EarSense: Earphones as a Teeth Activity Sensor

Jay Prakash
Singapore University of Technology and Design (SUTD), Singapore, & University of Illinois at Urbana Champaign (UIUC), USA
jay_prakash@mymail.sutd.edu.sg

Zhijian Yang
University of Illinois at Urbana Champaign (UIUC), USA
zhijian7@illinois.edu

Yu-Lin Wei
University of Illinois at Urbana Champaign (UIUC), USA
yulinlw2@illinois.edu

Haitham Hassanieh
University of Illinois at Urbana Champaign (UIUC), USA
haitham@illinois.edu

Romit Roy Choudhury
University of Illinois at Urbana Champaign (UIUC), USA
croy@illinois.edu

ABSTRACT

This paper finds that actions of the teeth, namely tapping and sliding, produce vibrations in the jaw and skull. These vibrations are strong enough to propagate to the edge of the face and produce vibratory signals at an earphone. By re-tasking the earphone speaker as an input transducer – a software modification in the sound card – we are able to sense teeth-related gestures across various models of ear/headphones. In fact, by analyzing the signals at the two earphones, we show the feasibility of also localizing teeth gestures, resulting in a human-to-machine interface. Challenges range from coping with weak signals, distortions due to different teeth compositions, lack of timing resolution, spectral dispersion, etc. We address these problems with a sequence of sensing techniques, resulting in the ability to detect 6 distinct gestures in real-time. Results from 18 volunteers exhibit robustness, even though our system – *EarSense* – does not depend on per-user training. Importantly, *EarSense* also remains robust in the presence of concurrent user activities, like walking, nodding, cooking and cycling. Our ongoing work is focused on detecting teeth gestures even while music is being played in the earphone; once that problem is solved, we believe *EarSense* could be even more compelling.

CCS CONCEPTS

• **Human-centered computing** → *Interaction techniques*; • **Information systems** → **Mobile information processing systems**; • **Computer systems organization** → **Embedded and cyber-physical systems**.

KEYWORDS

Earable, Teeth gestures, User Interface, Vibroacoustics, Headphones, Earphones

MobiCom '20, September 21–25, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7085-1/20/09…$15.00
https://doi.org/10.1145/3372224.3419197

ACM Reference Format:
Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: Earphones as a Teeth Activity Sensor. In *The 26th Annual International Conference on Mobile Computing and Networking (MobiCom '20), September 21–25, 2020, London, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3372224.3419197

1 INTRODUCTION

This paper explores the possibility of sensing teeth-and-jaw motion using commercial off-the-shelf earphones. The key idea stems from the fact that teeth produce vibrations when they tap, slide, or grind against each other. These vibrations travel through the jaw and skull bones as surface vibrations, reaching the outer ear where the earphone is located (Figure 1). The earphone speaker's diaphragm responds to these vibrations, and a weak electric signal travels back through the audio jack to the sound card. Re-purposing the sound card allows us to extract these vibration signals.

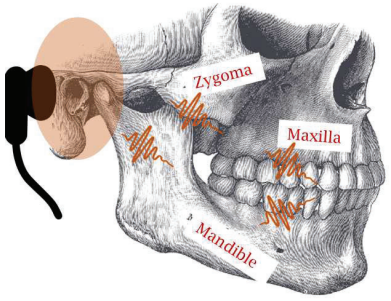

Figure 1: Vibrations generated by the teeth traverse as surface vibrations through the maxilla, mandible and zygoma bones, to ultimately arrive at the end of the jaw, near the ear.

In this paper, we present *EarSense*, a new approach to using earphones as a teeth activity sensor. *EarSense* accepts the extracted vibration signals as input and infers teeth activity such as teeth-tapping locations, sliding directions, grinding, etc. This enables a new form of contact-less user interface where a user can scroll, click, type, pause, etc. simply using his/her teeth. Unlike other contact-less user interfaces like hand or body gestures, eye trackers, or voice activated interfaces, *EarSense* is non-invasive, secure, and maintains the privacy of the user, i.e., the command cannot be heard or seen by anyone. Section 2 will envision the landscape of potential

applications, however, this paper focuses on laying the foundations around teeth-related vibration signals and inferences from them.

While there has been substantial interest around facial sensing in the medical, wearable and HCI communities [6, 9, 12, 16, 21, 25, 28], *EarSense* differs from past work in 3 key aspects. *First*, the signals we capture are very weak, distorted and polluted by various other sound and vibration signals. *Second*, unlike past work that have designed dedicated hardware and invasive devices for such forms of sensing, we capture the signal using everyday off-the-shelf earphones that people wear normally. *Finally*, our approach focuses on measuring and modeling signal propagation through teeth and jaw, improving our understanding of the underlying model in order to generalize our system. Past work, to the extent we have found, have leaned heavily on directly applying learning and classification of specific tasks, derived from a focused use-case or application [3, 9, 22, 28]. We believe this paper will shed light on the properties and behavior of vibrations inside the mouth, while also showing that even off-the-shelf earphones can derive useful information from them.

We define our problem as follows: *given two signal streams sensed by the left and right earphones, our goal is to reliably identify as many teeth related gestures.* We characterize a gesture as an $\langle action, location \rangle$ tuple, meaning that a particular teeth action is being performed at a specific location. For example, actions could be tapping or sliding, while locations could be front teeth, left teeth, top teeth, etc. Hence, a user may perform a ⟨ *downward slide, middle-teeth* ⟩ to scroll down her mobile phone screen, while a ⟨ tap, right teeth ⟩ could be a right click. Gestures can be defined at will, as long as users can perform these gestures without difficulty. Clearly, the finer the granularity at which we recognize actions and locations, the more gestures we can support.

To achieve reliable gesture recognition, we leverage 3 opportunities in extracting such ⟨ *action, location* ⟩ primitives. *First*, teeth actions exhibit diversity in their spectral properties because each action produces distinct physical forces and timing. *Second*, we can localize the actions using the relative time gap at which vibrations arrive at the left and right ear. This time gap is a function of the distance the vibration has traveled through the jaw/skull bone, which is in turn a function of the location. *Third*, we can leverage the structure of human mouth and teeth to further constraint the location of the action and improve the reliability of our detection.

Translating the above ideas into a real-world practical system requires addressing several challenges. (1) The distances inside the mouth are small, resulting in tiny time differences that are hard to estimate with the earphone's low sampling frequencies. This is particularly problematic for high frequency vibrations that travel fast, making them harder to distinguish.[1]. (2) The wide variation across individuals makes it difficult to adopt global models. To avoid per-user training, we need to process and identify patterns in the signal that can be robustly used for gesture identification. (3) Finally, human motion, such as walking, also produce vibrations in the body; these vibrations partly contaminate the earphone signals, polluting teeth gestures. The earphones speakers themselves also inject distortions due to various hardware artifacts. Any gesture recognition must cope with such interference.

This paper develops a fusion of geometric modeling and signal processing techniques to systematically address the above challenges. We implement *EarSense* on a simple platform of earphones connected to a sound card (we test 5 different headphones and 3 sound cards, including the most popular *Realtek* cards). Only the software of the sound cards is modified so that the speaker's vibrations are extracted as a regular *.wav* file. The modifications are minimal (<5 lines of code).

We evaluated *EarSense* on 18 volunteers (including an 8 year old child) and collected teeth-gesture data; no prior calibrations or trainings were performed. Our findings reveal that 6 gestures can be decoded with consistent robustness at >90% accuracy. We also found that users are not able to perform more than 8 gestures with their teeth, and are most comfortable with 6-7 gestures, implying that *EarSense* is close to the limits of viability. For a potential tooth-brushing application (where volunteers brushed their teeth), *EarSense* was able to localize the brush at 7 regions of the lower and upper teeth.

In sum, our contributions can be summarized as follows.

- We develop *EarSense*, a relatively new sensing modality that uses off-the-self earphones to sense teeth-activity.
- We propose techniques to identify, model, and localize gestures from teeth vibration signals; we cope with interference from other activities, and generalize to different users without the need for per-user training.
- We implement and evaluate *EarSense* on 18 users and 2 example applications: a private user interface for sending commands and a tooth-brushing monitor. We have not tackled the problem of teeth sensing while the earphone is playing audio sounds; we leave this to future work.

We begin the rest of the paper with a discussion on potential applications for *EarSense* followed by basic measurements, system design, evaluation, and future work.

2 APPLICATIONS

This section envisions a host of possible applications that can build on top of *EarSense*. While each application would bring unique challenges and opportunities, this paper is focused on enabling the core technical capability. Application specific customization would emerge with time.

■ Accessibility

Google's "Switch Access" for Android devices [1] is a growing platform for accessibility needs. Google's core idea is to use a separate device consisting of several large push-buttons. In accessibility mode, the Android screen sequentially highlights the click-able icons and the user is expected to press a button when the desired icon is highlighted (see video here [13]). *EarSense* can eliminate this additional "Switch" device by re-purposing the earphone. Teeth gestures should also offer faster/flexible navigation options to those who can perform them. Importantly, far more patients face challenges with moving their hands and fingers, compared to their teeth.

[1]Unlike RF signals or sound signals in the air, the speed of surface vibrations in the body is frequency dependent and is faster for higher frequencies [37].

Hence, *EarSense* is likely to find broader applicability than finger-controlled "Switch Access". For instance, navigation joy-sticks in wheel chairs could be combined with a *EarSense*-like interface.

■ **Health Sensing: Seizures and Dental Disorders**
An oncoming seizure or epilepsy can often be preceded by symptoms like teeth chattering, repetitive lip smacking, a certain body odor, etc. [7]. Today, such patients are often accompanied by dogs that are capable of smelling odors that patients are known to emit 30 – 45 minutes before the seizure [5]. *EarSense*'s ability to sense teeth-chatter could open new ways of early seizure detection, valuable especially when alert dogs are unavailable.

Various dental disorders, such as malocclusion, oclussion rehabilitation, and certain periodontal diseases require doctors to record the patient's teeth actions and vibrations. With *EarSense*, such visits to the clinic can be avoided. Perhaps everyday dental monitoring – such as proper chewing, brushing, and flossing habits – could also be facilitated by *EarSense*. The need to monitor such everyday behavior may be necessary in some cases, particularly for kids.

■ **Hands-free Interfaces**
In hospitals, doctors are often unable to perform basic touch-screen gestures since their gloves need to be removed first. Workers in factories, construction sites, and warehouses often have their hands occupied, thereby unable to perform simple machine operations; they may have to break the work flow to, say, move a lever left to right. Many of these cases are noisy sites, forcing workers to wear headphones or earmuffs. We believe *EarSense* can plug various "holes" in streamlining such work-flows.

■ **Two Factor Authentication (TFA)**
Surveys show that users still avoid TFA where possible (<10% and <20% of Gmail and Microsoft Office 365 users, respectively, use TFA [27, 30]). The main reason is the burden of finding the phone, unlocking the screen, and typing one time passwords (OTPs) retrieved from either SMS or in-application code generators. *EarSense* could reduce this burden by treating the earphone as the input interface to the phone. Instead of screen gestures on the phone, the user could perform teeth gestures that the earphone would forward to the paired smartphone. The rest of the TFA process can remain as is; the smartphone could forward the gesture to the authentication server, allowing the user to log-in. Figure 2 is a screenshot from a demo of TFA with *EarSense* – the web browser on the left shows the OTP on the screen; the right panel shows the user's face, while he is entering the OTP through teeth-gestures.

Finally, unlocking phones with PIN numbers (or swiping gestures) are susceptible to over-the-shoulder eavesdroppers. *EarSense* protects from such attacks since decoding teeth motions inside the mouth raises the bar for attacks. Thus, in conclusion, we envision *EarSense* as the supporting technology for a number of emerging applications. The exact productization process would need more customization and engineering.

3 BASICS AND MEASUREMENTS

We now lay out the basic properties, assumptions, and measurements to set up the research context for this paper.

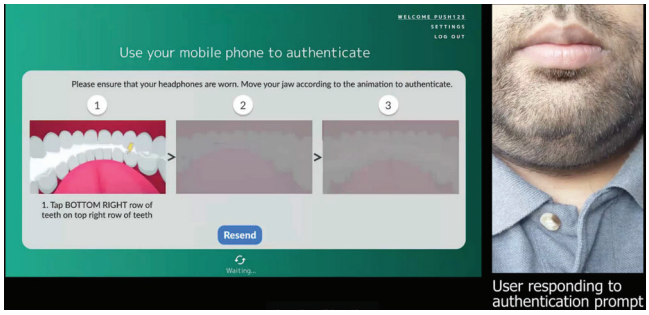

Figure 2: Screenshot on the left shows a web application presenting an OTP as part of a Two Factor Authentication (TFA) process. The right panel shows the user entering the OTP via teeth gestures.

(1) Defining the teeth gestures

When humans move their mouth/jaw, it is the lower jaw that moves. There are indeed 3 degrees of freedom (up/down, left/right, and in/out), but each of these motions is quite heavily restricted (especially left/right and in/out). This limits the number of gestures possible. Moreover, humans have limited control on their teeth, i.e., a user cannot tap only their upper and lower canine without the adjacent teeth touching each other. This implies that teeth gestures would have to be coarse-grained, especially if users must perform them without difficulty.

From user studies with 18 volunteers, we observed that only a few are able to perform 9 gestures as shown in Table 1, while the rest preferred 6-7. This scopes out our teeth gesture set and the 3 main gesture categories – namely Taps L/M/R, Slides L/R, and Slides U/D – are explained visually in Figure 3. Observe that Taps L/M/R and slides L/R are relatively easy to understand. Slide Up indicates that the lower teeth move upward while grazing the upper teeth, and Down is the vice versa. For both the sliding actions, multiple upper and lower teeth can touch each other at the same time. While performing experiments with volunteers, we demonstrated the teeth gestures using our hands, where fingers denoted the teeth.

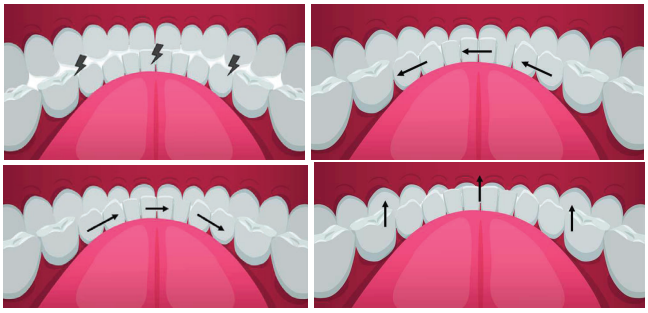

Figure 3: Visualizing teeth gestures: (a) taps left/middle /right (b) slide: left (c) slide: right and (d) slide: up

(2) Visualizing *EarSense* signals from earphones

Fig. 4 shows the time domain signal picked up by the left and right earphones when a user is tapping the left canine teeth, and then

Taps L/M/R	1. Tap extreme left side upper and lower teeth
	2. Tap left side upper and lower teeth
	3. Tap middle upper and lower teeth
	4. Tap right side upper and lower teeth
	5. Tap extreme right side upper and lower teeth
Slide L/R	6. Slide teeth from left to right
	7. Slide teeth from right to left
Slide U/D	8. Slide front lower teeth upwards
	9. Slide front lower teeth downwards

Table 1: Nine teeth gestures users could perform, however some users found it difficult to separate between (1,2) and (4,5), resulting in 7 "comfortable" gestures.

sliding them from left to right. High level observations are that tapping produces impulse-like signals in the time domain, while sliding generates weak but extended vibrations. Fig. 5 shows the frequency domain representations of the same tap and slide gestures. Unsurprisingly, the taps are far wider in bandwidth. Additionally, the signal travels through different teeth, and since their chemical compositions are quite different, they impact the frequencies differently, resulting in substantial spectral variations between the left and right ear. For slides, however, the action produces a narrow band signal. The effect of different teeth on this narrow band is far more uniform.

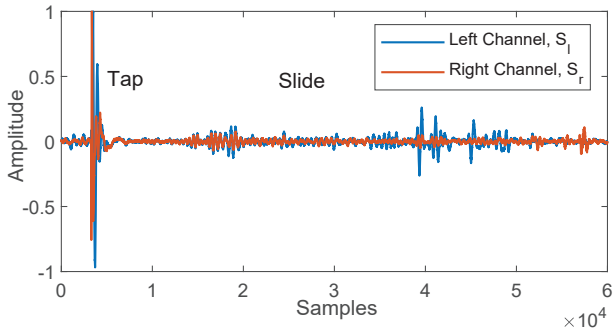

Figure 4: Signals at left and right channels of headphone for tapping left canines and sliding left to right.

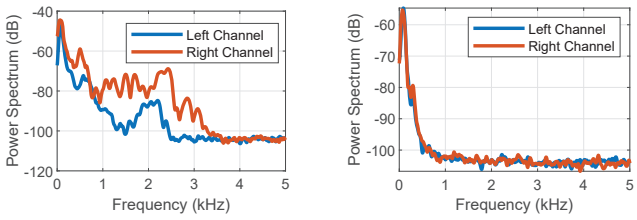

Figure 5: (a) Teeth tap/strike has a larger bandwidth while (b) sliding of teeth has narrower bandwidth (and more uniformity) across the left and right ear.

(3) Surface or air vibrations?

We intend to understand if the earphone speaker's diaphragm is vibrating due to solid-surface vibrations, or due to air vibrations travelling through the internal air-cavities, in the mouth and the ear, or through external air channel around the face. Fig. 6 compares the two cases by performing teeth gestures when the user is wearing the earphone, and then performing the same gestures when the earphone is not in contact with the human ear (i.e., there is a small air-gap between the earphone and the ear). Evidently, the signal is almost at the noise floor when the earphone is not in contact, proving that received signals are indeed surface vibrations.

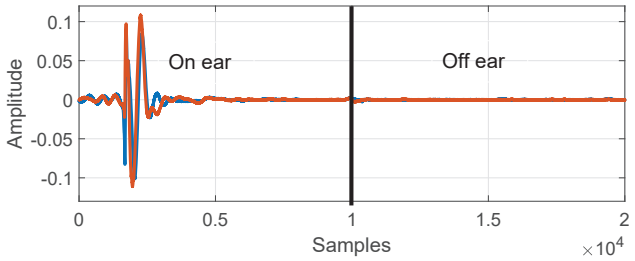

Figure 6: A small air gap between the earphone and the ear prevents sensing of the teeth vibrations.

Importantly, Fig. 7 shows the case when the user is speaking, and the earphone is worn on the ear. This suggests that human speech induces surface vibrations in the jaw and skull, which is also picked up by the earphone's speaker. This aligns with theories and experiments in literature [38]. Typically, for earphones/headphones to capture sounds over the air, the diaphragm has to be very close to the source.

(4) Can other activities interfere with *EarSense*?

Fig. 8 shows the spectrogram when a user was asked to perform various activities (e.g., nodding, walking, listening to ambient music, speaking, etc.) while wearing the earphone. Evidently, teeth gestures and self-speech exhibit far stronger energy than any of the other activities. Walking also produces some signals primarily because every strike of the foot on the floor creates vibrations that propagate through the human skeleton [31]. This vibration can interfere with the teeth signal. Importantly, since the foot-strike is also an impulse-like signal, it has a wide frequency footprint (evident from the graph). For other activities, including loud music, there is no perceptible signal above the noise floor.

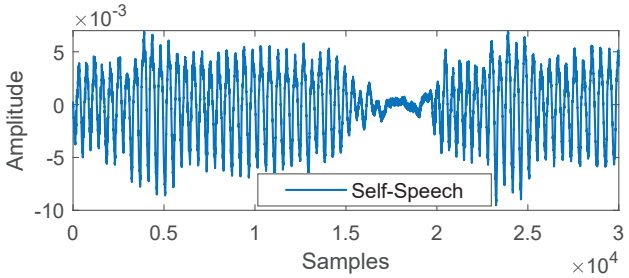

Figure 7: Human speech produces surface vibrations inside the human mouth, which is also picked up by the earphone speaker's diaphragm.

(5) Effect of earphone wearing positions

Every time the user wears the earphone, the position and orientation of the device may not be identical. If the teeth gesture signal is sensitive to this position/orientation, then *EarSense* may not be viable for practical applications. Fig. 9 shows the consistency of the

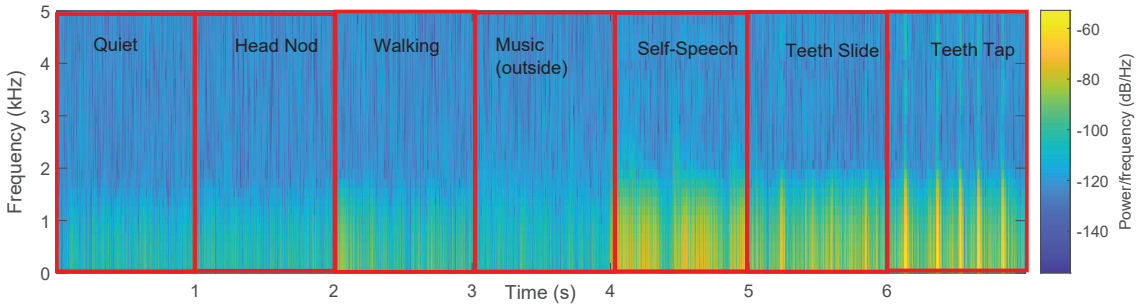

Figure 8: Spectrum of signals captured for different activities, starting from quiet, head nodding, walking, music from home speaker, self-speech, teeth slide, teeth tap.

detected signal when the same teeth gesture is performed across 5 different wearing positions. In this experiment, 5 different people were asked to mount the earphone on a single user, producing natural diversity across the wearing positions. The raw measurements are still reasonably consistent.

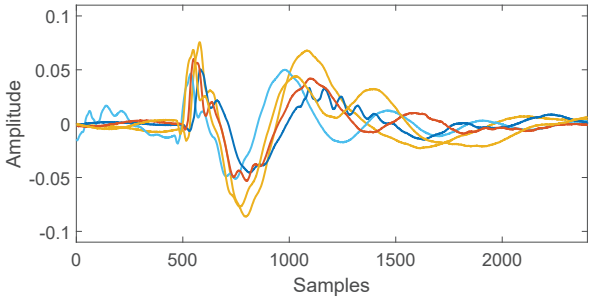

Figure 9: Teeth gesture signals are consistent across wearing positions, a requisite for robustness.

4 SYSTEM OVERVIEW

Fig. 10 presents the overall flow of *EarSense*'s processing pipeline. *EarSense* accepts two channels of vibration signals (as measured by stereo headphones) as input and returns a $\langle action, location \rangle$ pair as output. The input signals are first checked for teeth gestures (to separate from other activities such as walking, eating, or speaking). Once confirmed, the next step is to classify whether the gesture is tap or slide. This is not difficult since the spectral properties exhibit a distinct difference.

If a tap is detected, then *time difference of arrival* (TDoA) techniques are employed first to trilaterate the source location. In general, TDoA is a promising indicator because the ear that is closer to the tap location receives the vibra-acoustic signal earlier. But at times, i.e., in 35% of our sample space, users are unable to ensure that only the suggested teeth locations are tapped. There are softer interactions at other locations of jaws that corrupt the clean TDoA of captured signals. Hence, we use additional properties of spectral dispersion, and a third opportunity that we call "cheek waves" to boost the reliability of tap localization.

Now, if a slide is detected, a delay profile is derived using sliding windows and is matched to estimate if the slide was from left-to-right or the other way. Along similar lines, the presence of an impulsive surface wave followed by narrow-band body waves are used to classify between upward and downward slides.

As detailed in Section 5, we jointly exploit properties of the generated signals (using spatio-temporal analysis), basics of waves propagation, and geometric constraints in the motion of jaws to detect 7 teeth gestures in the form of $< action, location >$ pairs. Together, they form the building blocks for a teeth-based user interface, enabling different applications as discussed in Section 2.

5 SYSTEM DESIGN

This section expands on the individual modules outlined in Section 4. Let us denote the system inputs as S_l and S_r, corresponding to the two signals recorded at the left and right headphones.

5.1 Teeth Gesture or Not

The function of this module is to separate teeth gestures from all other interference and noise. As shown earlier, the headphone's diaphragm is almost behaving like a "conduction microphone", meaning that it only picks up surface vibrations (and hardly any air-borne signals). Through extensive testing, we observed that 3 physical motions register surface vibrations at the headphone: (1) self-speech, (2) walking, and (3) eating. Since *EarSense* assumes that the user would not gesture while eating, our task is to reliably recognize self-speech and walking. This entails 2 steps, namely pre-processing (to identify segments of activity) followed by activity recognition. Figure 11 shows the sequence.

Pre-Processing: We begin by computing the envelope $Env(\cdot)$ of the input signal. A sliding window on the envelope calculates the energy of the n^{th} window at the left and right channels as:

$$E_{l/r}(n) = \sum_{i=kn}^{kn+N} Env^2(S_{l/r}(i)) \quad (1)$$

where N is the size of each window, k is the step size for sliding the window. An activity is suspected only if both $E_r(n)$ and $E_l(n)$ are at least double the noise floor. Then, if such energy is observed for longer than $\tau_N = 0.25s$ – the minimum duration of any walking or eating action – then the whole high energy time segment is extracted for activity classification. We denote this segment as S'_r and S'_l.

Detecting Self-speech: Self-speech signals are generally easy to identify. The opportunity lies in the distinctive base + harmonic

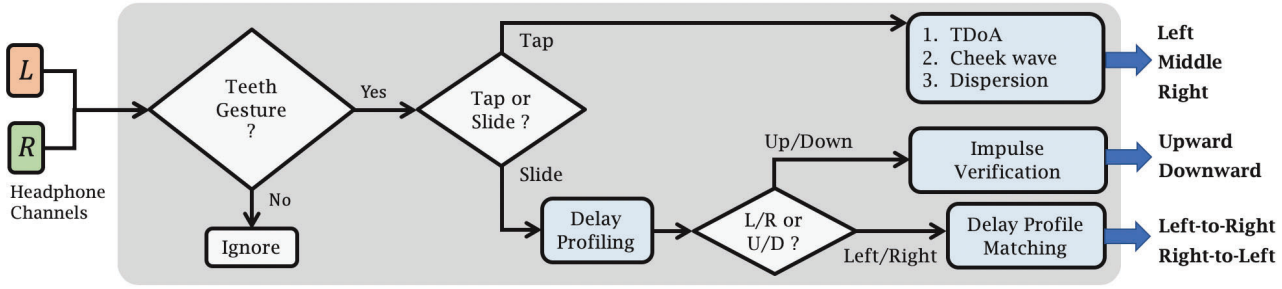

Figure 10: *EarSense* architecture: A .wav file consisting of 2 vibro-acoustic signals are accepted as input and teeth gestures (one of seven candidates) are presented as an output.

frequency patterns produced by a human voice. Specifically, the base frequency of human speech is within [80, 255] Hz [20]. If the vibrations are indeed speech, the peak in (S_r' and S_l') should be within the range of [80, 255] Hz, which in turn gives us the base frequency. Knowing the base frequency, we can compute the harmonics (since they are integer multiples of the base); thus, we calculate the energy sum, s, of the base and the first harmonic. The ratio of s to the energy of the entire segment E reveals the presence of speech signals. Non-speech signals do not exhibit such harmonic patterns in the low frequency bands.

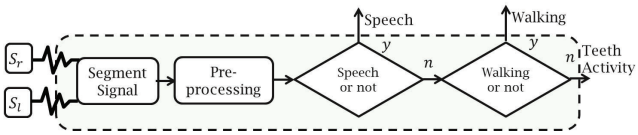

Figure 11: Flow of operations for teeth-gesture or not.

Detecting Walking: Human walk comprises of leg swings and heel strikes on the ground. Swinging does not induce vibrations at the headphone diaphragm, but heel strikes produce clear spikes. Though both leg and tooth induce impulse-like vibrations in the skeleton, tooth strikes involve enamel which is the hardest substance in human body (hardness of 5 on *Mohs scale* of hardness [8]). Steel achieves a value of 4 on the same scale [35]. On the other hand, leg strikes involve much softer surfaces, i.e., soft shoe cushion. Additionally, body acts as a filter to the traversing vibrations as well. Hence these two signals end up falling in quite different frequency buckets. We observe leg strikes concentrated in sub 100 Hz range, while teeth strike can go all the way up to 2 kHz. We also observe that sliding teeth also generates signals with substantial energy in frequencies above 100 Hz. These permit energy-based classification, especially in the 100-2000 Hz band. Specifically, we calculate ratio of the energy in the [100, 2000]Hz band, to the entire signal energy. This ratio reliably separates the human walking activity.

5.2 Teeth Tap or Slide

Once teeth gesture is detected, our goal is to separate the 2 types of gestures. The opportunity arises from the spectral properties of tap and slides. As shown in Figure 5, tap is an impulse-like force in the time domain, exhibiting a sharp and abrupt change in first and second orders of derivatives. On the other hand, sliding generates smoother and narrow-band signals of relatively longer duration.

Fig. 12 shows the sequence of operations. Unfortunately, Fourier analysis is not reliable in such cases due to sharp changes and potential discontinuities. Continuous wavelet transform (CWT) has been reported to be far more robust [10, 32, 36, 39]. CWT coefficients are larger near abrupt local changes in the signal, hence helps in identifying transients. We use Wavelet transformation over S_r' and S_l' to capture the energy span with high spatio-temporal resolution. Fig. 13 shows CWT for an example tap on left canine teeth, and an example slide from left to right.

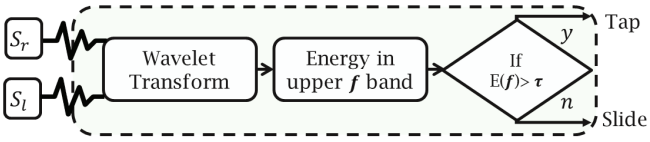

Figure 12: Classifying teeth tap vs. teeth Slide.

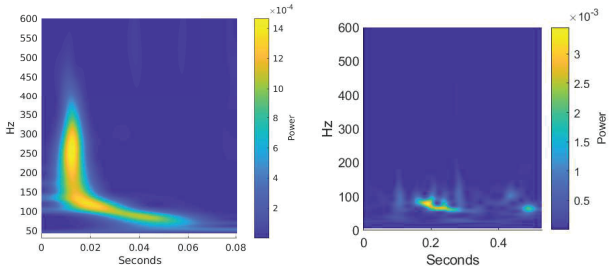

Figure 13: (a) Wavelet transform of hitting left canine. Such strikes/taps have energy in wider band and lower frequencies arrive later in time. (b) Slides from left to right exhibit different properties than taps.

The energy in frequencies > 200Hz, and the signal duration, are together analyzed to classify between taps and slides. A segment is declared a tap if the energy in the upper frequency band is on the higher side of the decision boundary at *either* the left or the right earphone, and signal duration is shorter than a threshold. This threshold is defined by computing the max of the tap-duration, and adding the variance to it, as a safety factor. This proves highly robust.

5.3 Tap Localization

Once a tap is identified, we intend to localize it in the granularity of left, middle and right. Time difference of arrival (TDoA) should be a

promising indicator for detecting the position of the tap, i.e., for the tap in Fig. 14, the user's left ear should receive a delayed copy of the right ear's recordings. TDoA can be written as $\frac{2d_{sc}}{V(f)}$, where d_{sc} is the distance of strike from center of the teeth. As such, sounds produced during occlusion (i.e., the closing of the lower jaw onto the upper jaw) travels through hard and soft tissues as a mixture of compressive, shear, and transverse waves at speeds up to 1000 m/s [11]. But, unlike localization on uniform surfaces, TDoA inside the mouth faces issues from heterogeneous chemical compositions even among symmetric teeth [18], location of the tongue, and other physiological factors. This precludes TDoA from being the only feature for robust localization.

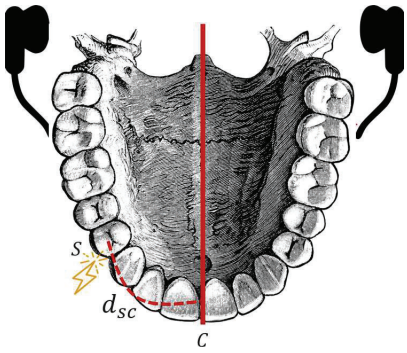

Figure 14: TDoA at left and right channels of a headphone is used for estimation of a strike at any location, S, on jaw with reference to the centre, C.

To this end, we extract 3 other features from the signals, $S_r^{'}$ and $S_l^{'}$, and perform a voting to classify tap location. These features are:

- Measure of acoustic dispersion
- CWT filter bands based TDoA
- Dominance of the cheek waves

Acoustic Dispersion: When teeth strike each other at any location, surface acoustic waves (SAW) are generated. A solid and hard surface like teeth is a dispersive medium and transmits SAW of different frequencies at different speeds, $V(f)$, [17, 37]. The propagation speed depends on the physical properties of the material (density, Young's modulus, thickness and Poisson ratio). Now, for a wide band signal generated from teeth taps, different frequency components arrive at the headphone's diaphragm at different times. This phenomenon of dispersion leads to asymmetric signals in the channels (left and right) of earphone if d_{sc} is non-zero. As the tap location approaches towards either ear, d_{sc} increases, the recorded signals witness greater asymmetry. The degree of asymmetry is a measure of spectral dispersion, hence a promising feature for tap localization.

Fig 15 shows the dispersion, $\lambda_1 < \lambda_2 ... < \lambda_6$, in tap signal as obtained at one of the channels. We propose to exploit dispersion and its dependence on d_{sc} using offsets between corresponding zero-crossing times at left, $S_l^{'}$, and right, $S_r^{'}$, channels for estimating the location of the tap. First, the instance of the burst of high frequency waves is calculated. Then all temporal instances, Tl_i, of zero amplitudes are extracted. Let $\mathbf{Z}_l(\mathbf{Z}_r)$ be the zero-crossing instances for $S_l^{'}(S_r^{'})$ and represented as:

$$\mathbf{Z}_l \leftarrow [Tl_1, Tl_2, Tl_3, ..., Tl_N]. \quad (2)$$

The zero-crossing difference (ZcD) is defined as:

$$\mathbf{ZcD} \leftarrow [Tl_1 - Tr_1, Tl_2 - Tr_2, Tl_3 - Tr_3, ..., Tl_N - Tr_N]. \quad (3)$$

If **ZcD** has majority of positive elements, left channel witnesses comparatively late arrivals of frequencies, it would indicate towards the tap on the right side of the teeth. The sum of first k elements in **ZcD**, should be a function of distance of separation, d_{sc}. If the $\sum$ is above a (+)ve threshold, right tap is reported; below a (-)ve threshold, left tap is reported; and within ϵ around 0, then middle tap is reported. We acknowledge that the individual differences across left and right teeth might lead to different numbers of crests and zero-crossing instances. We create ZcD of a length corresponding to the minimum of the two sides. These would have created challenges if we attempted a higher resolution of recognition.

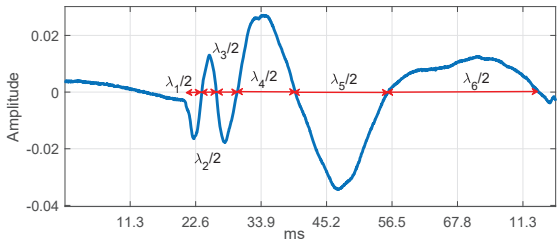

Figure 15: Tap signal: dispersion is clearly visible, high frequency signal comes earlier than lower ones.

Wavelet Transform based TDoA: Recall that frequencies travel at unequal speeds through solids, leading to different TDoAs [37]; this pollutes cross-correlation based TDoA approaches. To tackle this problem, we first filter the signal into narrow frequency bands and then compute TDoA for each band. Due to geometric constraints of the jaw, the distance of strike, $2d_{sc}$, is typically low ($\approx 4cm$) for both left and right strikes. Since high frequencies have larger $V(f)$, they arrive on almost same time at both sides of the ear and are used to detect the start of the taps. These can be seen as sharp dips (or oscillations) in the time domain. Hence, higher frequency bands do not lend themselves well to TDoA. We note that the time gap between the arrival of lower frequencies is much more, due to separation in the distance, than those corresponding to the initial burst of higher frequencies. *EarSense* finds early arrival of energy in CWT bands below 300 Hz for both $S_l^{'}$, and $S_r^{'}$. Further, it finds time difference between left and right channels for each of the bands and determines tap location like the ZcD-based method.

Cheek Waves: When performing a tap, just before the teeth strike each other, cheek muscles around the end of the jaw contracts and generates pulses. Fig 17 shows the weak cheek wave signals from the earphones. Though the contractions happen near both ears, it is dominant on the side of the tap. Although cheek waves are weak, they offer a valuable hint for localization.

We begin by identifying the teeth-strike window in both $S_r^{'}$ and $S_l^{'}$. Then a small window of t ms (before the start of teeth strike) is extracted out and analyzed for the presence of cheek waves. A confidence score is generated corresponding to differences in energy

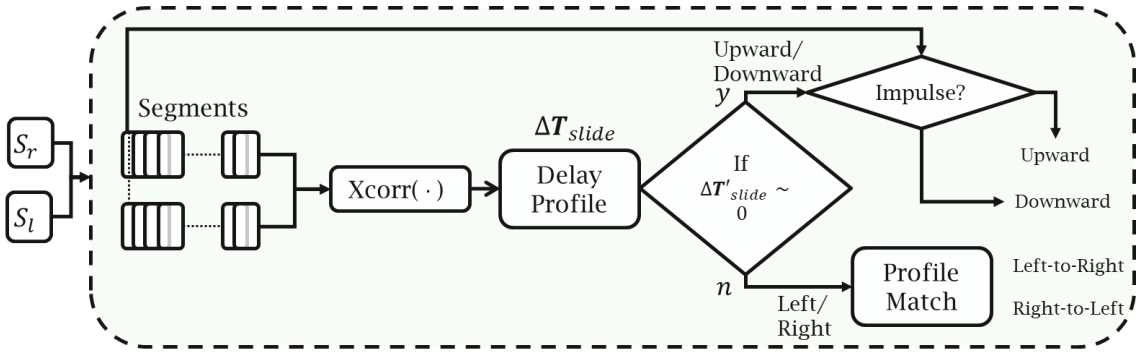

Figure 16: Slide type and localization: Left/Right or Up/Down.

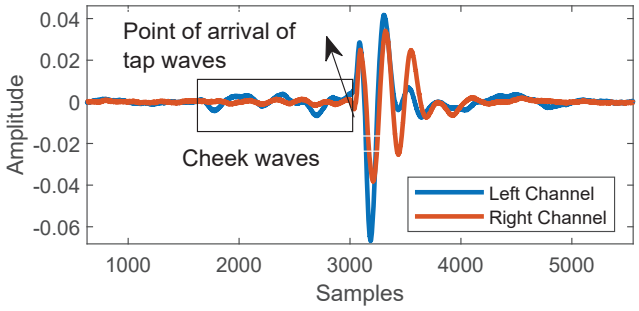

Figure 17: Cheek waves as an opportunity.

(between the left and right channels). Specifically, if the difference is near zero (i.e., within twice of the noise floor), then we declare the strike location as "middle". If the difference is greater (i.e., (+)ve), then we localize the tap as "left", and vice versa when the difference is (-)ve and its power is less than $2x$ of the noise floor.

Finally, the 3 parameters of ZcD, TDoA, and cheek waves energy, are together used to localize tap, via majority voting.

5.4 Teeth Slide Classification

Classifying between left, right, up, and down teeth-slides is rooted in temporal analysis of the relative delays between two earphone channels. Fig. 16 illustrates the overview of the operations. At a high level, left and right slides are classified based on the variation of delay differences over time. Upward and downward slides have similar delay profiles since they occur in the middle of teeth, hence, an impulse detection method is used (to be explained soon).

Delay Profiling: Horizontal or Vertical: A slide can be segmented into a sequence of windows at both channels. Each window gives a hint about the location of teeth interaction, i.e., slide is actually a time series of $< action, location >$ pairs, through the TDoA of signals on the two channels. We call the series of TDoA over different segments as the delay profile. By examining the delay profile, we show it is possible to tell horizontal and vertical slides apart.

Specifically, sliding teeth produces friction between the upper and lower jaw, which in turn generates body waves. The acoustic waves emanating from slides are comparatively narrow band, restricted around 50 – 200 Hz. The signals received by left and right channels are the source signal convoluted with the left and right channels of teeth, thus are delayed based on the swipe location. Dispersion is not evident here, and as can be seen in Fig. 18, they are correlated and temporally shifted, and thus delay can be directly calculated using cross-correlation.

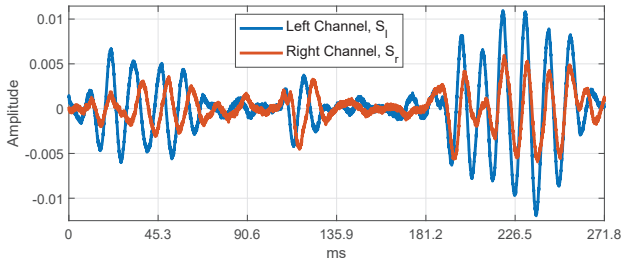

Figure 18: Slide signal: dispersion is not evident, delay profile can be obtained by cross-correlation.

A moving window calculates a sequence of delays between two channels, creating a profile. Here delay profile, ΔT_{slide}, is an array corresponding to delay calculated between S'_r and S'_l for each of the overlapping moving windows and whose i^{th} element is given as:

$$\Delta T_{slide}(i) = \arg\max_k(corr(S'_{l,i}(1:end), S'_{r,i}(k:end)). \quad (4)$$

Here $S_{l,i}$ and $S_{r,i}$ are given as $(i-1)win/2 : (i+1)win/2$ samples of S_l and S_r. The geometric constraints of the jaw present opportunities for classifying teeth slides. During a vertical slide, the teeth in contact do not change much. Hence $\Delta \mathbf{T}_{slide}$ is nearly constant and close to 0. Biases are possible because of potential deformations in some jaw anatomies, but this would be a constant. For horizontal slide, the contact part of the teeth will change greatly, leading to a non-constant delay profile, Thus, by looking at the variance of delay profile, if stable (i.e., nearly constant), we will report a vertical slide, otherwise report a horizontal slide.

If Vertical → Upward or Downward: We observe that up and down slides vary in the existence of dispersion at the start (i.e., when the signal arrives). Upward slides are impulsive and show

more dispersive characteristics at the start. This is identified from the spread of the spectrum, CWT coefficients, and ZcD in the first window.

If Horizontal → Leftward or Rightward: ΔT_{slide} obtained from the previous step, is compared against the expected set of delay templates (see Fig. 19).

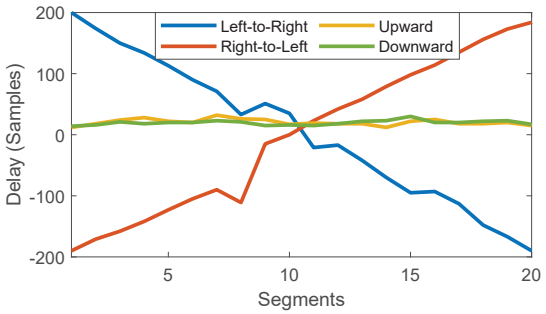

Figure 19: Relative delays for different slide directions

Ideally, a left-to-right slide has a decreasing trend because as the slide progresses, the left incisors slide first, followed by the middle, and then the right ones (observe that only the lower jaw is sliding while the upper jaw remains static). Hence, the delay is maximally (+)ve at the start and decreases to maximally (-)ve towards the end. Opposite trends are observed for right-to-left slide in Fig. 19. We note that not all people have very uniform teeth. Hence, local aberrations are observed, but on average the extrapolation of this trend generalizes well. With this end-to-end system design in place, we now move to the evaluation of *EarSense*.

6 IMPLEMENTATION

We have implemented *EarSense* on 5 different headphone/earphone models: Audio-Technica ATH-M30X, Sony MDRZX110, Bose solo3, Samsung EHS64, and RockPapa. The headphone was connected through the audio jack to 3 different sound cards: Realtek 892, Realtek 888, and Realtek 885. Based on preliminary analysis, we observed that the supra-aural headphones also capture the vibrations and can infer features like path delays, dispersion, and delay profiles. Fig. 20 shows an example of our simple experimental setup. The overall implementation has 3 main components:

(A) Re-purposing the Sound Card: Our goal is to convert speakers into a sensor for teeth vibrations and turn the sound card into a receiver of such signals. For this, we leverage the fact that most audio sound cards, like Realtek, provide an API to program audio ports for dual functionality [14, 34]. This API can be exploited to make output audio jack (socket) of the headphones/earphones function as an input.

The API provides access to audio pins that can be re-tasked. Figure 21 shows the architecture of the sound cards and the locations of the pins. Pins 14 and 15 (stereo L and R at PORT-E) are actually Analog I/Os and can be re-tasked as per user specifications. We reconfigure pins in software using a set of commands from the Audio kernel space. This provides an opportunity to dynamically re-configure the audio jacks to receive the vibration signals from the speaker.

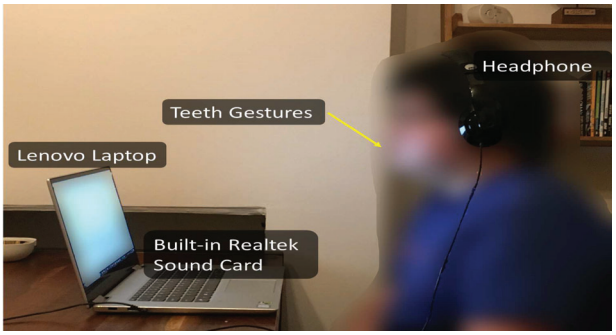

Figure 20: Setup: Headphone connected to laptop.

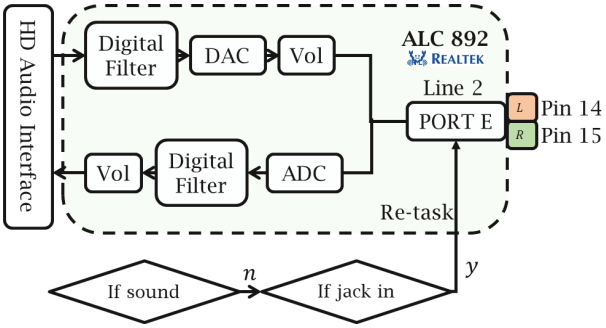

Figure 21: ALSA 892 configuration and re-tasking flow.

(B) Extracting and Processing the Signals: Our script first detects the presence of a device in the jack using an API function (`acpid`) [23]. Upon detection, it re-tasks the port using a suitable API (`hda_verb`) corresponding to the jack. If music or a call are not playing, our script ensures that headphones remain in the receiver mode, samples the signals, and exports to MATLAB. If an output stream is triggered, it toggles to transmit mode.

(C) Data Collection: We invited 18 volunteers to evaluate our system and asked them to perform 7 gestures: Taps (L, M and R) and slides (L to R, R to L, Up and Down) (more details in section 3). 14 participants were asked to perform each gesture 10 times. To cope with motion interference, 4 participants were deliberately asked to perform each of the 7 gestures while also performing other activities such as walking, cooking, cycling, walking, head-nodding, etc. The collected results were then fed to the *EarSense* algorithm.

7 RESULTS

7.1 Detecting Teeth Activity

The first step in *EarSense*'s pipeline is to detect teeth activity and distinguish it from other vibrations caused by walking or talking. Physical activity and speech, however, present as different signals levels at the speakers. Hence, we require separate thresholds to distinguish speech and walking from teeth activity. As a result, our system first distinguishes teeth activity from speaking and then from walking. For each case, we find the threshold that achieves the highest accuracy of detection. We then vary this threshold to test its stability.

Figure 22 shows the ROC curve of the teeth-or-not classifier. We maintain 94.0% true positives while false positives are still less than

10%. Figure 23 shows two bars representing the detection accuracy of teeth activity against speech and walking. The x-axis shows the normalized threshold where for each case, we normalize by the optimal threshold achieved. We vary the threshold by ±25% from the optimal and record the accuracy. The figure shows that the peak accuracy is 93.3% in the case of speech and 96.8% in the case of walking. If we choose the optimal threshold in both cases, we will achieve a final overall tooth activity identification accuracy of 90.2%.

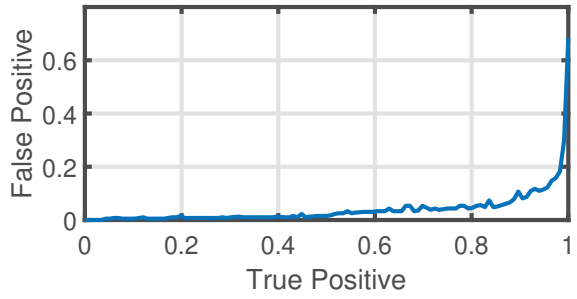

Figure 22: ROC curve for detecting teeth activity.

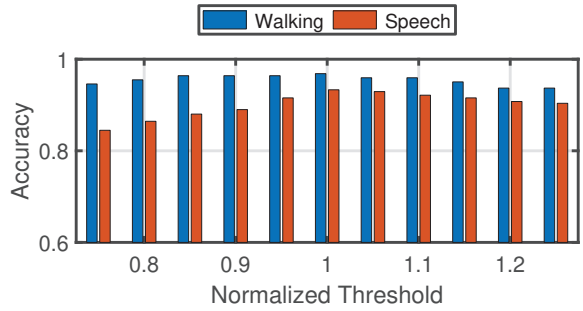

Figure 23: Sensitivity of the accuracy for distinguishing speech and walking signals from teeth activity.

7.2 Distinguishing Tap vs. Slide

Once teeth activity has been detected, the next step is to identify whether it is a "Tap" or a "Slide". Figure 24 shows the accuracy for each class, across different locations. The results are best for the middle tap, 97%, primarily because the intensity of vibration signals are stronger from this middle area. For slides, the accuracy of detecting left, right, upward or downward (U/D_{Slide}) slides are more than 90%. If we increase granularity, from 6 to 7 gestures, and look at the accuracy of detecting the upward and downward slides individually, the accuracy drops to around 73%. This is because it is hard for early users to slide upward or downward without tapping their teeth. However, we observed that as users gained more experience, they were less likely to tap their teeth while sliding, which improved accuracy.

7.3 Gesture Localization

After *EarSense* decides whether the teeth activity is a "Tap" or "Slide", it must localize the gesture. Recall that for "Tap" there are three locations: left, right and middle. For "Slide" there are four locations: up, down, left, right. For localizing the "Tap", we compare three techniques:

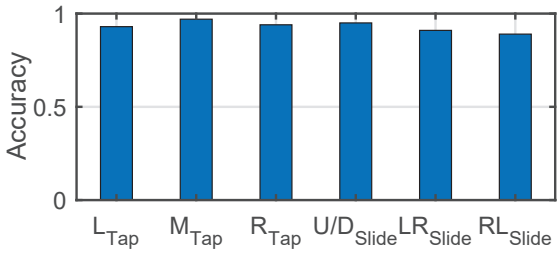

Figure 24: Accuracy of distinguishing Tap vs. Slide.

- **ZcD:** Zero Crossing Difference
- **TDoA:** Time Difference of Arrival
- ***EarSense*:** Joint Localization with "Cheek" Waves

Fig. 25 shows the accuracy of each of the three schemes for different locations. *EarSense* outperforms both **ZcD** and **TDoA**, which shows that using "Cheek" waves improves the accuracy of *EarSense*'s gesture localization. *EarSense*'s accuracy is above 93% and uniform across locations whereas **ZcD** and **TDoA** achieve higher accuracy for the middle tap location and their accuracy for left and right tap is below 80%.

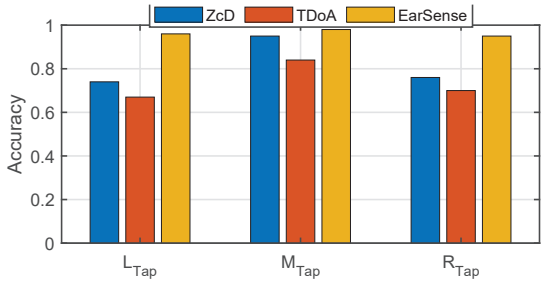

Figure 25: Tap localization using different methods.

Figures 26 and 27 show the accuracy of localizing the gestures against each gesture-location for each of the 14 users. The figures show that *EarSense* can accurately localize taps, right and left slides, and upward or downward slides with an accuracy larger than 90% and 80% and 85%, respectively. More importantly, the figures show that the accuracy is uniform across users. This shows *EarSense* is not over-fitting to some users and generalizes well without per-user training.

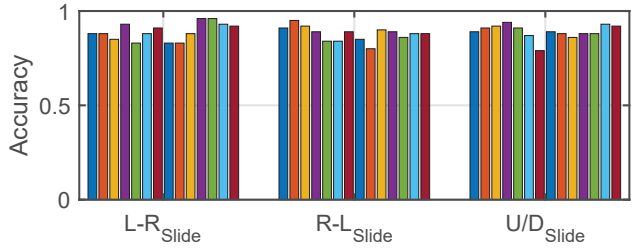

Figure 26: Slide localization accuracy across 14 users.

7.4 Overall Gesture Recognition

Fig. 28(a) shows the overall classification results across all 7 gestures and all users. Evidently, the accuracy exceeds 90% except when discriminating between up-down and down-up slides. The reduction in accuracy is partly a deficiency of our inference technique, but also stems from users adjusting their teeth (up/down)

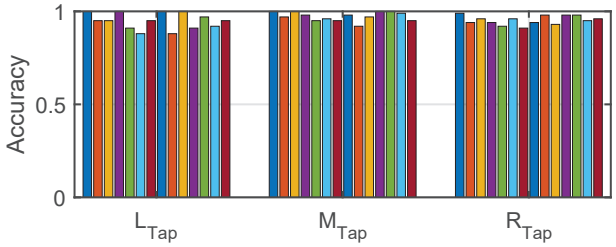

Figure 27: Tap localization accuracy across 14 users.

to perform the gestures correctly. In light of this, Fig. 28(b) shows the same overall results, but with 6 gestures now (i.e., combining up-down and down-up slides into a single U/D_{slide}). Naturally, results improve appreciably with a minimum of 88%. We believe this is an acceptable trade-off, i.e., sacrificing one gesture for a 17% improvement in worst-case classification.

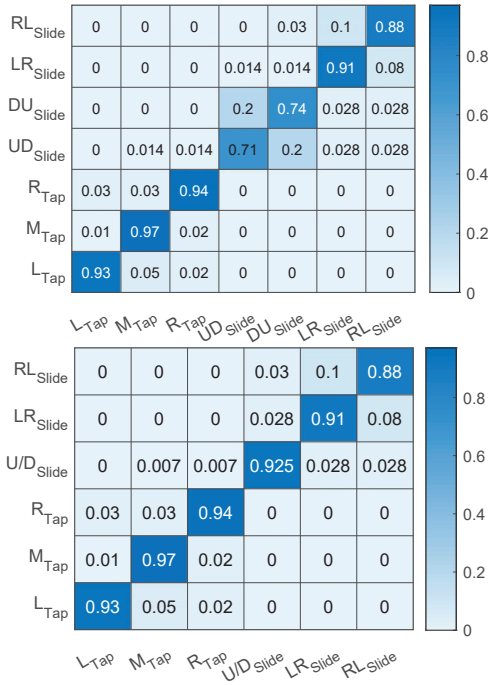

Figure 28: Confusion matrix across all users for (a) 7 gestures and (b) 6 gestures.

Finally, Fig. 29 shows the distribution of 6-gesture recognition accuracy, demonstrating the robustness of *EarSense*. Statistically, *EarSense* achieves a median accuracy of 0.98 and a 90th percentile accuracy of 0.895 across all users and gestures. This shows that *EarSense* is able to accurately and reliably recognize gestures.

7.5 Impact of Mobile Settings

We analyze whether other user-activities (e.g., walking, cycling, nodding) produce vibrations in some parts of the body that percolate into the earphone and interfere *EarSense*'s gesture recognition. Fig. 30 plots the accuracy for each of these activities, on a per-gesture basis. The performance is best for teeth taps, namely 85% for walking, and 90-93% for cooking, cycling, and nodding. This is

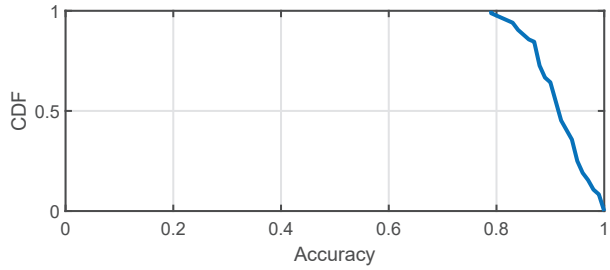

Figure 29: CDF of accuracy for all cases with 6 gestures.

because taps exhibit stronger SNR compared to slides. Walking suffers slightly because the vibration from the feet striking the ground pollutes the signal. Other activities produce weaker interference, but since teeth-slides produce weak SNRs, the benefits get offset.

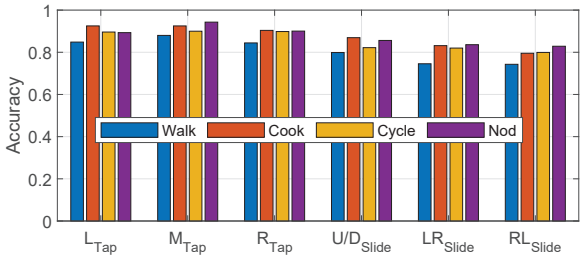

Figure 30: Gesture recognition accuracy against concurrent activities in mobile settings.

7.6 Tooth Brushing

We also evaluated the performance of *EarSense* for brushing teeth. Specifically, we localize the position of the tooth-brush to one of seven locations uniformly distributed on the jaw:

- R3/L3: Right/Left Molars (Extreme right/left)
- R2/L2: Right/Left Premolars
- R1/L1: Right/Left Canines
- M: Middle Incisors

Users were requested to use an electronic tooth brush to brush each position for 2 secs in random order. Unlike non-electric tooth brush, the hand does not oscillate rapidly while brushing. This negates the possibility of capturing unwanted interference. Also, the body acts as a filter to the vibrations of the toothbrush traveling through the skeleton, i.e., the arms-to-neck-to-ear channel has muscles and skin which largely attenuate the signal. Figure 31 shows the overall confusion matrix for 7 locations. The average localization accuracy is 89%. Most confusions happen with nearby locations. Only $< 3\%$ experiments show location error of two positions apart which is likely due to the different size and structure of the teeth across different humans.

8 LIMITATIONS AND DISCUSSION

We discuss a few open issues in our current version of *EarSense*.

Sensing while playing music: The current limitation with *EarSense* is that it cannot be used while music is being played in the earphone. This is because the speaker's diaphragm will anyway

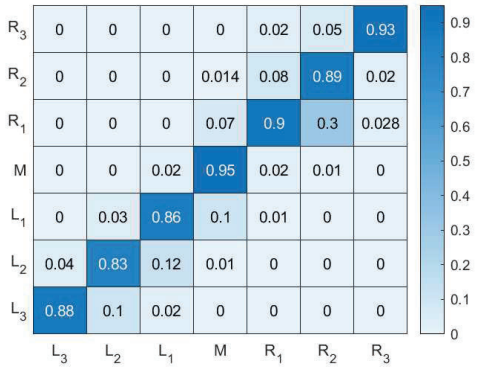

Figure 31: Confusion matrix of 7 tooth brush locations when the user is brushing with an electronic tooth-brush.

vibrate due to the music; the teeth vibrations would need to be isolated from the aggregate signal (a problem similar to RF full duplex cancellation). We leave this problem to future work, however, we note that startups like Nura [29] are solving similar problems to estimate the ear impulse response (so music can be customized to users). Moreover, it may be viable to make a head gesture to pause the music and switch the earphone to a receiver mode. Once control commands are given through *EarSense*, the music can resume.

Wireless earphones: Jack re-tasking was possible with wired earphones, however, may not be easy with wireless earbuds. We hope our research motivates earable-designers to provide such capabilities as APIs in upcoming releases. Companies are already exploring such ideas (e.g., Braggi's Dash can accept gestures such as a tap on cheeks [4]). We believe that a firmware update can enable earphone apps to toggle between modes, allowing for *EarSense*-like systems.

Leveraging In-ear microphones: We used the earphone's speaker (instead of the microphone) to sense vibrations. This is because the microphones typically face outwards to listen to the user's voice or for noise cancellation. New earbuds are emerging with *in-ear microphones* [33], i.e., microphones at the inner end of the earbud, designed for ear-sensing and better noise cancellation. The data from in-ear microphones are still not available (e.g., Samsung's APK is only for internal use). As such APKs become public, the microphone can be immediately useful for *EarSense*-like systems.

Longitudinal studies across larger populations: We understand that teeth compositions change over time; that can affect the gesture recognition quality. Also, 18 users may not be adequate to capture all variations across the human population. Clearly, a longer-term user-focused commitment is necessary to push *EarSense* towards deployment.

9 RELATED WORK

Related work around teeth and oral sensing can be classified under 2 umbrellas: (a) invasive and (b) non-invasive.

Invasive: Multiple projects have been studying the design and usability of invasive systems for hand-free gestures recognition. TongueBoard [22] from Google discusses the design of an invasive oral interface for recognizing non-vocalized speech. It uses the SmartPalate system consisting of an array of 124 capacitive touch sensors embedded in an oral mouthpiece, and is able to classify between 15 words/2 gestures. In a different work, a clench interface has been designed using an in-mouth pressure sensor [40] to facilitate biting based input interfaces. Similarly, [9] discusses the design of an intra-oral input interface. Clearly, these devices are critical for specific medical applications where carrying/implanting these devices are well justified. With *EarSense*, the motivation is to piggyback oral sensing on regular ear devices, and thereby being amenable to the masses.

Non-Invasive: TYTH [28] discusses a creative tongue-teeth localization approach using dedicated hardware. It senses changes in spatial features of neuro-muscular signals, EEG, ECG, and skin surface deformation (SKD) as the tongue moves around the mouth. SelfSync [19] exploits the presence of two or more smart devices for joint quantification of gestures. The gesture set is limited to 3; moreover, a user may not carry multiple devices. [3] is the closest to our work but differs by requiring a dedicated bone-conduction microphone; the gestures are restricted only to "clicks"; and relies on per-user training. [2] uses barometers inside earphones to capture changes in air-pressure inside the ear due to different facial gestures. [26] designs a dedicated circuit to adopt intercom like feature with COTS earphones.

In another line of work, [15] modifies a toothbrush by attaching a magnet to the handle. The orientation and motion of the toothbrush is captured by the magnetic sensor in the wristwatch, which aids in recognizing tooth-brushing gestures. Similarly, wrist-worn inertial sensors have been used to monitor brushing patterns [24]. Clearly, *EarSense* makes an attempt to achieve better gesture dictionaries with COTS (wired) earphones.

10 CONCLUSION

We demonstrate the feasibility of sensing teeth-gestures using off-the-shelf earphones. The intuition is that teeth taps and slides produce surface vibrations in the jaws, that ultimately reach the ears and create oscillations in the earphone speaker's diaphragm. By recording these vibrations in the sound card, and processing them via lightweight techniques, we are able to detect teeth-gestures. We reliably localize $6-7$ gestures under the assumption that the earphone is not playing any sound while the gestures are being performed. The natural next step is to extract such vibrations even when music is being played through the earphones – a topic of our ongoing research.

11 ACKNOWLEDGMENTS

We sincerely thank the anonymous shepherd and reviewers for their insightful comments and suggestions. We are also grateful to NSF (award numbers: 1918531, 1910933, 1909568, and 1719337), NIH (award number: 1R34DA050262-01), Google, and Samsung for partially funding this research.

REFERENCES

[1] 2020. About Switch Access for Android. Retrieved July, 2020 from https://support.google.com/accessibility/android/answer/6122836

[2] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. CanalSense: Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on*

User Interface Software and Technology (UIST '17). ACM, New York, NY, USA, 679–689. https://doi.org/10.1145/3126594.3126649

[3] Daniel Ashbrook, Carlos Tejada, Dhwanit Mehta, Anthony Jiminez, Goudam Muralitharam, Sangeeta Gajendra, and Ross Tallents. 2016. Bitey: An exploration of tooth click gestures for hands-free user interface control. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 158–169.

[4] Bragi. 2019. Headphones of tomorrow will have apps. Retrieved June 27, 2019 from https://cdn.shopify.com/s/files/1/0078/8955/1475/files/20190402_Bragi_Smart_Headphones.pdf

[5] Amélie Catala, Marine Grandgeorge, Jean-Luc Schaff, Hugo Cousillas, Martine Hausberger, and Jennifer Cattet. 2019. Dogs demonstrate the existence of an epileptic seizure odour in humans. *Scientific reports* 9, 1 (2019), 1–7.

[6] Jingyuan Cheng, Ayano Okoso, Kai Kunze, Niels Henze, Albrecht Schmidt, Paul Lukowicz, and Koichi Kise. 2014. On the tip of my tongue: a non-invasive pressure-based tongue interface. In *Proceedings of the 5th Augmented Human International Conference*. ACM, 12.

[7] Michael S Duchowny, Trevor J Resnick, Marcel J Deray, and Luis A Alvarez. 1988. Video EEG diagnosis of repetitive behavior in early childhood and its relationship to seizures. *Pediatric neurology* 4, 3 (1988), 162–164.

[8] David Fedak and Charles Baldwin. 2005. A comparison of enameled and stainless steel surfaces. In *Ceramic engineering and science proceedings*, Vol. 26. Wiley Online Library, 45–54.

[9] Pablo Gallego Cascón, Denys J.C. Matthies, Sachith Muthukumarana, and Suranga Nanayakkara. 2019. ChewIt. An Intraoral Interface for Discreet Interactions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 326, 13 pages. https://doi.org/10.1145/3290605.3300556

[10] L Gaul and S Hurlebaus. 1998. Identification of the impact location on a plate using wavelets. *Mechanical Systems and Signal Processing* 12, 6 (1998), 783–795.

[11] Sleiman R Ghorayeb and Teresa Valle. 2002. Experimental evaluation of human teeth using noninvasive ultrasound: echodentography. *ieee transactions on ultrasonics, ferroelectrics, and frequency control* 49, 10 (2002), 1437–1443.

[12] Mayank Goel, Chen Zhao, Ruth Vinisha, and Shwetak N. Patel. 2015. Tongue-in-Cheek: Using Wireless Signals to Enable Non-Intrusive and Flexible Facial Gestures Detection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 255–258. https://doi.org/10.1145/2702123.2702591

[13] Google. 2020. What is Switch Access on Android? Retrieved July, 2020 from https://youtu.be/rAIXE6ilRQ0

[14] David henningsson. 2018. Turn your mic jack into a headphone jack. Retrieved Aug 1, 2018 from http://voices.canonical.com/david.henningsson/2011/11/29/turn-your-mic-jack-into-a-headphone-jack/

[15] Hua Huang and Shan Lin. 2016. Toothbrushing monitoring using wrist watch. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*. ACM, 202–215.

[16] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[17] Hyosu Kim, Anish Byanjankar, Yunxin Liu, Yuanchao Shu, and Insik Shin. 2018. UbiTap: Leveraging Acoustic Dispersion for Ubiquitous Touch Interface on Solid Surfaces. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 211–223.

[18] Kewal Krishan, Tanuj Kanchan, and Arun K Garg. 2015. Dental evidence in forensic identification–An overview, methodology and present status. *The open dentistry journal* 9 (2015), 250.

[19] Juyoung Lee, Shaurye Aggarwal, Jason Wu, Thad Starner, and Woontack Woo. 2019. SelfSync: Exploring Self-synchronous Body-based Hotword Gestures for Initiating Interaction. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC '19)*. ACM, New York, NY, USA, 123–128. https://doi.org/10.1145/3341163.3347745

[20] Siyoung Lee, Junsoo Kim, Inyeol Yun, Geun Yeol Bae, Daegun Kim, Sangsik Park, Il-Min Yi, Wonkyu Moon, Yoonyoung Chung, and Kilwon Cho. 2019. An ultrathin conformable vibration-responsive electronic skin for quantitative vocal recognition. *Nature communications* 10, 1 (2019), 1–11.

[21] Richard Li and Gabriel Reyes. 2018. Buccal: Low-cost Cheek Sensing for Inferring Continuous Jaw Motion in Mobile Virtual Reality. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers (ISWC '18)*. ACM, New York, NY, USA, 180–183. https://doi.org/10.1145/3267242.3267265

[22] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In *Proceedings of the 10th Augmented Human International Conference 2019 (AH2019)*. ACM, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831

[23] Arch Linux. 2019. acpid. Retrieved Dec 11, 2019 from https://wiki.archlinux.org/index.php/Acpid

[24] C. Luo, X. Feng, J. Chen, J. Li, W. Xu, W. Li, L. Zhang, Z. Tari, and A. Y. Zomaya. 2019. Brush like a Dentist: Accurate Monitoring of Toothbrushing via Wrist-Worn Gesture Sensing. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*. 1234–1242. https://doi.org/10.1109/INFOCOM.2019.8737513

[25] Balz Maag, Zimu Zhou, Olga Saukh, and Lothar Thiele. 2017. BARTON: Low power tongue movement sensing with in-ear barometers. In *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 9–16.

[26] Hiroyuki Manabe and Masaaki Fukumoto. 2012. Headphone taps: a simple technique to add input function to regular headphones. In *MobileHCI '12*.

[27] Specops Software Marcus Kaber. 2018. Global survey reveals low adoption of multi-factor authentication for Office 365. Retrieved July, 2020 from https://venturebeat.com/2018/08/22/global-survey-reveals-low-adoption-of-multi-factor-authentication-for-office-365/

[28] Phuc Nguyen, Nam Bui, Anh Nguyen, Hoang Truong, Abhijit Suresh, Matt Whitlock, Duy Pham, Thang Dinh, and Tam Vu. 2018. Tyth-typing on your teeth: Tongue-teeth localization for human-computer interface. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 269–282.

[29] Nura. 2019. Nuraphone: How it works. Retrieved June 27, 2019 from https://www.nuraphone.com/pages/how-it-works

[30] Thuy Ong. 2018. Over 90 percent of Gmail users still donâĂŹt use two-factor authentication. Retrieved July, 2020 from https://www.theverge.com/2018/1/23/16922500/gmail-users-two-factor-authentication-google

[31] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. 2019. STEAR: Robust Step Counting from Earables. In *Proceedings of the 1st International Workshop on Earable Computing*. 36–41.

[32] Jin-Peng Qi, Qing Zhang, Ying Zhu, and Jie Qi. 2014. A novel method for fast change-point detection on simulated time series and electrocardiogram data. *PloS one* 9, 4 (2014), e93365.

[33] Samsung. 2019. In Ear. Retrieved Dec 12, 2019 from https://www.samsung.com/us/audio/headphones/in-ear/

[34] Hardware Secrets. 2018. Realtek 892. Retrieved Dec 11, 2019 from https://www.hardwaresecrets.com/datasheets/ALC892-CG_DataSheet_1.3.pdf

[35] M. Staines, W. H. Robinson, and J. A. A. Hood. 1981. Spherical indentation of tooth enamel. *Journal of Materials Science* 16, 9 (01 Sep 1981), 2551–2556. https://doi.org/10.1007/BF01113595

[36] Gilbert Strang and Truong Nguyen. 1996. *Wavelets and filter banks*. SIAM.

[37] Amir Sulaiman, Kirill Poletkin, and Andy WH Khong. 2010. Source localization in the presence of dispersion for next generation touch interface. In *2010 International Conference on Cyberworlds*. IEEE, 82–86.

[38] Satoru Tsuge, Takashi Osanai, Hisanori Makinae, Toshiaki Kamada, Minoru Fukumi, and Shingo Kuroiwa. 2008. Combination method of bone-conduction speech and air-conduction speech for speaker recognition. In *Ninth Annual Conference of the International Speech Communication Association*.

[39] Abhisek Ukil and Rastko Živanović. 2008. Adjusted Haar wavelet for application in the power systems disturbance analysis. *Digital Signal Processing* 18, 2 (2008), 103–115.

[40] Xuhai Xu, Chun Yu, Anind K. Dey, and Jennifer Mankoff. 2019. Clench Interface: Novel Biting Input Techniques. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 275, 12 pages. https://doi.org/10.1145/3290605.3300505