3-2022

Innovative human motion sensing with earbuds

Dong MA
Singapore Management University, dongma@smu.edu.sg

Andrea FERLINI

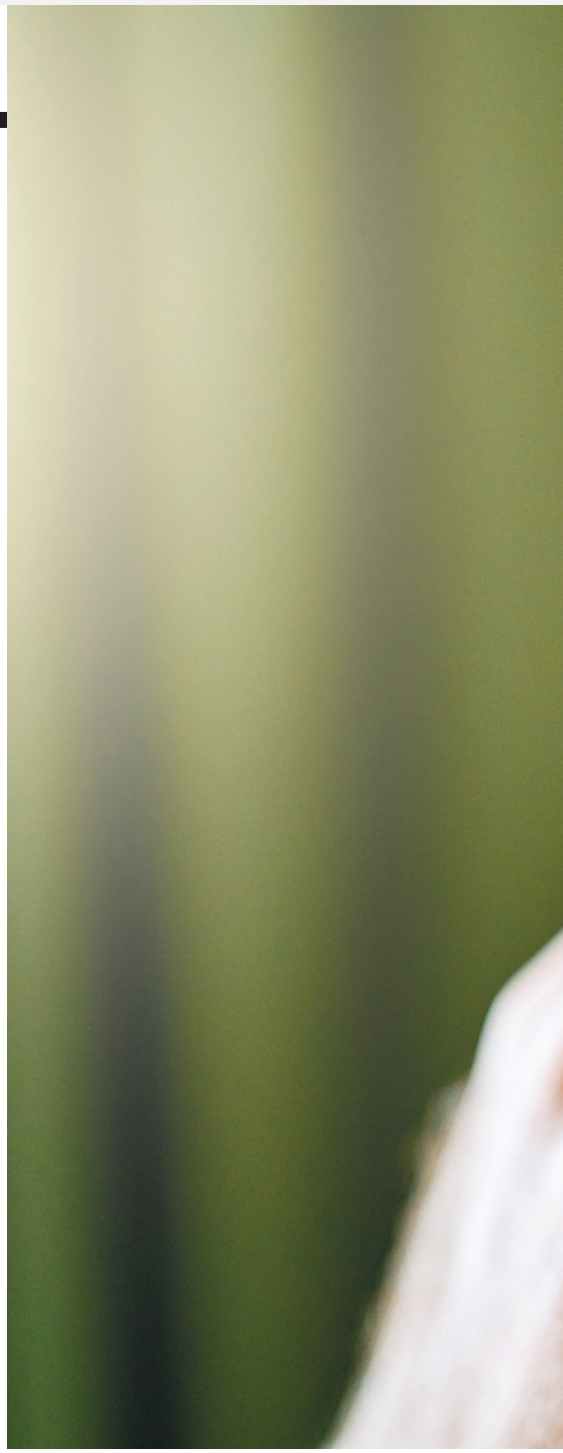
Cecilia MASCOLO

Citation

MA, Dong; FERLINI, Andrea; and MASCOLO, Cecilia. Innovative human motion sensing with earbuds. (2022). *ACM GetMobile: Mobile Computing and Communications.*. 25, (4), 24.
Available at: https://ink.library.smu.edu.sg/sis_research/7277

Dong Ma *Singapore Management University, Singapore*
Andrea Ferlini *and* **Cecilia Mascolo** *University of Cambridge, Cambridge, UK*

Editors: Nicholas D. Lane and Xia Zhou



INNOVATIVE HUMAN MOTION SENSING WITH EARBUDS

Excerpted from "OESense: employing occlusion effect for in-ear human sensing" from *MobiSys '21: Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services* with permission. https://dl.acm.org/doi/10.1145/3458864.3467680 ©ACM 2019

Earbuds, ear-worn wearables, have attracted growing attention from both industry and academia. This trend has witnessed manufacturers embedding multiple sensors on earbuds to enrich their functionalities. For example, Apple AirPods, Sony WF-1000XM3, and Bose QuietControl 30, have been equipped with accelerometers for tapping interaction or multiple microphones for noise cancellation. On the other hand, the research community regards earbuds as a powerful personal-scale human sensing and computing platform. By integrating sensors like PPG, barometer, and ultrasonic sensors, researchers have been devising a plethora of earable sensing applications, such as blood pressure monitoring [1], facial expression recognition [2], and authentication [3].

Compared to traditional wearables, earbuds possess two advantages for human sensing. First, the human ear is an ideal position to capture various neurological, cardiovascular, and dietary signs, which promise great sensing potential for health monitoring. Second, earbuds are worn in the upper part of the body, which not only complements the sensing scope of smartphones/smartwatches, but also is more robust to intensive body artifacts (e.g., hand swing) during motion detection [4].

Historically, researchers used Inertial Measurement Units (IMU), accelerometers in particular, to sense motion. Some examples are human activity recognition, eating habits monitoring, smoking gesture recognition, and gait analysis. However, as the human head has a high degree of freedom to move/rotate and it does not always move accordingly with the rest of the body, accelerometers on the earbud are affected by head movements. With collected data, we observe that (1) when there is no head movement, accelerometer can detect most intense (walk and run) and light (chew) activities, but fails to capture extremely weak signals like drink-induced vibrations; (2) head movements have minor impact on intense activities (walk

Photo, istockphoto.com

and run), while completely obfuscating the accelerometer readings of light activities as the magnitude of head movement is larger.

Besides accelerometers, microphones (external facing) have also been adopted to detect motion events (e.g., gestures recognition [5]. However, microphones achieve poor performance on motion detection. To validate this, we record the microphone data from an earbud when a subject performs different activities mentioned above. The collected data indicates that (1) compared to accelerometer, external-facing microphone shows less potential for motion detection (only run can be reliably detected). The reason is that the external microphone measures the air-conducted sound, which suffers from strong attenuation, therefore only motions producing relatively high volume can be detected; (2) in the presence of environmental sounds (e.g., music), the sensing signals are completely buried in the background music. Given that motion-induced sounds and background music are both audible and share most of the spectrum, it would be very challenging to filter such interference with signal processing techniques from the frequency domain.

We explore other alternatives for human motion sensing on earbuds. To achieve

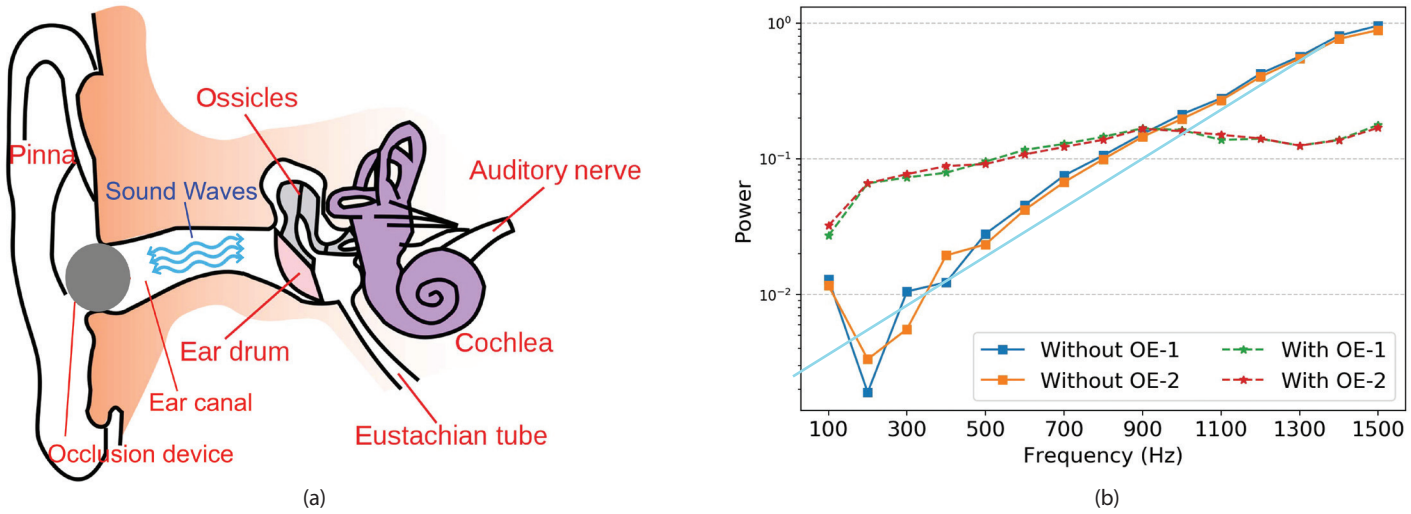


FIGURE 1. (a) Illustration of human ear anatomy and the occlusion effect. (b) Impact of occlusion effect (OE) on the frequency response. (1 and 2 denote two measurements, which suggests the response is highly consistent.)

reliable detection of both intense and light human motions, we present *OESense*, a novel acoustic-based in-ear human motion sensing system. OESense performs robust motion sensing based on two critical design choices. First, to tackle environmental noise, OESense leverages an inward-facing microphone to record motion-induced sounds inside the ear canal. As a result, most of the environmental noise is naturally suppressed. Further, acoustic signals are inherently immune to motion artifacts, like head movements. Second, to cope with the low SNR of traditional acoustic approaches, OESense exploits a phenomenon known as occlusion effect to enable the detection of both intense and light motions in the human ear canal. Concretely, when a motion stimulus is applied to the human body, the occlusion effect boosts low-frequency bone-conducted sounds (most human motions are in a few Hertz range) when the ear canal orifice is occluded.

We prototyped OESense with a pair of wired earbuds and a Raspberry Pi. Three applications have been selected as instances of intense, mixed, and light motion detection tasks: step counting, human activity recognition, and face-tapping gesture interaction. We evaluated our claims with 31 subjects, demonstrating the superior sensing performance of OESense over traditional motion sensing approaches. Our results show OESense obtains robust performance on the three applications under various scenarios. Moreover, we demonstrated that OESense is compatible with the original functionalities of the earbuds, such as playing music and picking up phone calls. More details are available in [11].

THE OCCLUSION EFFECT

When vibratory stimuli are applied on the human body, the generated sound will propagate to other parts of the body through bone conduction. Ordinarily, bone-conducted sounds induce the vibration of the ear canal wall, and the generated sounds will escape through the opening of the ear canal. However, when the ear canal is blocked, sounds are trapped and reflected back to the eardrum [6], as shown in Figure 1(a). Such occlusion increases the acoustic impedance of the ear canal opening at low frequencies, repurposing the ear canal as a low-pass filter [7]. Therefore, the low-frequency components of a bone-conducted sound will be enhanced in an occluded ear canal, defined as the occlusion effect [8]. A common instance of this is that people perceive echo-like sounds of their own voice when an object (like a finger) fills the outer portion of the ear canal.

Quantitatively, the occlusion effect can be denoted as the ratio between the sound pressure in the occluded ear canal and that in the open ear [9]. As measured in [10], it can boost the sound below 1000~Hz by up to 40~dB depending on the frequency. We also measure the impact of occlusion effect on the ear canal frequency response. We use the earbud speaker to transmit a single tone between 100-1500Hz (100Hz separation) and record the reflected sound with an inward-facing microphone. Figure 1(b) compares the frequency response with and without occlusion effect (completely blocking the ear canal opening). We can see that the blocked ear canal produces stronger response at frequencies below 900Hz, while the open ear canal gains much higher response at higher frequencies. In addition, we repeat the measurements twice (remove the earbud and wear it again) and the response is highly consistent, demonstrating the robust and reliable presence of the occlusion effect.

OESense SYSTEM

Leveraging the occlusion effect for human motion sensing promises three advantages. First, due to the occlusion of ear canal orifice, the inward-facing microphone mainly captures the bone-conducted sound in the ear canal and is less susceptible to environmental noises like traffic sounds and human speech. Second, given that most human-produced motions are in relatively low frequencies (a few Hertz), the amplification gain provided by the occlusion effect can improve the SNR of the sensing signal. Third, although earbuds are mainly used for delivery of sounds (e.g., music or phone calls) to the human ear, these sounds are usually in higher frequencies so sound delivery and human sensing (under 50Hz) can coexist without mutual interference.

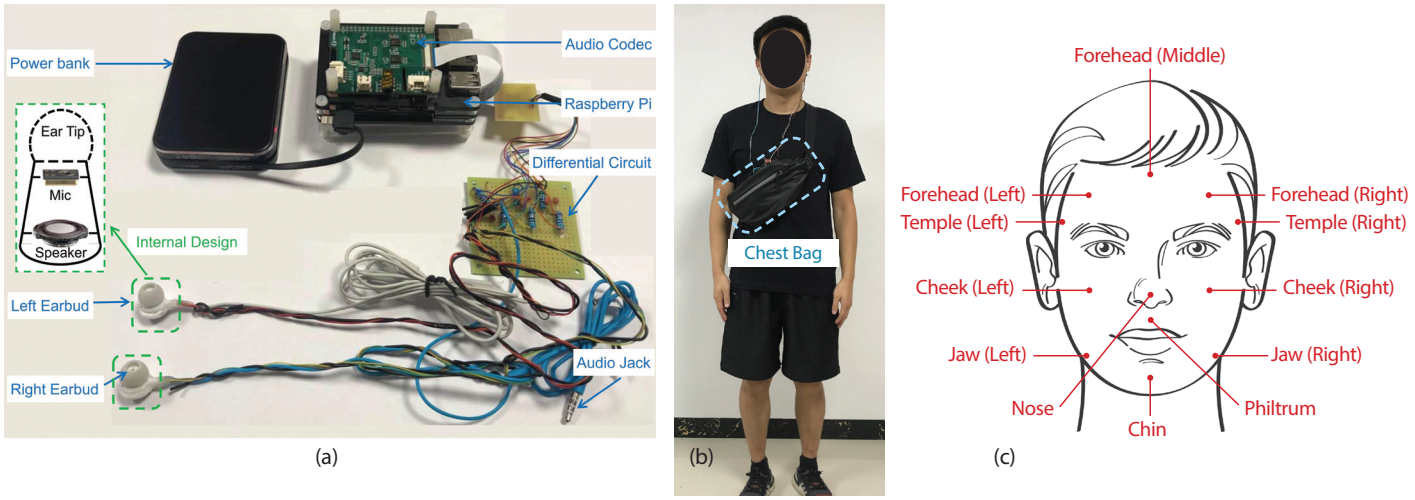


FIGURE 2. (a) The developed data recording prototype. (b) Illustration of a participant wearing the device. (c) Illustration of the designed tapping gestures.

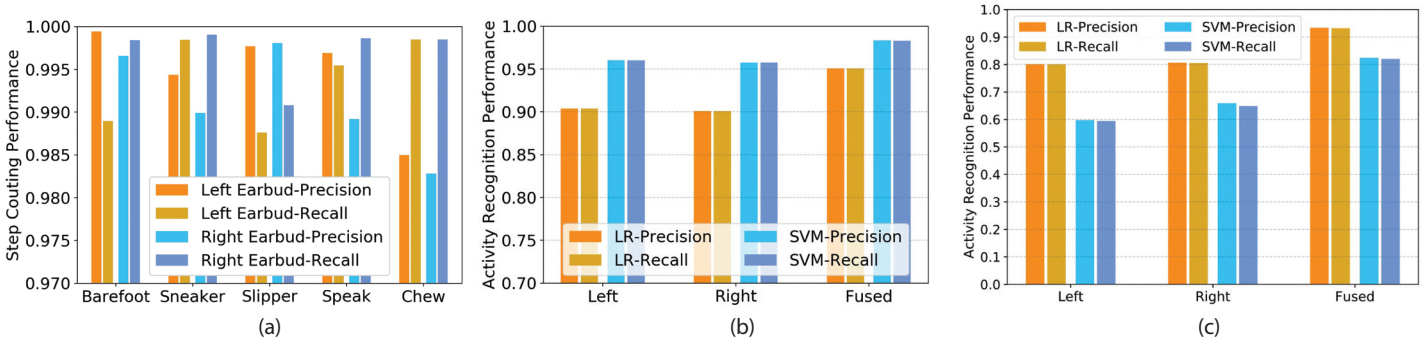


FIGURE 3. The sensing performance of (a) step counting, (b) activity recognition, and (c) tapping gesture recognition.

Applications

In this work, we explore the performance of OESense on three typical applications:

- Step counting (intense): Human walking involves big movements of the whole body and can be detected at different body positions (like foot, waist, and head).
- Human activity recognition (mixed): We select five activities including walking, running, being still, chewing, and drinking, which combines both intense body motions and weak surface vibrations.
- Face-tapping gesture interaction (light): Vibrations generated by tapping different parts of the human face propagate to the ear via different paths. The received signals present distinctive patterns, enabling the recognition of different tapping gestures. This could be a potential way to interact with earables in the future.

Sensing Pipeline

OESense utilizes customized signal processing techniques and machine learning for different applications. Collectively, the in-ear microphone signal is first processed with a low-pass filter (with 50Hz cut-off frequency) to eliminate environmental and human sounds.

For step counting, we then apply the Hilbert transform on the filtered signal to extract its upper envelope and lower envelope. Afterward, a low-pass filter (<5Hz) is performed on the two envelopes separately to smooth them. We apply peak detection on the smoothed envelopes, which outputs the time index and amplitude of each peak. To avoid over-counting (i.e., false positives), we further propose two strategies to filter the detected peaks: (1) the minimum peak interval between adjacent peaks is set to 0.3s as normal human walking frequency is lower than 3.3Hz, (2) the minimum peak amplitude is set to 0.3 times of average amplitude of all detected peaks. Any peak that fails to satisfy either one of the conditions will be omitted. Lastly, to combat the sporadic noise that only produces an upper peak or a lower peak, we count a step only when a pair of upper peak and lower peak is aligned, i.e., the time lag between them is shorter than 0.2s.

For activity recognition, the recorded audio stream is divided into small segments using the sliding window technique (with 50% overlapping ratio). For gesture recognition, we utilized the same approach of detecting steps to extract the gesture signal. Afterward, for each instance, we extract various frequency-based, structural, and statistical audio features, such as Mel-frequency cepstral coefficients (MFCC), chroma of short-time Fourier transform

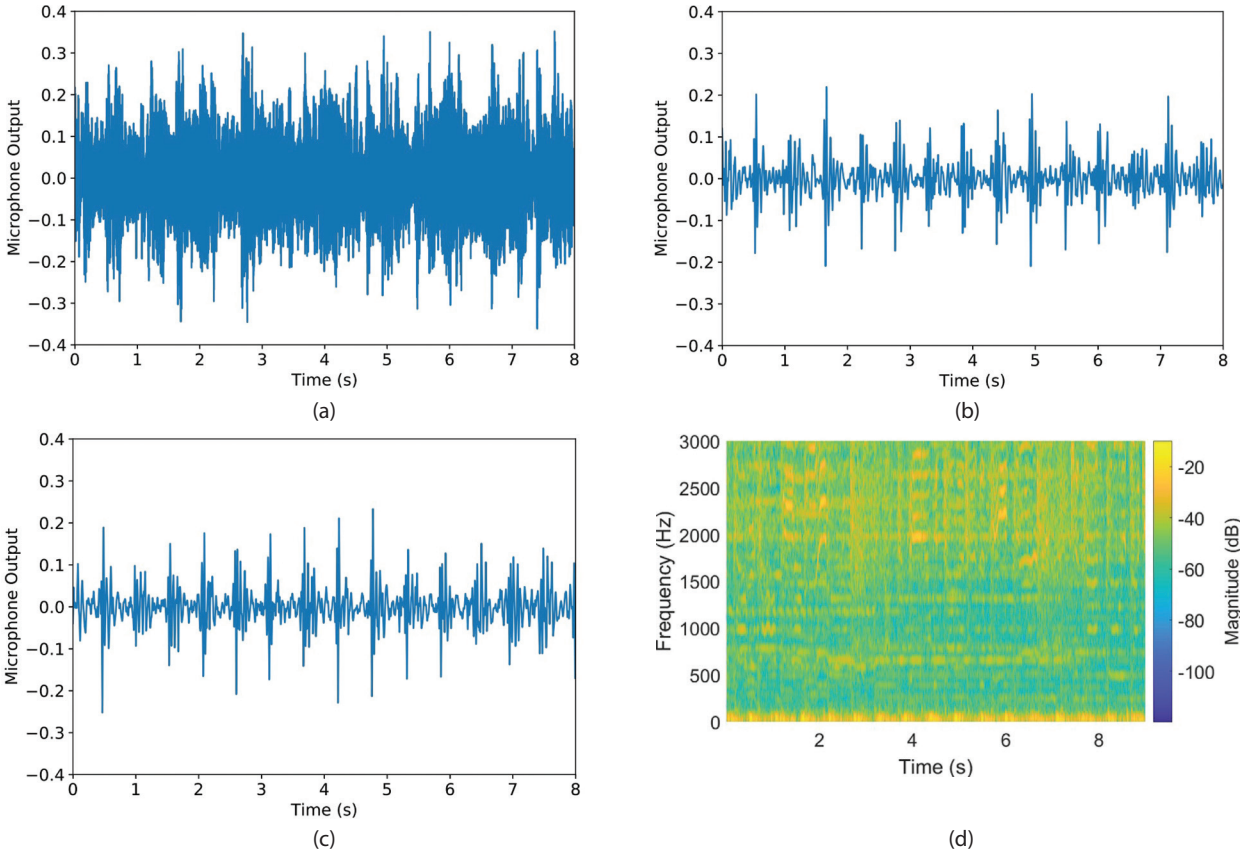


FIGURE 4. The (a) original signal and (b) low-pass filtered signal for participant walking during music playback, (c) low-pass filtered signal for the same participant walking without music playback, (d) spectrogram of the original signal.

(STFT), contrast of STFT, tonnetz, and root mean square error. Finally, the extracted features are fed into typical machine learning classifiers, such as Logistic Regression (LR) or Support Vector Machine (SVM), to identify different activities or gestures.

Hardware Prototyping

We prototyped OESense by adding an inward-facing microphone (SPU1410LR5H-QB) to a commercial earbud (MINISO Marvel earphones). As shown in Figure 2(a) (enclosed within a green dash box), we embedded the microphone at the front end of the earbud and moved the original speaker to the back end. Such design would optimize the SNR of the microphone. Then, we developed a data-logger for microphone data acquisition. To minimize noise, each microphone is connected to a differential circuit before sampled by an audio codec. We use ReSpeaker Voice Accessory HAT as the audio codec, which is controlled by a Python program running on a Raspberry Pi 4B. We sample the microphone data at 48~kHz. To avoid affecting the subjects' walking style, all the components are enclosed in a chest bag worn by them, as shown in Figure 2(b). 31 participants (16 males, 15 females, with an age of 26.6±5.8) were recruited to collect the above-mentioned activity data under various conditions. Particularly, for face-tapping gesture recognition, we selected 12 positions, as shown in Fig 2(c) on the human face as the interaction spots. Accordingly, twelve gestures are created by finger tapping (one-time) on each position. (Ethical approval for carrying out all the studies has been granted by the corresponding institution.)

EVALUATION

Sensing Performance

We applied the designed signal processing and machine learning on the collected dataset to evaluate the sensing performance of OESense. For step counting, we can observe from Fig 2(a) that step counting precision and recall are higher than 97.5% regardless of walking scenarios, demonstrating the superior performance of OESense on step counting. For human activity recognition, Fig 2(b) indicates that (1) SVM always achieves better performance than LR; (2) Left and right earbuds achieve similar performance; (3) The fused dataset obtains the highest recognition precision and recall both at around 98.3%. Such improvement might arise from the fact that the fused dataset gains from two sensing channels and is more resilient to signal distortions when one of the ear tips is loose. For tapping gesture recognition, we can see from Fig 2(c) that LR consistently achieves better performance, which might be because the features from different gestures are likely to be separated linearly. As expected, the fused dataset (93.2% recall) outperforms the two individual datasets (80.1% and 80.5% recall for left and right, respectively) on the 12 gestures, demonstrating the benefits of sensing with both earbuds. For more thorough performance analysis, please refer to [11].

Robustness

Given that the original functionality of earbuds is to deliver sounds (e.g., music and phone calls) to the human ear, a common question is whether these sounds (usually much higher volume) pollute the audio sensing signals. To investigate this, we asked one participant to walk while the earbuds play a song with the built-in speakers at an appropriate volume. Figure 4(a) illustrates the original signal collected from the left earbud and we can see it is dominated by the music. Figure 4(b) shows the low-pass filtered (<50~Hz) version of the signal, where the steps can be clearly observed. Then, without music playing, the subject walked another trace in the same condition. The signal after low-pass filtering is plotted in Figure 4(c). Visually, we can see that the two filtered versions have high similarity, and the step counts can be easily derived. We also quantify the similarity of signals from frequency domain using the structural similarity (SSIM, a well-known metric to compare similarity between two images [12]). Specifically, we first obtain the spectrogram of each signal using a short-time Fourier transform (STFT), and then calculating the SSIM index between two spectrograms (images). Our results show that the SSIM index between Figure 4(b) and Figure 4(c) is 0.95, suggesting that music playback has extremely limited impacts. To further explain how OESense combats human speech and music, we plot the spectrogram of Figure 4(a) in Figure 4(d). It is clear that music mainly resides in higher frequencies, while step-induced sounds are located in extremely low frequencies with a strong amplitude.

In terms of phone calls, the frequency range of human voice over telephony transmission is within 300-3400~Hz [13], so that it can be completely removed after the low-pass filtering. For low frequency noise in the environment (e.g., fan motion), OESense leverages the sealing of the ear canal, which serves as an additional layer of filter to further suppress the noise, making the internal microphone less vulnerable to external sounds.

LIMITATIONS AND FUTURE WORK

Leveraging the occlusion effect, the presented OESense system shows great sensing potential for both intense and light human activities. However, there are also several limitations. First, physical occlusion of the ear canal might lead to impaired awareness of the surrounding environment (e.g., traffic sounds) and incur safety issues. A possible solution is to imitate the transparency mode on AirPods Pro. Specifically, the external microphone can measure the outside sounds and replay the meaningful parts (like sirens) through the onboard speakers. Second, OESense was implemented with a Raspberry Pi, which is energy-expensive and cumbersome for mobile scenarios. Thus, further efforts (advanced audio chips, dedicated PCB development and wireless design) to implement OESense in an energy-efficient manner are required. Third, while we have demonstrated that OESense can detect the three sensing applications separately, whether it is able to run these concurrently remains unclear. ■

Acknowledgement

This work is supported by ERC through Project 833296 (EAR) and by Nokia Bell Labs through a donation. We thank D. Spathis for the insightful discussions, the anonymous shepherd and reviewers for the valuable comments, and the volunteers for the data collection.

Dong Ma is an assistant professor at the School of Computing and Information Systems, Singapore Management University. He earned his Ph.D. from the University of New South Wales, Australia. His research interests rotate around cyber-physical systems, including ubiquitous computing, pervasive sensing, vibration communication, energy harvesting and mobile healthcare.

Andrea Ferlini is a third-year Ph.D. student in Mobile Systems at the University of Cambridge. His main research interest is on earables sensing. After investigating the potential of earables for head tracking, magnetic sensing, in-ear-microphone-based HCI and authentication. He is focusing currently on in-ear physiological sensing.

Cecilia Mascolo is a professor of Mobile Systems in the Department of Computer Science and Technology, University of Cambridge, UK. She is the director of the Centre for Mobile, Wearable System and Augmented Intelligence. Her research interests are in mobile systems and data for health, sensor systems and networking and mobile data analysis.

REFERENCES

[1] Nam Bui, Nhat Pham, et al. 2019. eBP: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. *25th Annual International Conference on Mobile Computing and Networking*, 1–17.

[2] Toshiyuki Ando, Yuki Kubo, et al. 2017. CanalSense: Face-related movement recognition system based on sensing air pressure in ear canals. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 679–689.

[3] Yang Gao, Wei Wang, et al. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–24.

[4] Jay Prakash, Zhijian Yang, et al. 2019. STEAR: Robust Step Counting from Earables. *Proceedings of the 1st international workshop on earable computing*, 36–41.

[5] Xuhai Xu, Haitian Shi, et al. 2020. EarBuddy: Enabling on-face interaction via wireless earbuds. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

[6] Stefan Stenfelt. 2011. Acoustic and physiologic aspects of bone conduction hearing. *Implantable bone conduction hearing aids*, volume 71, 10–21. Karger Publishers.

[7] Roman Schlieper, Song Li, et al. 2019. The relationship between the acoustic impedance of headphones and the occlusion effect. *Audio Engineering Society Conference: 2019 AES International Conference on Headphone Technology.* Audio Engineering Society.

[8] Michael A. Stone, Anna M. Paul, et al. 2014. A technique for estimating the occlusion effect for frequencies below 125 hz. *Ear and hearing*, 35(1):49.

[9] Stefan Stenfelt and Sabine Reinfeldt. 2007. A model of the occlusion effect with bone-conducted stimulation. *International journal of audiology*, 46(10):595–608.

[10] Kévin Carillo, Olivier Doutres, et al. 2020. Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation. *The Journal of the Acoustical Society of America*, 147(5):3476–3489.

[11] Dong Ma, Andrea Ferlini, et al. 2021. OESense: employing occlusion effect for in-ear human sensing. *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 175-187.

[12] Mehul P. Sampat, et al. 2009. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401.

[13] D. Esteban, C. Galand, et al. 1978. 9.6/7.2 kbps voice excited predictive coder (vepc). In ICASSP '78. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Volume 3, 307–311.