HeadFi: Bringing Intelligence to All Headphones

Xiaoran Fan[★], Longfei Shangguan[‡], Siddharth Rupavatharam[†]
Yanyong Zhang[*], Jie Xiong[◇], Yunfei Ma[§], Richard Howard[†],
[★]Samsung AI Center, New York, [‡]Microsoft, [†]Rutgers University,
[*]University of Science and Technology of China, [◇]University of Massachusetts Amherst, [§]Alibaba Group

ABSTRACT

Headphones continue to become more intelligent as new functions (*e.g.*, touch-based gesture control) appear. These functions usually rely on auxiliary sensors (*e.g.*, accelerometer and gyroscope) that are available in smart headphones. However, for those headphones that do not have such sensors, supporting these functions becomes a daunting task. This paper presents HeadFi, a new design paradigm for bringing intelligence to headphones. Instead of adding auxiliary sensors into headphones, HeadFi turns the pair of drivers that are readily available inside all headphones into a versatile sensor to enable new applications spanning across mobile health, user-interface, and context-awareness. HeadFi works as a plug-in peripheral connecting the headphones and the pairing device (*e.g.*, a smartphone). The simplicity (can be as simple as only two resistors) and small form factor of this design lend itself to be embedded into the pairing device as an integrated circuit. We envision HeadFi can serve as a vital *supplementary* solution to existing smart headphone design by directly transforming large amounts of existing "dumb" headphones into intelligent ones. We prototype HeadFi on PCB and conduct extensive experiments with 53 volunteers using 54 pairs of non-smart headphones under the institutional review board (IRB) protocols. The results show that HeadFi can achieve 97.2%–99.5% accuracy on user identification, 96.8%–99.2% accuracy on heart rate monitoring, and 97.7%–99.3% accuracy on gesture recognition.

CCS CONCEPTS

• **Human-centered computing** → **Mobile computing**; • **Hardware** → **Emerging interfaces**.

KEYWORDS

Earable Computing, Wearable Devices, User Identification, Heart-rate Monitoring, Touch Gesture Control, Voice Communication

ACM Reference Format:
Xiaoran Fan[★], Longfei Shangguan[‡], Siddharth Rupavatharam[†], Yanyong Zhang[*], Jie Xiong[◇], Yunfei Ma[§], Richard Howard[†], . 2021. HeadFi: Bringing Intelligence to All Headphones. In *The 27th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '21), October 25–29, 2021, New Orleans, LA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3447993.3448624

[★]This work was completed while the author was at Rutgers University.
[*]Corresponding author.

ACM MobiCom '21, October 25–29, 2021, New Orleans, LA, USA

ACM ISBN 978-1-4503-8342-4/21/10.
https://doi.org/10.1145/3447993.3448624

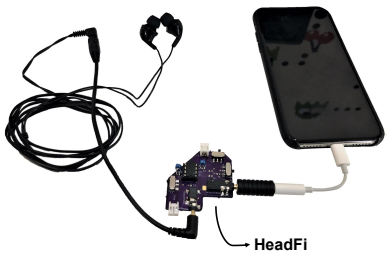


Figure 1: Illustration of HeadFi prototype. HeadFi works as a plug-in peripheral that connects a pair of headphones and a smartphone. It captures the minute voltage change on the headphones' drivers and offloads voltage readings to the smartphone for processing. HeadFi can be miniaturized and further embedded into a smartphone as an integrated circuit.

1 INTRODUCTION

Headphones[1] are among the most popular wearable devices worldwide, and are forecast to maintain the leading position in the coming years [8]. Recently, there has been a growing trend in bringing intelligence to headphones. For instance, Apple Airpods [18] and Samsung Galaxy Buds [17] put microphones in or near the ear to enable active noise cancellation and audio personalization. Motion sensing headphones such as Microsoft surface headphones [14] and BOSE QC35 headphones [11] leverage embedded sensors to enable on-ear touch control, allowing users to play or pause audio, and wake up voice assistant (*e.g.*, Siri, Alexa, and Cortana) through gestures. With miniature inertial sensors, headphones can now even pick up vital signs for respiration and heart rate monitoring [28].

Existing smart headphones all build upon advanced hardware components (mostly embedded sensors). However, statistics show that over 99% of consumer headphones shipped in 2019 are not equipped with embedded sensors, and over 43% of consumer headphones even lack a microphone [7, 48]. Thus, consumers have to purchase a new pair of smart headphones with embedded sensors to enjoy the sensing features.

In this paper, we ask the following question: *can we turn these non-smart headphones in hand into intelligent ones without redesigning the headphone or adding embedded sensors?* A positive answer would enable the consumers to enjoy smart features on their "dumb" headphones at a minimal cost. More importantly, it would also

[1]We use headphones to represent in-ear (🎧), supra-aural (*a.k.a.*, on-ear) and circumaural (*a.k.a.*, over-ear) (🎧) listening devices throughout the paper.

pave the way for realizing earable intelligence at an unprecedented scale by transforming the large amount of existing non-smart headphones into intelligent ones.

We try to answer this question by presenting the design and implementation of HeadFi—a low-power and low-cost peripheral that can be conveniently plugged into a device (such as one's smartphone) to enable a multitude of smart functionalities on non-smart headphones. Our solution serves as an alternative approach to providing smart features to headphone users. HeadFi differs from the existing smart headphone design in the following two key aspects. Firstly, it uses the headphones, in particular the pair of drivers[2] already inside a headphone, as a versatile *sensor* as opposed to adding auxiliary sensors to enable smart features. Secondly, it serves as a plug-in peripheral, connecting the headphones and the pairing device (e.g., a smartphone) in a non-intrusive manner.

HeadFi leverages the *coupling effect* between the headphones and the surroundings to enable new functionalities. For example, when a user wears a pair of headphones, the headphones, ear canal, and eardrum would couple together to form a semi-hermetic space that is extremely sensitive to pressure changes. A pressure change can be induced externally by a vibration of the headphones caused by a gentle touch. Similarly, internal physiological activities such as heartbeats cause repetitive deformation of blood vessels in the ear canal, altering the pressure inside the semi-hermetic space. As the voltage measured at the headphones is affected by these pressure changes (§2.1), we can thus leverage the voltage variations to detect the external and subtle internal physiological changes.

To realize this high-level idea, we need to address both technical and implementation challenges. From the technical point of view, the primary challenge comes from measuring the minute variation in voltage induced by the pressure change. The voltage measurement on the headphones is determined by both the audio input signal (*e.g.*, music) and the excitation signal. In practice, however, the excitation signals are weak and can easily be buried in the audio input signal that is orders of magnitude stronger (discussed in Section 2.2). From the usability point of view, our design should not break the appearance and the internal structure of the headphones. Besides, as the headphones are usually driven by mobile devices, our design should also be low power, incurring zero or negligible power consumption.

To address these challenges, we are inspired by a null measurement circuit design—*Wheatstone bridge*. Originally Wheatstone bridge was used to measure an unknown resistance by balancing the two arms of the bridge. In HeadFi, we re-purpose the Wheatstone bridge to cancel the strong interference of the audio input signals to measure the subtle variations in voltage caused by excitation signals. Specifically, the left and right two drivers of the headphones are connected to the two arms of the bridge using the headphones' stereo jacket. Once the bridge is balanced, its output voltage does not change with the variation of the audio input signal. On the other hand, the output voltage of this bridge still varies with the pressure change around the headphones, which is affected by the excitation signals such as the hand touch and the physiological activities.

[2]Different from computer hardware drivers, a headphone driver is a capacitive electronic component that drives the sound down to the ear canal.

Using Wheatstone bridge to detect subtle excitation signals provides multiple advantages over existing high-precision methods [30, 32, 40, 41, 63]. First, it provides a high measurement sensitivity as it is purely a passive circuit and thus less affected by thermal noises compared to active circuits. Second, the inherent differential circuit setup of this bridge cancels the strong audio input signals without any overhead. Third, it only consists of two simple, passive resistors. The simplicity of this design makes it easy to be miniaturized and embedded into mobile devices. To summarize, this paper makes the following contributions:

- We identify the feasibility of using the drivers already inside headphones to enable smart features. This idea potentially transforms existing non-smart headphones into smart ones at an unprecedented scale.
- We propose a simple yet effective circuit design to realize this idea. Our design uses purely passive components and costs extremely little (*i.e.*, <50 cents when fabricated at scale). We envision to integrate it into the pairing device (*e.g.*, a smartphone) in the future. Our measurement study shows HeadFi has little impact on sound quality of existing audio outputs (§2.4.3).
- We build a proof-of-concept prototype and conduct comprehensive experiments. These experiments involve 53 volunteers and 54 pairs of headphones with estimated prices ranging from $2.99 to $15,000. We further showcase four types of smart applications on non-smart headphones: user identification, touch based gesture control, physiological sensing, and voice communication without a microphone. We believe the potential of HeadFi is far beyond these.

While the current prototype of HeadFi is for wired headphones, the design can be easily extended to work with wireless headphones by putting the miniaturized circuit in between the amplifier and the Digital-to-analog converter (DAC). The rest of this paper is organized as follows: Section 2 presents the design and performance validation. We showcase four intelligent applications in Section 3–6. We discuss related works in Section 7 and potential improvement in Section 8. Conclusion follows in Section 9.

2 TRANSFORMING HEADPHONES INTO SENSORS

HeadFi employs the pair of drivers inside headphones as versatile sensors to realize the functionalities mentioned above. In essence, speakers and microphones are reciprocal in principle [1]. For headphones without a built-in microphone, an intuitive solution would be turning the speaker[3] into a microphone to capture these excitation signals. However, this solution does not work in our case due to the following two reasons. First, the sensitivity of speaker-converted microphone is inferior to purposely-built microphones as diaphragms in headphones are well-calibrated for playing sound as opposed to sound recording [50]. Second, the excitation signals are feeble and will be buried in the music signal that is orders of magnitude higher. Instead of converting the speakers into microphones, we explore the coupling effect between headphones and the surrounding environment and design a differential circuit to

[3]The driver is the key component of a speaker in headphones and therefore driver and speaker are used interchangeably throughout the paper.

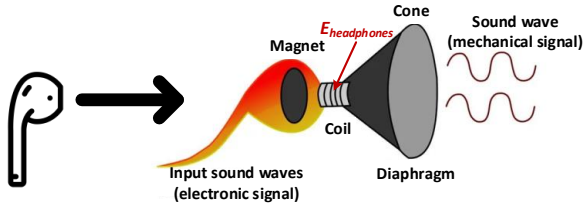


Figure 2: An illustration of headphones' working principle.

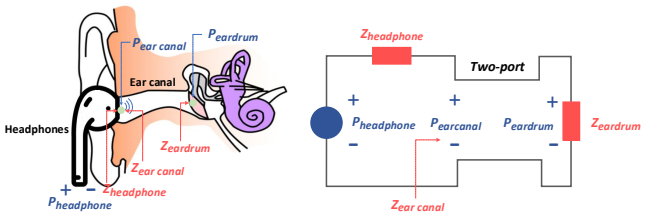


Figure 3: The ear structure (left) and the two-port Thevenin equivalent network (right).

capture the minute voltage variation.

2.1 Modeling the Coupling Effect

The drivers in headphones turn electrical energy into sound by vibrating the air through built-in magnets. We refer to the alternating voltage that travels through the headphones' voice coil as $E_{headphones}$. As shown in Figure 2, the Lorentz force inducted by the voltage variation pulls the voice coil back and forth, which then drives the diaphragm to push the air. In this way, the electrical signals are transformed into sound. Note that this process is reciprocal, *i.e.* the change of air pressure around the diaphragm of headphones also alters $E_{headphones}$.

The alternating voltage $E_{headphones}$ is determined by three factors: *i*) the electrical energy of the audio input signal (*e.g.*, music); *ii*) the equivalent impedance of the headphones' driver ($Z_{headphones}$); and *iii*) the air pressure at the headphones' diaphragm ($P_{headphones}$). Here we take the scenario when a user wears the headphones to illustrate the concept. When a user puts on the headphones, the headphones will cover the semi-closed inner ear of the user, as shown in Figure 3 (left). The headphones, ear canal, and eardrum then couple together to establish a pressure field that can be modeled by the two-port Thevenin equivalent network [44], as shown in Figure 3 (right). The definitions of the variables used in this model are listed in Table 1. The relationship between the impedance Z_x and the pressure P_x in this network can be modeled as follows:

$$\frac{P_{earcanal}}{P_{headphones}} = \frac{Z_{earcanal}}{Z_{earcanal} + Z_{headphones}} \tag{1}$$

From the above equation, we can see $Z_{headphones}$ varies with the Thevenin pressure $P_{headphones}$ and $P_{earcanal}$, as well as the ear canal's impedance $Z_{earcanal}$. These three factors are all affected by human-induced excitation signals. For instance, when a user touches the headphones' enclosure, this touch would drive the enclosure to vibrate and thus affects the Thevenin pressure $P_{headphones}$. Similarly, physiological activities such as breathing and heart beating would cause repetitive deformation of blood vessels in the ear canal and alter $P_{earcanal}$. Besides, the size and shape of the ear canal vary among individuals [44, 60]. Consequently, the ear canal's impedance $Z_{earcanal}$ differs from each other. As $E_{headphones}$ is linearly related to $Z_{headphones}$, we can therefore leverage $E_{headphones}$ to sense the human-headphone interactions and physiological activities. Note that the coupling effect still exists when the headphones are not worn (Section 2.4.4). In this case, the headphones are coupled with the surrounding environment.

2.2 Challenges

To realize this high-level idea, we face two fundamental challenges:

$P_{headphones}$:	Thevenin pressure of headphones.
$Z_{headphones}$:	The equivalent impedance of headphones.
$P_{eardrum}$:	Thevenin pressure of eardrum.
$Z_{eardrum}$:	The equivalent impedance of eardrum.
$P_{earcanal}$:	Thevenin pressure of ear canal.
$Z_{earcanal}$:	The equivalent impedance of ear canal.

Table 1: Definition of variables in Thevenin equivalent network.

How to measure the subtle variation of $E_{headphones}$ in an accurate and non-intrusive way? Adopting the general-purpose voltmeter to measure $E_{headphones}$ is usually inconvenient. Building a dedicated voltmeter, on the other hand, would inevitably add weight, size, and cost to the portable headphones. Even worse, the voltmeter's accuracy suffers from the strong magnetic interference of the working headphones [27]. Besides, the meter readings contain uncertainties due to the limited resolution and calibration offset.

How to capture the minute changes in $E_{headphones}$ caused by excitation signals in the presence of strong audio input signal? $E_{headphone}$ varies with both excitation signals and audio input signals. Unfortunately, the excitation signal can easily be buried in the audio signal, which is orders of magnitude larger. As shown in Figure 4, 5 and 6, the input audio signal (music) is in the order of hundred millivolts, while the voltage variation caused by a user's speech, most of the time, is less than one millivolt.[4] Measuring such a minute variation in voltage is challenging even in the absence of strong audio input signals because the measurement accuracy is related to the voltage value. For instance, measuring a change from 3.3 V to 3.2 V is less error-prone compared to a change from 0.1 V to 0 V, even though the amount of change is the same (0.1 V). This discrepancy is due to the nature of electronic circuits being more susceptible to noise and variations near 0 V.

As a mainstream approach, differential amplifiers have been employed to detect the minute change in voltage [12, 46, 62]. However, the stable circuit operation in these designs comes with their own design challenges. For instance, they all build upon bulky circuits and require the input and output loads to be well matched across frequencies [25]. Besides, these designs also suffer from strong noise from *i*) the pre-amplifier due to the thermal noise [24, 36] induced by the large input resistance, and *ii*) additional errors from the process of subtraction between two large numbers (the signal and the reference) to measure the small difference.

[4] The data is measured by the AKG K240s headphones.

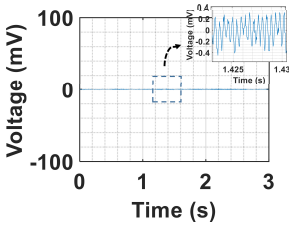


Figure 4: $E_{headphones}$ caused by talking to the headphones.

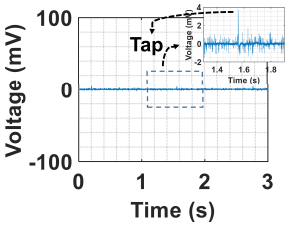


Figure 5: $E_{headphones}$ caused by tapping the headphones.

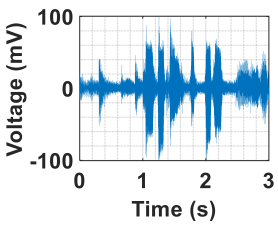


Figure 6: $E_{headphones}$ caused by playing a piece of music.

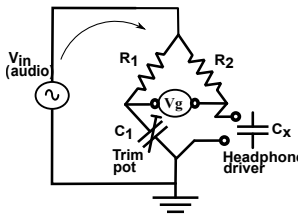


Figure 7: Null measurement in the Wheatstone bridge.

2.3 Null Measurement

We leverage a passive, null measurement circuit—*Wheatstone bridge*, to detect the minute variation of $E_{headphones}$.

Wheatstone bridge primer. A Wheatstone bridge consists of two voltage divider arms, each consisting of two simple resistors connecting the power source and ground terminal. Originally this bridge was used to measure an unknown resistance (as small as several milli-Ohms) by tuning the trimpot until the two arms reach to a balanced state (*i.e.*, the output voltage is zero) [35]. As shown in Figure 7, R_1 and R_2 are two identical bridge arm resistors. C_x is the unknown load and C_1 is the the trimpot. The trimpot C_1 is tuned until its impedance equals to the impedance of C_x, leading to a "balanced" bridge. In such a balanced state, the voltages on these two loads are the same, resulting in a zero voltage output ($V_g = 0$). Any minute change in the impedance of C_x would alter the voltage on this load and break the balance of the bridge, resulting in a non-zero voltage output (*i.e.*, $V_g \neq 0$).

Detecting minute voltage with the bridge. In HeadFi, we repurpose the Wheatstone bridge to cancel the strong audio input signals and measure the subtle changes in headphone impedance[5] caused by excitation signals. We replace the unknown load C_x in the bridge with the driver of the headphones. The audio input (*e.g.*, music signal) serves as the voltage supply V_{in} to this bridge. Once the bridge is balanced, the voltage output V_g becomes zero. The variation in audio input signal V_{in} does not break the bridge's balance. However, as we mentioned in Section 2.1, the excitation signals caused by human gestures and physiological activities inherently break the balance of the bridge and alter the voltage measured ($E_{headphones}$). More importantly, the Wheatstone bridge is super sensitive to the voltage variation at the headphones. Thus, we can leverage the variation in the voltage output of this bridge V_g to detect even very subtle excitation signals.

Using Wheatstone bridge to measure the variation of $E_{headphones}$ owns three key advantages over differential amplifier based designs: *i*) it is sensitive to minute voltage change and therefore it enables us to detect even very weak excitation signals. Following Kirchhoff's law, the two arms of the Wheatstone bridge contain only passive resistors (no capacitor or indicator) and therefore it achieves the lowest possible noise; *ii*) the differential circuit of Wheatstone bridge naturally cancels out the strong input audio signal without an extra overhead; *iii*) the simplicity of this design makes it easy to be miniaturized and embedded into a smartphone as an integrated circuit.

Balancing the Wheatstone bridge. To measure the minute change in $E_{headphones}$, it is important to balance the Wheatstone bridge first. The audio input signal is a wideband AC signal varying over the entire audible band from 20 Hz to 20 kHz. To balance the bridge over this audible band, the trimpot C_1 should be tuned to match C_x—the load of headphones' driver. Accordingly, C_1 should be an RLC type of circuit to match the driver's load. However, in practice, this balancing mechanism is not scalable since different headphones have dramatically different driver impedance values.

We instead leverage the symmetry nature of the drivers to solve this problem. The drivers of headphones come in a pair (*i.e.*, in both left side and right side of the headphones). To ensure a good user experience, each pair of drivers undergo a fine-grained calibration during manufacturing to ensure the impedance of the two drivers are exactly the same. Based on this intuition, HeadFi replaces the trimpot C_1 with the second driver in the headphones, which naturally balances the bridge without introducing any complex tuning circuits.

Physical interpretation of V_g. The pair of drivers in headphones are wired to be in-phase for coherent stereo AC signal. Note that when C_x and C_1 are replaced by the two drivers, the voltage measured at the left driver E_{left} and the right driver E_{right} come to the bridge are *phase inverted*. That is to say, the voltage output V_g of the bridge characterizes the difference of E_{left} and E_{right}: $V_g = E_{left} - E_{right}$. In some applications (*e.g.*, heartbeat, and breathing monitoring), the excitation signal is picked up by both drivers in the headphones. Hence, a critical question is whether the voltage variation caused by the excitation signals is canceled by the bridge without being detected, *i.e.*, $V_g = 0$. In practice, the excitation signals arrive at these two drivers usually through different paths. Hence, $E_{left} \neq E_{right}$. HeadFi can therefore still leverage this differential voltage measurement to detect the minute excitation signals. We believe the generalized concept of repurposing an actuator pair's response into sensing signals using the Wheatstone bridge can be applied in other domains such as robot control. For example, one can use the bridge to cancel the control signal and detect subtle mechanical vibration caused by collision [31].

Hardware implementation. Figure 8 shows the schematic of HeadFi. We prototype HeadFi on PCB board as a plug-in peripheral, connecting the headphones and the smartphone with two standard Stereo 3.5 mm jacks. The user can manually turn on/off HeadFi using the switch S_1, which allows the input audio signal to go through/bypass the bridge. R_1 and R_2 are two identical 50 ohm

[5]The voltage change is linearly related to the headphone impedance change.

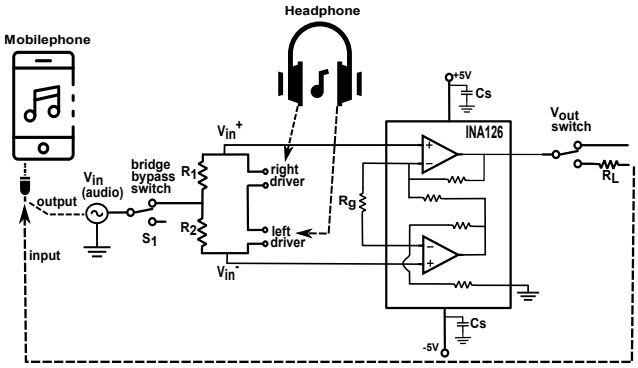


Figure 8: Schematic of HeadFi.

Figure 9: Part of the headphones used in our experiments.

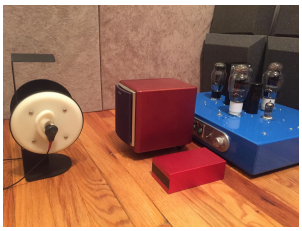


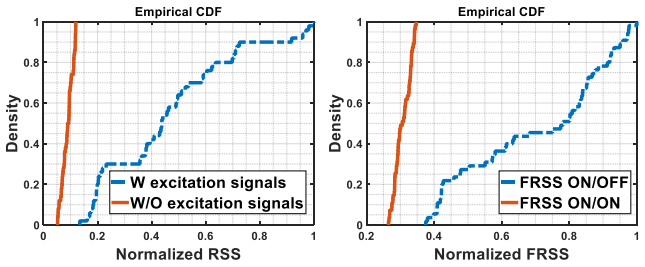
Figure 10: The setup for our benchmark experiments.

resistors. The output of this bridge is connected to a low-power amplifier, which can be replaced by the built-in amplifier in the smartphone. With this setting, the output voltage signal V_g will be automatically sent to the smartphone through the audio cable. However, the ADC in the smartphone does not sample signals coming from its audio jack unless it detects the presence of a microphone. Microphone detection is achieved by measuring the impedance of the device plugged into this audio jack. The impedance of a microphone is in the order of kilo-ohms. As long as a large impedance is detected, a microphone is considered to be detected. However, the output impedance of the amplifier in HeadFi is less than 100 Ω. We thus add a large resistor R_L (5 kΩ) to HeadFi to fool the smartphone as if a microphone exists.

Manufacturing cost. Our prototype consists of two passive resistors and an amplifier; hence its cost would be extremely low (< 50 cents) when fabricated at scale. The power consumption of this board, on the other hand, comes from the amplifier (*e.g.*, 0.2 mW), which can be further reduced by using the dedicated, low-power amplifier in the smartphone.

2.4 Benchmark evaluation

We conduct benchmark experiments to answer the following two questions: *i*) Is HeadFi sensitive enough to capture subtle voltage variation? *ii*) Does HeadFi affect the sound quality of the output audio signal? These benchmark experiments involve 54 pairs of different "dumb" headphones with price ranging from $2.99 to $15,000. Figure 9 shows a photo shot of the headphones involved in the experiments. The list of these tested headphones and their details are presented in Table 7 in the Appendix.



(a) CDF of the normalized RSS across 54 pairs of headphones. **(b) CDF of normalized FRSS across 54 pairs of headphones.**

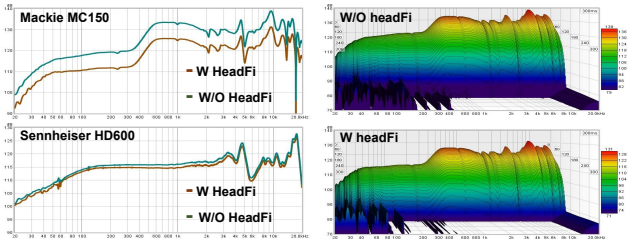
Figure 11: Evaluating the sensitivity of HeadFi on (a) direct excitation signals; and (b) indirect excitation signals.

2.4.1 Detection sensitivity on direct excitation signal. Most earable applications rely on the measurement of the direct excitation signal, *e.g.*, physiological activities (§4), touch-based gestures (§5), and human voice signals (§6). We now show the sensitivity of HeadFi is high enough to detect these direct excitation signals. We employ a Philips MC 175C speaker and multiple pairs of headphones for the benchmark experiment. The headphones are put on an E.A.R.S dummy head [15] 20 cm away from the speaker as shown in Figure 10. The speaker broadcasts a 1 kHz sinusoidal tone as the excitation signal. The volume of this signal is set to 60 dBA—a value close to the chat volume at 1 m away [13]. Note that even a subtle touch on the headphones produces a much stronger signal than this tone signal. HeadFi is connected to the headphones and "records" the RSS (received signal strength) of the excitation signal. We repeat this experiment on all 54 pairs of headphones and plot the empirical CDF of RSS measurements in Figure 11(a). For comparison, we also measure the RSS values when the speaker does not send any excitation signal. We observe the median value of the normalized RSS readings is around 0.09 in the absence of excitation signal. It jumps to 0.44 in the presence of weak excitation signals. The lowest RSS value in the presence of the excitation signal is 0.14, which is still higher than the maximum RSS value in the absence of the exciting signal. These results demonstrate HeadFi is sensitive enough to detect even minute excitation signals.

2.4.2 Detection sensitivity on indirect excitation signal. Some applications do not produce direct excitation signals. For example, to detect whether the user puts on the headphones or not, the smartphone itself needs to emit an acoustic signal. HeadFi then records the reflections of this signal to sense the surrounding environment. In this benchmark experiment, we program a smartphone to send out a chirp signal with its frequency changes linearly from 20 Hz to 20 kHz. HeadFi then records the RSS of the reflection signal. Note that as RSS value can only be obtained for a single frequency, for a frequency-varying chirp signal, we thus define a new metric *FRSS* by taking into account the responses over the entire chirp frequency band:

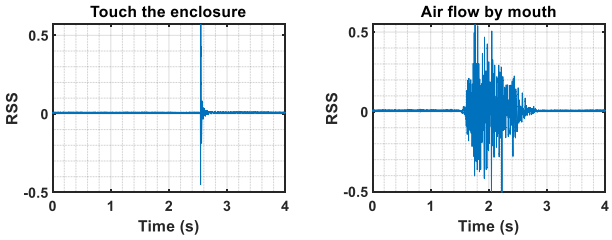
$$FRSS = \sum_{k=0}^{n-1} |X_1(k) - X_2(k)| \quad (2)$$

where $X_1(k)$ and $X_2(k)$ are the normalized outputs of the Discrete Fourier Transform (DFT) of the reflected chirp signal when the



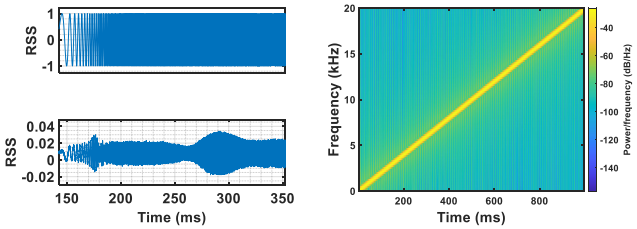
(a) The impact of HeadFi on headphones' frequency response (FR). (b) A snapshot of the cumulative spectral decay (CSD).

Figure 12: The impact of HeadFi on headphones. (a) The FR of a high- (Sennheiser HD600, $399.95) and low-quality (Mackie MC150, $49.0) headphones in the presence and absence of HeadFi, respectively. (b) A CSD snapshot of Mackie MC150.



(a) Tap the enclosure (b) Breathe air to the driver

Figure 13: Time domain signal when the headphones are placed on a table. (a) The enclosure is tapped by a finger and (b) breathing air to the driver.



(a) The transmitted and received chirp signal in time domain. (b) The spectrogram of the transmitted chirp signal.

Figure 14: An illustration of chirp signal. (a) The transmitted (top) and received chirp signal in time domain. (b) The spectrogram of a transmitted chirp whose frequency spans from 20 Hz to 20 kHz in one second.

headphones are ON and OFF the dummy head respectively. n is the number of DFT frequency bins.

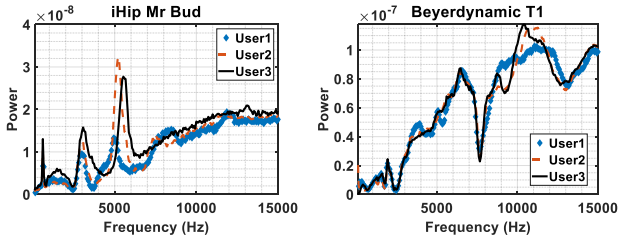
We first place a pair of headphones on the dummy head and record the output of HeadFi as shown in Figure 10. We then take the headphones off the dummy head and record the output of HeadFi again. We repeat this experiment 54 times by replacing the headphones each time. Figure 11(b) shows the CDF of the normalized signal difference when the headphones are ON and OFF the dummy head for all 54 pairs of headphones. For comparison, we also plot the difference when two measurements are both obtained with the headphones on the dummy head. We can see a clear gap between the two curves in Figure 11(b), indicating HeadFi can pick up the environment changes around the headphones. Note that the plot includes data from all 54 pairs of headphones. For a single pair of headphones, the gap is even larger. The result demonstrates that HeadFi is sensitive enough to capture the minute change of the indirect excitation signals. Our evaluation on user identification shows that HeadFi is sensitive enough even to differentiate two twin girls (§3.2.2) by profiling their unique ear canals.

2.4.3 Impact on sound quality of the output audio signals. One may concern that HeadFi contaminates the output signal (*e.g.*, music), since it wires the headphones and the pairing device as if it breaks the audio chain. We put two types of headphones on a MiniDSP E.A.R.S dummy head and measure the frequency response (FR) of these headphones in the presence and absence (for comparison) of HeadFi. Figure 12(a) shows the result. We observe the two FR curves show very similar patterns for the two headphones, indicating HeadFi does not affect the frequency response of the headphones. The gap between two FR curves indicates the electrical signal experiences a larger attenuation in the presence of HeadFi. As a result, the user will hear a slightly weaker sound but the signal quality is not affected. This is due to the extra voltage loss when the music signal passes through HeadFi. We further measure the cumulative spectral decay (CSD) of the low-quality Mackie MC-150 headphones to validate this observation. CSD is a standard metric for measuring the performance of the headphone driver. As shown in Figure 12(b), we observe two CSD snapshots exhibit very similar patterns, indicating HeadFi has a minimal impact on sound quality of the output signals. Another concern is that for low-end headphones (*e.g.*, MSRP < $2), the left and right drivers may not perfectly match, and two simple resistors might not be able to cancel the audio signal completely. To deal with this issue, one possible solution is to add an auto-balance RLC potentiometer in the nulling circuit, which can tune the value of the resistors on both sides to re-balance the bridge in HeadFi.

2.4.4 HeadFi can still work when headphones are not worn by a user. Note that HeadFi does not need to be worn by a user to work. The pressure field-based fine-grained sensing capability still exists when the headphones are not worn by a user. The coupling effect mentioned in Section 2.1 now appears between the headphones and their surroundings. Any external excitation from the environment can still change this pressure field and disturb the coupling. Therefore, HeadFi can still work when headphones are not worn by a user. To validate this, we place a pair of headphones on a table and conduct two experiments to demonstrate the sensing capability of HeadFi: (i) sense subtle finger touch and (ii) sense airflow induced by mouth. Figure 13 shows the time domain signal and we can see that HeadFi is able to detect these two types of external excitations.

3 USER IDENTIFICATION

We first demonstrate HeadFi can be used for user identification. The mainstream identification method – face recognition, does not work well in poor lighting conditions or when the user wears a mark. HeadFi can be leveraged to check the user identity and unlock the phone (pairing device) regardless of the lighting conditions. Face recognition also raises privacy concerns, whereas HeadFi can identify users without taking photos.



(a) iHip Mr Bud ($ 3.6) (b) Beyerdynamic T1 ($ 999)

Figure 15: Channel response of three persons characterized by (a) low-end and (b) high-end headphones.

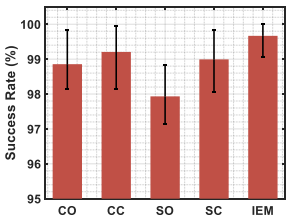


Figure 16: ON-OFF detection

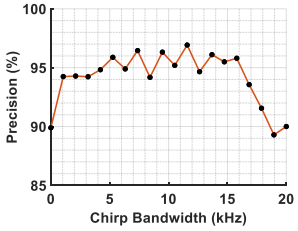
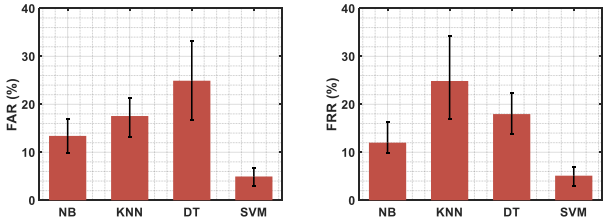


Figure 17: Precision test.



(a) FAR (b) FRR

Figure 18: The identification performance for four classifiers. (a) FAR results. (b) FRR results.

3.1 Signal Processing

Ideally, an identification service should be non-intrusive, *i.e.*, it should be triggered automatically as long as the user put on the headphones. As such, our design should be able to *i*) detect if the user puts on the headphones and *ii*) identify the user automatically.

Headphones ON-OFF detection. Our design is inspired by the *seashell resonance effect* [38]: when a seashell is clasped to the ear, the ambient acoustic noise will resonate within the cavity of the seashell and certain frequency noise will get amplified. One can thus hear ocean-tide-like sounds from the seashell. Similarly, once the user puts on her headphones, the headphones, ear canal, and eardrum establish a resonance chamber, amplifying the ambient acoustic noise. This amplified noise leads to a higher voltage signal output measured at HeadFi. Based on this observation, we use the RSS and its standard deviation (σ) for ON-OFF detection. These two values jump dramatically when the user puts on the headphones.

Identification. Since the headphones now can transmit and receive at the same time, we can now proactively probe the ear channel response using the headphones. Specifically, the smartphone sends a chirp signal through the headphones to profile the user's inner ear structure. The two drivers of the headphones receive echo signals that characterize the ear canal's channel response. Figure 14(a) and 14(b) show the chirp signals in time domain and frequency domain, respectively.

As HeadFi measures the voltage difference between the two drivers of headphones, one may wonder whether the channel response from the left ear cancels out that from the right ear. Interestingly, the ear-related physiological uniqueness not just exists between two users, but also between two ears of the same person [39, 42]. Hence the channel response measured at two ears would not be the same. Figure 15 shows the channel response measured by HeadFi on three different persons. We can see the channel responses are dramatically different in frequency bands higher than 3 kHz. This is because the physiological differences between human ears are in the scale of sub-centimeter level, which can be picked up by signals with a wavelength of sub-centimeter ($\geq$ 3 kHz). We further adopt a preamplifier (INA126) to control the output level. As a result, HeadFi can retrieve a clear echo even the excitation signal is weak.

Proof-of-concept. As a proof-of-concept, we use support vector machine (SVM), a light-weight classifier for user identification. Specifically, we collect multiple copies of the user's echo chirp as positive samples. We then collect the same amount of negative samples by putting the headphones on the E.A.R.S dummy head. Finally we train a binary SVM classifier and perform k-fold [22] cross-validation.

3.2 Experiment

The experiments involve 27 participants (7 females and 20 males), including one pair of identical twins. By default, we use the Jays U-JAYS supra-aural headphones (MSRP $ 19.99) as the testing device. The chirp duration is one second throughout the experiments. The participant is asked to put on and then take off the headphones each time we record an echo chirp. We record 50 echo chirps for each of the 25 participants and 100 echo chirps for each of the twins.

3.2.1 ON-OFF detection. We first evaluate the success rate of ON-OFF detection across 54 pairs of headphones. We further categorize the results into five groups based on headphones types and show the results in Figure 16. We observe that the success rate is consistently high (>97.93%) across all five types of headphones. In particular, IEM headphones achieve the highest success rate (99.8% on average) since this type of headphones go deeper into the ear canal and thus are less affected by noise.

3.2.2 User Identification. Next, we evaluate the performance of user identification. In each experiment, we adopt k-fold (k=5) cross validation to demonstrate the system performance. We adopt precision [29] as our evaluation metric. A high precision value indicates only the authorized users can successfully pass the verification. Figure 17 shows the precision under different chirp bandwidth settings. When the chirp bandwidth is relatively small (*e.g.*, < 4 kHz), we observe that the precision grows with increasing chirp bandwidth. The precision then fluctuates around 95% as we further increase the bandwidth to 15 kHz. It then drops to around 90% as the frequency bandwidth goes beyond 15 kHz. Such precision variation is due to the subtle changes during multiple rounds of putting on headphones: the sub-*mm* level changes can be captured by the

		Prediction		
		User One	User Two	Total
Ground-truth	User One	36,018(94.8%)	1,982(5.2%)	38,000
	User Two	1,831(4.8%)	36,169(95.2%)	38,000

Table 2: Confusion matrix for twin girls. The results are presented using k-fold cross validation.

Status	Sitting still	Moving head	Eating	Walking
FRR (%)	3.64	4.75	5.15	8.75

Table 3: Impact of human motions.

high-frequency (higher than 15 kHz) signal, which disturbs the user identification. Suggested by this study, we employ a frequency band from 100 Hz to 10 kHz as the default chirp bandwidth. We exclude the frequency band below 100 Hz because most mechanical movement-induced noise is in this frequency range.

Impact of different classifiers. Next, we evaluate the identification performance with four classifiers, Naive Bayesian (NB), k-nearest neighbors (KNN), decision tree (DT), and SVM. We investigate and report the false acceptance rate (FAR) and false rejection rate (FRR). As shown in Figure 18(a) and 18(b), SVM achieves the best performance for both FAR and FRR. We envision more advanced learning techniques such as DNN can be applied to further improve the identification performance.

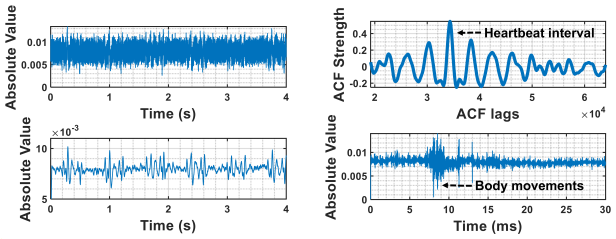
Differentiating twins. We further conduct user identification experiment on two 26-year old identical twin girls. Identifying twins is challenging because they share very similar physiological features. However, as suggested by the confusion matrix in Table 2, the identification performance for twins is comparable (95% success rate) to other individuals. Note we collected 100 echo chirps for each individual of the twins. Therefore we performed a total of 38000 classification tests for each individual in the k-fold cross-validation.

Impact of human motions. We conduct user identification when the subject is sitting still, moving her head, eating, and walking. The result is shown in Table 3. We observe that larger body movements undermine the user identification performance. In particular, HeadFi achieves the lowest false rejection rate when the subject is sitting still. The false rejection rate increases as the user starts to move, *e.g.* eating, walking, or moving her head. This is expected since the headphones are likely to move with the human motions and alter the channel response.

Long-term user identification. We further track one volunteer over two months and record the user identification performance over time. The result is shown in Table 4. We observe that the identification precision decreases gradually from 96.45% to 92.17% over two months. We suspect the reason behind this is the physiological characteristics of this subject change over time. For example, the fluid in the ear can alter the ear canal's frequency response, which impacts the user identification performance [26]. To validate this hypothesis, we conduct user identification after the shower and

Time	Reference	One Day	One Week	One Month	Two Months
Average Precision (%)	96.45	95.20	94.51	93.26	92.17

Table 4: Identification performance over time.



(a) The voltage output before and after filtering. **(b) The ACF plot and time domain interference.**

Figure 19: (a) heartbeat signal becomes clear after filtering. (b) ACF is adopted to calculate the heartbeat rate (top) and an example of time domain interference caused by body movements (bottom).

observe a 7% drop in identification accuracy.

4 PHYSIOLOGICAL SENSING

Next, we demonstrate the feasibility of applying HeadFi to detect subtle physiological signals. Vital physiological sign sensing plays a key role in human health monitoring. HeadFi can empower users to continuously and accurately monitor a variety of key physiological activities (*e.g.*, heartbeat rate) using their non-smart headphones. Below we take heartbeat monitoring as an illustrative example.

4.1 Signal Processing

Monitoring heartbeat is challenging due to the extremely weak excitation signal induced by the subtle blood vessel deformation in the ear canal. As shown in Figure 19(a) (top), such a minute excitation signal can be buried in the noise and interfered by user motions. To solve this challenge, we first pass the signal output from HeadFi through a low-pass filter with a very low cut-off frequency ($F_c = 24$ Hz) to remove the high-frequency noise introduced by the echoes of audio input signals and environment excitations. The result is shown in Figure 19(a) (bottom). We then leverage the auto-correlation function (ACF) to identify the periodicity which corresponds to the heartbeat rate:

$$r_{xx}(k) = \frac{1}{N-k} \sum_{n=0}^{N-1-k} x(n)x(n+k). \tag{3}$$

where $x(n)$ is a copy of the signal output from HeadFi and k is the lag. N is the length of the received signals. Figure 19(b) (top) shows an example of the auto-correlation output. The location of peak values reflects the time period of one heartbeat cycle. Blindly enumerating all choice of k in hopes of finding the peak is computationally intractable. It may also introduce false positives. We thus set the upper (U) and lower (L) bounds of k based on the possible heartbeat rate of human beings (35 - 200 bpm [59]). Our goal can be represented by the following function:

$$k^\star = \arg\max_{k \subseteq (L,U)} r_{xx}(k). \tag{4}$$

We then calculate the heartbeat rate using the equation $R_{BPM} = 60 \cdot \frac{F_s}{k^\star}$, where F_s is the sampling rate. In reality, however, body movements may also introduce strong excitation signals that can overwhelm the minute heartbeat signals, as shown in Figure 19(b) (bottom). We thus truncate the voltage output from HeadFi into

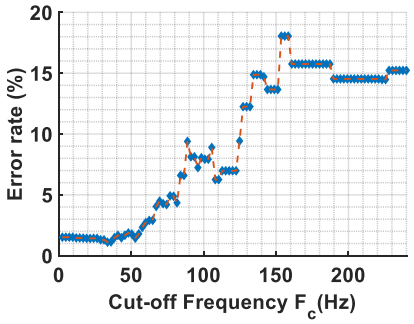
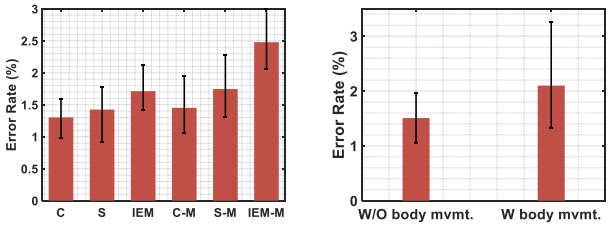


Figure 20: The heartbeat monitoring error rate is low when F_C is below 50 Hz.



(a) Impact of ambient noise. **(b) Impact of body movement.**

Figure 21: Error rate of the heartbeat rate estimation. (a) We measure the error rate both in the absence (the first three columns) and in the presence (the last three columns) of the audio input signal. (b) We measure the error rate both in the absence and presence of strong body movements.

windows and calculate R_{BPM} within each window. We then apply an outlier detection algorithm [54] to filter out those outlier estimations and average the remaining to obtain the heartbeat rate.

4.2 Experiment

In this section, we evaluate the performance of heartbeat rate monitoring. Each measurement lasts for 40 seconds. We truncate a recording session using a window size of four seconds, with an overlapping of two seconds. We measure the heartbeat rate of the participant in two conditions: *i*) with audio input signal on (*i.e.*, listening to the music during the testing); and *ii*) with audio input signal off. The ground-truth is obtained by a CONTEC CMS50D1A pulse oximeter [6]. We use error rate (ER) to measure the performance of our heartbeat monitoring: $ER = \frac{|R_{HF}-R_{PO}|}{R_{PO}}$, where R_{HF} and R_{PO} are the heartbeat rate reported by HeadFi and the oximeter, respectively.

Impact of the cut off frequency F_C. We first change the low pass filter's cutoff frequency from 2.4 Hz to 240 Hz and measure the error rate under each cutoff frequency setting. The participant listens to the music throughout the experiment. As shown in Figure 20, we observe that the error rate stays at a low level (below 2.0%) when the cut-off frequency is lower than 50 Hz. The error rate then grows to around 15% significantly as we increase F_C to 150 Hz. Suggested by this result, we set the cut-off frequency to 24 Hz, which empirically minimizes the error rate.

Heartbeat rate monitoring accuracy. We evaluate the accuracy of the heartbeat rate estimation using all 54 pairs of headphones. In particular, we categorize these headphones into three groups, namely, circumaural headphones (C), supra-aural headphones (S), and in-ear model (IEM). The result is shown in Figure 21(a). We observe that HeadFi achieves consistently low error rate across all three groups of headphones. Circumaural headphones (C) achieve the lowest error rate both in the absence (1.37%, C) and presence (1.42%, C-M) of audio input signals, followed by supra-aural headphones (1.40% and 1.68% in these two cases, respectively). HeadFi achieves the highest error rate for the IEM headphones: around 1.64% and 2.42% in the absence (IEM) and presence (IEM-M) of audio input signals, respectively. While the intrinsic reason behind this performance drop is unknown, one possible reason could be that IEM headphones have less contact area with skins and thus receive the weakest vibration signals compared to the other two types of headphones. The maximum error rate achieved by HeadFi is around 3%, which still satisfies the requirement (less than 5%) for commercial heartbeat monitoring systems [49]. These results demonstrate the feasibility of using HeadFi to measure user's heartbeat rate even in the presence of strong interference signals (*e.g.*, music).

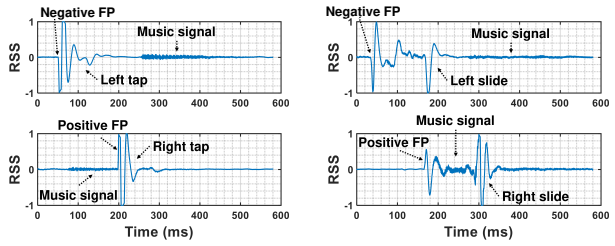
Impact of body movement. In this experiment, 27 participants (including 7 females and 20 males between 27 to 55 years old) are asked to put on/off the headphones occasionally during the testing, which brings in a strong interference signal. Figure 21(b) shows the error rate. We also show the error rate in the absence of body movement for comparison. We observe a slight increase (0.59% on average) in the error rate in the presence of body movements, while the overall error rate is still less than 3%, well below the requirement for commercial heartbeat monitoring systems (< 5%).

5 TOUCH-BASED GESTURE RECOGNITION

We next demonstrate the feasibility of transforming the enclosures of the non-smart headphones into virtual touchpads using HeadFi. The rationale behind this is that the variation in the output voltage V_g caused by different gestures manifests unique features in both spatial and temporal domains. Without loss of generality, we define four touch-based gestures: *i*) tapping the left enclosure – pause or play; *ii*) tapping the right enclosure – mute; *iii*) sliding on the left enclosure – volume up and *iv*) sliding on the right enclosure – volume down. Note that the gestures that can be supported by HeadFi are not limited to these four gestures.

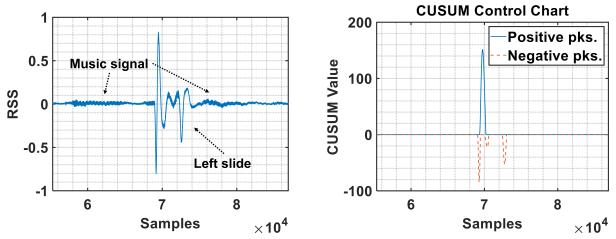
5.1 Signal Processing

Distinguishing left tapping and right tapping. We invite a volunteer to tap the left and right enclosure of one pair of headphones and record the RSS out of HeadFi. As shown in Figure 22(a), when there is a tap on the headphones, we can always observe multiple RSS peaks. In particular, when the user taps the left enclosure, there is a negative peak followed by a positive peak, as shown in Figure 22(a) (top). In contrast, the positive peak shows up ahead of the negative peak when the user taps the right enclosure (Figure 22(a) (bottom)). This is because the Wheatstone bridge measures the



(a) Tapping (b) Sliding

Figure 22: The voltage output signals V_g caused by different touch-based gestures. (a) tapping the left (top) and right (bottom) enclosure. (b) sliding on the left (top) and right (bottom) enclosure.



(a) Before applying CUSUM (b) After applying CUSUM

Figure 23: Output signal V_g before and after applying CUSUM.

differential voltage between the two drivers of headphones. Consequently, the excitation signals measured at the bridge are *phase inverted* for right tap and left tap. Note that the echoes of input music signal been recorded by HeadFi are orders of magnitude weaker and would not overwhelm peaks introduced by tapping gestures.

Distinguishing left sliding and right sliding. Similar to the tapping gestures, left and right sliding can also be easily distinguished based on the same principle. On the other hand, sliding gestures usually last longer than tapping in time domain, as shown in Figure 22(b). We can thus leverage the peak interval to distinguish them.

Algorithm. We adopt cumulative sum (CUSUM), a light-weight detection technique to capture these temporal features for gesture recognition. Specifically, we denote the output voltage samples by X_n. CUSUM associates each signal sample with a weight ω_n and then computes a value S_n with the following equations:

$$\begin{aligned} S_0 &= 0 \\ S_{n+1} &= max(0, S_n + x_n - \omega_n). \end{aligned} \tag{5}$$

This simple function, however, removes all negative peaks and thus cannot be directly used to distinguish left tapping and right tapping. Note that we build the second CUSUM function by replacing the *max* with a *min* operation in order to keep the negative peaks. The output voltage samples go through these two CUSUM functions (*max* and *min*) in parallel. Figure 23 shows the signal before and after applying the CUSUM operation, respectively. We observe that the impact of ambient music signals has been successfully removed after applying the CUSUM operation, leaving us only the peaks. We then determine left sliding/tapping or right sliding/tapping based on the following rule:

$$\begin{cases} t_1 \geq t_2 & left \\ t_1 < t_2 & right \end{cases} \tag{6}$$

where t_1 and t_2 are the starting time points of the first positive peak and first negative peak, respectively. We further define the duration of a gesture as the mean time between the first and the last non-zero CUSUM value. To distinguish tapping and sliding gestures, we measure the duration of different individuals and empirically set a threshold of 5000 samples (equivalent to 0.1*s* at the 48000 Hz sampling rate).

		Predicted Gesture				
		One	Two	Three	Four	Total
Ground-truth Gesture	One	297(99.0%)	1(0.3%)	2	0	300
	Two	2(0.6%)	297(99.0%)	1(0.3%)	0	300
	Three	0	1(0.3%)	297(99.0%)	2(0.6%)	300
	Four	1(0.3%)	0	1(0.3%)	298(99.3%)	300

Table 5: The recognition accuracy for the predefined 4 touch gestures without audio input.

		Predicted Gesture				
		One	Two	Three	Four	Total
Ground-truth Gesture	One	295(98.3%)	3(1.0%)	1(0.3%)	2(0.6%)	300
	Two	1(0.3%)	296(99.0%)	1(0.3%)	2(0.6%)	300
	Three	1(0.3%)	3(1.0%)	294(98.0%)	2(0.6%)	300
	Four	2(0.6%)	2(0.6%)	3(1.0%)	293(97.7%)	300

Table 6: The recognition accuracy for the predefined 4 touch gestures with music playing.

5.2 Experiment

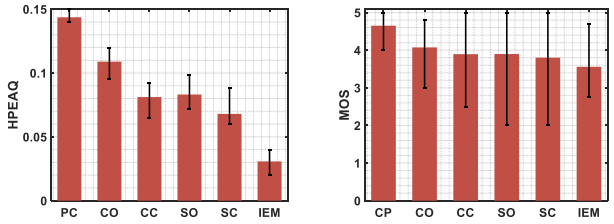
We use AKG K240s (MSRP 39.99$) headphones as the testing device. We repeat each gesture 300 times with the audio input signal on and off, respectively. The collected data are offloaded to a laptop for analysis. Table 5 and 6 show the confusion matrix of the classification result. The overall classification result is consistent across four gestures in both quiet (without audio input signals) and noisy (with audio input signals) conditions. We achieve 99% classification accuracy in the absence of the audio input signals. The classification result drops slightly to around 98% in the presence of audio input signals. This result demonstrates the feasibility of applying HeadFi to enable touch-based gesture control on the headphones. We would like to point out that we adopt the most straightforward detection algorithm (*i.e.*, CUSUM) here as a proof-of-concept. One can leverage advanced machine learning algorithms to further improve the detection performance and scale to more complex gestures.

6 VOICE COMMUNICATION

Last but not least, we demonstrate the feasibility of using HeadFi to enable full-duplex voice communication on those headphones without a built-in microphone. As discussed in Section 2.3, the human voice signals will not be canceled out by the bridge since the voice signals propagate to left and right headphone drivers through two complicated but independent channels determined by air, bones, tissue, *etc.*

The impact of echoes. One interesting issue that may exist with our design is the echo. [6] This is because during a voice call, HeadFi captures the voice from not just the HeadFi user side but also the

[6] Without loss of generality, here we assume the user on the other side is not using HeadFi for easier explanation.



(a) Objective evaluation. (b) Subjective evaluation.

Figure 24: (a) Objective and (b) subjective evaluations on the quality of voice call over different types of headphones.

other side at the headphone's diaphragm. Both captured voices will be transmitted to the other side. Thus, the other side may hear an echo of her own voice. Fortunately, this issue is already addressed by the service providers. To provide high-quality voice communication, service providers usually run sophisticated signal cancellation algorithms at the base station to remove echoes before transmitting the voice signals to the receiver [2]. Therefore, echoes would not be a problem and the evaluation results also confirm this. Next, we present our evaluation results in voice communication.

6.1 Experiment

Metric and experiment setup. Objective speech quality evaluation methods such as Perceptual Evaluation of Speech Quality (PESQ) and Perceptual Evaluation of Audio Quality (PEAQ) are widely used in telephony. These metrics are however not suitable for HeadFi since they are specifically designed for evaluating the degradation of audio signal caused by telephony network but not the degradation caused by end device. Motivated by PEAQ, we conduct an objective evaluation of the voice quality by correlating the voice recorded by HeadFi with the one recorded by the built-in microphone of a smartphone. We refer to this score as HeadFi PEAQ (HPEAQ). A higher HPEAQ manifests a higher similarity. In this experiment, a speaker transmits an acoustic chirp signal spanning from 300 Hz to 3 kHz.[7] We use both the built-in microphone in an iPhone 6 and HeadFi to record this echo. We then compute the HPEAQ value of these two signals to measure their similarity.

We repeat this experiment using 54 pairs of headphones and summarize the results in Figure 24(a). For comparison, we also record the audio using the embedded microphone in an HP X360 laptop and compute the HPEAQ value (termed as PC in the figure). Note that HPEAQ values here are far away from 1 mainly because the signal amplitudes are dramatically different. However, the frequency responses of the signals are very similar. We observe PC leads the board on the HPEAQ score, followed by CO, CC, SO, and SC. IEM headphones achieve the lowest HPEAQ score. This is because the IEM type of headphones go deeper into the ear and thus can only capture those signals propagated through human tissues and bones. In contrast, the over-ear (circumaural) and on-ear (supra-aural) types of headphones can capture both over-the-air and through-the-face-surface transmissions that get attenuated less.

Mean opinion score (MOS). Besides the objective evaluation, to better understand the HPEAQ discrepancy over the five types of headphones, we further conduct a subjective evaluation on the voice quality using MOS. MOS is another widely adopted metric in evaluating the Quality of Experience in telecommunication engineering. In our scenario, it represents the subjective opinion on the overall quality of a voice call with one side of the user using HeadFi. Each subject is asked to choose a score from a list to express his/her opinion on the quality of the voice call.

Our MOS survey involves 26 participants (6 females and 20 males) with ages ranging from 24 to 60 years old. We chat with each participant for a few minutes over the phone. During the process, we employ five different types of headphones to talk to each participant. These headphones include an AKG k701 circumaural open-back headphones (CO), a JVC HA-RZ910 circumaural closed-back headphones (CC), a Grado SR60 supra-aural open-back headphones (SO), a Jays U-JAYS supra-aural closed-back headphones (SC), and an iHip Mr. Bud In Ear Model (IEM). We plug them into HeadFi for voice calls. For comparison, we also employ the built-in microphone in an iPhone 6 (CP) for voice calls in each experiment. At the end of the call, we ask the participant to provide feedback on the sound quality by choosing a score defined below (based on ITU-T recommendations [9]):

Score	Explanation
1	impossible to communicate
2	very annoying, a lot of noise and breaks
3	annoying, some noise can be perceived
4	good, sound clear
5	perfect, like face to face conversation

Figure 24(b) shows the MOS distribution for the five tested headphones using our design and the reference built-in microphone in the smartphone. Note that the smartphone employs a dedicated audio amplifier and active noise reduction circuits in the microphone front-end [37, 47]; hence it achieves the highest average-MOS (4.8). We observe that three (CO, CC, and CP) out of these five headphones achieve consistently high average-MOS (around or above 4), indicating that participants feel the voice communication quality provided by our design is decent. The MOS of the remaining two headphones (SC and IEM) drops slightly below 4.

On the other hand, we observe that the subjective MOS exhibits a similar variation trend as the objective HPEAQ values across the five types of headphones (Figure 24(a)). However, the absolute values of MOS and HPEAQ are not linearly correlated. For instance, we see a significant HPEAQ drop on IEM headphones, while the MOS value on this type of headphones is pretty much the same as the other headphones. This is due to the non-linearity and complexity of the human auditory system discussed in the literature for decades [56, 57]. The most frequently mentioned negative feedback from our participants is the sound volume sometimes is a little bit low, and occasionally the background humming noise can be heard. This feedback is expected because the amplifier used in HeadFi (Texas Instruments INA126) is not optimized for audio quality. This issue can be addressed by using an amplifier designed purposely for audio processing.

[7] The frequency of human voice ranges from 300 Hz to 3 kHz [10].

7 RELATED WORK

Touch-based gesture control. The touch-based gesture control is usually realized by adding capacitive or resistive sensors into headphones. These sensors typically measure changes in resistance or capacitance to detect gestures such as touch. Many smart headphones including Microsoft Surface Headphones [14], Sony 1000XM3 headphones [4], Zealot B21 headphones [3], and Bose NC700 headphones [5] come with this function. Our system can serve as a supplementary solution to conveniently transform those "dumb" headphones into smart headphones.

Physiological sensing. There is a growing trend in embedding sensors in headphones for physiological sensing. Heartbeat rate, respiratory rate, and blood pressure can be monitored using electrocardiography (ECG), ballistocardiography (BCG), and photoplethysmography (PPG) [28, 51, 52, 58, 61]. Bui *et al.* adopted PPG sensors and developed an in-ear system to measure the blood pressure [23]. Anh *et al.* proposed to customize an in-ear sensor to measure the brain activities [45]. Rupavatharam *et al.* proposed to use the IMU in a pair of dedicated designed headphones to monitor jaw clenching [55]. Roddiger *et al.* developed a respiration rate monitoring system using the embedded IMU [53]. There is also an open-source multi-sensor integrated research platform, eSense, for earable computing research [16]. These sensing systems and platforms rely on dedicated sensors that add weight, require additional form factor design, and incur higher cost and power consumption.

User authentication. The unique physical structure of the ear canal can be used to authenticate users. High-frequency audio signals are bounced off the ear canal to serve as a unique feature for authentication. Arakawa *et al.* proposed to use the Mel-frequency cepstral coefficients (MFCC) instead of the frequency-domain transfer function to achieve a higher authentication accuracy [21]. Higashiguchi *et al.* proposed to use the built-in microphones in a cellphone to perform ear-related user authentications [34]. Akkermans *et al.* and Mahto *et al.* studied the feasibility of using inaudible pilot tones for user authentications [20, 43]. Gao *et al.* designed an ear-related user authentication system using commercially available headphones [33]. These techniques, however, rely heavily on auxiliary and application-specific sensors placed around ear canal that add size and even affect the quality of the audio output which is the primary function of headphones. HeadFi instead bring intelligence to "dumb" headphones without requiring extra sensors or compromising the output signal quality.

8 DISCUSSION

HeadFi leaves room for further investigations, as discussed below:

The effect on user experience. HeadFi converts the audio output into mono for sensing and such audio conversion may affect user experience in some, if not all applications, as we discussed below. *i*) Voice call. As stereo itself is not supported for voice calls, HeadFi does not affect the user experience on voice calls. *ii*) Music playing. The user experience can get compromised for stereotype music since the audio output will be converted into mono by HeadFi. *iii*) User identification. HeadFi has limited influence on user experience as user identification takes a very short time and happens occasionally in the time domain. *iv*) Physiological sensing. These applications such as heartbeat rate monitoring usually require continuous sensing. The user experience would thus be affected. To minimize such impact, we include a switch (S_1 in Figure 8) in HeadFi, allowing the user to turn on/off HeadFi as needed. A more comprehensive solution could be using a separate matching network to independently balance the left and right drivers rather than balancing them as a single pair. This allows the user to retain the stereo experience in the presence of HeadFi.

HeadFi vs. dedicated sensors. While adding dedicated sensors may achieve a better user experience in some applications, it has certain drawbacks as we discussed below. From the user's point of view, adding sensors to their headphones is not always feasible as it requires hardware modifications (*e.g.*, embedding sensors into the headphones) which may break the internal structure, layout, and circuit of the headphones. In contrast, HeadFi serves as a plug-in peripheral wiring the headphones and pairing device (*e.g.*, a smartphone) without a need of any hardware modification. On the other hand, from the manufacture's point of view, adding dedicated sensors to headphones usually incurs an extra cost to both headphone hardware and the assembly line. In contrast, HeadFi relies on low-cost hardware that is as simple as two resistors, making it a cost-effective solution. In addition, as most of the headphones owned by users or shipped to the market nowadays are still non-smart ones, HeadFi thus can serve as an important alternative solution to existing smart headphone design by turning those non-smart headphones on hand into smart ones, thereby paving the way for realizing earable intelligence at an unprecedented scale.

9 CONCLUSION

We have presented the design, implementation, and evaluation of HeadFi, a low-power peripheral to bring intelligence to headphones. HeadFi employs the pair of drivers inside headphones as a versatile sensor to enable new functionalities as opposed to adding embedded sensors. This design can potentially upgrade existing non-smart headphones into intelligent ones. We prototype HeadFi on PCB board and demonstrate the potential of HeadFi by showcasing four representative applications using 54 pairs of headphones.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers and shepherd for their insightful comments. We thank Dr. Lin Zhong for providing us insightful feedback on this work. We also thank Mr. Gefeng Wang for providing us testing headphones. This work is supported by 2030 National Key AI Program of China Grant No. 2018AAA0100500 and the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC001. Corresponding author: yanyongz@ustc.edu.cn

REFERENCES

[1] 2014. Gig Fix: Turn Your Headphones Into A Mic. Webpage.
[2] 2014. Voice Quality Enhancement and Echo Cancellation. .
[3] 2018. B21 Super Bass Wireless Bluetooth Headphone Stereo Touch Control Headset Noise Cancelling With Micro. Webpage.
[4] 2018. Wireless Noise Canceling Stereo Headset WH-1000XM3. Webpage.
[5] 2019. Bose Noise Cancelling Headphones 700. Webpage.
[6] 2019. CMS50D1A GEHP040AHUS pulse oximeter. Webpage.
[7] 2019. Earphones and Headphones Market Size, Industry Report. Webpage.
[8] 2019. Global unit sales of headphones and headsets from 2013 to 2017. Website.
[9] 2019. P.800.1 : Mean opinion score (MOS) terminology. Webpage.
[10] 2019. Voice frequency. Webpage.

[11] 2020. BOSE QC-35 Wireless Headphones.
[12] 2020. The Differential Amplifier. Webpage.
[13] 2020. How Sound Works. Webpage.
[14] 2020. Microsoft Surface Headphones. Website.
[15] 2020. MiniDSP E.A.R.S. Webpage.
[16] 2020. A Research Space for EARABLE COMPUTING. Webpage.
[17] 2020. Samgsung Galaxy Buds. Website.
[18] 2021. Apple Airpods. Website.
[19] 2021. Hifi Shark - Used, Second hand and Pre-owned Hifi. Webpage.
[20] Anton HM Akkermans, Tom AM Kevenaar, and Daniel WE Schobben. 2005. Acoustic ear recognition for person identification. In *AutoID*. IEEE.
[21] Takayuki Arakawa, Takafumi Koshinaka, Shohei Yano, Hideki Irisawa, Ryoji Miyahara, and Hitoshi Imaoka. 2016. Fast and accurate personal authentication using ear acoustics. In *APSIPA*. IEEE.
[22] Yoshua Bengio and Yves Grandvalet. 2004. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research* (2004).
[23] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. eBP: A Wearable System For Frequent and Comfortable Blood Pressure Monitoring From User's Ear. In *MobiCom*.
[24] Bruce Carter. 2009. *Op Amp noise theory and applications, 12.3.2 Thermal Noise*. Elsevier.
[25] Bruce Carter and Thomas R Brown. 2001. *Handbook of operational amplifier applications*. Texas Instruments Dallas, Tex, USA.
[26] Justin Chan, Sharat Raju, Rajalakshmi Nandakumar, Randall Bly, and Shyamnath Gollakota. 2019. Detecting middle ear fluid using smartphones. *Science translational medicine* 11, 492 (2019).
[27] John Clarke, Claudia D Tesche, and RP Giffard. 1979. Optimization of dc SQUID voltmeter and magnetometer circuits. *Journal of Low Temperature Physics* (1979).
[28] David Da He, Eric S Winokur, and Charles G Sodini. 2012. An ear-worn continuous ballistocardiogram (BCG) sensor for cardiovascular monitoring. In *EMBC*. IEEE.
[29] Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*. 233–240.
[30] Christian C Enz and Gabor C Temes. 1996. Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization. *Proc. IEEE* (1996).
[31] Xiaoran Fan, Daewon Lee, Yuan Chen, Colin Prepscius, Volkan Isler, Larry Jackel, H Sebastian Seung, and Daniel Lee. 2020. Acoustic collision detection and localization for robot manipulators. IROS.
[32] Joel Gak, Matías Miguez, Martín Bremermann, and Alfredo Arnaud. 2008. On the reduction of thermal and flicker noise in ENG signal recording amplifiers. *Analog Integrated Circuits and Signal Processing* (2008).
[33] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *IMWUT* (2019).
[34] Yutaka Higashiguchi, Yoshinobu Kajikawa, and Shunsuke Kita. 2017. A personal authentication system based on pinna related transfer function. In *ICBAKE*. IEEE.
[35] Karl Hoffmann. 1974. *Applying the Wheatstone bridge circuit*. HBM Germany.
[36] Texas Instruments. 2007. Noise analysis in operational amplifier circuits. *Application Report, SLVA043B* (2007).
[37] Thomas M Jensen, Vladan Bajic, and Andrew P Bright. 2016. Active noise cancellation using multiple reference microphone signals. US Patent 9,330,652.
[38] Liang-Ting Jiang and Joshua R Smith. 2012. Seashell effect pretouch sensing for robotic grasping. In *ICRA*. IEEE.
[39] Agnès Job, Paul Grateau, and Jacques Picard. 1998. Intrinsic differences in hearing performances between ears revealed by the asymmetrical shooting posture in the army. *Hearing research* (1998).
[40] Ron Kapusta, Haiyang Zhu, and Colin Lyden. 2014. Sampling circuits that break the kT/C thermal noise limit. *IEEE Journal of Solid-State Circuits* (2014).
[41] Ronald A Kapusta, Katsufumi Nakamura, et al. 2007. Methods and apparatus for reducing thermal noise. US Patent 7,298,151.
[42] F Laurain King and Doreen Kimura. 1972. Left-ear superiority in dichotic perception of vocal nonverbal sounds. *Canadian Journal of Psychology/Revue canadienne de psychologie* (1972).
[43] Shivangi Mahto, Takayuki Arakawa, and Takafumi Koshinak. 2018. Ear acoustic biometrics using inaudible signals and its application to continuous user authentication. In *EUSIPCO*. IEEE.
[44] Henrik Møller, Dorte Hammershøi, Clemen Boje Jensen, and Michael Friis Sørensen. 1995. Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society* (1995).
[45] Anh Nguyen, Raghda Alqurashi, Zohreh Raghebi, Farnoush Banaei-Kashani, Ann C Halbower, and Tam Vu. 2016. A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring. In *SenSys*.
[46] Viet Nguyen, Siddharth Rupavatharam, Luyang Liu, Richard Howard, and Marco Gruteser. 2019. HandSense: capacitive coupling-based dynamic, micro finger gesture recognition. In *SenSys*.
[47] Guy C Nicholson. 2013. Active noise cancellation decisions in a portable audio device. US Patent 8,515,089.
[48] Sean Olive, Omid Khonsaripour, and Todd Welti. 2018. A Survey and Analysis of Consumer and Professional Headphones Based on Their Objective and Subjective Performances. In *Audio Engineering Society Convention 145*. Audio Engineering Society.
[49] Alexandros Pantelopoulos and Nikolaos G Bourbakis. 2009. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* (2009).
[50] Melih Papila, Raphael T Haftka, Toshikazu Nishida, and Mark Sheplak. 2006. Piezoresistive microphone design pareto optimization: tradeoff between sensitivity and noise floor. *Journal of microelectromechanical systems* (2006).
[51] Ming-Zher Poh, Kyunghee Kim, Andrew Goessling, Nicholas Swenson, and Rosalind Picard. 2010. Cardiovascular monitoring using earphones and a mobile device. *IEEE Pervasive Computing* (2010).
[52] Ming-Zher Poh, Kyunghee Kim, Andrew D Goessling, Nicholas C Swenson, and Rosalind W Picard. 2009. Heartphones: Sensor earphones and mobile application for non-obtrusive health monitoring. In *ISWC*. IEEE.
[53] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl. 2019. Towards Respiration Rate Monitoring Using an In-Ear Headphone Inertial Measurement Unit. In *EarComp*.
[54] Peter J Rousseeuw and Annick M Leroy. 2005. *Robust regression and outlier detection*. John wiley & sons.
[55] Siddharth Rupavatharam and Marco Gruteser. 2019. Towards In-Ear Inertial Jaw Clenching Detection. In *EarComp*.
[56] Otto Stuhlman Jr. 1937. The nonlinear transmission characteristics of the auditory ossicles. *The Journal of the Acoustical Society of America* (1937).
[57] Frédéric E Theunissen, Kamal Sen, and Allison J Doupe. 2000. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience* (2000).
[58] Stefan Vogel, Markus Hülsbusch, Thomas Hennig, Vladimir Blazek, and Steffen Leonhardt. 2009. In-ear vital signs monitoring using a novel microoptic reflective sensor. *IEEE Transactions on Information Technology in Biomedicine* (2009).
[59] Joseph C Volpe Jr. 2008. Heart rate monitor for controlling entertainment devices. US Patent 7,354,380.
[60] Susan E Voss and Jont B Allen. 1994. Measurement of acoustic impedance and reflectance in the human ear canal. *The Journal of the Acoustical Society of America* (1994).
[61] Eric S Winokur, David Da He, and Charles G Sodini. 2012. A wearable vital signs monitor at the ear for continuous heart rate and pulse transit time measurements. In *EMBS*. IEEE.
[62] Myung-Gyoo Won, Jae-hoon Kim, and Jong-wook Park. 2006. Temperature sensing circuit for use in semiconductor integrated circuit. US Patent 7,107,178.
[63] Daniel Yum. 1991. Bandgap voltage reference circuit. US Patent 5,053,640.

A APPENDIX

A.1 List of headphones in experiments

Table 7 is the list of headphones used in our experiments. Note some headphones are discontinued. The estimated prices for the discontinued headphones are sourced from Hifi-Shark [19]. Take a departure from other consumer electronics like computers or cellphones, headphones are built to last, especially the traditional headphones. Many of our tested headphones are manufactured 20 years ago and they still work fine.

No.	Type	Name	Driver type	Estimated Price ($)	Year
01	Circumaural	Sennheiser HD800S	Dynamic	1,699.95	2015
02	Circumaural	Verum Audio Verum	Planar	349.00	2018
03	Circumaural	Sennheiser HD58x	Dynamic	150.00	2017
04	Circumaural	Sennheiser HD650	Dynamic	499.95	2003
05	Circumaural	Sennheiser HD540II	Dynamic	200.00	1986
06	Circumaural	Harman Kardon HARKAR NC	Dynamic	299.95	2013
07	Circumaural	JVC HA-RZ910	Dynamic	78.54	2014
08	Circumaural	Equation RP-21	Dynamic	99.95	2010
09	Circumaural	AKG K500	Dynamic	150.00	1991
10	Circumaural	AKG K701	Dynamic	449.00	2005
11	Circumaural	AKG K812	Dynamic	1,499.00	2013
12	Circumaural	Monoprie M565C	Planar	199.99	2018
13	Circumaural	Focal Elear	Dynamic	999.00	2017
14	Circumaural	Focal Elex	Dynamic	600.00	2017
15	Circumaural	Sennheiser HD580	Dynamic	200.00	1991
16	Circumaural	Sennheiser HD540G	Dynamic	400.00	1988
17	Circumaural	Hifiman Susvara	Planar	6,000.00	2016
18	Circumaural	Hifiman HE500	Planar	700.00	2010
19	Circumaural	Audeze LCD2	Planar	995.00	2010
20	Circumaural	Beyerdynamic T1	Dynamic	999.00	2009
21	Circumaural	Sony MDR R10	Dynamic	15,000.00	1989
22	Circumaural	AKG K240s	Dynamic	39.99	2001
23	Circumaural	Abyss AB-1266	Planar	6,000.00	2012
24	Circumaural	Abyss AB-1266 Phi	Planar	7,000.00	2017
25	Circumaural	Kennerton Thror	Planar	3,080.00	2018
26	Circumaural	Hifiman HE6	Planar	1,000.00	2010
27	Circumaural	Sennheiser HD580 Jubilee	Dynamic	500.00	1995
28	Circumaural	Sennheiser HD600	Dynamic	399.95	1997
29	Circumaural	AKG K872	Dynamic	1,499.00	2016
30	Circumaural	Audeze LCD3	Planar	1,945.00	2012
31	Circumaural	Sennheiser HD224	Dynamic	49.00	1977
32	Supra-aural	Koss Porta Pro	Dynamic	49.99	1984
33	Supra-aural	Grado SR60	Dynamic	69.00	1995
34	Supra-aural	Koss KSC75x	Dynamic	19.99	2004
35	Supra-aural	BlueAnt Embrace	Dynamic	200.00	2011
36	Supra-aural	Equation Audio RP-15MC	Dynamic	50.00	2004
37	Supra-aural	Audio-Technica ATH-OR7	Dynamic	200.00	2009
38	Supra-aural	Marshall Monitor II ANC	Dynamic	319.99	2019
39	Supra-aural	Spider PowerForce	Dynamic	50.00	2012
40	Supra-aural	Monoprice Pro	Dynamic	29.99	2012
41	Supra-aural	Musical Fidelity MF-100	Dynamic	199.00	2013
42	Supra-aural	AKG Q460	Dynamic	129.90	2010
43	Supra-aural	Beats Diamond Tears	Dynamic	349.95	2012
44	Supra-aural	Mackie MC-250	Dynamic	80.00	2019
45	Supra-aural	Jays U-JAYS On Ear	Dynamic	19.99	2017
46	In-ear	Beats Heartbeats	Dynamic	99.99	2012
47	In-ear	Philips Fidelio S2	Dynamic	149.00	2013
48	In-ear	Etymotic Research ER6i	Armature	149.00	2004
49	In-ear	SENFER DT6	Trio hybrid	35.00	2019
50	In-ear	Beats urBeats	Dynamic	99.99	2018
51	In-ear	KEF M200	Duo dynamic	200.00	2013
52	In-ear	Sennheiser HD405	Dynamic	35.00	2001
53	In-ear	iHip Mr Bud	Dynamic	3.60	2015
54	In-ear	Insten In-ear	Dynamic	2.99	2013

Table 7: List of headphones in our experiment.