HeadFi: A New Design Paradigm for Smart Headphones

Paper #196: 12 pages, plus references

Abstract

Headphones continue to grow more intelligent as new functions (*e.g.*, touch-based gesture control) appear. Such functions usually rely on auxiliary sensors (*e.g.*, accelerometer, gyroscope) that are available on many smart headphones. However, for those headphones that do not have such sensors, supporting these functions becomes a daunting task. This paper presents HeadFi, a design for bringing "smartness" to headphones. Instead of adding auxiliary sensors into headphones, HeadFi turns the pair of drivers that are readily available on all headphones into a versatile sensor to enable new applications spanning across mobile health, user-interfaces, and context-awareness. HeadFi works as a plug-in peripheral connecting the headphones and pairing device (*e.g.*, smartphone). The simplicity (can be as simple as only two resistors) and small form factor of this design lends itself to embedding into the pairing device as an integrated circuit. As a result, we envision HeadFi can serve as a vital *supplementary* solution to existing smart headphones design by directly transforming "dumb" headphones into intelligent ones. To show the feasibility of our design, 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. Experiment 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.

1 Introduction

Headphones[1] are among the most popular wearable devices worldwide, and are forecast to maintain the leading position in the coming years [7]. Recently, there has been a growing trend in bringing intelligence to headphones. For instance, Apple Airpods [14] and Samsung Galaxy Buds [28] put microphones in or near the ear to enable active noise cancellation. Motion sensing headphones such as Microsoft surface headphones [53] and BOSE QC35 headphones [18] leverage embedded sensors to enable on-ear touch control, allowing users to play or pause audio, and wake up the 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 [24].

Existing smart headphones all build upon advanced hardware units (mostly embedded sensors). However, statistics show that over 99% of consumer headphones are not equipped with embedded sensors, and over 43% of consumer headphones even lack a microphone [6, 44]. Consumers thus have to purchase a new pair of smart headphones with embedded sensors to enjoy these features.

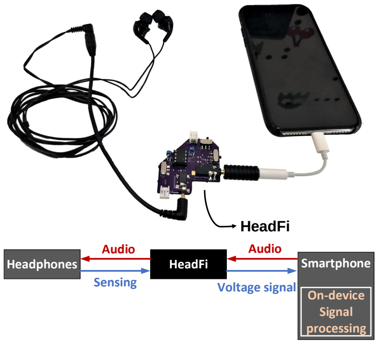


Figure 1: Illustration of HeadFi prototype. HeadFi works as a plug-in peripheral that connects a pair of headphones and a smartphone. It captures the minute voltage change on the headphones' drivers and offloads voltage readings to the smartphone for processing. HeadFi can be miniaturized and further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

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

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 the 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 headphones design in the following two key aspects. Firstly, it uses the headphones, in particular the pair of drivers inside,[2] as a versatile *sensor* to enable smart services as opposed to adding auxiliary sensors. Secondly, it serves as a plug-in peripheral, connecting the headphones and the pairing device (such as a smartphone) in a non-intrusive manner. By using this peripheral, any existing headphone, without driver modification, can enjoy an array of smart features, such as gesture control and physiological sensing.

[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.

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

HeadFi leverages the *coupling effect* between the headphones and the surroundings to enable new functionalities. Specifically, when a user wears a pair of headphones, the headphones, ear canal, and eardrum would couple together to form a semi-hermetic space and this space is extremely sensitive to pressure changes. A pressure change can be induced externally by a vibration of the headphone caused by an impulse as gentle as a touch. Similarly, internal physiological activities such as heart beats cause repetitive deformation of blood vessels in the ear canal which alter 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 this voltage variation to detect both the external changes and subtle internal physiological changes.

To realize this high-level idea, however, we need to address both technical and implementation challenges. From the technical point of view, the primary challenge comes from the difficulty in measuring the minute change in voltage induced by the pressure change. The voltage measurement on the headphones is affected by both the audio input signal (*e.g.*, music) and excitation signals caused by pressure change. In practice, however, the excitation signals are weak and can easily be buried in the audio input signal which is orders of magnitude stronger. From the usability point of view, our design should not break the appearance of the headphones as well as the internal structure and circuit of the headphones. In addition, as the headphones are usually coupled with a mobile device, our design should also be low-power, incurring zero or ignorable power consumption to the mobile device.

To solve these challenges, we are inspired by a null measurement circuit, Wheatstone bridge. Originally Wheatstone bridge was used to accurately measure an unknown resistance by producing zero voltage difference between the two parallel branches of the bridge. In HeadFi, we re-purpose the Wheatstone bridge to eliminate the strong interference of the audio input signals, and at the same time measure the subtle changes in headphones impedance caused by excitation signals. Specifically, we connect the two drivers of headphones to the two branches of the bridge using the headphones' audio cable to balance the bridge. Once the bridge is balanced, the output voltage does not change with the variation in the audio input signal. On the other hand, the output voltage of this bridge still changes with the impedance/pressure of the headphones, which is affected by the excitation signals from human gestures or physiological activities.

Using Wheatstone bridge to detect subtle excitation signals provides key advantages over existing high-precision methods [26, 27, 36, 37, 61]. First, it provides high measurement sensitivity as it is purely a passive circuit and thus less affected by thermal noises compared against active circuits; Second, the inherent differential circuit setup of this bridge cancels the overwhelming audio input signals without any overhead (*e.g.*, a sophisticated cancellation circuit); Third, it simply consists of two passive resistors, which 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 on headphones to enable smart applications. This can potentially transform existing non-smart headphones into smart ones.
- We propose a simple yet effect circuit design to realize this idea. Our design uses passive circuit components and costs extremely low (*i.e.*, <50 cents when fabricated at scale). The simplicity of this design shows a potential of integrating it into the pairing device (*e.g.*, smartphone). Our headphone measurement study demonstrates HeadFi would not affect the sound quality (§2.4.3).
- We build a proof-of-concept prototype and conduct comprehensive experiments with 53 volunteers using 54 pairs of headphones with prices ranging from $ 2.99 to $ 15,000. We further showcase four types of smart services on non-smart headphones: user identification, touch based gesture control, physiological sensing, and voice communication. We believe the potential of HeadFi is far beyond these.

While the current prototype of HeadFi is for wired headphones, the hardware 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). Our current prototype takes only one channel of the input acoustic signal, hence it will transform stereo sound into mono automatically. To cope with issue, we add a switch on HeadFi to support on-demand running of smart applications. 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. Conclusion (§8) follows.

2 Transforming Headphones to Sensors

HeadFi employs the pair of headphone drivers as versatile sensors to realize the functionalities mentioned above. While headphones and microphones are reciprocal in principle [1], an intuitive solution would be re-purposing headphones as microphones to capture these excitation signals. However, this does not work in our case due to the following reasons. The audio quality of re-purposed headphones is inferior compared to purpose-built microphones as diaphragms on headphones are well-calibrated for sound playing as opposed to sound recording [46]. Besides, the excitation signals recorded by re-purposed headphones are feeble and are likely to be buried in the input music signal, which is orders of magnitude higher. Instead of re-purposing headphones as microphones, we explore the coupling effect between headphones and surrounding environment and design a highly accurate circuit to capture the minute voltage variation.

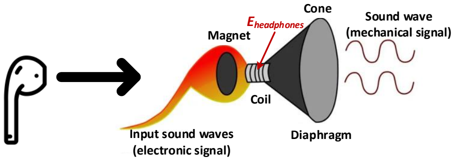


Figure 2: An illustration of headphones' working principle.

2.1 Modeling the Coupling Effect

Principles of headphone drivers. The drivers on headphones turn electrical energy into sound by using magnets to vibrate the air. We refer to the alternating voltage of an audio signal 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 air. In this way, the electrical signals are transformed into sound. Note that this principle 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 ($P_{headphones}$). When a user puts on her headphones, the headphones will cover the semi-closed inner ear of the user, as shown in Figure 3 (left). The headphones, the ear canal, and eardrum then couple together and establish a pressure field that can be modeled by the two-port Thevenin equivalent network [40], as shown in Figure 3 (right). The definition of variables in this model is listed in Table 1.

The relation between the impedance Z_x and the pressure P_x in this model can be represented by:

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

From the above equation, we can see $Z_{headphones}$ varies with the Thevenin pressure $P_{headphones}$, $P_{earcanal}$, and also the impedance $Z_{earcanal}$, all of which are affected by human-induced excitation signals. For instance, when a user touches the enclosure of her headphones, this touch gesture would drive the enclosure to vibrate and thus affects the value of Thevenin pressure $P_{headphones}$. Similarly, the physiological activities such as breathing and heart beating would cause repetitive deformation of blood vessels in the ear canal and thus alters $P_{earcanal}$. Also, the size and shape of the ear canal vary among each individual [40, 58]. Thus, the impedance of the ear canal $Z_{earcanal}$ differs from each other as well. As $E_{headphones}$ is affected by $Z_{headphones}$, we can thus leverage $E_{headphones}$ to sense these human-headphones interactions and physiological activities.

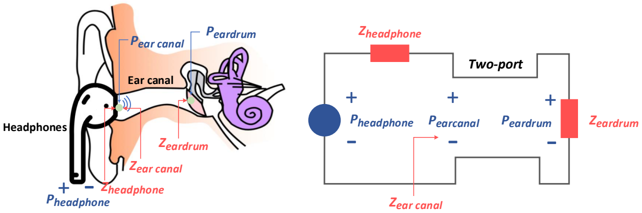


Figure 3: The ear structure (left) and the two-port Thevenin equivalent network (right).

$P_{headphones}$:	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.

2.2 Challenges

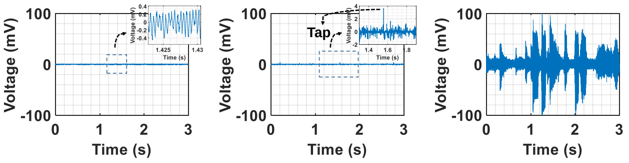
To translate this high-level idea into a working system, we face two fundamental challenges:

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, bulk, power, and cost to the portable headphones. Even worse, the accuracy of the voltmeter suffers from the strong magnetic interference of the working headphones [23]. What's more, the meter readings contain uncertainties inherently due to the limited resolution and calibration offset.

How to capture the minute changes in $E_{headphones}$ caused by excitation signals, particularly in the presence of strong audio input signal? $E_{headphone}$ varies with both excitation signals and audio input signals. Unfortunately, the excitation signal is easily buried in the audio signal, which is orders of magnitude larger. As shown in Figure 4, the input music signal is at the order of hundred millivolts, while the voltage variation caused by a user's speech, for most of the time, is less than one millivolt[3]. Measuring such a minute variation in voltage is still challenging even in the absence of strong audio input signals because the measurement accuracy scales with 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.

Typically differential amplifiers are used to detect minute changes in voltage [10, 42, 60]. These designs amplify the

[3] The data is measured by using an AKG K240s headphones.



(a) Voice **(b)** Gesture **(c)** Music

Figure 4: $E_{headphones}$ caused by different excitation signals. (a) Talking to the headphones; (b) tapping the headphones enclosure, and (c) playing a piece of music.

two input signals and then subtract them. However, the stable circuit operation in these designs comes with its own design challenges, including requiring dual (*i.e.* both positive and negative) voltage sources, requiring input and output loads to be impedance matched across frequencies, and complicated bulky circuits [21]. Besides, these designs also suffer from strong noises as the two input signals would have noise added by *i*) the pre-amplifier due to the thermal noise [20, 32] from the resistance of the sensor and the large input resistance, and *ii*) additional errors from the process of subtracting two large numbers (the signal and the reference) to measure the small difference.

2.3 Null Measurement

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

Wheatstone bridge primer. The Wheatstone bridge consists of two voltage divider arms, each consisting of two simple resistors in series-connected between the same voltage supply and ground terminal. Originally this bridge was used to accurately measure an unknown resistance (as small as several milli-Ohms) by producing zero voltage difference between the two parallel branches when balanced [31]. in this bridge circuit, R_1 and R_2 are two identical bridge arm resistors. C_x and C_1 are the unknown load and the trimpot, respectively. The trimpot C_1 is tuned until it equals to 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 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.*, $V_g \neq 0$).

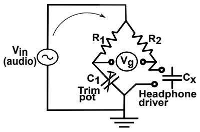


Figure 5: Null measurement in the Wheatstone bridge.

Detecting minute voltage with the bridge. In HeadFi, we re-purpose the Wheatstone bridge to eliminate strong audio input signals, and at the same time measuring the subtle changes in headphones impedance 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

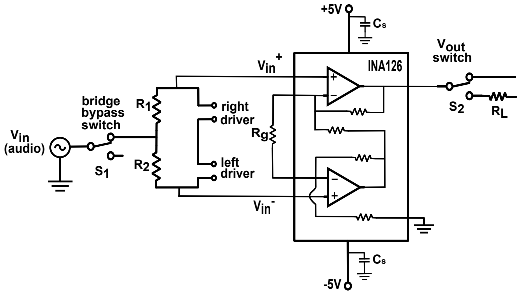


Figure 6: Schematic of HeadFi.

the bridge is balanced, the voltage output V_g becomes zero. The variation in audio input signal V_{in} would not break the balance of the bridge. However, as we mentioned in Section 2.1, the excitation signals caused by human gestures and physiological activities would alter the voltage measured at the headphones $E_{headphones}$, which inherently breaks the balance of the bridge. More importantly, Wheatstone bridge is super sensitive to the voltage variation at the headphones. We can thus 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 $E_{headphones}$ provides three key advantages: *i*) it provides high measurement sensitivity, which enables detecting the minute change in $E_{headphones}$ caused by excitation signals. It does the subtraction with Kirchhoff's law and purely passive real resistors. Thus the lowest possible noise; *ii*) the inherent differential circuit setup of Wheatstone bridge eliminates the more significant input audio signal on $E_{headphones}$ without any overhead; *iii*) it is a pure passive network consisting of two low-cost resistors, whereby introducing very little noise and making it easy to be miniaturized and embedded into a smartphone as an integrated circuit.

Balancing the bridge. To measure the minute change in $E_{headphones}$, it is essential to balance the Wheatstone bridge first. The audio input signal is a wide band 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 headphone's driver. C_1 thus should be an RLC type matching circuit supporting offline calibration (in accordance with the driver's load). In practice, however, this balancing mechanism is not scalable, since different headphones have dramatically different drivers.

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

Physical interpretation of V_g. The pair of drivers on headphones are wired to be in-phase for coherent stereo playback (AC signal). Once the trimpot C_1 is replaced by one of the pairing drivers, the voltages measured at the left driver E_{left} and the right driver E_{right} come to the bridge are *phase inverted*. Said differently, 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 will be picked up by both drivers on the headphones. Hence, a fundamental question is whether the voltage variation caused by excitation signals will be canceled by the bridge, *i.e.*, $V_g = 0$. In practice, the excitation signal arrives at these two drivers through different paths. Hence, $E_{left} \neq E_{right}$. HeadFi can still leverage this differential voltage measurement to detect and further differentiate the minute excitation signals.

Hardware implementation. Figure 6 shows the schematic of HeadFi. We prototype HeadFi on PCB board (Figure 1) 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 resistor. The output of this bridge connects to a low-power amplifier, which can be further 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 on 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 in this audio jack. The impedance of a microphone is in the order of thousands ohm. As long as a large impedance is measured, a microphone is considered to be detected. However, the output impedance of the amplifier on HeadFi is only less than 100 ohm. We thus include a large resistor R_L (5K ohm) on HeadFi to fool the smartphone as if a microphone exists.

Our prototype consists of two passive resistors and an amplifier; hence its cost would be extremely low (< 50 cents) when fabricating 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 on the smartphone.

Figure 7: Part of 54 pairs of non-smart headphones for experiments.

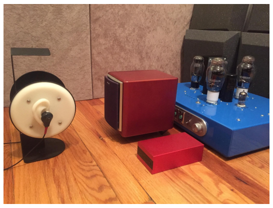
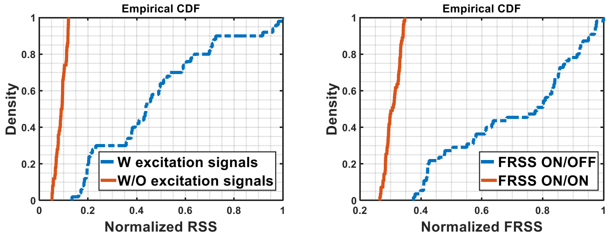


Figure 8: Experiment setup: a dummy head, a speaker, an amplifier, and a DAC.



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

(b) CDF of normalized FRSS across 54 pairs of headphones.

Figure 9: Evaluating the sensitivity of HeadFi on (a) direct excitation signals, and (b) indirect excitation signals.

2.4 Benchmark evaluation

We conduct experiment to answer the following two questions *i*) Is HeadFi sensitive enough to capture subtle voltage variation? *ii*) Will HeadFi affect the sound quality of an output signal? These benchmark studies involve 54 pairs of different "dumb" headphones with a price ranging from $2.99 to $15,000. Due to page limitation, we will provide an external link for the list of all headphones in the camera ready version. Figure 7 shows these testing devices.

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 the direct excitation signal. Specifically, we employ the Philips MC 175C speaker and one pair of headphones for the benchmark experiment. The headphones are put on an E.A.R.S dummy head [12] 20 cm away from the speaker as shown in Figure 8. The speaker broadcasts a 1 KHz sinusoidal tone signal at a weak 60 dBA volume[4]. Note that even a subtle touch on the headphone generates a much stronger signal than this tone signal. HeadFi (connecting to the headphones) then "records" the RSS (received signal strength) of this excitation signal. We repeat this experiment with all 54 pairs of headphones and plot the empirical CDF of RSS measurements in Figure 9(a). We also measure the RSS values when the speaker is not sending any signal. We observe the median value of the normalized RSS readings is 0.09 and 0.44 without and with the weak excitation signal respectively. 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 to minute excitation signals.

[4] This is close to the chat volume at 1 m away [11].

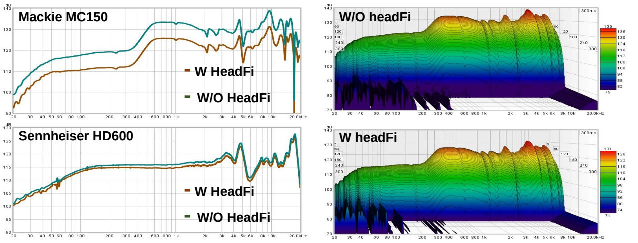
2.4.2 Detection sensitivity on indirect excitation signal.

Some applications do not produce a direct excitation signal. For example, to detect whether the user puts on the headphones or not, the smartphone itself needs to emit an acoustic signal and HeadFi then records the reflections of this signal as an indication of surrounding environment. In this benchmark experiment, we program a smartphone to send out a chirp signal with its frequency changing linearly from 20 Hz to 20 KHz. HeadFi then records the RSS of reflected 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 consideration of the responses over the entire chirp frequency band.

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

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 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 8. We then take the headphones off the dummy head and record the HeadFi' output again. We repeat this trail of experiments 54 times by replacing the headphones each time. Figure 9(b) shows the CDF of 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 when the headphones are ON the dummy head. We can see a clear gap between two lines in the figure, indicating HeadFi can pick up the environment change around the headphones. We also observe that the minimum difference between the headphones ON and OFF the dummy head is larger than the maximum difference when the headphones are always ON the dummy head. And this gap grows further for a single pair of headphones. This result clearly demonstrates that HeadFi is sensitive to the minute change of indirect excitation signals. Our evaluation on user identification further demonstrates that HeadFi is sensitive enough to differentiate two twin girls (§3.2.2) by profiling their unique ear canals.



(a) The impact of HeadFi on headphones' frequency response (FR).

(b) A snapshot of the cumulative spectral decay (CSD).

Figure 10: 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.

2.4.3 Impact on the sound quality.

One may fear HeadFi contaminates the output signal (*e.g.*, music), since it wires the headphones and the pairing device as if it breaks in the audio chain. Hence we put two types of headphones on a MiniDSP E.A.R.S dummy head and measure the frequency response (FR) of these headphone in the presence and absence (for comparison) of HeadFi. Figure 10(a) shows the result. We observe two FR curves show a very similar pattern for both headphones, indicating HeadFi does not affect the frequency response of the headphone itself. The gap between two FR curves indicates the electrical signal experiences more attenuation in the presence of HeadFi. As a result, the user will hear less loud sound. This is due to the extra voltage loss when the electric music signal passes through HeadFi. To validate this observation, we further measure the cumulative spectral decay (CSD) of of a low-quality Mackie MC-150 headphones. CSD is a standard metric on measuring the performance of the driver. As shown in Figure 10(b), we observe two CSD snapshots exhibit a very similar pattern, indicating HeadFi has minimal impact on the headphones and their output signal. One issue with our current design is that HeadFi automatically transforms stereo sound into mono, since it takes only one channel of an audio input. We feel this is not a serious problem as most applications execute in demand. In addition, other applications like voice communication itself is mono and thus work well with HeadFi. We put a switch (S_1 in Figure 6) on HeadFi to allow bypassing HeadFi if needed.

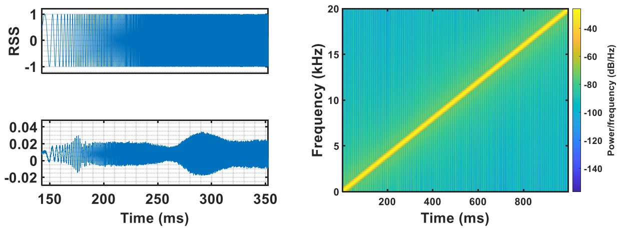
3 User Identification

We first demonstrate HeadFi can be used for user identification. The mainstream identification method – face recognition, does not scale well in poor light conditions. It also raises privacy concerns. HeadFi can be leveraged to check the user identity and unlock the phone (pairing device) when face recognition does not function well.

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* [34]: when a seashell is attached to the ear, the ambient acoustic noise will resonate within the cavity of the seashell, which amplifies the noise of a certain frequency. One can thus hear ocean-tide like sound from the seashell. Similarly, once the user puts on her headphones,



(a) The transmitted and received chirp in time domain.

(b) The spectrogram of the transmitted chirp.

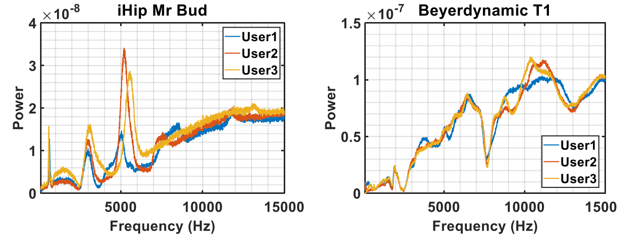
Figure 11: An illustration of chirp signal. (a) The transmitted (top) and received chirp signal in time domain. (b) The spectrogram of a transmitted chirp whose frequency spans from 20 Hz to 20 kHz in one second.

the headphones, ear canal, and eardrum together establish a resonance chamber, which amplifies the ambient acoustic noises. This amplified noise leads to a higher voltage signal output by HeadFi. Based on this observation, we use the RSS and its standard deviation (σ) to conduct ON-OFF detection. These two values jump to much higher levels when the user puts on her headphones.

Identification. Since the plugged headphones work in a quasi-full-duplex manner, 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 inner ear structure of the user. The two drivers of the headphones receive an echo signal that characterizes the channel response of the ear canal. Figure 11(a) and 11(b) show the chirp signals in time 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. However, the ear-related physiological uniqueness not just exists between two persons, but also between two ears of the same person [35, 38]. Hence the channel response measured at two ears would not be the same. Figure 12 shows the channel response measured by HeadFi on three different persons. We can see the channel response are dramatically different in frequency band 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 signal with a wavelength of sub-centimeters (≥ 3 kHz).

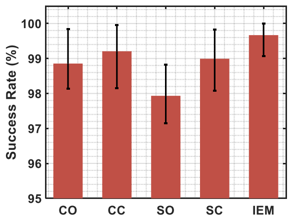
Proof-of-concept. As a proof-of-concept, we use support vector machine (SVM), a light-weight classifier for user Identification. Specifically, we collect multiple copies of the user's echo chirp as positive samples. We then collect the same amount of negative samples by putting the headphones on the E.A.R.S dummy head. Finally we train a binary SVM classifier and perform k-fold [17] cross-validation.



(a) iHip Mr Bud ($ 3.6)

(b) Beyerdynamic T1 ($ 999)

Figure 12: Channel response of three persons characterized by (a) low-end and (b) high-end headphones.



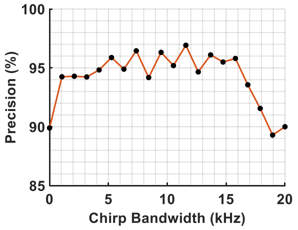


Figure 13: ON-OFF detection across headphones.

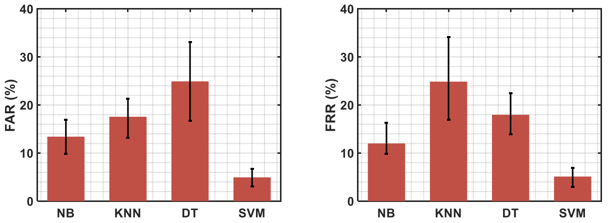
Figure 14: Precision over chirp bandwidth.

3.2 Experiment

The experiments involve 27 participants (7 females and 20 males), including a 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 take off the headphones and put it on again each time we record an echo chirp. We record 50 echo chirps for 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 headphone types and show them in Figure 13. We observe that the success detection 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 goes 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 mechanism to validate our system performance, where k is 5. We use precision [25] as our evaluation metric. A high precision value indicates only the authorized users can successfully pass the verification. Figure 14 shows the precision in 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, which aligns with the observation in Figure 12. The precision then fluctuates around 95% as we increase the bandwidth to 15 kHz. It then drops to around 90% as the frequency



(a) FAR.

(b) FRR.

Figure 15: We study the identification performances for four classifiers. (a) FAR results. (b) FRR results.

		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.

bandwidth goes beyond 15 kHz. We believe this is because there are subtle displacements during multiple rounds of putting on headphones and this sub-*mm* scale variation in displacement can be captured by the high frequency (larger than 15 kHz) signal, which disturbs the user identification. Suggested by this study, we employ 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 are in this frequency range.

Impact of different classifier. Next we evaluate the identification performances across 4 classifiers, Naive Bayesian (NB), *k*-nearest neighbors (KNN), decision tree (DT), and SVM. We conduct experiment with multiple subjects. We then investigate and report the false acceptance rate (FAR) and false rejection rate (FRR). As shown in Figure 15(a) and 15(b), SVM achieves the best performances for both FAR and FRR. We envision more advanced learning techniques (*e.g.*, DNN) can further improve the identification performances.

Differentiating twins. We further conduct user identification 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 performances for twins are 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 for the *k*-fold cross validation.

Impact of human motion. We conduct user identification when the subject is sitting still, moving her head, eating, and walking. From the result we observe that stronger body movements degrade the user identification performance. HeadFi achieves lowest false rejection rate when the subject is sitting still. The false rejection rate grows as user starts to move, *e.g.* eating, walking or moving her head. This is expected since the headphones is likely to move with human head and alter the channel response.

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 performances over time.

Status	Sit	Head	Eating	Walk
FRR (%)	3.64	4.75	5.15	8.75

Table 3: Impact of motions.

Long-term user identification We further track one volunteer over two months and observe the performance of our user identification service 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 is the physilogical characteristics on this subject change over time. For example, the ear fluid or other conditions can change the frequency response of the ear canal, which impacts the user identification performance. [22] To validate this hypothesis, we conduct user identification after the shower and observe a 7% drop on the identification accuracy.

4 Physiological Sensing

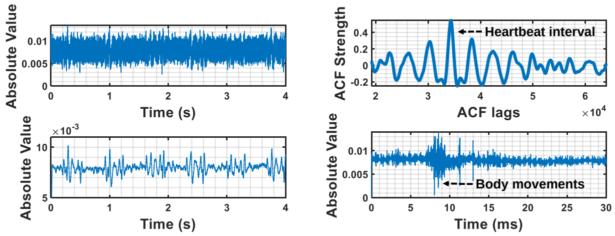
Next, we demonstrate the feasibility of using HeadFi to detect subtle physiological signals. Vital physiological sign sensing plays a key role in human health monitoring. HeadFi can empower users to continuously and accurately monitor a variety of key physiological activities (*e.g.*, heart rate) using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

4.1 Signal Processing

Detecting the heart rate is challenging due to the extremely weak excitation signal induced by the subtle blood vessel variation in the ear canal. As shown in the top figure of Figure 16(a), the signal can be buried by the noise. Also, this weak signal can be easily interfered by the headphone user's motions. To solve these challenges, 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 16(a) (bottom). We then leverage the auto-correlation function (ACF) to identify the periodicity which corresponds to the heart 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 signals output from HeadFi and k is the lag. N is the length of the received signals. Figure 16(b)



(a) The voltage output before and after filtering.

(b) The ACF plot and time domain interference.

Figure 16: (a): heartbeat signal becomes clear after filtering. (b): (top) we adopt ACF to calculate the heart rate and (bottom) we show an example time domain interference caused by body movement.

(top) shows an example of auto-correlation output. The sample index the peak value is located 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 an upper (U) and lower (L) bound of k based on the possible heart rate of human beings (35 -200 bpm [57]). Our goal can be represented by the following function:

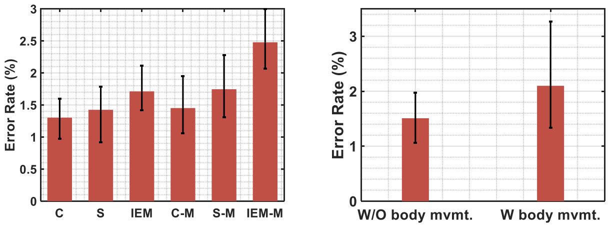
$$k^{\star} = \underset{k \subseteq (L,U)}{\arg\max}\, r_{xx}(k). \tag{4}$$

The heart rate can be calculated as $R_{BPM} = 60 \cdot \frac{F_s}{k^{\star}}$, 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 16(b) (bottom). We thus truncate the voltage output from HeadFi into windows and calculate R_{BPM} within each window. We then apply an outlier detection algorithm [50] to filter out those outlier estimations and average the remaining to obtain the heartbeat rate.

4.2 Experiment

Next, we evaluate the performance of heart rate monitoring. Each measurement lasts for 40 seconds. We truncate a recording session using a window size of four seconds, with two seconds overlapping. The participant measures her heart rate 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 [5]. We use error rate (ER) to measure the performance of our hear rate monitoring: $ER = \frac{|R_{HF} - R_{PO}|}{R_{PO}}$, where R_{HF} and R_{PO} are the heart rate reported by HeadFi and the oximeter, respectively.

Impact of the cut off frequency F_C. We first change the cut off frequency of the low pass filter from 2.4 Hz to 240 Hz and measure the error rate in each cut off frequency setting. The participant listens to the music throughout the experiments. We observe that the error rate stays in a low level (below 2.0%) when the cut-off frequency is lower than 50 Hz.



(a) Impact of ambient noise.

(b) Impact of body movement.

Figure 17: Error rate of heart rate estimation. (a): We measure the error rate both in absence (the first three columns) and 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.

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.

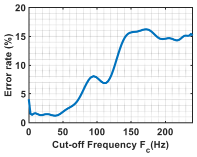
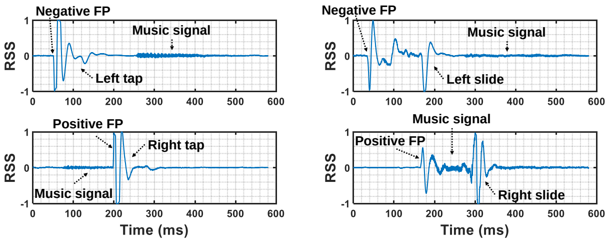


Figure 18: Impact of F_C.

Heart rate monitoring accuracy. We evaluate the accuracy of the heart rate estimated by HeadFi using all 54 pairs of headphones. We categorize these results into three groups based on the headphones been used: circumaural headphones (C), supra-aural headphones (S), and in-ear model (IEM). The result is shown in Figure 17(a). We observe HeadFi achieves consistently high performance 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 when connecting to 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 other two types of headphones. The maximum error rate achieved by HeadFi is around 3%, which satisfied the requirement (less than 5%) of many commercial heart rate monitoring systems [45]. These results demonstrate the feasibility of using HeadFi to measure user's heart rate even in the presence of strong interference signals (*e.g.*, music).

Impact of body movement. In this experiment, the participant puts on/off the headphones occasionally during the testing, which essentially brings in a strong interference signal. We then These experiments involve 27 participants including 7 females and 20 males with their age spanning from 27 to 55 years old. Figure 17(b) shows the error rate



(a) Tapping

(b) Sliding

Figure 19: 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.

in the presence of body movement. We also show the error rate in the absence of body movement for comparison. We observe a slight increase in the error rate (0.59% on average) in the presence of strong body movements, while the overall error rate is still less than 3%, well below the requirement of many commercial heart rate monitoring system (< 5%).

5 Touch-based Gesture Recognition

We next demonstrate the feasibility of transforming the enclosures of the non-smart headphones into virtual touchpads with HeadFi. The rationale behind is that the variation in the output voltage V_g caused by different gestures manifests unique features in both spatial and temporal domains.

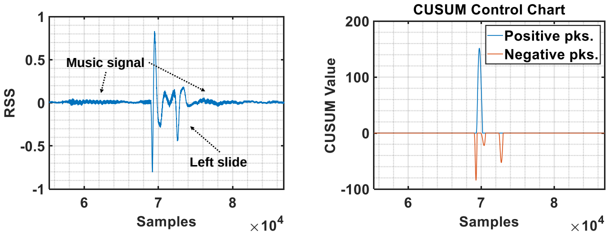
Without loss of generality, we define four touch-based gestures: *i*) tapping the left enclosure – pause or play; *ii*) tapping the right enclosure – mute; *iii*) sliding on the left enclosure – volume up and *iv*) sliding on the right enclosure – volume down. Note that the gestures can be supported by HeadFi are not limited to these four gestures.

5.1 Signal Processing

Our basic idea is to analyze the temporal features of the output voltage signals for gesture recognition.

Distinguishing left tapping and right tapping. We invite a volunteer to tap the left and right enclosure of her headphones and record the RSS of HeadFi' output. As shown in Figure 19(a), when there is a tap on the headphone, we can always observe multiple peaks. Taking a step further, when the user taps her left enclosure, there is a negative peak showing up on the RSS samples, followed by a positive peak, as shown in Figure 19(a) (top). In contrast, the positive peak shows up ahead of the negative peak when the user taps her right enclosure (Figure 19(a) (top)). This is because the Wheatstone bridge measures the differential voltage between the two drivers of headphones. Thus, the excitation signals measured at the bridge are *phase inverted* for right tap and left tap. Note that the echos 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



(a) Before applying CUSUM

(b) After applying CUSUM

Figure 20: Output signal V_g before and after applying CUSUM.

tapping, 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 19(b). We can thus leverage the peak interval to distinguish them.

Algorithm. We adopt cumulative sum (CUSUM), a lightweight 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 the negative peaks and keeps the the large positive peaks only. We thus build the second CUSUM function by replacing the *max* with a *min* operation in order to keep the large negative peaks. The output voltage samples go through these two CUSUM functions (*max* and *min*) in parallel. Figure 20 shows the signal before and after applying the CUSUM operation, respectively. We observe that the impact of ambient music signals has been wiped out after applying CUSUM, leaving us the peaks only. We then determine left sliding/tapping or right sliding/tapping applying 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 them among different individuals, and empirically set a threshold of 5000 samples (equivalent to 0.1s at 48000 Hz sampling rate).

5.2 Experiment

We use an 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 gesture data are offloaded to a laptop for analysis. Table 5 and 6 show the confusion matrix of the classification result. The

		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 being playing.

overall classification result is consistent across four gestures in both quiet (without audio input signal) and noisy (with audio input signal) conditions. We achieve 99% classification accuracy in the absence of the audio input signal. The classification result drops slightly to around 98% in the presence of audio input signal. This result demonstrates the feasibility of applying HeadFi to enable touch-based gestures 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.

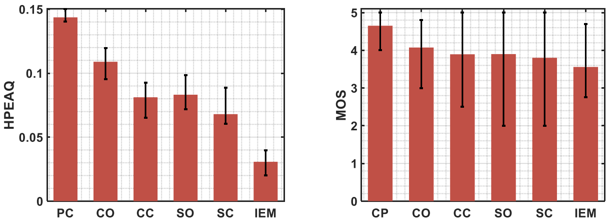
6 Voice Communication

Last but not least, we demonstrate the feasibility of using HeadFi to enable full-duplex voice communication on those headphones lacking a dedicated 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 (air, bones, tissue *etc.*).

The impact of echoes. One interesting issue which may exist with our design is the echo. This is because HeadFi captures the voices from not just the HeadFi user side but also the other side[5] at the headphone's diaphragm and both captured voices will be sent to the other side. Thus, the other side may hear an echo of the 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 [55]. Therefore, echoes would not prohibit the use of HeadFi. Our evaluation results also confirm this. Next, we present our 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 particularly designed for evaluating the degradation of audio signal introduced by telephony network, not the audio signals from the end-device. Motivated by PEAQ, we conduct an objective evaluation on the voice quality by correlating the voice recorded by HeadFi with the one recorded by the built-in microphone of a smartphone. We refer 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[6]. We use both the built-in microphone on an iPhone 6 and HeadFi to record this echo. We then compute the HPEAQ value of these two signals to measure theire similarity.



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

Figure 21: (a) Object and (b) subject evaluations on the quality of voice call over different types of headphones.

We repeat this experiment using 54 pairs of different headphones and summarize the results based on their categories in Figure 21(a). For comparison, we also record the audio using the embedded microphone on an HP spectre X360 laptop and compute the HPEAQ value (termed as PC in the figure). We observe PC leads the board on HPEAQ score, followed by CO, CC, SO, and SC. IEM headphones achieve the lowest HPEAQ score. This is because the IEM types of headphones go deeper into the ear and thus can only capture those excessively attenuated 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 attenuate less.

Mean opinion score (MOS). Besides the objective evaluation, to better understand the HPEAQ discrepancy over the five headphones, we further conduct an subjective evaluation on the voice quality by MOS experiments. MOS is another widely adopted evaluation 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.

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

[6] The frequency of human voice ranges from 300 Hz to 3 KHz [9].

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 the 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 to HeadFi for talking. For comparison, we also employ the built in microphone on an iPhone 6 (CP) for voice call 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 [8]):

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 21(b) shows the MOS distribution for the five tested headphones with our design and the reference smartphone with a built-in microphone. Note that the smartphone employs a dedicated high-end audio amplifier and active noise reduction circuits in the microphone front-end [33, 43]. 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 descent. 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 across five types of headphones with the objective HPEAQ values (Figure 21(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 human auditory system discussed in literature for decades [52, 54]. 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 audio grade amplifier.

7 Related Work

We discuss related works in this section.

Touch-based gesture control. The touch-based gesture control is usually realized by adding capacitive or resistive sensors into headphones. Many smart headphones such as Microsoft Surface Headphones [53], Sony 1000XM3 headphones [3], Zealot B21 headphones [2], and Bose NC700 headphones [4] come with this function. It enables users to send audio related commands by simply performing predefined gestures around the headphones.

Physiological sensing. There is a growing trend on putting sensors onto headphones for physiological sensing. For instance, monitoring bioinformation such as electrocardiography (ECG), ballistocardiography (BCG), and photoplethysmography (PPG) for heart rate, respiratory rate, and blood pressure monitoring [24, 47, 48, 56, 59]. Bui *et al.* adopted PPG sensors and developed an in ear system to measure the blood pressure [19]. Anh *et al.* proposed to customize an in-ear sensor to measure the brain activities [41]. Rupavatharam *et al.* proposed to use the IMU in a specifically designed headphones to monitor jaw clenching [51]. Roddiger *et al.* developed a respiration rate monitoring system using the embedded IMU [49]. There is also an open source multi-sensor integrated research platform, eSense, for earable computing research [13]. HeadFi differs from these designs on the way of designing the sensing platform.

User authentication. Leveraging the unique physical structure of the ear canal to authenticate users has been well investigated in the past years. Arakawa *et al.* proposed to use the mel-frequency cepstral coefficients (MFCC) instead of frequency domain transfer function for higher authentication accuracy [16]. Higashiguchi *et al.* proposed to use the built-in microphones on a cellphone to perform ear related user authentications [30]. Akkermans *et al.* and Mahto *et al.* studied the feasibility of using inaudible pilot tone for user authentications [15, 39]. Gao *et al.* designed an ear related user authentication system using commercially available headphones [29].

8 Conclusion

We have presented the design, implementation, and evaluation of HeadFi, a new design paradigm for smart headphones. HeadFi employs the pair of drivers on headphones as a versatile sensor to enable new functionalities as opposed to adding embedded sensors. This unique feature 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 earable applications using 54 pairs of headphones.

References

[1] Gig fix: Turn your headphones into a mic. Webpage, 2014.

[2] B21 super bass wireless bluetooth headphone stereo touch control headset noise cancelling with micro. Webpage, 2018.

[3] Wireless noise canceling stereo headset wh-1000xm3. Webpage, 2018.

[4] Bose noise cancelling headphones 700. Webpage, 2019.

[5] Cms50d1a gehp040ahus pulse oximeter. Webpage, 2019.

[6] Earphones and headphones market size, industry report. Webpage, 2019.

[7] Global unit sales of headphones and headsets from 2013 to 2017. Website, 2019.

[8] P.800.1 : Mean opinion score (mos) terminology. Webpage, 2019.

[9] Voice frequency. Webpage, 2019.

[10] The differential amplifier. Webpage, 2020.

[11] How sound works. Webpage, 2020.

[12] Minidsp e.a.r.s. Webpage, 2020.

[13] A research space for earable computing. Webpage, 2020.

[14] Apple Airpods. Website.

[15] Anton HM Akkermans, Tom AM Kevenaar, and Daniel WE Schobben. Acoustic ear recognition for person identification. In *AutoID*. IEEE, 2005.

[16] Takayuki Arakawa, Takafumi Koshinaka, Shohei Yano, Hideki Irisawa, Ryoji Miyahara, and Hitoshi Imaoka. Fast and accurate personal authentication using ear acoustics. In *APSIPA*. IEEE, 2016.

[17] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 2004.

[18] BOSS QC-35 Wireless Headphones. Website.

[19] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *MobiCom*, 2019.

[20] Bruce Carter. *Op Amp noise theory and applications, 12.3.2 Thermal Noise*. Elsevier, 2009.

[21] Bruce Carter and Thomas R Brown. *Handbook of operational amplifier applications*. Texas Instruments Dallas, Tex, USA, 2001.

[22] Justin Chan, Sharat Raju, Rajalakshmi Nandakumar, Randall Bly, and Shyamnath Gollakota. Detecting middle ear fluid using smartphones. *Science translational medicine*, 11(492), 2019.

[23] John Clarke, Claudia D Tesche, and RP Giffard. Optimization of dc squid voltmeter and magnetometer circuits. *Journal of Low Temperature Physics*.

[24] David Da He, Eric S Winokur, and Charles G Sodini. An ear-worn continuous ballistocardiogram (bcg) sensor for cardiovascular monitoring. In *EMBC*. IEEE, 2012.

[25] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[26] Christian C Enz and Gabor C Temes. Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization. *Proceedings of the IEEE*, 1996.

[27] Joel Gak, Matías Miguez, Martín Bremermann, and Alfredo Arnaud. On the reduction of thermal and flicker noise in eng signal recording amplifiers. *Analog Integrated Circuits and Signal Processing*, 2008.

[28] Samsgung Galaxy Buds. Website.

[29] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. Earecho: Using ear canal echo for wearable authentication. *IMWUT*, 2019.

[30] Yutaka Higashiguchi, Yoshinobu Kajikawa, and Shunsuke Kita. A personal authentication system based on pinna related transfer function. In *ICBAKE*. IEEE, 2017.

[31] Karl Hoffmann. *Applying the Wheatstone bridge circuit*. HBM Germany, 1974.

[32] Texas Instruments. Noise analysis in operational amplifier circuits. *Application Report, SLVA043B*, 2007.

[33] Thomas M Jensen, Vladan Bajic, and Andrew P Bright. Active noise cancellation using multiple reference microphone signals, May 3 2016. US Patent 9,330,652.

[34] Liang-Ting Jiang and Joshua R Smith. Seashell effect pretouch sensing for robotic grasping. In *ICRA*. IEEE, 2012.

[35] Agnès Job, Paul Grateau, and Jacques Picard. Intrinsic differences in hearing performances between ears revealed by the asymmetrical shooting posture in the army. *Hearing research*, 1998.

[36] Ron Kapusta, Haiyang Zhu, and Colin Lyden. Sampling circuits that break the kt/c thermal noise limit. *IEEE Journal of Solid-State Circuits*, 2014.

[37] Ronald A Kapusta, Katsufumi Nakamura, et al. Methods and apparatus for reducing thermal noise, 2007. US Patent 7,298,151.

[38] F Laurain King and Doreen Kimura. Left-ear superiority in dichotic perception of vocal nonverbal sounds. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 1972.

[39] Shivangi Mahto, Takayuki Arakawa, and Takafumi Koshinak. Ear acoustic biometrics using inaudible signals and its application to continuous user authentication. In *EUSIPCO*. IEEE, 2018.

[40] Henrik Møller, Dorte Hammershøi, Clemen Boje Jensen, and Michael Friis Sørensen. Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society*, 1995.

[41] Anh Nguyen, Raghda Alqurashi, Zohreh Raghebi, Farnoush Banaei-Kashani, Ann C Halbower, and Tam Vu. A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring. In *SenSys*, 2016.

[42] Viet Nguyen, Siddharth Rupavatharam, Luyang Liu, Richard Howard, and Marco Gruteser. Handsense: capacitive coupling-based dynamic, micro finger gesture recognition. In *SenSys*, 2019.

[43] Guy C Nicholson. Active noise cancellation decisions in a portable audio device, August 20 2013. US Patent 8,515,089.

[44] Sean Olive, Omid Khonsaripour, and Todd Welti. 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, 2018.

[45] Alexandros Pantelopoulos and Nikolaos G Bourbakis. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*.

[46] Melih Papila, Raphael T Haftka, Toshikazu Nishida, and Mark Sheplak. Piezoresistive microphone design pareto optimization: tradeoff between sensitivity and noise floor. *Journal of microelectromechanical systems*, 2006.

[47] Ming-Zher Poh, Kyunghee Kim, Andrew Goessling, Nicholas Swenson, and Rosalind Picard. Cardiovascular monitoring using earphones and a mobile device. *IEEE Pervasive Computing*, 2010.

[48] Ming-Zher Poh, Kyunghee Kim, Andrew D Goessling, Nicholas C Swenson, and Rosalind W Picard. Heartphones: Sensor earphones and mobile application for non-obtrusive health monitoring. In *ISWC*. IEEE, 2009.

[49] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl. Towards respiration rate monitoring using an in-ear headphone inertial measurement unit. In *EarComp*, 2019.

[50] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2005.

[51] Siddharth Rupavatharam and Marco Gruteser. Towards in-ear inertial jaw clenching detection. In *EarComp*, 2019.

[52] Otto Stuhlman Jr. The nonlinear transmission characteristics of the auditory ossicles. *The Journal of the Acoustical Society of America*.

[53] Microsoft Surface Headphones. Website.

[54] Frédéric E Theunissen, Kamal Sen, and Allison J Doupe. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*.

[55] Voice Quality Enhancement and Echo Cancellation. .

[56] Stefan Vogel, Markus Hülsbusch, Thomas Hennig, Vladimir Blazek, and Steffen Leonhardt. In-ear vital signs monitoring using a novel microoptic reflective sensor. *IEEE Transactions on Information Technology in Biomedicine*, 2009.

[57] Joseph C Volpe Jr. Heart rate monitor for controlling entertainment devices, April 8 2008. US Patent 7,354,380.

[58] Susan E Voss and Jont B Allen. Measurement of acoustic impedance and reflectance in the human ear canal. *The Journal of the Acoustical Society of America*.

[59] Eric S Winokur, David Da He, and Charles G Sodini. A wearable vital signs monitor at the ear for continuous heart rate and pulse transit time measurements. In *EMBS*. IEEE, 2012.

[60] Myung-Gyoo Won, Jae-hoon Kim, and Jong-wook Park. Temperature sensing circuit for use in semiconductor integrated circuit, September 12 2006. US Patent 7,107,178.

[61] Daniel Yum. Bandgap voltage reference circuit, October 1 1991. US Patent 5,053,640.