EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface

Rujia Sun
rs923@cornell.edu
Cornell University
Ithaca, NY, USA

Xiaohe Zhou
xz256@cornell.edu
Cornell University
Ithaca, NY, USA

Benjamin Steeper
bds238@cornell.edu
Cornell University
Ithaca, NY, USA

Ruidong Zhang
rz379@cornell.edu
Cornell University
Ithaca, NY, USA

Sicheng Yin
s2234521@ed.ac.uk
University of Edinburgh
Edinburgh, UK

Ke Li
kl975@cornell.edu
Cornell University
Ithaca, NY, USA

Shengzhang Wu
sw994@cornell.edu
Cornell University
Ithaca, NY, USA

Sam Tilsen
tilsen@cornell.edu
Cornell University
Ithaca, NY, USA

François Guimbretière
fvg3@cornell.edu
Cornell University
Ithaca, NY, USA

Cheng Zhang
chengzhang@cornell.edu
Cornell University
Ithaca, NY, USA

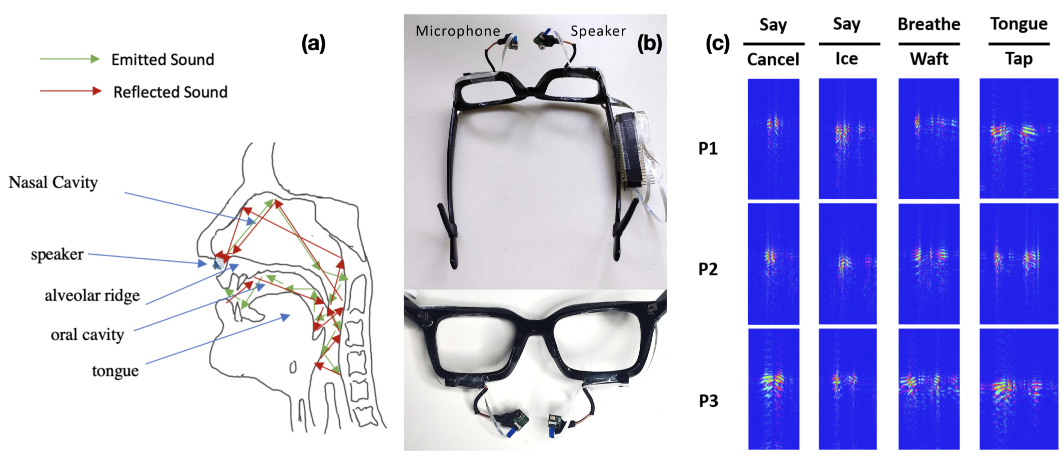

Figure 1: EchoNose system. (a) Illustration of the sensing principles of EchoNose. (b) The EchoNose hardware device. EchoNose is installed on commodity glass-frames with a pair of miniature speaker and microphone. (c) Illustration of echo profiles of different gestures from different participants.

ABSTRACT

Sensing movements and gestures inside the oral cavity has been a long-standing challenge for the wearable research community. This paper introduces EchoNose, a novel nose interface that explores a unique sensing approach to recognize gestures related to mouth, breathing, and tongue by analyzing the acoustic signal reflections inside the nasal and oral cavities. The interface incorporates a speaker and a microphone placed at the nostrils, emitting inaudible acoustic signals and capturing the corresponding reflections. These received signals were processed using a customized data processing

ISWC '23, October 08–12, 2023, Cancun, Quintana Roo, Mexico

ACM ISBN 979-8-4007-0199-3/23/10…$15.00
https://doi.org/10.1145/3594738.3611358

and machine learning pipeline, enabling the distinction of 16 gestures involving speech, tongue, and breathing. A user study with 10 participants demonstrates that EchoNose achieves an average accuracy of 93.7% in recognizing these 16 gestures. Based on these promising results, we discuss the potential opportunities and challenges associated with applying this innovative nose interface in various future applications.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; *Gestural input*.

KEYWORDS

Nose Interface; Tongue Gestures; Breathing Patterns; Silent Speech; Acoustic Sensing

ACM Reference Format:

Rujia Sun, Xiaohe Zhou, Benjamin Steeper, Ruidong Zhang, Sicheng Yin, Ke Li, Shengzhang Wu, Sam Tilsen, François Guimbretière, and Cheng Zhang. 2023. EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface. In *Proceedings of the 2023 International Symposium on Wearable Computers (ISWC '23), October 08–12, 2023, Cancun, Quintana Roo, Mexico.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594738.3611358

1 INTRODUCTION

Mouth, breathing and tongue gestures have been demonstrated as alternative hands-free interaction methods in many scenarios. For instance, silent speech can be used as an alternative interface to voiced speech [7, 16], while tongue gestures has been associated with applications in accessibility [5]. However, movements and gestures related to the mouth and tongue inside the oral cavity have been highly challenging to track for wearable devices because these movements happen inside and can not be easily captured using most wearable sensors. As a result, many prior works have to place sensors (e.g., magnets [1], capacitive sensors [8, 11]) inside the oral cavity or on the skin of the face (e.g., EMG [13]) to recognize the tongue or mouth gestures, which can be an inconvenient wearing experience for users in real-world settings.

In this paper we present EchoNose, a novel non-contact nasal interface that explores the feasibility of a new sensing approach utilizing active and inaudible acoustic signals propagating within the nasal and oral cavities to recognize 16 gestures associated with mouth, breathing, and tongue movements. It requires only a microphone and speaker positioned outside the nostrils. The *sensing principle* of EchoNose is based on the connection between the nasal and oral cavities, allowing acoustic signals emitted from the nostrils to travel through the nasal cavity into the oral cavity. Different tongue or mouth postures result in distinct shapes within the oral cavity, which reflect the acoustic signals back to the nasal cavity in complex multi-path echoes, as illustrated in Figure 1(a). These echoes can be captured by the microphone on the other nostril, enabling the processing and learning of tongue postures. To evaluate this novel sensing method, we implemented EchoNose on an off-the-shelf glass-frame, as shown in Figure 1(b). Using this glasses prototype, we conducted a user study with 10 participants to assess its performance, which demonstrated that EchoNose can recognize 16 gestures with an average accuracy of 93.7%. These gestures include 8 silent speech phrases, 4 breathing gestures, and 4 tongue gestures within the oral cavity. To the best of our knowledge, EchoNose represents the first wearable sensing method that explores active-acoustic sensing within the nasal and oral cavities. The promising results of this study provide proof of concept for this innovative sensing approach. We further discuss the challenges and opportunities associated with the future applications of EchoNose.

2 RELATED WORK

Capturing tongue gestures is highly challenging due to the tongue being inside the mouth, making it difficult to obtain accurate tongue movement information. Many previous works aim to recognize tongue gestures by instrumenting the tongue or face using magnetic sensors [12], EMG sensors [13], or optical sensors [6]. For instance, Tongue-in-cheek [5] designed a headset with 10 GHz wireless signals to recognize six tongue gestures. To recognize silent speech, researchers have explored various methods to capture tongue or mouth movements during speech, such as using an ultrasonic probe under the cheek [9], a wearable camera on a necklace [15], a capacitive sensor inside the mouth [11], or an EMG sensor on the face [2]. A recent work, EchoSpeech [16], implemented a glass-frame with active acoustic sensing to detect skin deformations for silent speech recognition. To detect breathing gestures, prior studies have utilized a chest-mounted vest to capture breathing patterns on the chest [14], a microphone on a headset [3] or under the mouth [18] to detect the breathing sounds.

In comparison to these prior works, EchoNose offers several significant contributions. No previous studies have explored the use of acoustic signal reflection inside the nasal cavity to recognize mouth and tongue gestures and movements within the oral cavity. Furthermore, previous works such as EchoSpeech [16] could only recognize one type of these three gestures. The novel sensing principle presented in EchoNose can recognize all three types of gestures (silent speech, tongue gestures, and breathing patterns) using a single affordable and non-contact sensing unit attached to a glass-frame form factor. The findings from this paper will enable the wearable technology community to explore this innovative sensing principle in other potential applications, which will be discussed in the subsequent sections.

3 SYSTEM DESIGN AND IMPLEMENTATION

3.1 Hardware Prototype

We implemented EchoNose on an off-the-shelf glass-frame, as depicted in Figure 1 (b). To affix the speaker and microphone to the glass-frame, we designed two 3D-printed pieces dedicated to housing the microphone and speaker separately. One end of each 3D-printed piece was securely attached to the glass-frame using hot glue, while the other end was connected to a flexible copper wire. The speaker was positioned on the left side, while the microphone was positioned on the right side of the glass-frame. The speaker and the microphone were placed beneath the user's nostril comfortably and non-intrusively, without causing any pain. The flexible copper wire allowed us to adjust the position of the microphone and speaker according to the unique facial and nose shapes of each user, ensuring that the sensors would not contact skin. The microphone and speaker were connected to a microcontroller (Teensy 4.1) mounted on the left leg of the glass-frame. The microcontroller controlled the sensors for emitting and receiving acoustic signals, which were subsequently stored on an SD card and analyzed offline.

3.2 Acoustic Data Processing and Deep Learning

To sense the activities inside the nasal and oral cavity, we apply Frequency Modulated Continuous Wave (FMCW) based acoustic sensing. This FMCW-based acoustic sensing method has been demonstrated effective in tracking the object position [17] and also detecting the deformations on the body [10]. Following similar methods [10], the speaker emits FMCW signals with a frequency range between 20kHz-24kHz, which is inaudible to most humans. The microphone sampled at 50kHz to capture the reflected acoustic signals. These received acoustic signals were cross-correlated with the emitted FMCW signal to obtain the strength of echoes that traveled different distances in different paths, which can be called as an *echo frame*. Connecting echo frames at adjacent timestamps formed an *echo profile*, which represented the strength of echoes traveling different distances over a certain period of time. To remove the echoes caused by the static objects in the environment, we calculated the differential echo profiles by subtracting the previous echo frame from the current one. These processed differential echo profiles containing rich information about the movements were sent to the customized machine learning pipeline to recognize gestures.

After the calculation of echo profiles, the gestures were represented as distinct visual patterns in echo profiles, as illustrated in Figure 1 (c). We then designed a customized deep learning model to distinguish these gestures. We selected ResNet-18 as the encoder thanks to its wide success in image comprehension. We then employed a fully-connected decoder with Cross Entropy Loss to classify 16 gestures. An Adam optimizer was used with an initial learning rate of 2×10^{-4}. The batch size was set to 5 and the model was trained for 300 epochs.

4 EVALUATION

In this section, we present our user study to provide a preliminary examination of how EchoNose works in recognizing silent speech, tongue gesture and breathing patterns.

4.1 Gesture Sets

We evaluated a total of 16 gestures in three groups. The first group consisted of eight silent speech phrases: **Up, Down, OK, Cancel, Previous, Next, Fire, and Ice**. We chose the first six phrases to satisfy the basic needs of menu-based applications. The last two phrases were specifically designed to activate gaming functions, such as generating fire or ice from a virtual avatar. The second group comprised four tongue gestures: **Left cheek, Right cheek, Tap, and Back Swipe**, which were inspired by the work of [5] and involved the tongue touching the sides of the cheeks, as well as tapping or swiping the ceiling of the oral cavity. These gestures were designed to simulate scenarios requiring directional control, as they are popular applications for tongue gestures. The third group consisted of four breathing gestures, inspired by prior work [14]: **Waft, Gale, Gust, and Calm**. "Calm" refers to holding one's breath, while "Waft", "Gale", and "Gust" all involve exhaling forcefully. "Waft" is a gentle and short exhalation with an open mouth, "Gale" is a strong and long exhalation, and "Gust" is a strong and short exhalation. It is worth mentioning that in [14], a chest-mounted band in addition to a VR headset was used to recognize these gestures, whereas EchoNose can recognize the same gestures with just an attachment to glasses. We will further discuss the application of EchoNose to other head-mounted wearable form factors, such as VR headsets, in the subsequent sections.

4.2 Procedures

We recruited 11 participants (3 male, 8 female, age from 21 to 29, 1 participant's data was lost due to hardware malfunction) for this user study approved by Cornell's Institional Review Borad (IRB). At beginning of the study, we demonstrated the form factor and helped participants to get familarized with the procedures. Then, a researcher helped the participant wear the device and adjusted the position of the microphone and speaker to ensure both were beneath the nostrils without contacting.

The user study contained 18 sessions for each participant. The first two were used for participants to practice, while the later 16 sessions for each user were used for evaluating the performance of EchoNose. Within each session, the participant was asked to perform the three groups of gestures in the following sequence: silent speech, breathing pattern and tongue gestures. Within each gesture group, participants repeated each gesture for 4 times in random order. A monitor was placed in front of the participant to display the name of the gesture that needs to be performed. The participant used the keyboard to mark the start and end of the performed gesture, which were saved for later processing. In real-world settings, a user can frequently take off and put on glasses, which can lead to shifts in wearing positions. To evaluate whether the remounting of the device can impact the performance of EchoNose, participants were asked to take off and put on the glasses between sessions. At the end of the user study, each participant was given a questionnaire to collect their basic demographic information (e.g., age, gender). Due to hardware failure, we lost the data from one participant. Another 3 sessions were re-collected due to interruption in the session, excessive self-reported wrong gestures, and glasses falling down the nose. In total, we collected 160 valid sessions from 10 participants including 10240 gesture instances.

4.3 Results

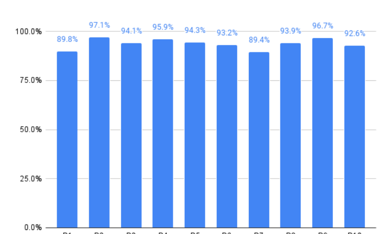

Figure 2: Performance of EchoNose evaluated on 10 participants.

Of the 18 sessions conducted, the first 2 were used to familiarize participants with the procedures and gestures, which were excluded from the evaluation. We performed 8-fold cross-validation on the remaining 16 sessions, using 14 sessions for training and 2 for testing each time. Figure 2 shows the results of all 10 participants. The performance was consistent across all participants, averaging 93.7% with a standard deviation of 2.6%.

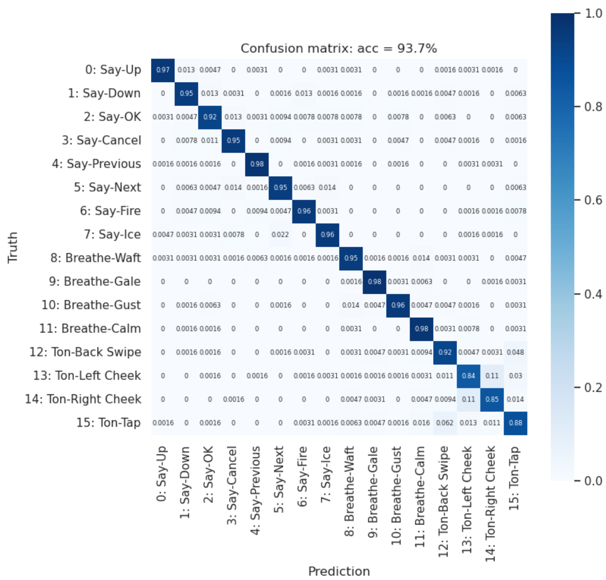

Figure 3: Confusion matrix of 16 gestures recognition results. Ton is an abbreviation for tongue.

We analyzed the performance and presented the confusion in Figure 3. The confusion matrix was highly symmetric. It is evident that most confusions came from tongue gestures. The top confusions were: Right Cheek → Left Cheek (11%), Left Cheek → Right Cheek (11%), Tap → Back Swipe (6.2%), and Back Swipe → Tap (4.8%). The tongue gestures were more difficult to perform since all movements were inside the oral cavity. In addition, the acoustic signals needed to travel much further through a winding path. We believe that these factors contribute to the distribution of errors.

We also examined EchoNose's user-dependency by training a leave-one-participant-out (LOPO) model. With this user-independent model, new users do not need to provide any training data before using the system. However, we observed significant variance in the user-independent performance on different participants, ranging from 38.3% (P6) to 77.8% (P9) with an average of 65.3% and standard deviation of 11.4%. While the performance was significantly better than random guesses, it was significantly worse than the user-dependent evaluation. This indicates that EchoNose is still a user-dependent system.

The performance shows the promising potential of EchoNose as an interface that integrates silent speech, breathing patterns, and tongue gestures as a unified system.

5 DISCUSSION

The goal of this paper is to present the proof-of-concept of this novel sensing approach to the wearable community which lowers the barrier of tracking gestures inside the oral cavity. To further help other researchers apply this system on their applications, we discuss the opportunities, challenges, and immediate next steps of applying EchoNose at scale.

Our current prototype is built on a glass-frame, featuring a design with a speaker and a microphone positioned under the user's nose. While this system is less obtrusive and non-contact compared to previous designs, we acknowledge that the appearance of this form factor may not be socially pleasant or acceptable to many users.

Figure 4: Hardware prototype attached to VR headset (Oculus Quest 2).

However, we believe that EchoNose can prove its versatility by extending compatibility to other popular head-mounted devices, such as Virtual Reality (VR) Headsets, face masks, and headphones like the Dyson Zone headphones [4]. In the case of VR headsets, we can seamlessly attach the necessary hardware to the underside, taking advantage of the larger size and less frequent use, making the additional components inconspicuous and acceptable. For face masks, we can embed the sensors beneath the nose, capitalizing on users' familiarity with daily mask-wearing during the pandemic. As for headphones, EchoNose can smoothly integrate with the Dyson Zone and similar models, ensuring the microphone and speaker remain concealed without significantly altering their appearance. However, it is essential to consider that the mask-type form factor may impact acoustic reflection patterns and mouth movements, necessitating further investigation in future studies.

As a proof of concept, EchoNose still has limitations and issues pending future investigation. Our gesture set is quite limited within each of the three categories, especially for tongue gestures and breathing patterns. Future directions include expanding gesture sets, increasing the number of users for user-independent models and accounting for the influences of environmental or human noises in real-life application.

In addition, the performance of our approach may be influenced by the condition of the user's nose. For example, if the user has a runny nose or nasal congestion, it is unclear how these factors will impact the reflection of acoustic signals within the oral and nasal cavities. Apparently, further research is needed to assess the extent to which the performance and accuracy of EchoNose are affected in these extreme scenarios.

Furthermore, there are still challenges to be addressed before EchoNose can be deployed. Currently, EchoNose requires participants to manually press the space bar to indicate the start and end of each gesture. For a practical system, automatic detection is necessary. It is possible to use energy-based segmentation such as demonstrated in SpeeChin [15] or end-to-end segmentation in EchoSpeech [16] to enable automatic segmentation. In addition, EchoSpeech also demonstrates that it is possible to implement a wireless and real-time system deployed on a smartphone. With future advancements in the computational strength in embedded systems, it might even be possible to deploy the system directly into devices such as AR glasses or VR headsets.

6 CONCLUSION

With EchoNose, we demonstrate the feasibility of recognizing mouth, breathing, and tongue gestures with a non-contact nose interface. EchoNose uses a miniature speaker at one nostril to inject acoustic signals into the nasal and oral cavities. Through reflection and diffraction, the signals travel inside the cavities, carrying information about their deformations before finally reaching the other nostril and being captured by a miniature microphone. With a user study of 10 participants, we demonstrate that EchoNose can recognize 16 gestures involving speech, tongue, and breathing with 93.7% accuracy. We believe that EchoNose demonstrates promise in various future applications.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2239569. We would like to thank Boao Chen for generating Figure 1(a). We would like to thank all participants, reviewers and the Cornell Information Science Department for supporting this project.

REFERENCES

[1] Abdelkareem Bedri, Himanshu Sahni, Pavleen Thukral, Thad Starner, David Byrd, Peter Presti, Gabriel Reyes, Maysam Ghovanloo, and Zehua Guo. 2015. Toward Silent-Speech Control of Consumer Wearables. *Computer* 48, 10 (2015), 54–62. https://doi.org/10.1109/MC.2015.310

[2] Ho-Seung Cha, Won-Du Chang, and Chang-Hwan Im. 2022. Deep-learning-based real-time silent speech recognition using facial electromyogram recorded around eyes for hands-free interfacing in a virtual reality environment. *Virtual Reality* 26, 3 (2022), 1047–1057. https://doi.org/10.1007/s10055-021-00616-0

[3] Yeqing Chen, Yulong Bian, Wei Gai, and Chenglei Yang. 2020. Adaptive Blowing Interaction Method Based on a Siamese Network. *IEEE Access* 8 (2020), 115486–115500. https://doi.org/10.1109/ACCESS.2020.3004349

[4] Dyson. 2023. *Zone™ air-purifying headphones | Dyson US.* Retrieved May 26, 2023 from https://www.dyson.com/headphones/zone

[5] 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* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 255–258. https://doi.org/10.1145/2702123.2702591

[6] Takuma Hashimoto, Suzanne Low, Koji Fujita, Risa Usumi, Hiroshi Yanagihara, Chihiro Takahashi, Maki Sugimoto, and Yuta Sugiura. 2018. TongueInput: Input Method by Tongue Gestures Using Optical Sensors Embedded in Mouthpiece. In *2018 57th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*. 1219–1224. https://doi.org/10.23919/SICE.2018.8492690

[7] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: "Hearing" Your Silent Speech Commands In Ear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 57 (jul 2022), 28 pages. https://doi.org/10.1145/3534613

[8] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Alex Olwal, Jun Rekimoto, and Thad Starner. 2021. Mobile, Hands-Free, Silent Speech Texting Using SilentSpeller. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 178, 5 pages. https://doi.org/10.1145/3411763.3451552

[9] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[10] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-Power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[11] 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* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831

[12] Hangue Park, Mehdi Kiani, Hyung-Min Lee, Jeonghee Kim, Jacob Block, Benoit Gosselin, and Maysam Ghovanloo. 2012. A Wireless Magnetoresistive Sensing System for an Intraoral Tongue-Computer Interface. *IEEE Transactions on Biomedical Circuits and Systems* 6, 6 (2012), 571–585. https://doi.org/10.1109/TBCAS.2012.2227962

[13] Makoto Sasaki, Kohei Onishi, Dimitar Stefanov, Katsuhiro Kamata, Atsushi Nakayama, Masahiro Yoshikawa, and Goro Obinata. 2016. Tongue interface based on surface EMG signals of suprahyoid muscles. *ROBOMECH Journal* 3, 1 (April 2016). https://doi.org/10.1186/s40648-016-0048-0

[14] Misha Sra, Xuhai Xu, and Pattie Maes. 2018. BreathVR: Leveraging Breathing as a Directly Controlled Interface for Virtual Reality Games. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173914

[15] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2022. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 192 (dec 2022), 23 pages. https://doi.org/10.1145/3494987

[16] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-Obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 852, 18 pages. https://doi.org/10.1145/3544548.3580801

[17] Yuzhou Zhuang, Yuntao Wang, Yukang Yan, Xuhai Xu, and Yuanchun Shi. 2021. ReflecTrack: Enabling 3D Acoustic Position Tracking Using Commodity Dual-Microphone Smartphones. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 1050–1062. https://doi.org/10.1145/3472749.3474805

[18] Daniel Zielasko, Neha Neha, Benjamin Weyers, and Torsten W. Kuhlen. 2017. A reliable non-verbal vocal input metaphor for clicking. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. 40–49. https://doi.org/10.1109/3DUI.2017.7893316