Beyond Audio: Towards a Design Space of Headphones as a Site for Interaction and Sensing

Payod Panda
Microsoft Research
Cambridge, UK
payod.panda@microsoft.com

Molly Jane Nicholas
University of California Berkeley
Berkeley, USA
molecule@berkeley.edu

David Nguyen
University of California Irvine
Irvine, USA
dvnguye5@uci.edu

Eyal Ofek
Microsoft Research
Redmond, USA
eyal.ofek@gmail.com

Michel Pahud
Microsoft Research
Redmond, USA
mpahud@microsoft.com

Sean Rintel
Microsoft Research
Cambridge, UK
serintel@microsoft.com

Mar Gonzalez-Franco
Microsoft Research
Redmond, USA
margonzalezfranco@gmail.com

Ken Hinckley
Microsoft Research
Redmond, USA
kenh@microsoft.com

Jaron Lanier
Microsoft
Berkeley, USA
jalani@microsoft.com

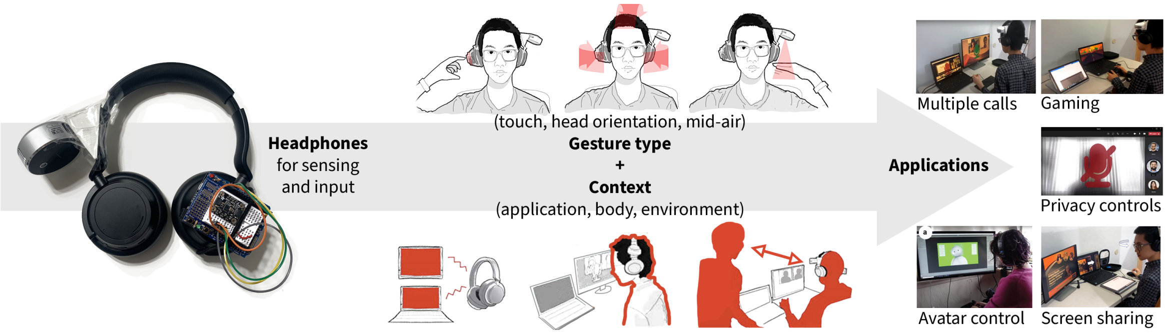


Figure 1: We explore interaction with headphones as a wearable sensor-enhanced input peripheral–as opposed to an output device with audio control functionality. Rooting user gestures within their context give rise to several application possibilities.

ABSTRACT

Via Research through Design (RtD), we explore the potential of headphones as a general-purpose input device for both foreground motion-gestures as well as background sensing of user activity. As a familiar wearable device, headphones offer a compelling site for head-situated interaction and sensing. Using emerging sensing modalities such as inertial motion, capacitive touch sensing, and depth cameras, our implemented prototypes explore sensing and interaction techniques that offer a range of compelling capabilities.

User scenarios include context-aware privacy, gestural audio-visual control, and co-opting natural body language as context to drive animated avatars for "camera-off" scenarios in remote work–or to co-opt (oft-subconscious) head movements such as dodging attacks in video games to enhance the gameplay experience.

Drawing from literature and other frameworks, we situate our prototypes and related techniques in a design space across the dual dimensions of (1) type of input (touch, mid-air, or head orientation); and (2) the context of user action (application, body, or environment). In particular, interactions that combine multiple inputs and contexts at the same time offer a rich design space of headphone-situated wearable interactions and sensing techniques.

CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**; *Interaction design process and methods.*

KEYWORDS

headphones, wearables, sensing, research through design, design space

ACM Reference Format:
Payod Panda, Molly Jane Nicholas, David Nguyen, Eyal Ofek, Michel Pahud, Sean Rintel, Mar Gonzalez-Franco, Ken Hinckley, and Jaron Lanier. 2023.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DIS '23, July 10–14, 2023, Pittsburgh, PA, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9893-0/23/07…$15.00
https://doi.org/10.1145/3563657.3596022

Beyond Audio: Towards a Design Space of Headphones as a Site for Interaction and Sensing. In *Designing Interactive Systems Conference (DIS ’23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3563657.3596022

1 INTRODUCTION

While existing commodity headphones largely focus on audio-visual (A/V) consumption–with integrated controls for volume level, muting, and other audio functions–the presence of microphones and even inertial motion sensors on some units (typically for spatial audio support [5, 10]) hints at richer possibilities for headphone-situated input, interaction, and wearable sensing.

As a wearable device, headphones travel with the user from device to device and from one usage scenario to another, offering semantically rich cues for non-verbal communication associated with head motion and orientation [9]. With the addition of a few pragmatic sensors, headphones thus offer a compelling way to capture the naturally-occuring vocabulary of user activity, including subtle head movements, the lean-towards or lean-back motions of upper body posture, as well as hand gestures on (or in mid-air proximity of) the headphones themselves.

In this way sensor-enhanced headphones offer designers an opportunity to (re-)consider notions of the user’s context, activity, and proximal hand gestures–enabling rich interactions beyond the status-quo, button-pushing type of interactions with headphones. For instance, one of our techniques re-interprets the common gesture of lifting the headphones’ earpiece (e.g. to listen and attend to a co-located colleague nearby) to implicitly mute the microphone and audio output to enhance the digital-audio experience. This shifts complexity from the user to the system, co-opting a natural user behavior and giving it a dual-purpose, context-dependent digital meaning. Further, by augmenting existing user actions, such an approach reduces the number of explicit gestures that the user has to learn, control, enact, and remember.

Adding new interactions to headphones introduces unique challenges. A clear example is the use of speech interfaces. Headphones are well positioned to receive input from a microphone, however there are many cases where such an interface is unsuitable, for instance when it interferes with a conversation. Another challenge is that headphones are not visible to the wearer, so interactions need to be managed eyes-free, mainly using proprioceptive positioning, audio, and haptic interactions. A further challenge (and opportunity) of head tracking is the range of head motions from intentional gestures such as nodding to indicate understanding, through partially-conscious actions such as redirecting one’s gaze to a different device, to fully subconscious natural head movements coincident to body posture or other user activity.

By interpreting such headphone signals in a context-appropriate manner, as well as considering new multi-modal interactions enabled through wearable sensors situated at various locations on a modified headphone, our work explores a design space of such possibilities. Our headphone prototype uses various combinations of an IMU, on-device buttons and inputs, and a LiDAR sensor to capture user signals. We demonstrate the value of capturing user input through a head-worn device through several prototype applications that use the headphones in multi-device and cross-platform scenarios. For instance, our system provides context-aware privacy by blurring a user’s video and muting their microphone in a video call when they disengage from the video call and have a side conversation. Our system responds to socially recognized gestures. For instance, cupping one’s hands near their ear signals the desire to hear better, and our system responds by increasing the audio volume. Another prototype automatically switches the window that is being shared in a video call as the user switches their attention between several displays and devices.

Through this paper, we make the following contributions to the DIS community:

- First, the design space discussed in section 4. This design space takes different kinds of input into account, while proposing different contexts that might be relevant to understand. The insight that each interaction could use multiple types of input and sense multiple contexts at the same time opens up a vast design space for future exploration around headphones that is also likely relevant to other wearables.
- Second, an annotated portfolio of functional prototypes and potential applications for their use in subsection 3.2. Our prototypes demonstrate how sensing input from headphones can enhance interactions in multi-device and cross-platform scenarios from the workplace to gaming.

Rather than a traditional report on a study, this paper is structured to follow our Research through Design process in order to answer the four criteria for evaluating interaction design research within HCI [83]. We show the *relevance* and *novelty* of our work by grounding it in current literature (section 2). We then document our *process* by discussing our design methodology (section 3) and sharing our functional prototypes (subsection 3.2). Finally, we discuss the *extensibility* of our work by suggesting a design space for sensor-augmented headphones, grounded in existing literature as well as our learning from building and experiencing our prototypes (section 4). We then discuss the ramifications for future work in section 5 and section 6, and conclude the paper (section 7).

2 BACKGROUND

Our work builds upon previously explored *interactions with headphones*, and uses *materialist design techniques* to situate it within the contexts derived from frameworks of *peripheral interaction* design and design within the *social environment*.

2.1 Headphone-based Interaction Design

Prior research exploring interactions mediated by visible or invisible head-wearable devices [17, 44, 50, 58] has remained focused on media control [17]. While some headphones already include on-device sensors, they tend to be focused on media experiences and controlling headphone-related data streams (e.g. volume) [14, 52, 53]. For instance, embedded motion sensors in the Apple AirPods Max allow simulation of a surround sound setup [5]. Similarly, Jawbone and other headset manufacturers auto-pause or auto-mute media playback based on device posture (e.g. placing headset around one’s neck). Additionally, most of these implementations are decontextualized interactions where the gestures and ear interactions do not leverage the context within which the user performs the task [50]. We seek to expand the design space for headphones, articulating the value of using headphones as a more generalized site for interaction and sensing, moving beyond audio control.

Researchers have also explored diverse input methods for headphones. For instance, media playback can be controlled via taps on earcups [52, 53] or touch sensors [14]. Wired headphones may use gestures such as tug and twist on the headphone cable to control audio playback [64, 70]. External accessories can also be paired with headphones to provide input [78]. We also augment the headphones with additional sensors, and explore interactions that can leverage multiple contexts at the same time. This is explained in depth in section 4.

2.2 Peripheral Interaction

Interacting with computing technology typically demands focused attention through input devices such as keyboards and touch screens [7]. These interactions also tend to be *reactive* (i.e. initiated by the user) rather than *proactive* (i.e. initiated by the interactive system) [38]. However, many of our everyday interactions happen in the periphery–for instance, drinking coffee (in the periphery) while reading a book (with focused attention). Researchers have used the periphery of user attention in human-computer interaction, introducing concepts like calm technology [80], ambient information systems [66] and peripheral displays [56]. However, these explorations primarily explored peripheral *perception* rather than peripheral *interaction*. Our ecosystems of interactive technology demand increasing amounts of focused attention from the user [6, 80]. It would thus be beneficial to offload some of these interactions from the center of attention to the periphery reducing cognitive load [75] and improving focus.

Prior work on the related concepts of foreground and background attention [15] has explored the value of considering the user's *context* during device usage [34]. This can be accomplished by using sensing techniques to capture 'natural' interactions [34]. Sensing the user's context can allow systems to be more *proactive* rather than always relying on active user input [38]. Headphones can leverage both foreground and background interactions: the foreground is the user's direct interaction with the *device* and the background is their interaction with the *environment*. Earlier studies of foreground and background / peripheral interactions primarily examined single device implementations (e.g. [1, 20, 33, 34, 48, 56]). We expand on this by exploring multi-device ecosystems, and by describing a design space rooted in the contexts in which interactive technologies are used.

2.3 Devices in Social Environments

Interactive devices, in particular wearables and peripherals, are often used in social environments. Wearing headphones in public is now considered socially acceptable [17]. We wear them for video calling in shared offices and commuting via public transportation. Dagan et al. [18] have identified two specific areas of value for designing interactions for devices in social environments: *augmenting existing social signaling*, and *proactive intervention in the social situation*. In our exploration, the social context within the environment plays a major role, and we primarily explore this through social signaling. Headphones are now familiar enough for there to be some socio-cultural norms for their usage. This makes headphones a particularly appropriate wearable peripheral to augment existing social signaling with new sensors and interaction techniques.

Understanding spatial relationships between users and devices is another important aspect that can be leveraged for implementing interactions. Hall's [30] notion of proxemics can be used to understand people's spatial relationships to each other and digital devices, and has been used to generate a variety of interaction techniques, especially in multi-device ecosystems [8, 27, 54]. For instance, Li et al. [49] used different cultures' kissing greetings for contextual awareness and [21, 36, 45] explored the importance of visible body gestures for both communicative purposes and individual activities. We use proxemics in the development of our design space, particularly around context awareness (subsection 4.2).

Our goals closely align with those of ubiquitous or pervasive computing, embracing the value of adding sensors and computation to objects throughout the environment, such that they become effectively invisible [28, 79]. Our designs seek to leverage the semantic meaning in the environment, and use existing gestural interactions which are so familiar as to be 'invisible' and yet imbued with meaning. Headphones enable the leverage of this socio-cultural awareness for designing interaction techniques. However, rather than construct a series of interactive objects, we take a high-level approach and identify the design space that can motivate future work by ubiquitous computing researchers.

3 METHODOLOGY

We used a Research through Design methodology [23, 72, 83] to explore the potential of headphones as input devices. We engaged in a material-centric design practice [22, 39, 40, 81] in order to identify and categorize common patterns in the way we interact with headphones. By "allowing material properties to guide our design" [19], this method enabled us to begin to identify the advantages of using gestures grounded in existing familiar interactions. Our exploration primarily centers around the use of headphones in the workplace [3, 76] and for gaming [2].

In a survey of the literature, we identified several patterns of cross-device interactions that would benefit from the addition of augmented headphones, including: controlling one device from another [69], redistributing an application across multiple devices (*integrating* [13]), and beginning a task on device and continuing it on another (*migrating* [13]). In addition to support from literature, these behaviours also seemed likely to trigger interesting interactions with the headphones while being commonly encountered throughout the day in a collaborative work environment. A previous gesture study elicited near-ear gestures from participants [17]. However, this study focused on specific interactions with a mobile phone device, and as such this was substantially different from our intended use cases.

Maintaining the material-centric approach (using headphones as the material), to study headphone usage in these scenarios six of the authors recorded themselves interacting with a pair of wireless headphones. The video recordings captured natural behaviours while wearing headphones in a variety of settings. Each author recorded until they had captured at least 20 minutes, and had seen all three of the following behaviours: multi-device usage, a real-world interruption, and changing tasks (including switching between devices). Our goal with this design process was to get a relevant *seed* for further exploration and discussion, rather than an exhaustive set of interactions for our chosen scenarios. Some of these behaviours are re-enacted and captured in Figure 2.

Through this activity, we identified some repeated gestures and interaction patterns, which shaped several of our designs (see subsection 3.2). For example, one common pattern occurred during interruptions: as a person approaches the headphone user, participants tended to lift the headphones off their head rather than quickly scramble to find the digital or physical mute button. This gesture of lifting the earcup away from the ears indicates the intent of listening to the person near you (see Figure 8). Through this gesture, the interaction design minimizes the need to search for a button and simultaneously signals conversation acceptance to the nearby person in a socially acceptable and familiar way. We also explored gestures that were inspired by commonly understood gestures in the researchers' cultures. For instance, cupping one's ear to indicate the desire to hear louder, or blocking the mouth to indicate being quiet. Our materialist approach allowed us to defamiliarize ourselves with headphones, a commonly-used apparatus, and reconsider, re-envision, and reconceptualize its role in our technologically-mediated lives.

Evaluating systems, tools, and toolkits is notoriously difficult [37], sometimes even considered harmful [26]. Beyond usability evaluations, there are a variety of strategies that can be used to assess toolkit effectiveness [47]. In this paper, we primarily focus on an *evaluation by demonstration* [47]. Our described usage scenarios demonstrate a subset of the envisioned application space. subsection 3.2 represents an annotated portfolio [25, 74, 83], meant to embody our design space (section 4). Together, these convey the decisions we made and the philosophy we developed throughout the project [74, 83]. Additionally, annotating our portfolio of prototypes allows us to step away from individual designed artefacts, look holistically, and derive a design space.

3.1 Early Prototypes and Technical Implementation

With the initial gestures identified through our previous exercise, we augmented an existing pair of wireless headphones with additional sensors and input widgets. In order to explore different combinations of widgets and sensors, we built a hot-swappable magnetic mount that allowed us to easily swap out components for others (Figure 3). These early prototypes allowed us to sense the user's head orientation through an IMU, and receive input through widgets like buttons and rotary encoders, but not mid-air gestures such as cupping one's ear.

In our subsequent prototypes, we incorporated a LiDAR sensor mounted above the earcup (see Figure 4), that allowed us to sense mid-air gestures around the ear down until the shoulder region. This followed the recommendation from Chen et al. [17], who recommended sensors that could track hands and fingers while covering the entire region around the ear and below. We trained a deep convolutional neural network with a MobileNet v2 architecture [68] to recognize different gestures based on the view of the hand from the earcup. Figure 5 shows some of the gestures we can detect with our prototype.

Our physical prototypes intentionally utilized low-fidelity materials such as cellotape, cardboard, and breadboards. The low fidelity nature of these materials encouraged exploration, iterations, and rapid prototyping. An early prototype used a magnetic hot-swappable mount with a cardboard base (Figure 3), which we upgraded to a breadboard attached to the headphones for quick swaps between electronic components (Figure 4). The last prototype, even though it used high tech components like an Intel RealSense LiDAR unit, still felt appropriately low-fidelity due to the use of cellotape attaching it to the earcup.

3.2 Annotated Portfolio of Applications and Interactions

In this section, we present the portfolio of our functional prototypes of augmented headphones to explore interactions. Within the RtD method of annotated portfolios, an annotation is any textual description that accompanies a design artefact [24]. Instead of talking about each prototype individually, we look at the portfolio holistically by *annotating* the portfolio with the interaction qualities that we perceived were embedded in the portfolio [11]. Annotating our portfolio of prototypes in such a way allows us to (1) show the benefits of using headphones as an input device for enhancing interaction across devices and platforms.

3.2.1 *Context-aware privacy.*

Typical interfaces for video calling systems involve binary choices (such as toggling sharing of audio and video) controlled by on-screen controls [65]. Conversational flow, which is already challenging [51, 67, 73], is additionally disrupted by the need to search for on-screen controls. A conversation in the physical world contains many social, environmental, and physical cues that keep all participants aware of the receptiveness of others, such as gaze, body language, and events in the environment. Our system can give participants on the call more awareness of these social dynamics of the conversation.

For example, headphones worn in video calls can detect a sudden rotation of the head from the screen to a new location. Persistently looking away from the primary screen is regarded as a meaningful disruption of the user's attention from the call. As a result, our prototype blurs the user's video and mutes their microphone to protect their privacy as they have a conversation outside the context of the call (see Figure 6). At the same time, the blurred video notifies other attendees that the user is temporarily away from the call. When the user returns their attention to the screen, the system removes the video blur, and reactivates their microphone.

Figure 7 shows a user in two separate video calls at the same time. By blurring the video feed of the session that is not in focus, the headphone-wearer's attention is communicated to other remote participants. Each set of remote participants is aware that the headphone-wearer is currently talking to the other session, without any need for the user to consciously select a button. We use a similar design to enable in-game communication where the player can choose to communicate with either an individual teammate or all teammates with a turn of the head (see Figure 7b). By combining head-worn sensors with context-aware application controls, we can expand the richness of these remote experiences, and minimize friction associated with video calling experiences.

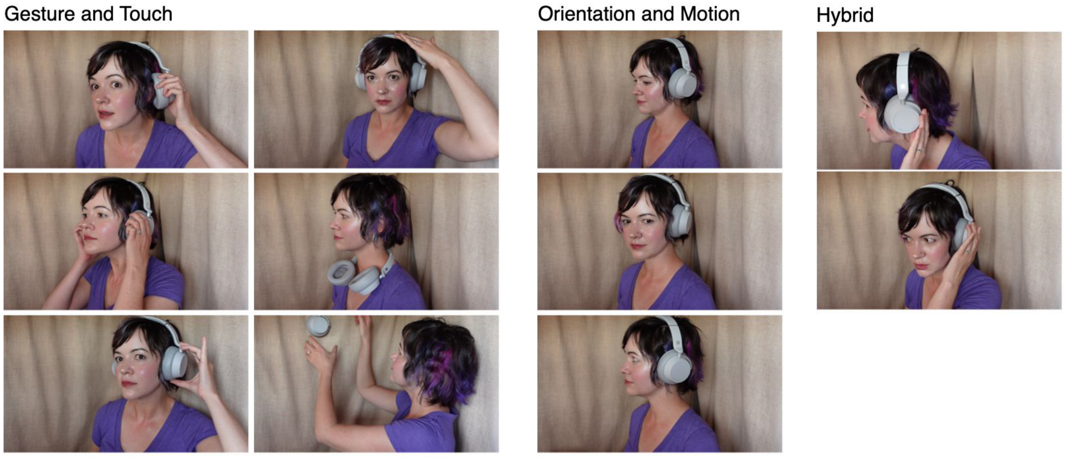


Figure 2: A re-enactment by a co-author of some of the recurring behaviours we observed through our elicitation process. We utilized behaviours like these as a starting point for designing the prototype applications we demonstrate in subsection 3.2.

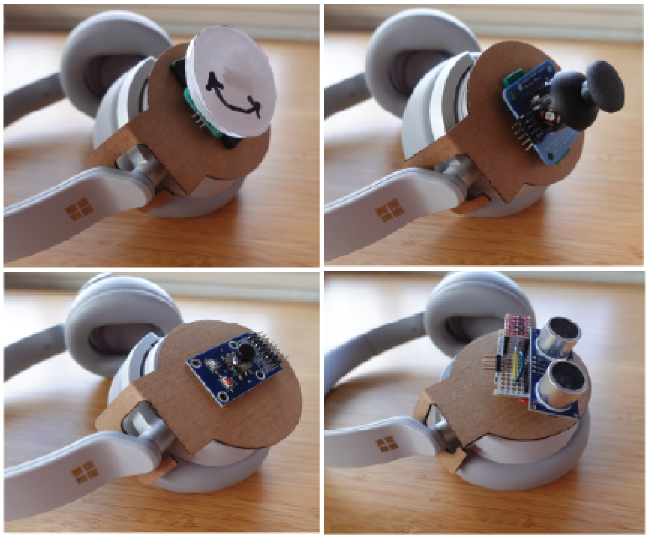


Figure 3: An early prototype showing modular customizable hardware through a hot-swappable magnetic mount.

The headphones use an IMU to detect the wearer's head turn. Additionally, the physical layout of the space and the currently active applications inform the interaction, allowing our headphones to augment the signal of turning one's head. Crucially, this movement would be performed *even if the user weren't wearing our augmented headphones*–we simply sense the already-occurring behaviours unrelated to the headphones. In the language of Dagan et al., this represents an augmentation of existing social signaling [18].

3.2.2 *Gestural audio visual control.* Currently users share media by clicking or touching specific controls in applications. Instead of using on-screen elements, our augmented headphones can enable interactions such as cupping the ear toward the audio source (See Figure 8b). Such a gesture could increase the audio volume while

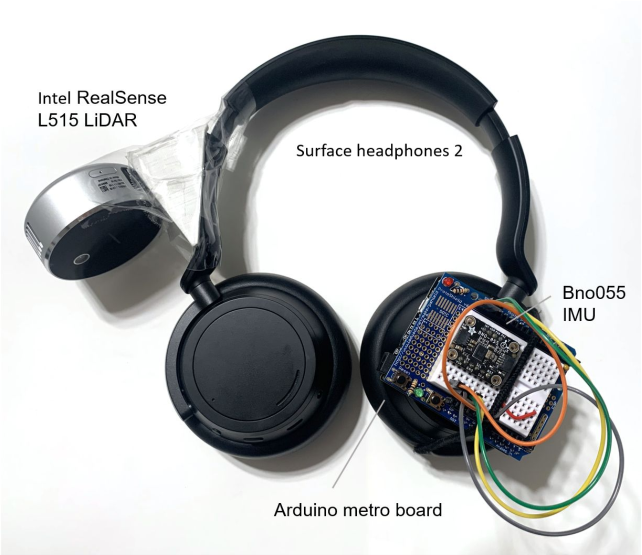


Figure 4: Wireless headphones equipped with an IMU (right) and a LiDAR (left).

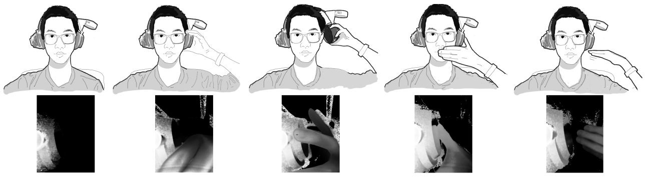


Figure 5: Selected gestures we used for our prototype system. The bottom shows LiDAR images of those gestures. (From left: default mode, cupping the earcup, raising the earcup, a mouth cover, a 'cut' signal.

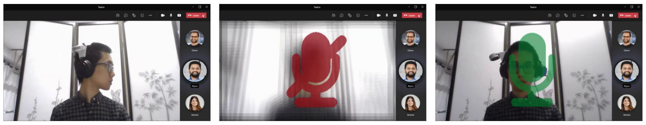


Figure 6: Context-aware privacy control. Left: A user disengages from a video call, reacting to a local conversation. Middle: After a set time, the system recognize a persisting attention in a direction different than any known device (*Environment Context*). As a result the video is blurred, the microphone is muted, and other users on the call are notified. Right: When the user returns to look at the screen, their video and microphone returns to normal.

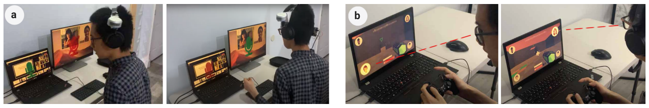


Figure 7: The headphones follow the head-pose of the user, and automatically manage the sharing of video and/or audio between multiple private chats, while communicating availability to other participants in either a) video calling or b) gaming scenarios.

reducing ambient noise. Additionally, the familiar gesture is also understood as a social signal by participants to mean "I can't hear you well", taking into account both the socio-cultural and digital contexts.

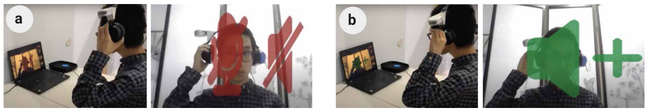


Figure 8: a) Lifting the earcup (a common gesture to attend to an in-person disruption) mutes the sound and the microphone to maintain privacy. b) Cupping the earcup (as one might do naturally to indicate they cannot hear) increases the volume of the system.

3.2.3 Redirect Input/Output. As ownership and use of multiple devices increases, researchers have begun to design experiences across multi-device ecosystems (see [13] for an excellent overview). However, most of these approaches look primarily at interaction design across core computing devices such as tablets, smartphones, laptops, and desktop monitors [29, 31, 55]. Peripheral devices remain underutilized in this space. Here, we identify additional design opportunities where augmented headphones can support cross-device experiences in unique and compelling ways.

As we discovered during our design exercise, people who are using multiple devices at once frequently look towards the device with which they are interacting. By capturing head orientation information, our augmented headphones can enable more rapid, convenient, and intuitive transitions from one device to another. For example, our headphones use head orientation as a proxy of the user's attention. If the user is attending to a second monitor or device that was previously granted screen sharing privileges, the system will automatically change the user's stream to show the slide (Figure 9 top row). If the user is returning their attention to the participants, for example to answer a question, the stream will change back to show the presenter's face. By removing the challenge of manually updating their video stream, the user may now easily manage multiple sources of content from writing and sketching on tablets (Figure 9 bottom row) to a camera viewing a physical project in the room, white board, or a work desk and more.

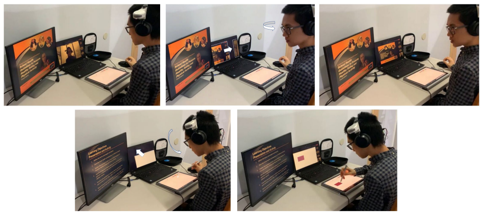


Figure 9: A user gives a presentation via a video calling tool. As they look at different devices, their streamed video updates to show the relevant source to the call participants.

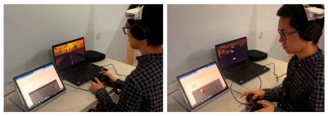


Figure 10: Augmented headphones can be used to manage other peripherals. In the picture to the right, the input from a game controller is redirected to the non-gaming device (the tablet).

A similar interaction can be used to stream audio to the headphones from multiple devices, or stream input to another device: as the user rotates their head from a game they are playing, the headphones may switch from the computer game audio to the audio of the TV in the room. Alternatively, the mouse or gamepad controller's input may be sent to a different device with just a turn of the head. In Figure 10, the user is playing a game using a game controller. By looking at a different device, the user may transfer the game controller input to the other non-gaming device in order to answer a phone call or change the music audio in the room without moving their hands off the game controller.

For these interactions, the sensors enabled us to sense the physical location of the devices in relation to the headphone wearer, and the current digital applications being used. This interaction additionally relies on sensing the head orientation to manage the input shift.

3.2.4 Embodied Peripheral Interactions. Peripherals worn on the body can opportunistically capture semi-conscious behaviour. Intentionally meaningful head gestures such as head nods and shakes may convey a complete message or combine with talk and other gestures, and may be part of foreground messaging or backchanneling [42, 43]. These may be observed by both the system (*application* *context*) and by other people (socio-cultural *environment context*). Such body movement also occurs in other scenarios, including gameplay. Augmenting wearable peripheral devices with sensors to capture such movement makes them capable of supporting much richer interactive designs.

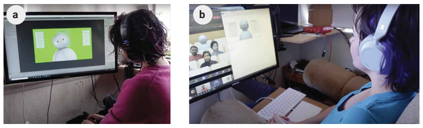


Figure 11: a) These headphones use an IMU to track natural body movement and map it onto avatar movement. b) Using an avatar that does not require a camera during a video call.

For example, in video calls, people may prefer to participate without the webcam on for a variety of reasons, such as privacy concerns [12, 16], preserving bandwidth, enabling full mobility [61], or to enable participation by neurodiverse people that may prefer not to be observed [84]. However, lack of video has a major impact on self-representation, presence, and rapport in video calling [41, 77]. While camera input can be modified to support partial or full face occlusion [57, 63], and filters to 'improve' appearance, full 3D avatars may replace one's real appearance [35]. For good or for ill, all such systems require a camera to be available, turned on, and facing the user for the entirety of an engagement. Background blurring [82] and full blurring [60] may also help with privacy, but risk inadvertent exposure, and again require a camera to be available.

Instead of relying on the camera for tracking at all, input from head-worn sensors could be used to capture and generate a range of signals to generate a visual representation of a participant. Users could join in an audio-only mode but participate more equally with other visually-represented participants with an avatar representation [35]. Our prototype provides audio-only participants in a video call with virtual avatars. We use audio-driven lip-movement, and head orientation from the IMU, and map them to the visual animation of the avatar's head. This provides a continuous, ever-changing, and personal stream of animation to the avatar (see Figure 11a, b). Here, we sense the user's head and body movements to augment the digital context of a video call.

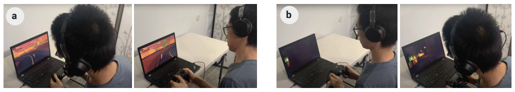


Figure 12: Leaning while wearing the headphone IMU impacts game play actions such as a) swerving the car to the side and b) peaking around the corner of a building.

Using the same orientation sensors, leaning while controlling a virtual vehicle (something many players do unconsciously while playing a racing game) could swerve the car left or right (see Figure 12). A system that uses these alternate input modalities could also support unique game mechanics, such as interdependence between video game players. This is similar to the Xbox co-pilot controller system where two players control a single character. Rather than using the Xbox joystick, rotation of the head left and right or up/down could map to character movement or remotely connect multiplayer experiences. For example, imagine a gaming system that requires remotely connected players to coordinate their movement as they both control the same character (a riff on the compelling game designs by [4]). Players could unlock new interactions within existing games, such as leaning the head to the side to peek around the corner while staying under cover (see Figure 12b), tilting the player's upper body to avoid projectiles flying toward her, balancing a bike during a turn, leaning forward to accelerate and backward to break and many more. These unique game control methods may additionally increase accessibility by allowing body leaning as a game mechanic instead of requiring fine finger control, and may enable single hand or even hands-free gaming.

4 DERIVING A DESIGN SPACE FOR HEADPHONE INTERACTIONS

Developing and experiencing these prototypes, and then creating annotations of interaction qualities allowed us to reflect upon their nature [71], and begin to abstract away the core concepts that can be used to create and expand upon interactions with headphones. With the insight from this reflection and supporting it with previous work, we define a design space for designing interactions with headphones that has two dimensions:

1. The type of *input* used to enable the interaction, and
2. The *context* within which the user executes the action.

4.1 Type of Input Gesture

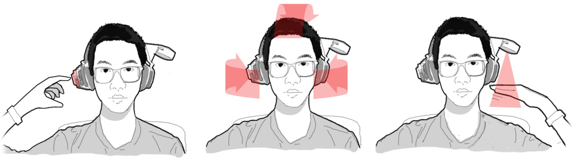


Figure 13: Types of gestures used in this work. Left: hardware widget on the headphones Middle: Sensing head gaze and right: hand pose near the earcup.

Some prior work has attempted to categorize user inputs for headphone interactions: Chen et al. [17] propose a comprehensive taxonomy of input gesture types by grouping them by locale, complexity, and form, each having multiple types (such as mid-air, touch-based, simple, compound etc.). However, while this taxonomy is theoretically sound, it is rather complex for practical application. A simpler taxonomy was proposed by Lissermann [50], who categorize interactions into touch, grasp, and mid-air gestures. However, both of these taxonomies focus on hand-based gesture inputs, and ignore hands-free operation made possible by sensors like an IMU. Based upon these frameworks and taking orientation-based sensors like IMUs into account, we propose categorizing user input into (see Figure 13):

1. **Touch-based gestures** [GES-TC]: These input gestures require physical touch from the user, and use tangible input

on headphones like buttons, knobs, and touch sensors (Figure 13 left). Such controls are common on existing commercial headphones, and are the most widely explored type of input in existing literature [14, 44, 52, 53, 64, 70]. These gestures are typically carried out intentionally by the user, and thus the interactions triggered by such gestures will be in the foreground [34].

2. **Mid-air gestures** [GES-MA]: These gestures are performed mid-air by the user, and are sensed by sensors like a LIDAR or proximity sensor. In their gesture elicitation study, Chen et al. [17] found that 58% of user-generated gestures were mid-air. Even so, such controls are not seen in commercial headphones, are rare in the research literature [50, 58], and are a rich area for exploration. People use hand gestures around the head to convey different social messages, such as showing a finger in front of the mouth to signal silence, or cupping the mouth with a palm to symbolise a private message. Sensing such gestures can enable designers to tap into cultural gestures that might be easier for the user to learn (Figure 13 right). These gestures are usually carried out intentionally by the user (foreground), but less frequently than touch-based gestures. With repeat usage, some of these gestures (like cupping one's ear) might move into the background [34].
3. **Head orientation** [GES-OR]: The user's head orientation can indicate the direction of the user's attention, and can be sensed by an IMU. It may also used for cultural behaviors such as "Yes" and "No" gestures (Figure 13 middle). Some commercial headphones sense the user's head orientation, but are currently limited to using this for spatial audio output [5]. Head orientation is underexplored in research literature exploring headphone interactions, since previous explorations have largely focused on foreground interactions. However, sensing the head orientation can be a powerful way to enable background interactions, or enriching foreground interaction by also sensing the context of the interaction. Looking at various devices or content during an interaction happens naturally and without explicit intent in the background [34].

4.2 Context of Input Gesture

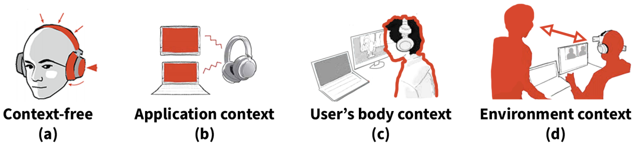


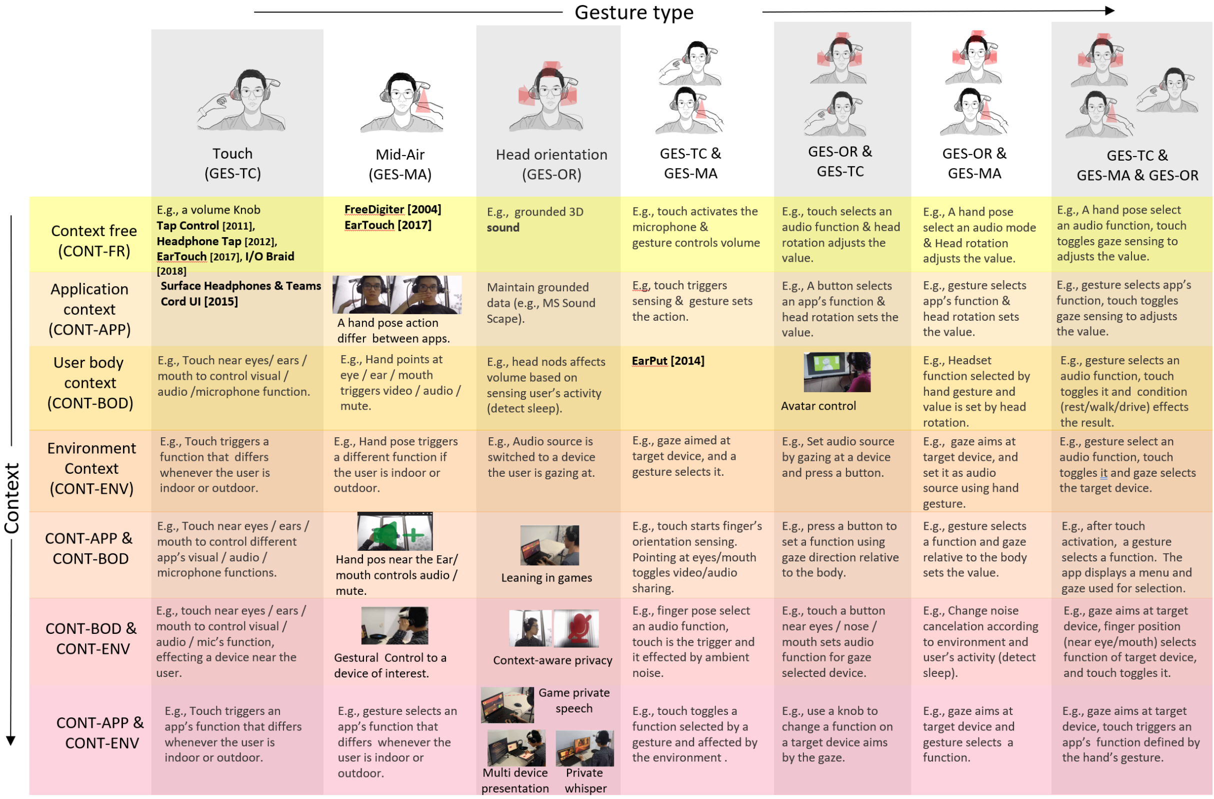
Figure 14: Context-free headphones have fixed functionality. (left) Different rings of context provide the application with more knowledge to fit the action, and the ability to have more varied input.

The gesture elicitation study by Chen et al. [17] found that an overwhelming majority (80.6%) of the 868 gestures created "naturalistically" by their participants for interaction around the ear were dependent on the context of the interaction. Thus, considering the context of the user action is of vital importance while designing interactions for headphones. Based upon previous headphone- and ear-based interaction design as well as frameworks of proxemics [30], peripheral interaction and the social context, we propose the following contexts for sensing the user's input gesture (see Figure 14):

1. **Context-free** [CONT-FR]: Context free gestures produce a similar result regardless of the active application, what the user is doing, or the user's social or physical environment. Most actions implemented in current headphones are context-free, such as changing volume. Constraining the interaction language to only support context-free gestures limits the number of actions the headphones can support.
2. **Application** [CONT-APP]: The application that the user is interacting with while using a gesture forms the first contextual layer. For instance, the same input gesture of turning a knob may control level of noise cancellation for audio listening applications, or visual quality for a media application.
3. **User's body** [CONT-BOD]: Next, we contextualize the gesture by either sensing the location of the gesture, or the state of the user's body. The human body is a semantically rich space, and several cultural gestures that enable social signalling are intrinsically tied to their location on the body (particularly the head)–such as cupping one's ear indicates the desire to be able to hear better. Similarly, an IMU can sense when the user is nodding their head or dozing off, which can be used to alter the effect of a gesture.
4. **Environment** [CONT-ENV]: Finally, we define the environment to include the *physical* (e.g. other devices, furniture) as well as the *social* (e.g. other people around the user, office vs. home environment) contexts. Leveraging this information can unlock powerful and seemingly magical interactions for the user. For instance, pressing a button while looking at a phone might answer an incoming call, and pressing the same button while looking at a light might turn that light on or off.

4.3 Locating Prototypes on the Design Space

Our proposed design space is shown in Figure 15. A major insight from developing the design space and populating it with our own prototypes was that neither the gesture types nor the contexts of use are mutually exclusive. Indeed, some of our prototypes, like multi-device presentation, made use of both the application context (to determine that we were in a video call while sharing our screen), as well as the environment context (to understand where the user's screens were physically located). Similarly, controlling an avatar might use both the body context (to detect gestures such as nods) and the app context (to detect being in an application with avatars), as well as multiple kinds of inputs (head orientation to move the head, and touch input to trigger emotions). Combining these dimensions may be a powerful way to ease the learning curve for functionality (see subsection 5.4), as well as to facilitate different levels of intentionality and context-awareness (see section 5).



Gesture type

Context

	Touch (GES-TC)	Mid-Air (GES-MA)	Head orientation (GES-OR)	GES-TC & GES-MA	GES-OR & GES-TC	GES-OR & GES-MA	GES-TC & GES-MA & GES-OR
Context free (CONT-FR)	E.g., a volume Knob **Tap Control** [2011], **Headphone Tap** [2012], **EarTouch** [2017], **I/O Braid** [2018]	**FreeDigiter [2004]** **EarTouch [2017]**	E.g., grounded 3D **sound**	E.g., touch activates the microphone & gesture controls volume	E.g., touch selects an audio function & head rotation adjusts the value.	E.g., A hand pose select an audio mode & Head rotation adjusts the value.	E.g., A hand pose select an audio function, touch toggles gaze sensing to adjusts the value.
Application context (CONT-APP)	**Surface Headphones & Teams Cord UI [2015]**	A hand pose action differ between apps.	Maintain grounded data (e.g., MS Sound Scape).	E.g, touch triggers sensing & gesture sets the action.	E.g., A button selects an app's function & head rotation sets the value.	E.g., gesture selects app's function & head rotation sets the value.	E.g., gesture selects app's function, touch toggles gaze sensing to adjusts the value.
User body context (CONT-BOD)	E.g., Touch near eyes/ ears / mouth to control visual / audio /microphone function.	E.g., Hand points at eye / ear / mouth triggers video / audio / mute.	E.g., head nods affects volume based on sensing user's activity (detect sleep).	**EarPut [2014]**	Avatar control	E.g., Headset function selected by hand gesture and value is set by head rotation.	E.g., gesture selects an audio function, touch toggles it and condition (rest/walk/drive) effects the result.
Environment Context (CONT-ENV)	E.g., Touch triggers a function that differs whenever the user is indoor or outdoor.	E.g., Hand pose triggers a different function if the user is indoor or outdoor.	E.g., Audio source is switched to a device the user is gazing at.	E.g., gaze aimed at target device, and a gesture selects it.	E.g., Set audio source by gazing at a device and press a button.	E.g., gaze aims at target device, and set it as audio source using hand gesture.	E.g., gesture select an audio function, touch toggles it and gaze selects the target device.
CONT-APP & CONT-BOD	E.g., Touch near eyes / ears / mouth to control different app's visual / audio / microphone functions.	Hand pos near the Ear/ mouth controls audio / mute.	Leaning in games	E.g., touch starts finger's orientation sensing. Pointing at eyes/mouth toggles video/audio sharing.	E.g., press a button to set a function using gaze direction relative to the body.	E.g., gesture selects a function and gaze relative to the body sets the value.	E.g., after touch activation, a gesture selects a function. The app displays a menu and gaze used for selection.
CONT-BOD & CONT-ENV	E.g., touch near eyes / ears / mouth to control visual / audio / mic's function, effecting a device near the user.	Gestural Control to a device of interest.	Context-aware privacy	E.g., finger pose select an audio function, touch is the trigger and it effected by ambient noise.	E.g., touch a button near eyes / nose / mouth sets audio function for gaze selected device.	E.g., Change noise cancelation according to environment and user's activity (detect sleep).	E.g., gaze aims at target device, finger position (near eye/mouth) selects function of target device, and touch toggles it.
CONT-APP & CONT-ENV	E.g., Touch triggers an app's function that differs whenever the user is indoor or outdoor.	E.g., gesture selects an app's function that differs whenever the user is indoor or outdoor.	Game private speech Multi device presentation Private whisper	E.g., touch toggles a function selected by a gesture and affected by the environment .	E.g., use a knob to change a function on a target device aims by the gaze.	E.g., gaze aims at target device and gesture selects a function.	E.g., gaze aims at target device, touch triggers an app's function defined by the hand's gesture.

Figure 15: Locating our prototypes (images), existing work (bold) and envisioned examples (gray) in the design space. By generating this morphological design space, we identify opportunities for future systems, and motivate the implementation of designs to address each square.

5 DISCUSSION

Naturalistic motion of the head is a complex input to parse (especially combined with hand gestures) from an experiential standpoint. We believe that there are three aspects for triggers that bear on gesture type, location, and patterns that need to be considered when exploring this design space.

5.1 Intentionality of user action with respect to the device

From an engineering standpoint, there is value in considering intentionality as a binary categorization variable, where actions performed by the user are either unintentional or intentional within the context of a given task. This is needed to build an automatic system that avoids false positives, errors, and unintended outcomes. However, a more nuanced perspective of intentionality as a continuum allows for a desired outcome which could be the result of an unintended action. As an example, looking away from an active video call might be a spontaneous reaction from the user, unintentional within the context of headphone usage, however intentional within the social context. Understanding such an action within the context of the current application (video calling) can allow the designer to create desirable outcomes from nuanced gestures. Such things are hard to design for, but by considering such a spectrum, a designer might explore more flexible methods for error handling.

5.2 Meaningfulness of observed motion

Intentionally meaningful head gestures such as head nods and shakes may convey a complete message or combine with talk and other gestures, and may be part of foreground messaging or backchanneling [42, 43]. These may be observed by the system and, in communication scenarios, by other people, and thus their use to trigger action or not needs to be carefully considered. The continuum shades along to subconscious motions which may be culturally specific and important (e.g. South Asian 'head bobbles' and the like [59]), and thus careful consideration needs to be made of whether or not to ignore or smooth these out, or the threshold between them and intentionally meaningful gestures that the designer wants to use as triggers. Finally, blended into and out of intentionally meaningful and subconscious motions, there are the autonomic motions of balancing the head on the shoulders, head motions that follow shoulder motions, and so on. Again, the threshold between these and subconscious but important motions may be complex.

Kuno et al. [46] have suggested that there will be a need to explore how to combine intentionally meaningful, subconscious, and autonomic motions to trigger a range of comfortable and intuitive experiences. For example, if head motion observed from a headphone IMU is used to drive an avatar in video call or video game, there may be value in relaying some of the unintentional head motion so that the avatar appears more authentically human [9] (or to make robots more human-like [32]), as well as the clearly intentional and meaningful head gestures such as nods and shakes. Further, the wearer of headphones may be involved in multiple tasks at the same time. We need to guard against bringing both intentional and unintentional head motions from one task to the other, while also allowing for some flexibility. For instance, head movements incurred from moving around the house are likely unintentional within the context of a video call.

5.3 Context-awareness

Making headphones aware of their context of use can unlock a range of experiences triggered by naturalistic movements. For example, turning 90 degrees away from an in-progress video call and speaking to someone not on the video call may trigger muting of the microphone and reducing noise-cancellation (see Figure 6). To enable such context-aware system control, the headphones need to be aware of the applications in and out of focus, proximity and orientation to other active devices, and whether behavioural use is relevant to foreground or background applications on nearby devices. This has implications for cross-device interactions, and how headphones fit into evolving device ecologies.

5.4 Balancing Usefulness and Ease-of-Use

One can add more functionality disregarding the context by adding more input buttons. However, such input can only control the headphone device itself. Additionally, adding more buttons might increase the effort for learning and remembering the eyes-free interface. Our designs seek to create a balance between simple usage and increased functionality by minimizing the number of input gestures. We achieve this by incorporating the application, user's body, and the environmental context into our design considerations, and by utilizing some widely-known cultural gestures. For example, blocking the eye can be used to toggle the use of the camera in a video call.

6 LIMITATIONS AND FUTURE WORK

The gestures described in our paper represent a subset of gestures that people may engage in during everyday cultural interactions, and do not necessarily represent absolute best gestures. These gestures may be unnatural in different cultures and scenarios, and their implementation in actual software should be done with care. For instance, looking away from a screen for a short time should not mute the user in a video call, nor change the screen share. One way to mitigate accidental interactions like this would be to have a time threshold–only interactions clocked for a certain time duration (which might be different for different applications) should trigger the interaction. Designers may decide that an alternate set of gestures would better combine to produce a different set of features, more appropriate for both their cultural setting and application goals.

One of the reasons for choosing these gestures was the possibility to sense them using simple prototypes that might be feasible for commercial use, such as a solid state LiDAR used in Apple iPhone Pro models as an aid for the camera. However, there are interesting new sensors that maybe used to generate an even richer gallery of inputs. Ultra Wide Band (UWB) sensors enable broadcasting small packets of data to nearby devices without pairing, determining the proximity to these devices. Capacitive sensing may be enhanced to sense the hand near the vicinity of the headphones and not just touch, enabling better classifications of hand gestures. Small cameras maybe positioned to scan the environment around the user's head, and electrodes positioned along the headband may provide a sense of brain activity. Any additional sensor can expand the understanding of the context of the user actions and gestures, and be a base for future research. However, designers and engineers are strongly advised to consider the ethical implications of using some of these sensors in devices meant for everyday use.

Even though we used an exploration around headphones to arrive at our design space, the considerations of the input gesture and the context of use regarding the application, user's body, and the environment are also extensible to other kinds of wearables. In future work, we would like to explore this design space to inform design decisions with other kinds of wearable (and non-wearable) devices.

Finally, this work was done using prototypes built in a lab during COVID-19 social distancing. As such, developing robust prototypes and conducting ecologically-valid experimentation were not possible. Future work on the viability of this design space will be necessary.

7 CONCLUSION

Headphones represent a unique and underutilized design opportunity (e.g., see design space in Figure 15). They are a widely used and socially accepted consumer product, but usage so far has primarily been limited to (1) sound-related controls and (2) eyes-free tangible interfaces. In this paper, we propose an expanded design space for interaction modalities that can occur through headphones, and described designs that take advantage of the fact that headphones sit on the head and are used during diverse activities in different contexts. For example, our designs explore the benefits of incorporating different types of gesture (tangible, mid-air, and head orientation) into headphone control. This additionally allows us to consider gestures which aren't typically associated with directly controlling headphones (such as lifting an earcup), and use them to define contextually relevant application behaviour. The use of a semantically-rich location such as the head provides the potential to learn more about user behavior, and condition application behavior to better fit the context. In this way, headphones can be understood as leveraging existing behaviour [62] to mediate digital experiences in diverse scenarios. We hope this paper will result in additional exploration of headphones as a potentially rich site for interaction and sensing.

In this paper, we have explored design opportunities and challenges related to using headphones as a wearable site for interaction. We articulated a design space for contextualized headphone interactions in the modern everyday environment, grounded within frameworks from existing literature and influenced by the availability of new technology. We constructed functioning prototypes of several potential applications, and used these to ground a discussion about the design space. We hope this work will encourage designers to re-envision headphones as a more general input/output wearable. More broadly, we also hope that reporting on the details of our Research through Design (RtD) processes demonstrates how a materialist design philosophy can explore how to shape future experiences in myriad ways.

REFERENCES

[1] 2012. Exploring peripheral interaction design for primary school teachers. In *Proc. of Int. Conf. on Tangible, Embedded and Embodied Interaction, TEI 2012.* 245–252. https://doi.org/10.1145/2148131.2148184

[2] 2020. *Earphone and Headphone Market Demand, Size, Share| Forecast 2030.* Technical Report. Market Research Future. https://www.marketresearchfuture.com/reports/earphone-headphone-market-7628

[3] 2020. *Headset Statistics for 2021 - UC Today.* Technical Report. UC Today. https://www.uctoday.com/endpoints/headset-statistics/

[4] Kaho Abe and Katherine Isbister. 2016. Hotaru: the lightning bug game. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* 277–280.

[5] Apple. 2020. Apple introduces AirPods Max, the magic of AirPods in a stunning over-ear design. https://www.apple.com/ng/newsroom/2020/12/apple-introduces-airpods-max-the-magic-of-airpods-in-a-stunning-over-ear-design/

[6] S. Bakker, E. Van Den Hoven, and B. Eggen. 2015. Peripheral interaction: characteristics and considerations. *Personal and Ubiquitous Computing* 19, 1 (may 2015), 239–254. https://doi.org/10.1007/s00779-014-0775-2

[7] S. Bakker, E. Van Den Hoven, D. Hausen, A. Butz, T. Selker, and B. Eggen. 2014. Peripheral interaction: Shaping the research and design space. In *Conference on Human Factors in Computing Systems - Proceedings.* Association for Computing Machinery, 99–102. https://doi.org/10.1145/2559206.2560470

[8] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic interaction: designing for a proximity and orientation-aware environment. In *ACM International Conference on Interactive Tabletops and Surfaces.* 121–130.

[9] S. M Boker, J.F. Cohn, B.-J. Theobald, I. Matthews, T. R Brick, and J.R. Spies. 2009. Effects of damping head movement and facial expression in dyadic conversation using real–time facial expression tracking and synthesized avatars. *Philosophical Trans. of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3485–3495.

[10] Bose. 2019. Bose Frames original version. https://www.bose.com/en_us/products/frames.html

[11] John Bowers. 2012. The logic of annotated portfolios: Communicating the value of 'research through design'. *Proceedings of the Designing Interactive Systems Conference, DIS '12* (2012), 68–77. https://doi.org/10.1145/2317956.2317968

[12] J. R. Brubaker, G. Venolia, and J. C. Tang. 2012. Focusing on Shared Experiences: Moving beyond the Camera in Video Communication. In *Proc. of the Designing Interactive Sys. Conf. (DIS '12).* ACM, 96–105. https://doi.org/10.1145/2317956.2317973

[13] F. Brudy, C. Holz, R. Rädle, C.-J. Wu, S. Houben, C.N. Klokmose, and N. Marquardt. 2019. Cross-device taxonomy: Survey, opportunities and challenges of interactions spanning across multiple devices. In *CHI 2019.* 1–28.

[14] V. Buil, G. Hollemans, and S. van de Wijdeven. 2005. Headphones with Touch Control. In *MobileHCI '05.* ACM, 377–378. https://doi.org/10.1145/1085777.1085877

[15] W. Buxton. 1995. Integrating the periphery and context: A new taxonomy of telematics. In *Proc. of graphics interface*, Vol. 95. 239–246.

[16] F.R. Castelli and M.A. Sarvary. [n. d.]. Why students do not turn on their video cameras during online classes and an equitable and inclusive plan to encourage them to do so. *Ecology and Evolution* n/a, n/a ([n. d.]). https://doi.org/10.1002/ece3.7123 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.7123.

[17] Y.-c. Chen, D.-y. Huang, B.-y. Chen, Y.-C. Chen, C.-Y. Liao, S.-w. Hsu, D.-Y. Huang, B.-Y. Chen, and Exploring User. 2020. Exploring User Defined Gestures for Ear-Based Interactions. https://doi.org/10.1145/1122445.1122456

[18] E. Dagan, Márquez S.E., Altarriba B.. F., M. Flores, R. Mitchell, and K.e Isbister. 2019. Design framework for social wearables. In *Proc. 2019 Designing Interactive Sys. Conf.* 1001–1015.

[19] L. Devendorf, J. Lo, N. Howell, J.L. Lee, N.-W. Gong, M.E. Karagozler, S. Fukuhara, I. Poupyrev, E. Paulos, and K. Ryokai. 2016. " I don't Want to Wear a Screen" Probing Perceptions of and Possibilities for Dynamic Displays on Clothing. In *CHI 16.* 6028–6039.

[20] Matjaz Divjak and Horst Bischof. 2009. Eye blink based fatigue detection for prevention of computer vision syndrome. In *Proceedings of the 11th IAPR Conference on Machine Vision Applications, MVA 2009.* 350–353.

[21] T. Erickson and W.A. Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM TOCHI* 7, 1 (2000), 59–83.

[22] Y. Fernaeus and P. Sundström. 2012. The Material Move How Materials Matter in Interaction Design Research. In *DIS '12.* ACM, 486–495. https://doi.org/10.1145/2317956.2318029

[23] C. Frayling. 1993. Research in Art and Design. *Royal College of Art Research Papers* 1, 1 (1993).

[24] Bill Gaver and John Bowers. 2012. Annotated portfolios. *Interactions* 19, 4 (2012), 40–49. https://doi.org/10.1145/2212877.2212889

[25] W. Gaver. 2012. What should we expect from research through design?. In *Proceedings of CHI '12.* 937–946.

[26] S. Greenberg and B. Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings CHI '08.* 111–120.

[27] S. Greenberg, N. Marquardt, T. Ballendat, R. Diaz-Marino, and M. Wang. 2011. Proxemic interactions: the new ubicomp? *interactions* 18, 1 (2011), 42–50.

[28] A. Greenfield. 2004. *Everyware: The Dawning Age of Ubiquitous Computing.* New Riders Publishing.

[29] J. Grudin. 2001. Partitioning digital worlds: focal and peripheral awareness in multiple monitor use. In *CHI 01.* 458–465.

[30] E.T. Hall. 1966. *The hidden dimension.* Vol. 609. Garden City, NY: Doubleday.

[31] P. Hamilton and D.J. Wigdor. 2014. Conductor: enabling and understanding cross-device interaction. In *CHI 14.* 2773–2782.

[32] T. Hashimoto and H. Kobayashi. 2009. Study on natural head motion in waiting state with receptionist robot SAYA that has human-like appearance. In *2009 IEEE Workshop on Robotic Intelligence in Informationally Structured Space.* 93–98. https://doi.org/10.1109/RIISS.2009.4937912

[33] M. Heijboer, E. van den Hoven, B. Bongers, and S. Bakker. 2016. Facilitating peripheral interaction: design and evaluation of peripheral interaction for a gesture-based lighting control with multimodal feedback. *Personal and Ubiquitous Computing* 20, 1 (feb 2016), 1–22. https://doi.org/10.1007/s00779-015-0893-5

[34] K. Hinckley, J. Pierce, E. Horvitz, and M. Sinclair. 2005. Foreground and background interaction with sensor-enhanced mobile devices. *ACM TOCHI* 12, 1 (2005), 31–52.

[35] Wang H.L. and K. Sengupta. 2001. Manipulation of remote 3D Avatar through facial feature detection and real time tracking. In *IEEE ICME 2001.* 857–860. https://doi.org/10.1109/ICME.2001.1237857

[36] J. Hollan, E. Hutchins, and D. Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM TOCHI* 7, 2 (2000), 174–196.

[37] D.R. Olsen Jr. 2007. Evaluating user interface systems research. In *Proceedings of UIST '07.* 251–258.

[38] W. Ju and L. Leifer. 2008. The design of implicit interactions: Making interactive systems less obnoxious. *Design Issues* 24, 3 (2008), 72–84. https://doi.org/10.1162/desi.2008.24.3.72

[39] H. Jung and E. Stolterman. 2010. Material probe: exploring materiality of digital artifacts. In *Proc. of int. conf. on Tangible, embedded, and embodied interaction.* 153–156.

[40] H. Jung and E. Stolterman. 2012. Digital Form and Materiality: Propositions for a New Approach to Interaction Design Research. In *Nordic CHI '12.* ACM, 645–654. https://doi.org/10.1145/2399016.2399115

[41] S.a Junuzovic, K. Inkpen, J. Tang, M. Sedlins, and K. Fisher. 2012. To see or not to see: A study comparing four-way avatar, video, and audio conferencing for work. In *GROUP '12.* 31–34. https://doi.org/10.1145/2389176.2389181

[42] A. Kendon. 2002. Some uses of the head shake. *Gesture* 2 (Dec. 2002), 147–182. https://doi.org/10.1075/gest.2.2.03ken

[43] A. Kendon. 2004. *Gesture: Visible action as utterance.* Cambridge University Press.

[44] T. Kikuchi, Y. Sugiura, K. Masai, M. Sugimoto, and B.H. Thomas. 2017. EarTouch: Turning the ear into an input surface. In *MobileHCI 2017.* https://doi.org/10.1145/3098279.3098538

[45] S.R. Klemmer, B. Hartmann, and L. Takayama. 2006. How bodies matter: five themes for interaction design. In *Proc. conf. on Designing Interactive systems.* 140–149.

[46] Y. Kuno, T. Ishiyama, S. Nakanishi, and Y. Shirai. 1999. Combining Observations of Intentional and Unintentional Behaviors for Human-Computer Interaction. In *CHI 99.* ACM, 238–245. https://doi.org/10.1145/302979.303051

[47] D. Ledo, S. Houben, J. Vermeulen, N. Marquardt, L. Oehlberg, and S. Greenberg. 2018. Evaluation strategies for HCI toolkit research. In *Proceedings of CHI '18.* 1–17.

[48] Juyoung Lee, Hui Shyong Yeo, Murtaza Dhuliawala, Jedidiah Akano, Junichi Shimizu, Thad Starner, Aaron Quigley, Woontack Woo, and Kai Kunze. 2017. Itchy nose: Discreet gesture interaction using EOG sensors in smart eyewear. In *Proceedings - International Symposium on Wearable Computers, ISWC*, Vol. Part F1305. 94–97. https://doi.org/10.1145/3123021.3123060

[49] R. Li, J. Lee, W. Woo, and T. Starner. 2020. Kissglass: Greeting gesture recognition using smart glasses. In *Proc. of the Augmented Humans Conf.* 1–5.

[50] R. Lissermann, J. Huber, A. Hadjakos, S. Nanayakkara, and M. Mühlhäuser. 2014. EarPut: Augmenting ear-worn devices for ear-based interaction. In *OzCHI 2014.* 300–307. https://doi.org/10.1145/2686612.2686655

[51] P. Luff, C. Heath, H. Kuzuoka, J. Hindmarsh, K. Yamazaki, and S. Oyama. 2003. Fractured ecologies: creating environments for collaboration. *Human-Computer*

Interaction 18, 1 (June 2003), 51–84. https://doi.org/10.1207/S15327051HCI1812_3

[52] H. Manabe and M. Fukumoto. 2011. Tap Control for Headphones without Sensors. In *UIST 11*. ACM, 309–314. https://doi.org/10.1145/2047196.2047236

[53] H. Manabe and M. Fukumoto. 2012. Headphone Taps: A Simple Technique to Add Input Function to Regular Headphones. In *MobileHCI '12*. ACM, 177–180. https://doi.org/10.1145/2371664.2371703

[54] N. Marquardt, R. Diaz-Marino, S. Boring, and S. Greenberg. 2011. The proximity toolkit: prototyping proxemic interactions in ubiquitous computing ecologies. In *UIST '11*. 315–326.

[55] N. Marquardt, K. Hinckley, and S. Greenberg. 2012. Cross-device interaction via micro-mobility and f-formations. In *UIST 12*. 13–22.

[56] T. Matthews, T. Rattenbury, and S. Carter. 2007. Defining, designing and evaluating peripheral displays: An analysis using activity theory. *Human-Computer Interaction* 22, 1-2 (2007), 221–261. https://doi.org/10.1080/07370020701307997

[57] V. Mehta, P. Gupta, R. Subramanian, and A. Dhall. 2021. FakeBuster: A DeepFakes Detection Tool for Video Conferencing Scenarios. *arXiv preprint arXiv:2101.03321* (2021).

[58] C. Metzger, M. Anderson, and T. Starner. 2004. FreeDigiter: A contact-free device for gesture control. In *ISWC*. 18–21. https://doi.org/10.1109/iswc.2004.23

[59] A.L. Molinsky, M.A. Krabbenhoft, N. Ambady, and Y.S. Choi. 2005. Cracking the Nonverbal Code: Intercultural Competence and Gesture Recognition Across Cultures. *Journal of Cross-Cultural Psychology* 36, 3 (May 2005), 380–395. https://doi.org/10.1177/0022022104273658

[60] C. Neustaedter, S. Greenberg, and M. Boyle. 2006. Blur Filtration Fails to Preserve Privacy for Home-Based Video Conferencing. *ACM Trans. Comput.-Hum. Interact.* 13, 1 (March 2006), 1–36. https://doi.org/10.1145/1143518.1143519

[61] C. Neustaedter, J. Procyk, A. Chua, A. Forghani, and C. Pang. 2020. Mobile Video Conferencing for Sharing Outdoor Leisure Activities Over Distance. *Human–Computer Interaction* 35, 2 (2020), 103–142. https://doi.org/10.1080/07370024.2017.1314186 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2017.1314186.

[62] M. Nicas and D. Best. 2008. A study quantifying the hand-to-face contact rate and its potential application to predicting respiratory tract infection. *Journal of Occup. Environ. Hyg.* 5, 6 (2008), 347–352. https://doi.org/10.1080/15459620802003896

[63] S.Y. Oh, J. Bailenson, N.e Krämer, and B. Li. 2016. Let the avatar brighten your smile: Effects of enhancing facial expressions in virtual environments. *PloS one* 11, 9 (2016), e0161794.

[64] A. Olwal, J. Moeller, G. Priest-Dorman, T. Starner, and B. Carroll. 2018. I/O Braid: Scalable touch-sensitive lighted cords using spiraling, repeating sensing textiles and fiber optics. In *UIST '18*. 485–497.

[65] Payod Panda, Molly Jane Nicholas, Mar Gonzalez-franco, Kori Inkpen, Eyal Ofek, Ross Cutler, Ken Hinckley, and Jaron Lanier. 2022. AllTogether: Effect of Avatars in Mixed-Modality Conferencing Environments. In *2022 Symposium on Human-Computer Interaction for Work (CHIWORK 2022)*. ACM, Durham, New Hampshire, USA. https://doi.org/10.1145/3533406.3539658

[66] Z. Pousman and J. Stasko. 2006. A taxonomy of ambient information systems: Four patterns of design. In *Proc. of Workshop on Advanced Visual Interfaces*, Vol. 2006. 67–74. https://doi.org/10.1145/1133265.1133277

[67] B. Saatçi, R. Rädle, S. Rintel, K. O'Hara, and C.N. Klokmose. 2019. Hybrid Meetings in the Modern Workplace: Stories of Success and Failure. In *Int. Conference on Collaboration and Technology*. Springer, 45–61.

[68] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of IEEE CVPR '18*. 4510–4520.

[69] S. Santosa and D.J. Wigdor. 2013. A field study of multi-device workflows in distributed workspaces. In *Proc. ACM int. joint conf. on Pervasive and ubiquitous computing*. 63–72.

[70] P. Schoessler, S.-w. Leigh, K. Jagannath, P. van Hoof, and H. Ishii. 2015. Cord UIs: Controlling Devices with Augmented Cables. In *TEI '15*. ACM, 395–398. https://doi.org/10.1145/2677199.2680601

[71] D.A. Schön. 1983. *The reflective practitioner : how professionals think in action.* 374 pages.

[72] J. S.E. Zimmerman and J. Forlizzi. 2010. An analysis and critique of Research through Design: towards a formalization of a research approach. In *proceedings of the Conference on Designing Interactive Systems*. 310–319.

[73] L.M. Seuren, J. Wherton, T. Greenhalgh, and S.E. Shaw. 2021. Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics* 172 (2021), 63 – 78. https://doi.org/10.1016/j.pragma.2020.11.005

[74] P.J. Stappers and E. Giaccardi. 2017. Research through design. In *The encyclopedia of human-computer interaction*. The Interaction Design Foundation, 1–94.

[75] J. Sweller. 2011. Human cognitive architecture: Why some instructional procedures work and others do not. In *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues*. https://doi.org/10.1037/13273-011

[76] UC Today. 2022. *UC Spending Will Continue to Rise Say Firms*. Technical Report. UC Today. https://www.uctoday.com/unified-communications/ucaas/uc-spending-will-continue-to-rise-say-firms/

[77] Susie W. 2017. Using internet video calls in qualitative (longitudinal) interviews: some implications for rapport. *Int. Journal of Social Research Methodology* 20, 6 (2017), 613–625. https://doi.org/10.1080/13645579.2016.1269505 arXiv:https://doi.org/10.1080/13645579.2016.1269505

[78] M. Weigel and J. Steimle. 2017. DeformWear: Deformation Input on Tiny Wearable Devices. *Proc. of Interact. Mob. Wearable Ubiquitous Tech.* 1, 2, Article 28 (June 2017), 23 pages. https://doi.org/10.1145/3090093

[79] Mark Weiser. 1991. The Computer for the 21 st Century. *Scientific american* 265, 3 (1991), 94–105.

[80] M. Weiser and John S. Brown. 1997. The Coming Age of Calm Technology. In *Beyond Calculation*. 75–85. https://doi.org/10.1007/978-1-4612-0685-9_6

[81] M. Wiberg, H. Ishii, P. Dourish, A. Vallgårda, T. Kerridge, P. Sundström, D. Rosner, and M. Rolston. 2013. Materiality Matters—Experience Materials. *Interactions* 20, 2 (March 2013), 54–57. https://doi.org/10.1145/2427076.2427087

[82] C. Zhang, Y. Rui, and L. He. 2006. Light Weight Background Blurring for Video Conferencing Applications. In *2006 International Conference on Image Processing*. 481–484. https://doi.org/10.1109/ICIP.2006.312498

[83] Forlizzi J.and Evenson S Zimmerman, J. 2007. Research through design as a method for interaction design research in HCI. In *CHI 2007*. 493–502.

[84] A. Zolyomi, A. Begel, J.F. Waldern, J. Tang, M. Barnett, E. Cutrell, D. McDuff, S. Andrist, and M. R. Morris. 2019. Managing Stress: The Needs of Autistic Adults in Video Calling. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 134 (Nov. 2019), 29 pages. https://doi.org/10.1145/3359236