Earable Computing: A New Area to Think About

Romit Roy Choudhury
University of Illinois at Urbana Champaign (UIUC)
USA

ABSTRACT

This position paper argues that earphones hold the potential for major disruptions in mobile, wearable computing. The early signs are positive and the industrial wheels are in motion. However, whether earphones truly become a disruptive new platform, or stop at being a useful accessory, could depend on whether we – the mobile computing researchers – deliver. If we do, tomorrow's earphones will run augmented reality via 3D sound, will have Alexa and Siri whisper just-in-time information, will track our motion and health, will make authentications seamless, and much more. The leap from today's earphones to "earables" could mimic the transformation we saw from basic-phones to smart phones. On the other hand, if we are unable to provide some of the disruptive building blocks, tomorrow's earphones may saturate in its capabilities. We believe this is an important juncture in time where the mobile computing research community has an opportunity to shape the future. This paper aims to discuss this landscape, including some challenges, opportunities, and applications.

CCS CONCEPTS

• **Computer systems organization → Embedded and cyber physical systems; wearable sensors and devices**.

KEYWORDS

Hearables, earphones, wearable, sensing, signal processing, embedded systems, acoustics, AR/VR, security, healthcare, localization.

ACM Reference Format:
Romit Roy Choudhury. 2021. Earable Computing: A New Area to Think About . In *The 22nd International Workshop on Mobile Computing Systems and Applications (HotMobile 2021), February 24–26, 2021, Virtual, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3446382.3450216

1 INTRODUCTION

This paper uses the term "earables" to refer to wearable devices around the ear and head, including earphones, hearing aids, and electronics-embedded glasses. Given their proximity to the ears, these earables are expected to interact with humans mostly through acoustics, i.e., listening to voice commands and whispering back information. A recent survey [1] reports that the "earables" market is seeing a rapid uptick in investments, with strong projections into the future. New features are beginning to roll out and innovative products are on the horizon. Let us briefly survey this active landscape; this should serve as useful context for identifying the research space around earable computing.

Hardware: A serious push is evident on the hardware front. Multimodal sensing, native digital computing, and various communication capabilities are getting embedded into the earphones. For instance, Apple, Samsung, Qualcomm have all added microphone arrays [2] to enable beamforming, MIMO, and signal separation. Samsung introduced in-ear microphones in their earbuds [3], aimed at sensing acoustic signals from inside the ear canal. Apple embeds dual accelerometers in their Airpods, designed to sense surface vibrations on the face/ear when a user is speaking. Their patents [4] describe that such surface vibrations help the device determine when its own user is talking to Siri, as opposed to someone else nearby. IMU sensors are also being used for understanding human gestures, such as walking, running, and exercising in the gym [5]. On the computing side, Qualcomm has released QCS400 [6] with *native support* for low energy wake-word detection and far-field automatic speech recognition (ASR). Oticon and Dolby [7, 8] are embarking on ambitious attempts to adopt brain-signal interfaces, to play music based on the user's mood. Bose has introduced Frames that are capable of native DSP and motion tracking of the human head; Amazon has released "Echo Frames", eye-glasses loaded with Alexa and edge-computing capabilities. There is clear evidence that the earable hardware platform is poised for rich innovation in the next few years.

Software: Software and computing are also on the rise. Waverly Labs [9] is promising near-real time language translations across 20 languages, supplemented with microphone arrays to beamform to the speech from both parties in the conversation. Nura (a startup) and NEC labs are pursuing a radical idea of shining sounds into the ear (ideally music), learning the ear-canal's impulse response, and personalizing music to the user's frequency response curve [10]. Bragi is envisioning pass-through earbuds [11] that would selectively allow some ambient sounds to be audible; they also use sensors to support a simple vocabulary of head gestures for controlling music and other operations. The hearing aids market is warming up as well [8], making renewed attempts to amplify only the sound of interest to the user. Finally, Oculus, Dolby, Bose and several others are keen on 3D sound [12] and acoustic augmented reality [13], enabling the user to listen to object annotations (say information about a painting) as he/she turns her head towards it.

Of course, it is prudent to question whether such visions will hold over time. *Or could they be hypes, like many others, that become viral for a while and then fade away due to various misalignments with reality?*

HotMobile 2021, February 24–26, 2021, Virtual, United Kingdom

ACM ISBN 978-1-4503-8323-3/21/02…$15.00
https://doi.org/10.1145/3446382.3450216

While it is difficult to accurately predict such trends, we make 3 general observations in favor of earables.

[1] Screens and keyboards are the primary interface between humans and mobile devices today. Although immensely successful, one issue with screens and keyboards is their intrusion into human attention. Reading and typing on mobile screens both require a cognitive context-switch; the speed of information exchange is also not very high, especially with typing. In contrast, speaking and listening are cognitively more seamless and arguably faster, and becoming viable at this point due to speech recognition, NLP, and conversational AI. Earables are well positioned to leverage this opportunity, enabling a new modality of hands-free, less-disrupting information exchange.

[2] Phones, watches, and fitbits have ushered the revolution around "quantified self" [14, 15], allowing humans to sense and measure themselves at fine granularities. Earables may become the gateway to sensing the *head, face, and upper body*. It may now be possible to understand user speech, analyze and control sounds we hear, sense eating and drinking habits, track gazing directions, and even sense biomarkers from around the brain, ears, and mouth. Valuable data from this region of the "self" was largely missing till this point.

[3] An earable platform already sits well in society. Humans have been using earphones for music and phone calls for more than a decade now, and the popularity is growing rapidly with wireless earbuds and wireless charging. Enriching this platform is less likely to bring new surprises to humans, unlike say, the Google Glass. This reduced uncertainty in adoption can be an important driver for research and innovation in this field.

While the above arguments are promising for earables, there are obviously hurdles and challenges along the way. But first, are there technical show-stoppers?

It is difficult to exhaustively foresee all the pitfalls in the roadmap. We discuss some of the foundational make-or-break factors, but none appear to be an immediate show-stopper.

■ **Energy** is a prime concern given that microphones operate at a non-trivial energy floor of ≈3-5 mWatt [16]. Continuously listening for voice commands and sensing human activities can easily drain the small battery. However, piezoelectric microphones and IMUs are being ushered in as energy-efficient front-ends to microphones – these MEMS sensors would triage signals and wake-up the full microphone stack only when necessary. Further, newer hardwares are offering native support for speech and signal processing, lowering the power consumption baselines appreciably. Apple airpods are already running Siri, although for 4 – 5 hours. We believe a push for energy is critical and a prime research topic for our OS/embedded systems community.

■ **Discomfort:** Continuously wearing an earphone is uncomfortable; blocking the ear for long durations may even be harmful since it prevents air circulation into the ear canal [17], increasing chances of infections. This might pose a serious road-block to creating an always-wearable device. However, hollow or "open-ear" designs are already under consideration in the academic and start-up communities [11, 18]. It may even be possible to design earables that mechanically toggle between "open" and "closed" modes. Imagine a camera-shutter that closes automatically to block external noise and provide high quality music, phone call, and noise cancellation. At other times, the shutter remains open for improved comfort and modest-quality sound experience (such as in glasses).

■ **Privacy:** With always-ON microphones in earphones, acoustic privacy is indeed a concern. Eavesdropping into ambient conversation will become easy; other inaudible attacks can be launched. While privacy erosion is not unique to earables, we believe there is opportunity to mitigate the problem. For example, microphones need not be ON all the time; piezo-electric or IMU sensors can recognize the user's voice and wake-up the microphone on-demand. Apple airpods are already at the preliminary stages of such IMU-based capabilities [19]. More ideas will emerge with time.

■ **Health and Radiation:** One could question if the RF radiation, due to bluetooth communication between earphones and smartphones, is a cause of concern. Today's wireless earbuds also face this question, but with day-long use of the earphones, the exposure to radiation could be greater. We do not have a good answer to this question, except that Google Glass was able to develop antenna solutions that beamform the signal away from the head. This cleared various levels of FCC approvals for the launch of Google glasses. We believe that such capabilities can be brought to earphones as well, but we admit this is point of critical research.

Assuming the above are not show-stoppers, the rest of the paper explores topics that could usher new/important capabilities into earphones. Some topics address today's pain-points while others envision opportunities for enabling new applications. Obviously, these topics are hardly exhaustive – many important problems are omitted due to limited expertise of the author and space restrictions (e.g., exciting progress in healthcare [20–23], hardware miniaturization and neural networks [24, 25], edge computing from earphones [26], HCI, brain computer interfaces (BCI), etc.). This paper should rather be viewed as an assortment of few problems to whet the community's appetite; to invite them to think, shape, and guide this emerging problem space.

2 SOME RESEARCH PROBLEMS

2.1 Speech Recognition at Low SINR Regimes

Many users prefer to speak softly to their earphone's voice assistants, especially in public places. Moreover, everyday environments like airports, cafes, and gyms, are noisy. This implies that voice commands would be recorded at low SINRs at the earphones. In winter, hats and and scarfs will further attenuate voice signals to the ear. A clear challenge is to decode speech at this low-SNR regime, known to be a classical/difficult problem in signal processing [24, 25].

While deep learning is the first (and perhaps appropriate) reaction to this problem, we see potential opportunities in sensing and signal processing. Specifically, when a user speaks, we have observed that the earphone-IMU senses surface vibrations from the face and skull, essentially because the throat produces and radiates out such vibrations. More importantly, IMUs remain unpolluted by ambient noise

since air vibrations are not strong enough to produce a response in these MEMS devices. This begs the natural question:
Can IMU and microphone signals be harnessed jointly to achieve speech recognition at low SINR?

The problem is not trivial. Figure 1(a) shows the microphone signal recorded from the earphone when the user gives a voice command; Figure 1(b) shows the corresponding IMU signal. Observe that the IMU signal is confined to a much narrower bandwidth of 500Hz (compared to 20kHz for the microphone), causing heavy *aliasing*. In addition, the speech vibrations arrives to the IMU over an unknown channel (i.e., through throat muscles, jaw bones, and tissues) that is likely non-linear and hard to estimate. Yet, the ambient noise that pollutes the microphone's spectrogram is entirely absent in the IMU data – a truly valuable opportunity.

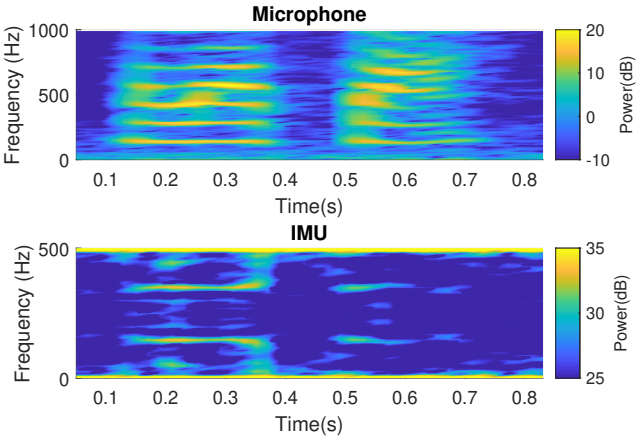

Figure 1: (a) 44kHz microphone recording of user speech (b) IMU's recording of same speech at 500Hz.

In this context, Figure 2 formulates the research problem in terms of the inputs and outputs. To decode low SINR speech (i.e., whisper decoder), the first input is the microphone signal, composed of the true voice signal $V(t)$ arriving over the air channel h_{air}, and added with ambient noise, $N_{background}$. The second input is the same voice signal $V(t)$, arriving over the solid body channel $h_{solid}(t)$, and then aliased due to sub-Nyquist sampling. The challenge for the "Whisper Decoder" algorithm is to estimate voice signal $\hat{V}(t)$ such that it can pass the automatic speech recognition (ASR) test.

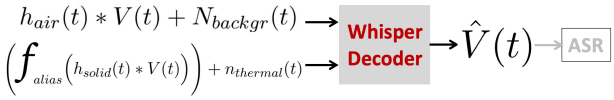

Figure 2: Formulation of joint (mic. + IMU) speech decoding.

2.2 Augmented Reality with 3D Sound

Humans have a natural ability to recognize the 3D direction, θ, from which a sound arrives to their ears. This is possible because, at a high level, the sound reaches the two ears at different times and the brain processes this *time difference of arrival* (TDoA), Δ, to infer the 3D angle. Now, if two earphones play a given sound with a time-separation Δ, the brain gets tricked into believing that the signal arrived from direction θ. These are called "binaural sounds" and is being actively studied for gaming and VR experiences [27].

3D binaural sounds can underpin various applications in next generation earables. As one example, Alice could ask her earphone's Siri to navigate her to a specific section of a museum, or to a platform in a railway station. Siri could simply respond by saying "follow me" but in a binaural mode, i.e., the voice appears to arrive from the direction in which Alice should walk. Thus, earables can provide a virtual voice escort that "hand-holds" Alice all the way to her destination [28]. Such services might be particularly valuable to the blind or elderly, who struggle to navigate complex environments.

Unfortunately, designing personalized binaural sounds is non-trivial. Briefly, this is because Δ is not a 1:1 function of θ, i.e., multiple arrival directions produce the same TDoA of Δ at the ears[1]. How can humans tell the difference between θ_i and θ_j then, when those angles produce the same Δ? This is where the "pinna" – the soft external part of the ear – plays an important role. The pinna scatters the impinging signals, creating distinct micro-echoes for different incoming angles (Figure 3). Additionally, the signals also bend (i.e., diffract) differently around the curvature of the head and nose, further breaking the ambiguities between θ_i and θ_j. When these head and pinna-filtered signal patterns arrive into the ear-drum, the brain is able to map them uniquely to the actual θ.

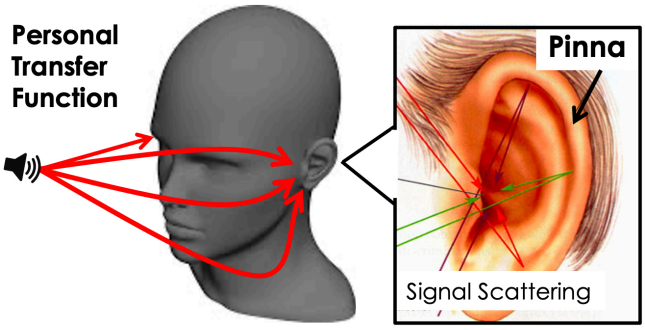

Figure 3: Simplified visualization of signal diffraction and scattering around the head and pinna.

Clearly, to produce accurate binaural experiences, earphones would need to synthesize this net effect of the pinna and head – together called the *head related transfer function* (HRTF). Today's technologies painstakingly create a coarse-grained, global HRTF from an artificial human model, and use this template across all users. Obviously, human pinnas and head/face shapes vary widely, hence using a global HRTF template is inadequate for precise binaural applications. This leads to the key research question:
Can personalized HRTFs be estimated for any user without visiting a special acoustic facility?

One possibility is for a user to visit an open space with minimal environmental echoes; wear an earphone with both microphones turned ON; play a known sound $x(t)$ from every 3D direction, θ_j; and record the corresponding signal y_{θ_j}. Then, the HRTF for each θ_j, denoted $HRTF_{\theta_j}$, can be estimated through deconvolution. Clearly, this is highly burdensome, not only in terms of the effort

[1]To understand this, consider a 2D hyperbola passing through the head, and the two ears being the center and focus of the hyperbola. From all points on this hyperbola, the Δ at the two ears is identical.

but also in estimating ground truth for each θ_j. An open question is to design practical methods for estimating personal HRTFs.

2.3 Motion and Activity Tracking

The motion sensors on earphones – namely the 2 IMUs, one on each side – presents an opportunity for tracking various aspects of human motion. Tracking head gestures (and correspondingly the gazing direction of the user) has already seen progress in recent papers [29–31]. More challenging problems relate to subtle motions that indirectly manifest on the IMU [32]. We pose 3 such problems as follows:

1. Is it possible to **sense breathing patterns of the user**, based on the faint periodic motion of the chest?
2. Can IMUs sense eating versus drinking, and if so, further **classify what the user might be eating?**
3. Finally, is it possible to **infer facial expressions**, given that facial muscle-motions manifest as IMU signals.

■ **Breathing:** Consider the problem of breathing pattern detection. It is clearly challenging since the inflation and deflation of the human lungs produces an extremely weak vertical oscillation of the head. With cheap IMUs in earphones, the oscillation signal is likely below the noise floor [23]. Nonetheless, it is also true that human breathing patterns are continuously present, and their frequencies do not span an excessive range. Moreover, in-ear microphones may be able to partially hear the breathing sound. With such information as initial conditions, and using weak signal-detection techniques (such as those used in astronomy), it may not be entirely infeasible. We believe this is a worthwhile challenge; an earphone or eyeglass can become a vehicle for continuous heart rate monitoring at minimal power consumption.

■ **In-mouth motion:** The lower jaw in the skull, called the *mandible*, is hinged at the back of the skull (below the ear), and moves up and down when a human chews food. The teeth also come in contact with each other and produce vibrations that propagate through multiple surface pathways to the ears [33]. The IMU picks up the aggregate of all these signals and exhibits variation based on the type and locations of the activity. Can we then infer what the user is eating and drinking?

Of course, deep learning is one possible approach. Users can train a neural network while they eat different types of food, and after a training phase, the earphone can begin to track food/drink consumption. On the other hand, modeling approaches are viable too. A great deal is now known about the anatomy and muscle/tissue structure, and by developing foundations on how vibrations travel through bones, muscles, and tissues, it may be possible to model the received signal at the IMU. If such models are feasible, then the action of the mouth and teeth can be formulated as an optimization problem. The objective function would be the loss score between the model and the measurement, and the constraints would be boundary conditions on the model parameters. Finally, the sounds of eating and drinking are also audible from the in-ear microphone, so acoustic signals can be valuable supplements to IMU. A multimodal solution to such a food tracking problem could draw a close to a long-standing important problem.

■ **Facial Expression:** Finally, the hardest of the 3 problems is to infer facial expressions from the IMU. The opportunity arises from the observation that multiple facial muscles stretch/contract for distinct facial emotions, which then manifest into motions picked up by the IMU [29]. Additionally, no acoustic signals are available, so the task is completely in a low dimensional motion space, and even polluted by head-motions that often accompany emotions (e.g., laughing with a head oscillation, or smiling and nodding simultaneously). The good news, however, is that even some basic progress in this direction can be useful – we envision that some outcomes of the study can influence security mechanisms, such as earphone authentication through facial gestures. If Bob uses Alice's headphones to snoop and listen to her emails, he would have to mimic Alice's facial expressions for authentication. The challenge is to produce adequate bits of entropy from facial expressions to thwart such attacks.

2.4 Security and Authentication

Earphones read out emails, messages, calendar events to the user, however, the user must first authenticate by entering her password/PIN number into her phone. This is non-ideal since wireless earphones allow users to move about without having the phone close at hand. A standalone authentication method is necessary on earphones, such as voice fingerprints. Unfortunately, voice synthesis has advanced significantly through deep learning – today it is easy for Bob to synthesize Alice's voice and use it as a replay attack. Given voice radiates over the air, it is not difficult for Bob to record Alice's voice without her permission, which makes such replay attacks easier than other forms of fingerprinting. Thus, the research question of interest is:
Can earphones support standalone authentication, freeing up the user from accessing the phone.

One opportunity relates to understanding the vibrations that IMUs pick up from the earphones when the user is speaking. Recall this vibration is sourced at the larynx of the throat, and conducts through jaw bones, muscles, and tissues, to ultimately reach the ears (Figure 4). We conjecture that this channel of conduction – a rich multipath channel – is quite unique to individuals; at the least, there are reasonable bits of entropy. Moreover, this channel is completely private to Alice. Thus, one can conceive an authentication method that characterizes this bone-muscle channel and uses it to enrich voice fingerprinting. Said differently, given the IMU records a signal $y_{imu}(t)$ for a human speech $v(t)$, can the corresponding body channel $h_{body}(t)$ be estimated from the standard equation:

$$y_{imu}(t) = v(t) * h_{body}(t) + n_{imu}(t)$$

where $n_{imu}(t)$ is the noise floor at the IMU.
The problem is non-trivial for the following reason. Observe that classical channel estimation techniques would have required accurate knowledge of the source signal, $v(t)$[2]. In this case, the source waveform is the voice $v(t)$, which is not accurately known; instead the microphone only records a modified version $y_{mic}(t) = v(t) * h_{air}$. The air-channel h_{air} – the air pathway from the mouth

[2]With known $v(t)$, the received signal $y(t)$ can be de-convolved with $v(t)$ to estimate the channel $h(t)$.

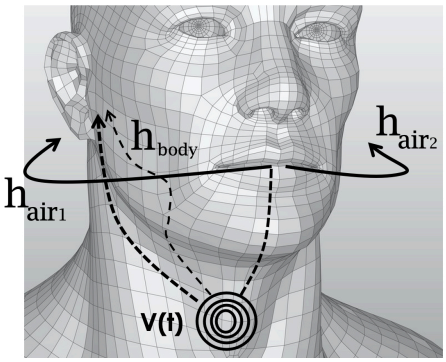

Figure 4: The air and body channels that conduct voice signals as air and surface vibrations.

to the ears – is difficult to measure, posing a challenge in estimating $v(t)$, ultimately affecting $\hat{h}_{body}(t)$.

Despite the challenge, emerging algorithmic techniques may be applicable here. These include (1) blind channel estimation (BCI), where the two earphone microphones may be able to pick up the same voice signal through relatively diverse channels h_{air_1} and h_{air_2}. Similarly, the two IMUs may also pick up y_{imu_1} and y_{imu_2} that are not identical since the left and right side of the skull/face may exhibit differences (e.g., the teeth structures are not symmetric). This spatial diversity may prove adequate for converging on h_{body}. (2) Perhaps neural networks can be applied here to train on an ensemble of words, ultimately leading to a robust estimate of the body channel. (3) Finally, the challenge can be eased by requiring the user to speak a specific password; since this voice waveform would be more predictable, the body channel can be estimated for only this word as $h_{body}^{password}$.

2.5 Blind Acoustic Beamforming

Pass-through earphones embed outward-facing microphones that listen to ambient sounds and replay into the ears. This allows environmental awareness (e.g., while walking on the roads, or for responding to family members at home while listening to music or news). Today, this replayed sound suffers from poor quality primarily due to a low-power microphone with no discriminative capabilities. This is a fundamental problem with hearing aids as well – the user is not able to "tune in" to one specific sound source, say the person speaking from the other side of a restaurant table.

The central research constraints are three-fold: (1) The earphone has limited microphones; adding more microphones is difficult and also not much profitable given the long wavelength of audible sound. (2) Unlike WiFi and BLE, the acoustic source signals are unknown [34]; this precludes channel estimation, thereby ruling out a host of beamforming algorithms. (3) The head moves frequently and the adaptive algorithms need to quickly adapt, a computational burden. This motivates the key research question:
Can blind beamforming/MIMO algorithms be designed with few (≈ 2) microphones.

Solving these problems is crucial if users would be expected to wear the earphones daylong. Even with hollow earphones, accurate ambient sound is critical for noise cancellation.

One line of attack on these problems relates to *exploiting the slow propagation speed of sound.* If the direct sound path and the reflected echoes arrive with larger time separation, there is opportunity for advanced signal processing [35]. The other opportunity pertains to utilizing the IMU motion sensors to track the movement of the head, and prime the beamforming algorithms with this (side-channel) information. Finally, if there is coordination between the left and right earphones, perhaps they can together form a bigger microphone array. Given that BLE signals travel much faster than sound, the left-right earphones could coordinate over BLE in real-time, while the sound samples are arriving relatively slowly (Figure 5). This opens entirely new architectures for various types of acoustic signal processing on earphones.

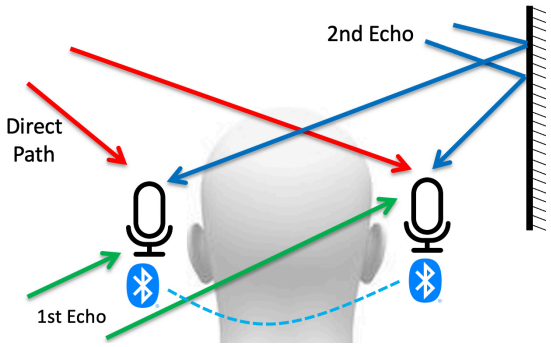

Figure 5: BLE coordinated beamforming in earbuds.

2.6 Many More Problems ...

Problems discussed thus far are biased by the our own backgrounds and interests. This subsection touches upon broader problems in the space of earables.

■ **Healthcare on earables:** The ear canal is the one of the "inlets" into the human body and innovations are already in progress at the intersection of earables and healthcare. Authors in [36, 37] are demonstrating the feasibility to measure blood pressure from inside the ear-canal, detect stages of sleep, and even quantify pain. Smartphones are able to detect in-ear infections by shining acoustic signals from the phone speaker [38]; authors in [33] show opportunities around in-mouth teeth sensing through earphones . We envision these as only the tip of the iceberg — bio-electrical, optical, PPG, EMG, ECG, and even neural/brain sensors may all lend themselves to *continuous sensing* through worn earphones. **One general challenge with continuous sensing is that mobility of the body-worn device interferes with the sensing process, yielding noisy/erroneous signals.** Chest-worn ECG and smartwatch PPG measurements, for instance, are known to suffer with mobility, posing as a key hurdle to continuous health monitoring. Given the head's relative stability, earables may offer natural benefits. The low-intensity (and IMU-measurable) head motions can perhaps be algorithmically filtered out from the bio-sensor data.

■ **Two Factor Authentication (2FA):** In a large part of the developing world, the smartphone is the only computing device for users. Two factor authentication (2FA), almost a standard security practice today, does not apply to users with a single device. The natural question is: **Why can't earphones serve as the second device for 2FA?** Even for the very low-end wired earphones, can

certain hardware fingerprints (detected through the audio jack of the smartphone) enable 2FA? If so, then the 2FA challenge could be easy and reliable; a user would only need to plug her earphone into the audio jack of the smartphone, and the smartphone would authenticate the user. Zooming out, what is the new space of opportunities (and attack vectors) that earphones bring, given that it has multiple sensors, is close to the human head, and can easily communicate/whisper to the user.

■ **Hardware architecture for earables** With the vision of an earable app store in the future, an important question is: **will today's hardware architecture mostly suffice, or is new clean-slate thinking necessary?** In particular, given the smaller form-factor, battery constraints, and speech-based real-time interaction, will the architectural challenges depart significantly from prior wearables, or will they be engineering tweaks? In fact, what does it take for earphones to run daylong as a stand-alone device, almost like a smartphone without a visual display. Perhaps a benchmark suite for earables would be a starting point to characterize the needs for an earable hardware platform. Understanding common math kernels, and the peak versus average workloads, might start offering insights into this question.

■ **Glasses as Open-ear Earables:** The earables vision includes glasses as well, where microphones, speakers, and other electronics are embedded in the temples (or the side-sticks) of the eyeglass. This presents a distinct advantage that the ear is not blocked, permitting day-long comfort; moreover, a large part of the older generation is used to wearing eye-glasses all the time, hence piggybacking on this "platform" extends additional advantages. One issue however is that the quality of sound will degrade (compared to an earphone), not only because of the interference from ambient sounds, but also because the sounds from the glass to the ear-drum undergoes channel-related distortions. Google glass had embraced bone conduction, where the idea is to propagate vibrations through the skull bones to ultimately produce the desired sound in the cochlea. In both cases, understanding the air or bone channel can critically improve the quality of experience. However, both air and bone-conduction face the following research question: **in the absence of an electronic microphone in the inner-ear, how can the channel to the ear-drum be measured from the external air/bone speaker.**

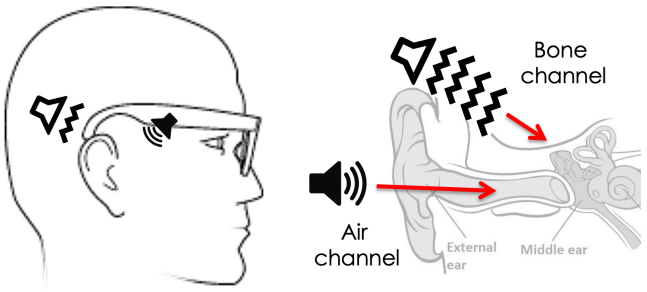

Figure 6: Bone and air conduction from a glass: the bone/air channels to the ear-drum are unknown.

One line of attack is the following: imagine the glass playing carefully designed sounds to the user, and the user giving feedback on the quality of the audible sound (perhaps through a rating app in her smartphone). Through each round of user feedback, the glass could perform a "gradient descent" (GD) and updates its estimate of the channel. The GD can be formulated in the standard form as:

$$maximize\ f(\bar{h} * \bar{x}) \qquad s.t. \quad h_i \in S, \quad \|\bar{x}\|^2 \leq B \tag{1}$$

where $f(.)$ is the coarse-grained user feedback function, h is the channel, x is the known sound played from the speaker, and S is some constraint set (e.g., sparse) on the channel, and B is bound on power. Unfortunately, h is unknown, so the standard GD update (i.e., Equation 2) does not apply since $z = \bar{h} * \bar{x}$ is unknown.

$$\bar{z}(k+1) = \bar{z}(k) - \alpha \nabla f(\bar{z}(k)) \tag{2}$$

However, it may be possible to measure ∇f at each round by precoding the sound with h^{-1} and measuring the partial derivatives for each h_i. Ultimately, the hope is that GD can be performed on h^{-1}. Of course, this becomes complicated, even if we discount the excessive burden of giving so many feedback. Hence, the research questions are **(1) how should h be estimated with such human-feedback functions, and (2) how can feedback iterations be minimized without compromising on the h estimate.**

3 CLOSING THOUGHTS AND CAVEATS

■ Acoustics is perhaps the only frequency band in which humans can both transmit to, and receive from, machines (via speaking and listening). Yet, this band has remained largely under-utilized primarily because speech recognition and natural language processing (NLP) were not ready. With rapid advances in those fields, voice and hearing interfaces are gaining popularity. Earables, we believe, are poised to harness this paradigm shift.

■ Of course, the 2-way voice interactions with machines are not a replacement for screens. High-bandwidth visual data consumption – from pictures and videos – will remain irreplaceable. However, a good fraction of the content exchange is short, quick, and amenable to speaking and listening (e.g., browsing, emails, podcasts, etc.). In fact, with spatial sounds, perhaps it would even be possible to "browse" through hearing. Consider a bank webpage with menu items on the left and account balances on the right. An earphone could play two parallel voices from the corresponding directions, reading out the webpage information. As the user turns her head towards one direction, that voice could become louder while the other dims down. In sum, harnessing the inherent spatial properties of sound could narrow the gap between audio and visuals.

■ The landscape of "earables" spans many areas, including hardware, sensing, signal processing, OS/embedded systems, ML, HCI, etc. One of the obvious weaknesses of this paper is its inadequate coverage of this landscape. Nonetheless, the hope is that even an "incomplete" paper could sow preliminary seeds of thought, with the hope that they would germinate on the fertile grounds of our collective research community.

ACKNOWLEDGMENTS

Thanks to shepherd Tam Vu, NSF, NIH, my students Zhijian Yang, Yu-Lin Wei, Jay Prakash, Liz Li, and Sheng Shen, and my colleagues, Haitham Hassanieh, Fahim Kawsar, and Rakesh Kumar, for various ideas and thoughtful conversations that influenced this article.

REFERENCES

[1] Au.D Lindsey Banks, "The complete guide to hearable technology in 2019," May 2019.

[2] Headphonesty, "10 best earbuds with microphone that are great for calls," June 2021.

[3] Samsung, "Samsung Gear IconX: Go phone-free," June 2019.

[4] Sorin V. Dusan, Esge B. Andersen, Aram Lindahl, and Andrew P. Bright, "System and method of detecting a user's voice activity using an accelerometer," 2016.

[5] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury, "Stear: Robust step counting from earables," in *Proceedings of the 1st International Workshop on Earable Computing*, 2019, pp. 36–41.

[6] Qualcomm, "With Qualcomm QCS400 SoCs, immersive audio meets on-device ai," June 2019.

[7] Poppy Crum, "Hearables will monitor your brain and body to augment your life," May 2019.

[8] N. Goff, Jesper Jensen, and Susanna Løve, "An introduction to opensound navigator," Tech. Rep., 2016.

[9] Waverly, "Goodbye language barriers," June 2019.

[10] Nura, "Nuraphone: How it works," June 2019.

[11] Bragi, "Headphones of tomorrow will have apps," April 2019.

[12] Thomas Bible, "Binaural audio for narrative vr," May 2016.

[13] James Peckham and Sharmishta Sarkar, "Bose frames review: Much more than just premium sunglasses," June 2019.

[14] Melanie Swan, "Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0," *Journal of Sensor and Actuator Networks*, vol. 1, no. 3, pp. 217–253, 2012.

[15] Deborah Lupton, *The quantified self*, John Wiley & Sons, 2016.

[16] B. Priyantha, D. Lymberopoulos, and J. Liu, "Littlerock: Enabling energy-efficient continuous sensing on mobile phones," *IEEE Pervasive Computing*, vol. 10, no. 2, pp. 12–15, April 2011.

[17] J.P. Mobley, C. Zhang, S.D. Soli, C. Johnson, and D. O'Connell, "Pressure-regulating ear plug," Oct. 13 1998, US Patent 5,819,745.

[18] Sheng Shen, Nirupam Roy, Junfeng Guang, Haitham Hassanieh, and Romit Roy Choudhury, "Mute: Bringing iot to noise cancellation," in *ACM Sigcomm*, 2018.

[19] Mikey Campbell, "Apple working on voice-recognizing headphones with built-in accelerometer, beamforming mics," April 2014.

[20] Earable Inc., "Bio-sensing earbuds," 2019.

[21] Nam Bui, Tam Vu, and et.al., "ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear," in *MobiCom*. ACM, 2019.

[22] Nhat Pham, Tam Vu, and et. al., "Wake: a behind-the-ear wearable system for microsleep detection," in *MobiSys*. ACM, 2020.

[23] Tobias Röddiger, Daniel Wolffram, David Laubenstein, Matthias Budde, and Michael Beigl, "Towards respiration rate monitoring using an in-ear headphone inertial measurement unit," in *Proceedings of the 1st International Workshop on Earable Computing*, 2019, pp. 48–53.

[24] Nicholas D. Lane. Abhinav Mehrotra and et. al., "Iterative compression of end-to-end asr model using automl," in *INTERSPEECH*, 2020.

[25] Haochen Sun Xiao Zeng, Kai Cao and Mi Zhang., "Sharpear: Real-time speech enhancement in noisy environments," in *Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 2017.

[26] Satyanarayanan M. et. al. Chen, Z., "An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance," in *SEC*. IEEE/ACM, 2017.

[27] Muhammad Huzaifa, Rishi Desai, Xutao Jiang, Joseph Ravichandran, Finn Sinclair, and Sarita V. Adve, "Exploring extended reality with illixr: A new playground for architecture research," 2020.

[28] Zhijian Yang, Yu-Lin Wei, Sheng Shen, and Romit Roy Choudhury, "Ear-ar: indoor acoustic augmented reality on earphones," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[29] Seungchul Lee, Chulhong Min, F. Kawsar, and et. al., "Automatic smile and frown recognition with kinetic earables," in *Augmented Human International Conference*. ACM, 2019.

[30] Andrea Ferlini, Alessandro Montanari, Cecilia Mascolo, and Robert Harle, "Head motion tracking through in-ear wearables," in *Proceedings of the 1st International Workshop on Earable Computing*, 2019, pp. 8–13.

[31] Meera Radhakrishnan and Archan Misra, "Can earables support effective user engagement during weight-based gym exercises?," in *Proceedings of the 1st International Workshop on Earable Computing*, 2019, pp. 42–47.

[32] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li, "Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 95–108.

[33] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy. Choudhury, "Earphones as a teeth activity sensor," in *MobiCom*. ACM, 2020.

[34] Yu-Lin Wei and Romit Roy Choudhury, "Angle-of-arrival (aoa) factorization in multipath channels," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[35] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury, "Voice localization using nearby wall reflections," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[36] Pham Nhat, Dinh Tuan, Raghebi Zohreh, Kim Taeho, Bui Nam, Nguyen Phuc, Truong Hoang, Banaei-Kashani Farnoush, Halbower Ann, Dinh Thang, and Tam Vu, "Wake: a behind-the-ear wearable system for micro-sleep detection," in *MobiSys*. ACM, 2020.

[37] Nam Bui, Nhat Pham, Jessica Jacqueline, Vu Tam, and et. al., "ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear,," in *MobiCom*. ACM, 2019.

[38] Rajalakshmi Nandakumar Randall Bly Shyamnath Gollakota Justin Chan, Sharat Raju, "Detecting middle ear fluid using smartphones," in *Science Translational Medicine*. AAAS, 2019.