# *CELIP*: Ultrasonic-based Lip Reading with Channel Estimation Approach for Virtual Reality Systems

Yongzhao Zhang
zhangyongzhao@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, Shanghai, China

Yi-Chao Chen*
yichao@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, Shanghai, China

Haonan Wang
wanghaonan@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, Shanghai, China

Xingyu Jin
jinxingyu@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, Shanghai, China

## ABSTRACT

We developed an ultrasonic-based silent speech interface for Virtual Reality (VR). As more and more customized devices are proposed to enhance the immersion and experience of VR, our system can be used to improve the capability of interactions between users and the systems, while retaining the possibilities of using various customized devices and avoiding some limitations of traditional speech recognition. By employing the channel estimation techniques with ultrasonic waves, we can derive movement characteristics of users' lips, which can be used to fine-tune existing speech recognition models and augmented by vast open-sourced speech datasets. Moreover, we use the speech interface to guide the initialization of customized models for new users, so that they can easily have the access to our system. A two-stage experiment has been conducted and the results show that our system can achieve 90.8% command-level accuracy and 1.3% word-error-rate in sentence-level accuracy.

## CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**.

## KEYWORDS

Silent Speech Interface; Acoustic Sensing; Virtual Reality

**ACM Reference Format:**
Yongzhao Zhang, Yi-Chao Chen, Haonan Wang, and Xingyu Jin. 2021. *CELIP*: Ultrasonic-based Lip Reading with Channel Estimation Approach for Virtual Reality Systems. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3460418.3480163

*Corresponding author


*UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA*

ACM ISBN 978-1-4503-8461-2/21/09…$15.00
https://doi.org/10.1145/3460418.3480163

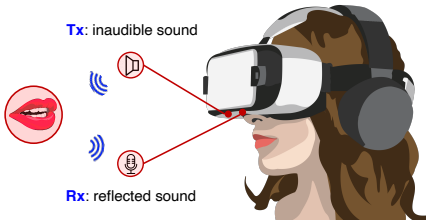


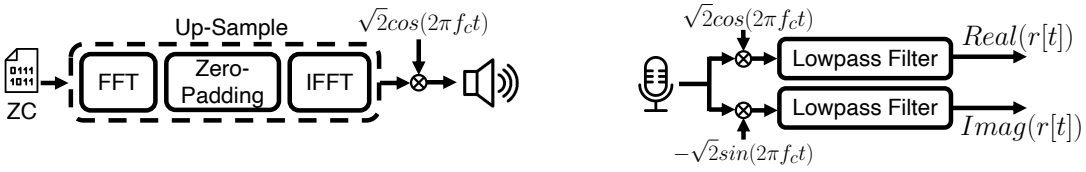

**Figure 1: The usage of *CELIP* on VR headsets.**

## 1 INTRODUCTION

In recent years, the development of Virtual Reality (VR) technology endows a host of applications in fields like entertainment, gaming, education, occupational training, medical operations, etc., which facilities the growing demand of immersive and realistic VR experiences. From this point of view, researchers developed various equipment to enable more realistic feedbacks, such as tactile [18], weight [5], force [30] and haptic [24], etc. To achieve this goal, some customized hardware, like a bow [30] and hand tools [5, 18, 24], are designed to simulate such feedbacks. However, on the other hand, the use of such devices will limit the users' ability to interact with the VR system. Different from the conventional game controller, the customized hardware will increase the immersive and realistic experiences, but decrease the capability of regular interactions.

Speech Interface (SI) is a natural way to solve the problems. Previous works confirm the potential of integrating SI with VR technology [10, 12, 31], which could bring several benefits like easy-to-use to most class of users and more immersive and enjoyable experience. Nevertheless, the implementation of SI in VR may encounter 4 limitations. First, the presence of voice commands in VR gaming may be annoying when the users are very engaged in playing games. Second, the voice recognition performance may downgrade because of mutual interference when two or more users are in the same space. Next, due to the needs of gameplay, strategies and instructions in competitive games should be invisible to the opponents. Finally, introducing SI in VR may raise privacy concerns, because others can easily understand what the users are doing.

To address above issues, we propose *CELIP*, a Silent Speech Interface (SSI) for VR system. As shown in Fig. 1, *CELIP* actively emits ultrasonic waves through an ordinary speaker, which then




(a) Transmitting ultrasonic waves.

(b) Receiving reflected signals.

**Figure 2: Transmitter and receiver design.**

are reflected by the user's lips and received by the microphone. The user's lip movements will change the channel states. By estimating the Channel Impulse Response (CIR) of the received signals, we can continuously monitor the movement characteristics and further infer the user's utterances. Users can interact with VR systems equipped with *CELIP* by mouthing their lips instead of vocalizing the utterances. To illustrate how *CELIP* works, we designed two application scenarios. The first scenario is similar to the game as described in BreathVR [26], which is a competitive game that involves two players facing each other. Each player gets a ball when it is their turn and then throws the ball at the opposing player. The goal is to get the ball over the opponent to score. The difference is that, while the ball is flying towards the opponent, the other player can manipulate the ball with $2 \times 4 \times 5 = 40$ commands activated by *CELIP*, which are the combination of 2 operations, *i.e.* 'accelerate' and 'decelerate', 4 directions, *i.e.* 'forward', 'backward','left' and 'right', and 5 levels, *i.e.* from 1 to 5. In comparison, BreathVR only supports 4 simple manipulations due to the limited sensing capability. The second scenario is similar to the ordinary voice interaction, where users use several pre-defined long sentences to interact with the VR system.

During the development of *CELIP*, we encountered several challenges. First, how to achieve sentence-level lip reading and reduce the amount of training data in the implementation of *CELIP* for VR. Different from the conventional voice recognition task, ultrasonic-based lip reading has, to the best of our knowledge, no open source datasets. Therefore, it is challenging to training and tuning a sentence-level recognition network with very limited amount of training data. Second, since the ultrasonic-based lip reading task is user diversity [34], how to adapt a model to unseen users with less effort when they start using *CELIP* in VR.

To solve the first challenge, we proposed to use movement characteristics called CIR profile, which is a single channel time series containing the features from the whole bandwidth. Thus, we can directly fine-tune the open-sourced pre-trained speech recognition model on our ultrasonic-based dataset to facilitate the model training and improve the performance. For the next problem, we noticed that the ultrasonic signal and the speech signal are distributed in different frequency bands, so we can easily separate them with conventional digital filters. Therefore, speech recognition system and *CELIP* can be running at the same time, so we can use the speech recognition system to guide the initialization of *CELIP* for new users.

Our contributions include:

- To the best of our knowledge, we are the first to use CIR profile derived from channel states to indicate the movement characteristics of lips, which is single channel time series data with real values.
- We draw on model design and numerous data sets in the field of speech recognition to facilitate the model training and improve the performance of *CELIP*.
- We use the speech interface to guide the initialization of customized model to make users have easier access to the VR systems equipped with *CELIP*.

## 2 METHOD

### 2.1 Basics of Signal Model and Channel Estimation

Assume the transmit signal is $x[n]$. After the acoustic waves passing through the signal channel, the received signal can be modeled as:

$$y[n] = h[n] * x[n] \tag{1}$$

where $h[n]$ denotes the Channel Impulse Response (CIR). The intuition behind *CELIP* is that lip movements will cause the change of channel states, which will result in the change of CIR. If we can consecutively measure $h[n]$, we can detect the contextual information of lip movements, then infer the speech utterances.

We can apply Fourier Transform on both side of E.q. 1 and obtain

$$y[\omega] = h[\omega] \cdot x[\omega]$$

Generally, because $x[\omega]$ has limited bandwidth, we can not directly divide it to the left side. However, if we multiply the conjugation of $x[\omega]$ on both sides and perform inverse Fourier transform, it is equivalent to apply circular correlation on the time domain [15]. Then the equation becomes:

$$Rxy[n] = Rxx[n] * h[n]$$

where $Rxy[n]$ and $Rxx[n]$ represent the circular cross-correlation of $x[n]$, $y[n]$ and circular auto-correlation of $x[n]$, $x[n]$. That is, $Rxy[n]$ is the convolution of $Rxx[n]$ and the channel state. Moreover, if the transmit signal has a good auto-correlation property, $Rxx[n]$ is a *sinc* function [16] and the convolution of a *sinc* function and the channel states can be considered as a good estimation of the true channel states.

## 2.2 Signal Design and CIR Profile Extraction

We choose Zadoff-Chu (ZC) sequence as our training sequence, because it is widely used in communication systems and is known to have optimum properties for correlation and channel estimation [25]. The ZC sequence is defined by:

$$ZC[n] = e^{-j\frac{\pi u n(n+1)}{Nzc}}$$

where $u$ and $N_{zc}$ is the root and length of ZC sequence, respectively. In practice, we need to transmit the sequence in inaudible band, which is normally above $17kHz$ for adults [29, 32, 34]. Therefore, we up-sample the ZC sequence with a frequency domain method [28] to limit the bandwidth to $B = 6kHz$. Next, we up-convert the signal to the pass-band by multiplying the carrier signal with a central frequency $f_c = 20kHz$, before we transmit the signal through a speaker. By doing so, the frequency of the acoustic waves is in the range of $17kHz$ to $23kHz$, which is beyond the capability of human hearing and below the sampling range for most microphones and speakers. The procedure is show in Fig. 2(b). For the received signal, we first perform down-conversion to convert it from pass band to base band, followed by a low pass filter to remove the high frequency components, as show in Fig. 2(a).

Next, we perform cross-correlation on each period to get the channel estimation $h_t[n]$, where $t = \{0, T, 2T, \dots\}$. $h_t[n]$ are complex numbers where both the amplitude and phase are crucial to indicate the CIR. Then we compute the different between the adjacent periods to remove the static components. Thus, we only consider the changing part of CIR which is caused by the lip movement. However, complex numbers are not widely adopted in current deep learning architecture, especially in speech recognition. Hence, we shift the difference of CIR from $[-\frac{B}{2}, \frac{B}{2}]$ to $[0, B]$ in frequency band by multiplying $e^{j\pi Bt}$. After that, we use the real part as our CIR profile, denoted by $c_t[n]$, which is real valued and contains all the information about the channel states and the Doppler features. By concatenating $c_t[n]$, we can obtain the feature series $c[n]$ indicating the movement characteristics of our lips. Fig. 3 shows an example of a sentence with 7 Chinese characters, which means 'how's the weather today' in English. Interestingly, the second and the third characters are the same, but with slightly different stress. Therefore, the corresponding parts of CIR profile have similar shape, but with different intensities.

## 2.3 Utterance Recognition

SoundLip [33] and EchoWhisper [7] proposed that the Doppler features of ultrasonic sine waves provides the potential for lip reading. They simultaneously monitor the Doppler shift of multiple sine waves in the inaudible frequency range. As a result, the features extracted from the Doppler shifts are composed of multiple input channels, in which the condition is different from traditional speech recognition. It means that the improvements in speech recognition field play a very small role in the evolving of ultrasonic-based lip reading field, including the model design and numerous audio datasets for pre-training. Also, Endophasia utilized the 2D-image profiles extracted from the ultrasonic signals to indicate lip movements, which borrows some techniques from computer vision tasks [11] for command-level classification.

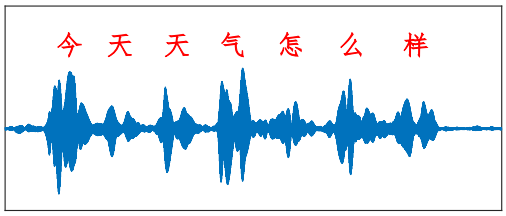
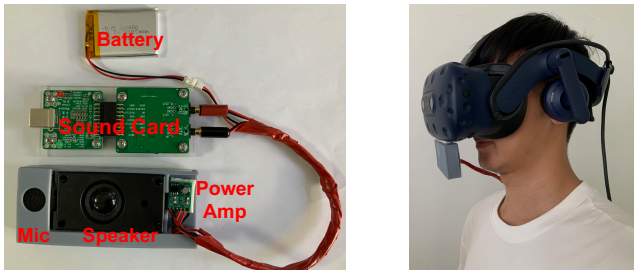



**Figure 3: An example for CIR profile.**

Different from previous ultrasonic-based lip reading works in recent years, the CIR profile proposed in *CELIP* is real valued time series data with only 1 input channel while covering the features from the whole bandwidth. It shares many similarities with ordinary voice signals and retains all the Doppler features and channel features used by the previous works. *CELIP* uses the CIR profile consecutively measured from the channel states, to indicate the movement characteristics of users' lips. It can be classified by a typical seq2seq model [3] with standard parameter settings in speech recognition, as well as other classical models [4, 8, 9]. Also, the models pre-trained on the numerous speech recognition datasets can be used to reduce the large data size required by [7, 33, 34], which will bring a lot of manual labor for researchers and users. Therefore, we use a typical pre-trained LAS model [3] for utterance recognition, which is composed of a pBLSTM-based listener and an attention-based speller. Please refer to [3] for more information about the model structure, parameter settings and training scheme.

## 2.4 Model Initialization for New Users

As described in Endophasia, the performance of a well-trained ultrasonic-based lip reading model for unseen users has only around 40% accuracy, which is not satisfactory to be applied in actual use. Therefore, a customized model should be fine-tuned with the data collected from the target user. However, when we integrate *CELIP* to VR systems, we find the data collection procedure for new users will affect the practicability of the system and reduce the user's willingness to use it to a certain extent. Because the voice signal and the ultrasonic signal are in different frequency bands, we can easily extract the voice signal and use a mature speech recognition model to guide the initialization of the customized model. This process can be automatically completed when the user uses the VR system, without requiring the user to intentionally go through the data collection stage. Benefit from the use of CIR profile, collecting three samples for each utterance can achieve a good performance. When the number of a certain utterance reaches this standard, we will automatically update the customized model in the background. In addition, if the existence of the corresponding utterance in the voice signal is detected in the subsequent use, we can further strengthen the robustness of the model.




(a) (b)

**Figure 4: The prototype headset integrated with *CELIP*.**

## 3 EXPERIMENT SETUP

### 3.1 Prototype Design

We developed a prototype to implement *CELIP* as shown in Fig. 4. We designed a 3D printed box to fix a microphone, a speaker and a power amplifier. Then we connect the microphone and the speaker to an external sound card and the power amplifier is powered by a battery. The sound card sends the data stream to the computer via a USB interface. After that, we stick the 3D printed box on the HTC Vive headsets. Note that all the components in our prototype already exist in COTS VR headsets, so that *CELIP* can be easily integrated into a headset by changing the position of a mic and a speaker. For the first scenario, the ball game is set in the 3D environment downloaded with the Sci-Fi Laboratory Pack 2. For the second scenario, we simulate the interaction mechanism of voice recognition in VR operation.

### 3.2 Experiment Design

A total of 6 subjects participated in our experiments and all of them are college students in the age from 22 to 26. Our experiments were conducted in two stages with Mandarin. In the first stage, the subjects were first invited to play the ball game with ordinary speech recognition interface in 3 groups, while collecting the CIR profiles in the background. The game time of each group was about 20 minutes. Then, the subjects were asked to read 70 given sentences according to the prompts, where a total of 37 different words are contained. Each sentence contains 6.7 words on average and was read 3 times, and each subject spent around 50 minutes to finish the task. Note that, during the practical usage, users only need to collect samples for a few sentences before using *CELIP*, instead of finishing all the 70 sentences. Next, we trained a customized model for each subject with leave-one-user-out scheme. That is, in order to train a model for a target subject, *i.e.* the unseen user, we use the data set from the remaining users to train the model, and use the data from the target user to fine-tune the model. In the second stage, the subjects were invited to repeat the tasks in the first stage with *CELIP*. Then we count the accuracy of command recognition in the ball game and report the word error rate (WER) of the long sentence recognition.

## 4 RESULTS

### 4.1 Accuracy of Commands in Scenario 1

For the ball game, a word error will make the entire command have a completely different meaning, so it is meaningless to calculate the word error rate in this case. Therefore, we directly record whether the recognized command is correct. In the second stage of our experiments, the subjects used a total of 336 commands, of which the recognition results of 305 commands were consistent with the subject's expectations, and the accuracy was around 90.8%.

### 4.2 WER in Scenario 2

WER is a widely used criterion in the scope of continuous speech recognition. It measures the minimum operations of substitution, deletion, and insertion to convert hypothetical sentences into the reference:

$$WER = \frac{S + D + I}{N}$$

where $S$ is the number of substitutions, $D$ is the number of deletions, $I$ is the number of insertions, and $N$ is the number of words in the reference. The average WER over the 6 subjects is 1.3%. The 70 sentences have a total of 469 words, which are repeated for 3 times, so 1407 of total words are considered for each subject. It means that, for each subject, only $1407 \times 0.013 = 18.2$ words need either substitution, deletion or insertion.

## 5 RELATED WORK

Several Silent Speech Interfaces (SSIs) with bulky/invasive sensor deployment or non-invasive wearable devices have been proposed. By invasively implanting/placement of sensors in the human body, researchers have proposed solutions to recognize brain activity in the voice motor cortex [2], or use intraoral magnetic beads [14] or capacitive touch sensors [19] to capture the tongue and jaw exercise. The inconvenience caused by these invasive solutions prevents their applications in VR. In order to provide a more practical and affordable solution, other studies have designed schemes to recognize speech by using alternative sensors (e.g., EEG [23], sEMG [20], ultrasound imaging [17]) to detect tongue, facial, throat movements, and microphones attached to the skin to hear non-audible murmurs (NAM) [13, 21, 22] or put close to the front of the mouth to capture whisper-like tiny voice while ingressive breathing [6]. Although these solutions are non-invasive, they require special sensors to be installed on the human body. Such schemes will increase the difficulty for users to use the VR devices and reduce their willingness to use the SSIs.

Some works have proposed the use of camera-based techniques for developing SSIs, *e.g.* LipNet [1], Lip-Interact [27], so that no special sensors are required. However, an additional camera needs to be installed towards the user's mouth to capture the user's lip movements. Also, the performance of camera-based schemes will be significantly affected by the lighting condition, when the user's mouth is covered by the headsets or when the light in the room is very dim. In this respect, the microphones and speakers used in active ultrasonic detection are not sensitive to directivity, and the cost is very low, and the acoustic waves are not sensitive to the light condition. Therefore, it has great potential to be applied in the VR systems to develop SSIs.

## 6 CONCLUSION

In this paper, we present *CELIP*, a sensing technique enabling silent speech interface in VR systems by utilizing ordinary sensors equipped in most COTS headsets, *i.e.* speakers and microphones. *CELIP* perceives users' lip movements by actively emitting ultrasonic waves and infer the users' utterances by analysing the CIR profile, which is derived from the channel states. As generating acoustic-based CIR profile requires only a speaker and a microphone, *CELIP* offers a non-invasive, robust and low-cost approach to develop silent speech interfaces for VR systems. By using CIR profile, we can use the conventional speech recognition model and datasets to facilitate the training stage and improve the performance. Also, we use the ordinary speech recognition interfaces to guide the initialization of customized model to improve the usability of *CELIP*. We designed a prototype of *CELIP* and conducted a two-stage experiment. The results show that *CELIP* can achieve 90.8% command recognition accuracy in the ball game and 1.3% WER in the long sentence recognition task, while imposing only a little effort for new users to access the interface.

## ACKNOWLEDGMENTS

This work is supported by NSFC (61936015, U1736207, 62072306), Startup Fund for Youngman Research at SJTU, and Program of Shanghai Academic Research Leader (20XD1402100).

## REFERENCES

[1] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2017. LipNet: End-to-End Sentence-level Lipreading. *arXiv: Learning* (2017).

[2] Jonathan S. Brumberg, Alfonso Nieto-Castanon, Philip R. Kennedy, and Frank H. Guenther. 2010. Brain-computer Interfaces for Speech Communication. *Speech Commun.* 52, 4 (April 2010), 367–379. https://doi.org/10.1016/j.specom.2010.01.001

[3] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211* (2015).

[4] C. Chiu and Colin Raffel. 2018. Monotonic Chunkwise Attention. *ArXiv* abs/1712.05382 (2018).

[5] Inrak Choi, Heather Culbertson, Mark R Miller, Alex Olwal, and Sean Follmer. 2017. Grabity: A wearable haptic interface for simulating weight and grasping in virtual reality. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 119–130.

[6] Masaaki Fukumoto. 2018. SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. ACM, New York, NY, USA, 237–246. https://doi.org/10.1145/3242587.3242603

[7] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–27.

[8] Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711* (2012).

[9] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[10] Susumu Harada, Jacob O Wobbrock, and James A Landay. 2011. Voice games: investigation into the use of non-speech voice input for making computer games more accessible. In *IFIP Conference on Human-Computer Interaction*. Springer, 11–29.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Daniel Hepperle, Yannick Weiß, Andreas Siess, and Matthias Wölfel. 2019. 2D, 3D or speech? A case study on which user interface is preferable for what kind of object interaction in immersive virtual reality. *Computers & Graphics* 82 (2019), 321–331.

[13] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication* 52, 4 (2010), 301 – 313. https://doi.org/10.1016/j.specom.2009.12.001 Silent Speech Interfaces.

[14] Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, and Sergey I. Rybchenko. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication* 55, 1 (2013), 22 – 32. https://doi.org/10.1016/j.specom.2012.02.001

[15] B Hunt. 1971. A matrix theory proof of the discrete convolution theorem. *IEEE Transactions on Audio and Electroacoustics* 19, 4 (1971), 285–288.

[16] Elias Kellner, Bibek Dhital, Valerij G Kiselev, and Marco Reisert. 2016. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magnetic resonance in medicine* 76, 5 (2016), 1574–1581.

[17] 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 (CHI '19)*. ACM, New York, NY, USA.

[18] Pascal Knierim, Thomas Kosch, Valentin Schwind, Markus Funk, Francisco Kiss, Stefan Schneegass, and Niels Henze. 2017. Tactile drones-providing immersive tactile feedback in virtual reality through quadcopters. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 433–436.

[19] 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)*. ACM, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831

[20] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline. 2017. Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (Dec 2017), 2386–2398. https://doi.org/10.1109/TASLP.2017.2740000

[21] Yoshitaka Nakajima, Hideki Kashioka, Nick Campbell, and Kiyohiro Shikano. 2006. Non-Audible Murmur (NAM) Recognition. *IEICE - Trans. Inf. Syst.* E89-D, 1 (Jan. 2006), 1–4. https://doi.org/10.1093/ietisy/e89-d.1.1

[22] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 5. V–708. https://doi.org/10.1109/ICASSP.2003.1200069

[23] Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. 2017. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of Neural Engineering* 15, 1 (nov 2017), 016002. https://doi.org/10.1088/1741-2552/aa8235

[24] Chaeyong Park, Jinhyuk Yoon, Seungjae Oh, and Seungmoon Choi. 2020. Augmenting Physical Buttons with Vibrotactile Feedback for Programmable Feels. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 924–937.

[25] B.M. Popovic. 1992. Generalized chirp-like polyphase sequences with optimum correlation properties. *IEEE Transactions on Information Theory* 38, 4 (1992), 1406–1409. https://doi.org/10.1109/18.144727

[26] 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*. 1–12.

[27] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. ACM, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[28] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 591–605.

[29] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 82–94.

[30] Tzu-Yun Wei, Hsin-Ruey Tsai, Yu-So Liao, Chieh Tsai, Yi-Shan Chen, Chi Wang, and Bing-Yu Chen. 2020. ElastiLinks: Force Feedback between VR Controllers with Dynamic Points of Application of Force. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1023–1034.

[31] Yannick Weiß, Daniel Hepperle, Andreas Sieß, and Matthias Wölfel. 2018. What user interface to use for virtual reality? 2d, 3d or speech–a user study. In *2018 International Conference on Cyberworlds (CW)*. IEEE, 50–57.

[32] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th annual international conference on mobile systems, applications, and services*. 15–28.

[33] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.

[34] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.