 
8:     end for
9:   end for
10: end procedure

```

Power Efficiency: To test the efficacy of our proposed time-domain approach, we implemented both our lookup table-based cosine multiplication approach and the Fourier transform-based frequency domain approach (using the popular KissFFT [28] library) on our representative Cortex M4F processor. We note that KissFFT and the time domain approaches take 1.060 secs and only 30 msec, respectively, to perform frequency shifting on a 10 second audio segment. Given the constant baseline current drawn by either approach,

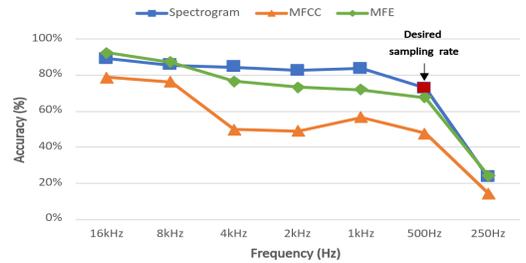


Fig. 3: Accuracy vs. Sampling Rate for Feature Extractors - Spectrogram, MFCC, MFE

the reduced latency translates into a corresponding $\sim 30x$ reduction in the computation energy.

Anonymization Effectiveness: To test whether randomized frequency shifting can help mask the personalized vocal traits of a user, we tested the speaker identification accuracy of a state-of-the-art *Wave2Vec 2.0* multi-class classifier model, on both the original and corresponding random-frequency shifted 15 samples of the VoxCeleb speech segments. We observed that pitch shifting indeed leads to a dramatic 90%+ drop in classification confidence for the majority of users (to an average of 11.8% from 97.7% without such shifting), with the confidence falling below 60% for all but one user.

Comprehensibility: Finally, to understand the effects of pitch shifting on comprehension (audio transcription), we studied the Word Error Rate (WER) for a state-of-the-art, open-source speech-to-text conversion model, “whisper-1,” for both the original and pitch-shifted audio data. We used 10 minute data segments, captured from a group of 15 participants (described later in Section VI) and computed the WER, using the *jiwer* Python library, for f_{shift} values = {25Hz, 50Hz, 100Hz, 200Hz, 400Hz}. Figure 4 illustrates that the average WER increases as the pitch shifting frequency rises, with the WER increase remaining negligible till 100 Hz. Given a typical maximum WER tolerance threshold of 0.25 [29], we determine that a pitch shifting range of [0, 200Hz] offers a good balance between anonymization and comprehensibility.

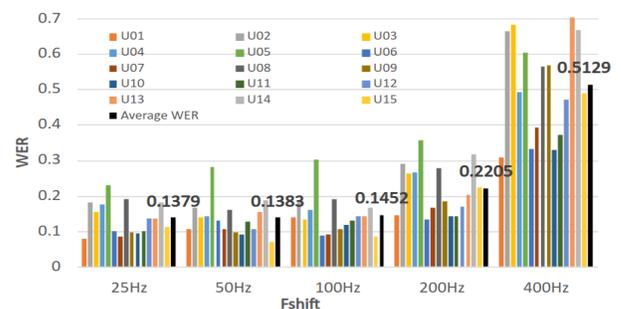


Fig. 4: WER vs Pitch Shifting Frequency

V. PA2BLO: HARDWARE & FIRMWARE

We now describe our prototype implementation of *PA2BLO* (illustrated in Figure 5), which utilizes a supercapacitor based energy source and a low-power micro-controller for ultra-low power, on-demand, processing of relevant audio

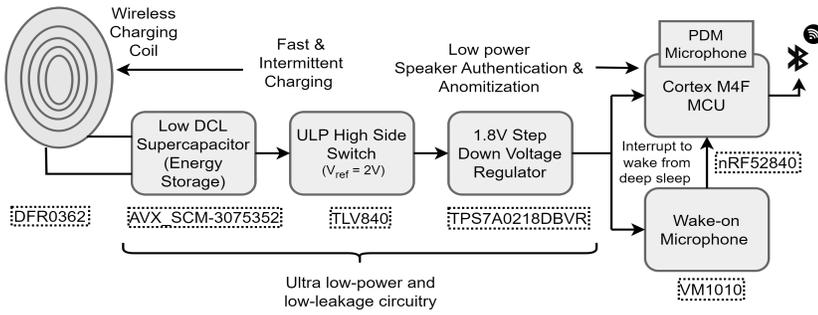
Fig. 5: *PA2BLO* Hardware Architecture

TABLE I: Supply Current: Specified vs Measured

Hardware Component	Supply Current	Measured Supply Current
TLV840	120nA	152nA
TPS7A0218DBVR	25nA	148nA
VM1010	10uA in Zero Power Listening (ZPL) 83uA in Normal mode.	13.28uA in ZPL 87.12uA in Normal mode.
nRF52840 Sense	5uA (Deep sleep)	19uA (30uA peak)

streams. Table I tabulates the ideal (datasheet-based) vs. real-world (measurement-based) supply current for several of the components listed below.

Wireless Resonant Coil for Energy Harvesting: To rapidly harvest energy in contactless fashion while ensuring an adequate hardware form factor, we utilize a DFR0362 wireless resonant coil system. The transmitting coil uses the Raspberry Pi’s 5V, 2.5A adapter, whereas the receiving coil is connected to our *PA2BLO* system, delivering 5V output with current= 1A, with 90% energy transfer efficiency as long for coil separation ≤ 1 cm. In addition, the receiving coil has a thickness of 2.3 mm and an inner diameter of 10mm, allowing it to be easily integrated into a badge-like wearable device. The receiver hardware is also compatible with NFC or Qi-based wireless chargers, as long as the coils are aligned.

Storage Supercapacitor: To support low-leakage, long duration storage of the harvested energy but with ultra-rapid charging, we experimentally selected a 15F supercapacitor (Part Number: AVX_SCM-3075352, with a 5.5V rating that is compatible with the 5V output of the resonant coil) for our *PA2BLO* device. This component allowed us to support both (a) rapid charging, with the supercapacitor being fully charged with a single 1-minute charging cycle and (b) adequate energy storage = 156.5J, which allowed the *PA2BLO* badge to achieve a full day (24 hour) standby duration. *Alternate feasible storage options are discussed in Section VIII.*

Low values of DC leakage and shunt resistance help prolong the system’s operating lifetime, by minimizing the leakage current during the idle periods of operation. Measurements with a multimeter show (see Fig 6) that the supercapacitor voltage, with no loading, dropped from 4.99 to 3.79V after 24 hours and eventually to 3.5V over a duration of 72 hours. Using the relationship $I = C \times \frac{dV}{dT}$ (I = leakage current, C = capacitance and $\frac{dV}{dT}$ denotes voltage change over time), we obtained an ultra-low leakage current value of 86 μA , which corresponds to the product dataset (84 μA at 72 hours).

Ultra-Low Power High Side Switch (HSS): The *PA2BLO*

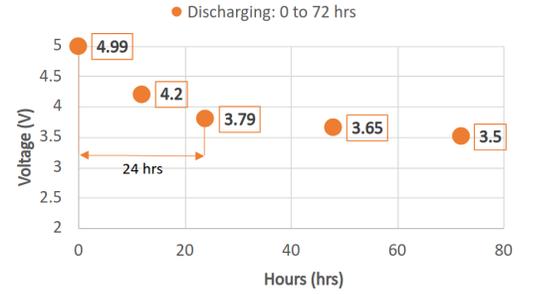


Fig. 6: AVX_SCM-3075352 Discharging Cycle

badge then uses a high-side switch to ensure that the processing pipeline remains disconnected until the supercapacitor is charged and reaches a designated operating voltage threshold V_{TH} . Through experimental studies, we selected the TLV840 high-side switch, with $V_{TH}=2V$, which aligns with the operational voltage of 1.8V for the subsequent components, including the wake-on microphone and the microcontroller. We experimentally verified that this switch has an ultra-low quiescent current ($I_Q=152$ nA, which agrees with the datasheet specification of 120 nA), which ensures a negligible loss of energy while the badge is in idle/standby mode.

Step-down Voltage Regulator: We then employ a step-down voltage regulator (the TPS7A0218DBVR with an operating range of 1.5-6V and an output drive current of 200mA) to convert the output voltage of the HSS, lying within a (2V, 5V) range, into a stable 1.8V output (voltage fluctuation= $\pm 1\%$) that is needed for the subsequent stages (Wake-on Microphone & ULP Micro-controller). Measurements on the TPS7A0218DBVR reveal a practical supply current of 148 nA, which is enough to ensure low-power operation of *PA2BLO*.

Wake-on-audio Receiver Microphone: *PA2BLO* uses a low-fidelity, Vesper VM1010 wake-on microphone to effectively ensure that the micro-controller remains in deep sleep mode during periods of low ambient noise, being activated only when the ambient sound level exceeds 65 dB SPL (typical of conversational human sound). The VM1010 wake-on microphone features two modes: Zero Power Listening (ZPL) and Normal Mode. Within our system, we exclusively utilize the Zero Power Listening mode (by setting the microphone’s mode pin to HIGH state), which results in an exceptionally low, ZPL power consumption rate of 10 μA .

ULP Micro-controller: We selected the Seed Studio XIAO nRF52840 micro-controller to provide ultra-low power (ULP) processing of the audio stream. The XIAO nRF52840 is equipped with an ARM Cortex M4F processor, which provides a deep sleep mode for ultra-low current consumption. The micro-controller includes 1MB of Flash Memory and 256 KB of SRAM, an operating voltage=1.8V and supports a compact thumb-sized, form factor (21 x 17 mm). The datasheet specifies a deep sleep current= 5 μA for the nRF52840. However, the actual deep sleep current varies significantly based on the associated firmware code. After optimizing the deep sleep code of the nRF52840 chipset, we observed a

deep sleep current of $19\mu\text{A}$, accompanied by $35\mu\text{A}$ spikes (see Table I). The XIAO nRF52840 is equipped with an on-board Pulse Density Modulation (PDM) microphone for acoustic sensing; the microphone signal is converted into Pulse Code Modulation (PCM) for digital audio processing on the nRF52840 chipset. The micro-controller also contains BLE 5.0 with supported data rates up to 2 Mbps. We set the transmission power of this radio to -4dBm , thereby reducing power consumption while supporting a range of ~ 10 meters.

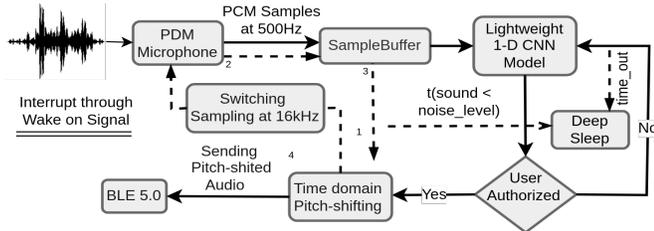


Fig. 7: Firmware Pipeline on Seed Studio nRF52840

A. PA2BLO Firmware & Computation Pipeline

To implement *PA2BLO*'s processing pipeline, we leverage on the Edge Impulse framework [30], which facilitates the export of a C++ library for integration with the nRF52840 firmware. Figure 7 shows the firmware flow on the nRF52840 chipset. The micro-controller chip is woken up by an external interrupt from the Vesper VM1010 Microphone. During the initial pre-authentication stage, it initializes the PDM microphone to receive data at a sampling rate of 500Hz using Direct Memory Access (DMA), bypassing the ARM Cortex-M4F processor. The received data is then stored in the Sample Buffer, which can hold up to 256 int8 samples.

Once the data is stored in the Sample Buffer, the 1-D CNN-based inference pipeline for speaker authentication is executed on 1 second-long segments of captured audio data. The inference process includes the computation of spectrogram coefficients over each one second audio sample, followed by the execution of an individual-specific, *custom-tuned* (to recognize only the designated, *authorized* user) 1-D CNN-based single-class (speaker vs. non-speaker) classifier. If the classifier confidence levels exceed a customizable threshold (set by default to be 0.6), the user is deemed to be authenticated. The 1-D CNN classifier model consumes 15.4 KB of peak RAM usage and 72.9 KB of Flash memory. This final model was derived after experimenting with varying model architecture and hyperparameters (e.g., different neurons, kernels, and dropout rates), so as to within the nRF52840's constraint of 1 MB flash and 256 KB SRAM.

If the model fails to authenticate the user after a certain number of segments (timeout), the nRF52840 re-enters its deep sleep mode. However, if the user is authenticated, the PDM microphone settings are switched to 16 KHz, and the user voice data is captured for pitch-shifting purposes. For an efficient pitch-shifting implementation on the nRF52840, we apply pre-computed arrays of cosine function values (across

the desired frequency range of $[0, 200\text{Hz}]$), stored in flash memory, on every 256-sample window.

The time-domain pitch-shifted data is then buffered and transmitted over BLE 5.0 with a Tx power of -4dBm (customizable), which ensures a reliable transmission range of ~ 10 meters. Because the micro-controller permits only single threaded operation, it cannot execute wireless data transmission and audio capture concurrently but must switch between them. To ensure that the BLE-based transfer does not result in unacceptably high losses of intermittent speech data, the audio data is thus transmitted in chunks rather than continuously. Once the mean sound level in the buffered data, consisting of 1 sec segments, falls below the minimum threshold (indicating the likely end of the speech segment), the nRF52840 sets up the wake-up pin in the Vesper microphone and re-enters its deep sleep mode. Figure 8 shows pictures of the prototype *PA2BLO* device, which users wore during the User Study (described in Section VII).

VI. SYSTEM-LEVEL EVALUATION

We now present results of experimental studies that help evaluate the efficacy of our *PA2BLO* smart badge in terms of system-level metrics, such as authentication accuracy and operational lifetime. These studies are conducted using a new audio dataset, captured in a controlled environment using the *PA2BLO* prototype, where 15 participants (10 males and 5 females) are recruited and asked to read aloud snippets of text for 10 minutes each. Each user's audio recording was captured, using the Seed nRF52840 MCU's microphone, at both low (500 Hz) and high quality (16 KHz) sampling rates.

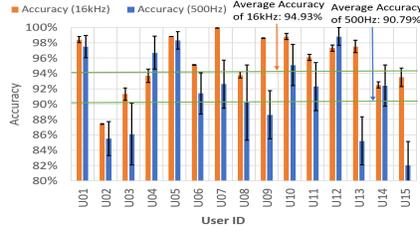
A. Authentication Accuracy: Single-class Models

Given *PA2BLO*'s goal of supporting only a single authorized user per badge, we first train and deploy a single (one-vs.-all) class classifier for each of the 15 users. This is in contrast to Section IV-A, where we developed a multi-class model that was evaluated using the externally recorded VoxCeleb dataset. As before, the 1-D CNN model was trained by splitting each audio recording into independent samples of 1 second duration, with an 80-20 training-vs.-test split. An individual i 's model is trained and test with a balanced dataset, consisting of equal number of (a) samples from user i and (b) samples from the other 14 participants.

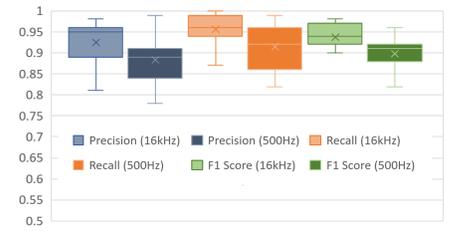
Figure 9a plots the authentication accuracy values (mean and 95% confidence bars across multiple test data splits) for each user, at both 500 Hz and 16 KHz sampling rates. We see that the mean accuracy using 500 Hz audio samples (90.79%) is only modestly lower than that obtained using 16 KHz audio data (94.93%). For U04 and U12, the mean accuracy at 500 Hz is marginally higher than at 16 KHz, but is evaluated to be statistically indistinguishable (p -values > 0.2) via a t-Test. We note, also, that the accuracy of such a person-specific single-class classifier is significantly higher than that of the common multi-class classifier (72.61%), reported in Section IV-A. We also computed the mean precision and recall of the classifiers, for 1 sec samples across all 15 users, to be 0.891 and 0.914,



Fig. 8: PA2BLO Prototype



(a) Accuracy



(b) Precision, Recall and F1 Scores

Fig. 9: Authentication Accuracy: 15 Users (16kHz and 500Hz)

respectively. Using a majority voting-based user authentication scheme over a 3-second voice segment translates to false positive and negative rates lower than 2%.

B. Energy Consumption & Active Lifetime

We now compute and analyze the energy consumption of our PA2BLO prototype over one active voice cycle, which consists of the following five distinct stages of data capture, processing and transmission: (1) Speaker Authentication; (2) Real-time Audio Pitch Shifting; (3) BLE Buffering of Pitch-shifted Data; (4) BLE Transmission; and (5) Return to Deep Sleep after BLE Transmission and Timeout.

Fig. 10a plots the power (current) drain of stages 1 and 2, during which the nRF52840 board is switched ON, and Speaker Authentication and (subsequently) audio pitch shifting is executed. The average current consumption for Stage 1 and 2 is 7.55 mA and 7.04 mA respectively, and the average current across both stages is 7.06mA over a 16-second time interval. Similarly, Figure 10b displays Stages 3 and 4 of the active cycle, which together result in an average current drain of 7.75 mA. Fig. 10c presents the current consumption of the nRF52840 while transitioning into deep sleep, with an average of 19 μ A (and peak value of 35 μ A). The average current consumption, over all five active stages of the processing pipeline, equals 7.405mA.

We can use these observations to estimate the total *active operational lifetime* of the PA2BLO prototype. Given an operating voltage $V_{op}=1.8V$, an average *active* current consumption $I_{average}=7.405mA$, the total active energy consumed for an entire hour’s worth of speaker voice is: $E_{active \text{ per hour}} = V_{op} \times I_{average} \times T = 47.98J$. Moreover, the energy dissipated due to leakage current in the *idle* period over an entire day can be computed as:

$$E_{idle \text{ cycle}} = \int_{t=0}^{24} V_{source}(t) \times (I_{leakage} + I_{supply}) dt = 42.62J \quad (2)$$

In addition, the total change in the 15F supercapacitor stored energy, as it drops from the fully-charged voltage $V_1 = 4.99V$ to the cut-off voltage $V_2 = 2$, can be computed as: $\Delta E = \frac{1}{2} \times C \times (V_2^2 - V_1^2) = 156.5J$. As a point of reference, we note that our PA2BLO badge effectively operates, using a single charge, with total energy that is $\frac{1}{15}^{th}$ of a typical CR2032 coin cell battery (2,400J).

We note that the total energy available, over a 24 hour period, for active audio capture, processing and transmission

is thus $(156.5-42.62)= 113.88J$. Accordingly, the maximum *active* duration of PA2BLO, for an operational lifetime of 24 hours, is $(\frac{113.88}{47.98})= 2.37$ hours. In other words, *our PA2BLO badge, charged in contactless fashion for just one minute, is capable of capturing and transmitting ≈ 2.5 hours of authenticated speech snippets over a 24 hour period.* Of course, the actual energy drain will vary based on the workload characteristics (i.e., frequency and duration of individual speech segments). We empirically studied the time to drain different PA2BLO prototypes when subject to pre-recorded 40 and 80 sec speech segments (by an authorized user), once every 2 mins. We observed that the PA2BLO prototype lasted for an average of ~ 3 and ~ 2.35 hours, corresponding to an *active voice duration* of 1.1 and 1.5 hours respectively.

VII. USER STUDY

To assess the performance characteristics and usability of PA2BLO under more ‘naturalistic’ usage, we also conducted a user study involving six participants who were paired into 3 separate ‘interaction dyads’. Each individual in a pair wore their personalized PA2BLO badge, set up to perform *voice logging* (for possible offline analysis). Each pair was assigned a set of collaborative tasks (e.g., creating an itinerary for an upcoming trip) that required them to engage in unscripted, interactive conversations over a session lasting 30 mins. To capture ground truth, we additionally (a) used a smartphone-based recorder to capture and manually annotate the entire session’s conversation trace, and (b) a separate ‘reference’ nRF52840 micro-controller that captured the audio and mimick-ed the authentication performed on the wearable badges. On analyzing the ground truth data, we observed that:

- the *active speech* duration had a distribution of mean (μ)= 10.98 mins and std. dev. (σ)= 1.326 mins.
- individuals spoke for relatively short periods before pausing: the duration of each active speech segment, across all 6 users, had mean (μ)= 6.12 secs and std. dev (σ)= 12.9 secs, with a speaker generating mean= 77.33 distinct segments over the 30 minute session.

Energy drainage: We captured and plotted (in Figure 11) the energy drain of each badge’s super-capacitor by ensuring that it was fully charged (4.98V) at the beginning of a session and then measuring its residual voltage (observed to be $\mu=4.1V$, $\sigma=0.09V$) at the end of each session. This translates into an overall mean energy consumption of 59.98J, which can be extrapolated to derive a mean maximum session lifetime of

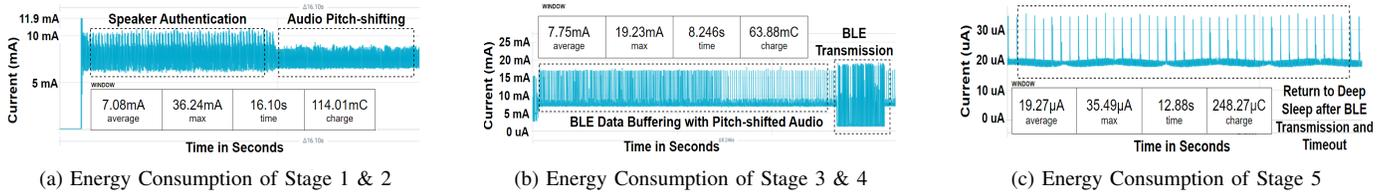


Fig. 10: Energy Consumption of the System in the Active Cycle

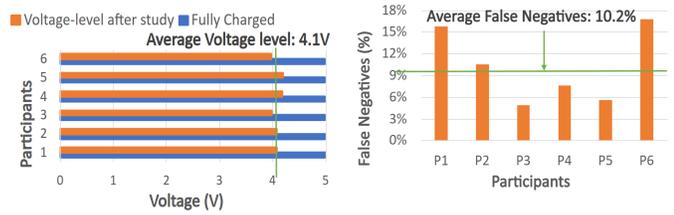


Fig. 11: Supercapacitor Voltage Drain

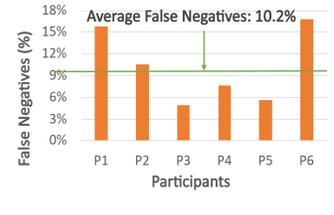


Fig. 12: Conversation False Negative Rates

$\frac{156.5}{59.98} = 1.4$ hours. (Note that this is very likely a conservative underestimate, as the supercapacitor drainage was observed to be non-linear, with a sharp drop in the beginning.)

Voice Authentication & Capture Accuracy: We observed that *PA2BLO*'s 500-Hz authentication module, executing on each smart badge, resulted in zero false positives—i.e., an individual's smart badge never captures the voice segments generated by the other participant. However, the false negative rate, plotted in Figure 12 across the 6 individual badges, was seen to have $\mu = 10.2\%$, $\sigma = 4.18\%$. $\approx 90\%$ of individual voice segments are correctly authenticated and captured by the *PA2BLO* prototype, demonstrating its feasibility in longitudinal monitoring of an individual's speech-related context.

Badge Usability: Each participating individual was also asked a series of survey questions based on the System Usability Scale (SUS) [31] to obtain their subjective assessment of the *PA2BLO* badge and technology. Participants reported an average SUS score= **86.4** (minimum=68), which indicates that the users reported the smart badge to be "highly usable" (SUS scores above 68 are considered to be good) and something that they would be very willing to wear during their daily lives.

VIII. DISCUSSION

Non-Capacitive Storage & Charging: Our choice of the 15F supercapacitor was driven by our desire to support super-fast (≤ 1 min) charging for daily use (with higher instantaneous current compared to a rechargeable battery [32]); it is *not* a core feature of *PA2BLO*. We note that while our current supercapacitor component and resonant coil harvester are bulky, thin film capacitors with energy storage density of $\approx 125\text{J}/\text{cm}^3$ are now technically feasible, and can thus offer 1000J+ energy storage (6x of our *PA2BLO* prototype) in typical badge form factors. In addition, our resonant coil harvester (chosen simply because it is compatible with wireless chargers based on the Qi standard) can also be replaced by a more space-efficient two-contact charging module. Alternately, we can also utilize a more conventional rechargeable smartwatch Li-ion battery—

e.g., a fully charged 900mWh battery ($\sim 3,200\text{J}$) of an Apple SmartWatch can support 4 hrs/day of active voice capture and transfer for a remarkably long period of 16.5 days.

Battery-less Operation via Energy Harvesting? It is worth investigating if *PA2BLO*'s optimized operation is amenable to truly battery-less, energy harvesting-based operation. We note that [33] has demonstrated how wearables can be powered with a flexible solar cell design, measuring $12.7 \times 64\text{mm}$ and worn around the human wrist, under constant illumination from average indoor lighting (i.e., 500 lux). Such a wearable harvests $150\mu\text{W}$, with resulting power density of $18.5\mu\text{W}/\text{cm}^2$; this translates into a harvestable power of 1.08 mW assuming a surface area of $65\text{mm} \times 90\text{mm}$ on a badge. Given *PA2BLO*'s avg. power consumption of 13.32mW, we thus require 12.3 secs of ambient light harvesting to support 1 sec of voice capture, translating into duty cycle=7.5%. In other words, the current *PA2BLO* design can be easily modified to utilize ambient indoor light harvesting to support approx. 108 mins (~ 1.5 hours) of intermittent, active operation daily.

IX. CONCLUSION

We have introduced the design and processing pipeline of *PA2BLO*, a wearable badge that is capable of ultra-low power capture and transmission of an authenticated individual's voice commands or speech interactions. We demonstrated two innovative capabilities that allow *PA2BLO* to support both low power operation and speaker de-identification: (a) an individualized speaker authentication model that achieves authentication F1-score=0.97 (over 3 sec audio segments) with ~ 14 mW power, and (b) a randomized pitch shifting module that achieves $\sim 30\text{x}$ power and latency reduction by avoiding expensive frequency domain operations. Through a variety of studies, we show that *PA2BLO* can support an operational lifetime of 1 day, with active voice periods of 1.5-2 hours, after being charged in contactless fashion for just 1 minute.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation, Singapore under its NRF Investigatorship grant (NRFNRFI05-2019-0007). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. We also express our gratitude to our shepherd, Dr. Sara Khalifa, and the other reviewers for their detailed feedback and constructive suggestions.

REFERENCES

- [1] C. Shen, B.-J. Ho, and M. Srivastava, "Milift: Efficient smartwatch-based workout tracking using automatic segmentation," *IEEE Transactions on Mobile Computing*, vol. 17, no. 7, pp. 1609–1622, 2018.
- [2] S. Sen, V. Subbaraju, A. Misra, R. Balan, and Y. Lee, "Annapura: An automated smartwatch-based eating detection and food journaling system," *Pervasive and Mobile Computing*, vol. 68, p. 101259, 2020.
- [3] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd, "Earbit: Using wearable sensors to detect eating episodes in unconstrained environments," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, sep 2017. [Online]. Available: <https://doi.org/10.1145/3130902>
- [4] M. Radhakrishnan, D. Rathnayake, O. K. Han, I. Hwang, and A. Misra, "Erica: Enabling real-time mistake detection & corrective feedback for free-weights exercises," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, ser. SenSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 558–571.
- [5] W. Xu, H. Yang, J. Chen, C. Luo, J. Zhang, Y. Zhao, and W. J. Li, "Washring: An energy-efficient and highly accurate handwashing monitoring system via smart ring," *IEEE Transactions on Mobile Computing*, pp. 1–14, 2022.
- [6] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa, "Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 2, jun 2020. [Online]. Available: <https://doi.org/10.1145/3397313>
- [7] M. Liu, S. Suh, B. Zhou, A. Gruenerbl, and P. Lukowicz, "Smartbadge: A wearable badge with multi-modal sensors for kitchen activity recognition," in *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, ser. UbiComp/ISWC '22 Adjunct. New York, NY, USA: Association for Computing Machinery, 2023, p. 356–363. [Online]. Available: <https://doi.org/10.1145/3544793.3560391>
- [8] R. Peixoto, J. Ribeiro, E. Pereira, F. Nunes, and A. Pereira, "Designing the smart badge: A wearable device for hospital workers." *EAI*, 8 2018.
- [9] W. Huang, Y.-S. Kuo, P. Pannuto, and P. Dutta, "Opo: A wearable sensor for capturing high-fidelity face-to-face interactions," in *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 61–75. [Online]. Available: <https://doi.org/10.1145/2668332.2668338>
- [10] N. Farina, G. Sherlock, S. Thomas, R. Lowry, and S. Banerjee, "Acceptability and feasibility of wearing activity monitors in community-dwelling older adults with dementia," *International Journal of Geriatric Psychiatry*, vol. 34, 01 2019.
- [11] S. Ömer Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 2018, pp. 10040–10050.
- [12] V. H. Tran, A. Misra, J. Xiong, and R. K. Balan, "Wiwear: Wearable sensing via directional wifi energy harvesting," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2019, pp. 1–10.
- [13] M. M. Sandhu, S. Khalifa, K. Geissdoerfer, R. Jurdak, and M. Portmann, "Solar: Energy positive human activity recognition using solar cells," in *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2021, pp. 1–10.
- [14] J. Hester and J. Sorber, "The future of sensing is batteryless, intermittent, and awesome," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: <https://doi.org/10.1145/3131672.3131699>
- [15] A. Bhatia, Y. Ahuja, S. Agarwal, and A. Parnami, "Exploring the design space of badge based input," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. ISS, pp. 1–20, 2020.
- [16] N. Pham, T. Dinh, Z. Raghebi, T. Kim, N. Bui, P. Nguyen, H. Truong, F. Banaei-Kashani, A. Halbower, T. Dinh *et al.*, "Wake: a behind-the-ear wearable system for microsleep detection," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, 2020, pp. 404–418.
- [17] S. Nirjon, J. Gummeson, D. Gelb, and K.-H. Kim, "Typingring: A wearable ring platform for text input," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 227–239.
- [18] P. Schilk, N. Polvani, A. Ronco, M. Cernak, and M. Magno, "In-ear-voice: Towards milli-watt audio enhancement with bone-conduction microphones for in-ear sensing platforms," in *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, 2023, pp. 1–12.
- [19] W.-Y. Choi, D. Ahn, S. B. Pan, K. I. Chung, Y. Chung, and S.-H. Chung, "Svm-based speaker verification system for match-on-card and its hardware implementation," *ETRI journal*, vol. 28, no. 3, pp. 320–328, 2006.
- [20] R. Liu, R. Rawassizadeh, and D. Kotz, "Toward accurate and efficient feature selection for speaker recognition on wearables," in *Proceedings of the 2017 Workshop on Wearable Systems and Applications*, 2017, pp. 41–46.
- [21] H. Lu, A. Bernheim Brush, B. Priyantha, A. K. Karlson, and J. Liu, "Speakersense: Energy efficient unobtrusive speaker identification on mobile phones," in *Pervasive Computing: 9th International Conference, Pervasive 2011, San Francisco, USA, June 12-15, 2011. Proceedings 9*. Springer, 2011, pp. 188–205.
- [22] Q. Cao, N. Weber, N. Balasubramanian, and A. Balasubramanian, "Deqa: On-device question answering," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 27–40. [Online]. Available: <https://doi.org/10.1145/3307334.3326071>
- [23] Y. Bai, N. Garg, and N. Roy, "Spidr: ultra-low-power acoustic spatial sensing for micro-robot navigation," in *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, ser. MobiSys '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 99–113.
- [24] M. Faghani, H. Rezaee-Dehsorkh, N. Ravanshad, and H. Aminzadeh, "Ultra-low-power voice activity detection system using level-crossing sampling," *Electronics*, vol. 12, no. 4, 2023.
- [25] J. Hester and J. Sorber, "Flicker: Rapid prototyping for the batteryless internet-of-things," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: <https://doi.org/10.1145/3131672.3131674>
- [26] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.
- [27] L. Tavi, T. Kinnunen, and R. González Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Communication*, vol. 140, pp. 1–10, 2022.
- [28] mborgerding, "kissfft," <https://github.com/mborgerding/kissfft>, commit = 8f47a67f595a6641c566087bf5277034b64f24d, 2021.
- [29] Y. Park, S. Patwardhan, K. Visweswariah, and S. C. Gates, "An empirical analysis of word error rate and keyword error rate." in *Interspeech*, vol. 2008, 2008, pp. 2070–2073.
- [30] S. Hymel, C. Banbury, D. Situnayake, A. Elium, C. Ward, M. Kelcey, M. Baaijens, M. Majchrzycki, J. Plunkett, D. Tischler, A. Grande, L. Moreau, D. Maslov, A. Beavis, J. Jongboom, and V. J. Reddi, "Edge impulse: An mlops platform for tiny machine learning," 2023.
- [31] J. Brooke, "Sus: A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, 11 1995.
- [32] P. Simon, Y. Gogotsi, and B. Dunn, "Where do batteries end and supercapacitors begin?" *Science*, vol. 343, no. 6176, pp. 1210–1211, 2014. [Online]. Available: <https://www.science.org/doi/abs/10.1126/science.1249625>
- [33] P. Jokic and M. Magno, "Powering smart wearable systems with flexible solar energy harvesting," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–4.