JawSense: Recognizing Unvoiced Sound using a Low-cost Ear-worn System

Prerna Khanna*
pkhanna@cs.stonybrook.edu
Stony Brook University
New York, USA

Tanmay Srivastava*
srivastava.tanmay111@gmail.com
Indian Institute of Technology
Gandhinagar, India

Shijia Pan
span24@ucmerced.edu
University of California
Merced, USA

Shubham Jain
jain@cs.stonybrook.edu
Stony Brook University
New York, USA

Phuc Nguyen
vp.nguyen@uta.edu
University of Texas at Arlington
Arlington, USA

ABSTRACT

This paper explores a new wearable system, called *JawSense*, that enables a novel form of human-computer interaction based on unvoiced jaw movement tracking. *JawSense* allows its user to interact with computing machine just by moving their jaw. We study the neurological and anatomical structure of the human cheek and jaw to design *JawSense* so that jaw movement can be reliably captured under the strong impact of noises from human artifacts. In particular, *JawSense* senses the muscle deformation and vibration caused by unvoiced speaking to decode the unvoiced phonemes spoken by the user. We model the relationship between jaw movements and phonemes to develop a classification algorithm to recognize nine phonemes. Through a prototyping implementation and evaluation with six subjects, we show that *JawSense* can be used as a form of hands-free and privacy-preserving human-computer interaction with 92% phoneme classification rate.

CCS CONCEPTS

• **Human-centered computing** → *Accessibility*; *Accessibility systems and tools*;

KEYWORDS

Unvoiced sound recognition, Wearable devices, Accelerometer sensing

ACM Reference Format:
Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. *JawSense*: Recognizing Unvoiced Sound using a Low-cost Ear-worn System. In *The 22nd International Workshop on Mobile Computing Systems and Applications (HotMobile '21), February 24–26, 2021, Virtual, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3446382.3448363

*Both authors contributed equally to this research.

HotMobile '21, February 24–26, 2021, Cyberspace

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3446382.3448363

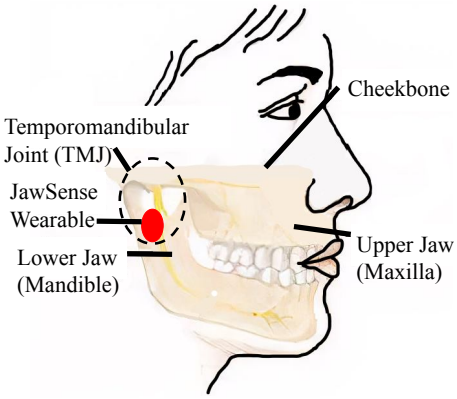

Figure 1: *JawSense*: concept and motivation

1 INTRODUCTION

Speech and touch-typing are the two most common input modalities for smart devices [10]. However, they pose challenges in many scenarios where touch-typing restricts hands-free operation and speech commands are not desirable.

For example, people with vision, speech, or motor neuron disabilities often rely on additional devices to interact with their computing machines like the Switch Access for Android [3]. Carrying an additional switch and navigating through it overloads many commands on few buttons, further restricting the scope of interaction. Deaf-blind people rely on Tadoma to gauge what the other person is speaking by tracking speech articulators' movement and sensing the vibrations [19]. This technique is not preferred as people do not prefer to touch each other's face and invade their personal space.

Hands-free operation is desired for users whose hands are fully occupied (e.g., a pianist wants to turn music sheet, a VR/AR user wants to type username/password, a user wants to answer message during a meeting, etc.) or those who have difficulty in coordinating their hands (e.g., Parkinson, Amyotrophic Lateral Sclerosis (ALS), and quadriplegic patients). In addition, with the COVID'19 pandemic, healthcare facilities require contactless interaction with equipment and smart devices. Speech commands are not preferred in public settings when sensitive information is involved as they can be eavesdropped, posing a breach in privacy [1]. Also, speech commands can not be used in noisy environments. In recent years,

various input modalities have been proposed to cater to accessibility needs and provide hands-free operation [8]. The use of teeth gestures in these works is not as intuitive as using language based commands.

In this paper, we present *JawSense*, a hands-free, socially acceptable, privacy-preserving, and intuitive input sensing modality.

It recognizes the unvoiced phonemes based on jaw movement patterns. Different phonemes involve different manner in which the jaw moves and the muscles around it contract and retract. We use a low-cost accelerometer to capture signals from articulators like jaw and cheeks to identify the unvoiced phoneme. Phonemes are the building blocks for words, thus recognizing phonemes can lead to potential sentence recognition. Our contributions are as follows:

- Identifying the best sensor placement location for single sensing modality for reliably capturing jaw movement and developing a socially acceptable wearable. Our prototype can be retrofitted to headphones/ earphones.
- Proposing a new jaw sensing technique for hands-free, privacy-preserving, and intuitive interaction with computing devices.
- Devising an algorithm to isolate unvoiced commands from voiced commands via spectrum area selection.
- Evaluating our system with six participants with real-world experiments. *JawSense* achieves an accuracy of 92% in detecting nine phonemes.

The rest of the paper is organized as follows. Section 2 discusses the domain knowledge and core intuition that enables *JawSense*. Next, Section 3 presents the system design. Then, Section 4 introduces the implementation, setup and evaluation. We discuss the future directions in Section 5 and conclude in Section 7.

2 BACKGROUND AND CHALLENGES

Human speech is a sequence of different sounds. A meaningful sound helps to distinguish between different words. For instance, mat and pat are different words distinguished by the sounds, */m/* and */p/*, respectively. */m/* and */p/* are known as **phonemes**. Phonemes are the smallest sound acting as building blocks for words in any language. The human voice generation mechanism has three subparts: the lung, the vocal cord, and articulators. The lung produces the air pressure required to vibrate the vocal cord. The vocal cord vibrates to produce audible vibrations, and articulators articulate sounds coming from the larynx. The jaw has often been deemed an articulator due to its involvement in speech.

As different phonemes are articulated differently, we utilize motion signals captured by a low-cost sensor from articulators like the jaw and cheeks for identifying phoneme. We use an accelerometer to measure the motion due to its simplicity, low-cost, and comparatively lower power consumption than other motion sensors.

We conduct a set of in-lab experiments to assess the feasibility of detecting a phoneme using a single accelerometer. In particular, we use the accelerometer signals from the jaw and cheeks. We chose nine of the most spoken phonemes across 451 languages [21] as listed in Table 1. We placed an accelerometer on the lower portion of the temporomandibular joint (TMJ), as shown in Figure 1. The TMJ acts like a sliding hinge connecting the skull and the lower jaw, permitting the jaw to move up and down, and in lateral direction.

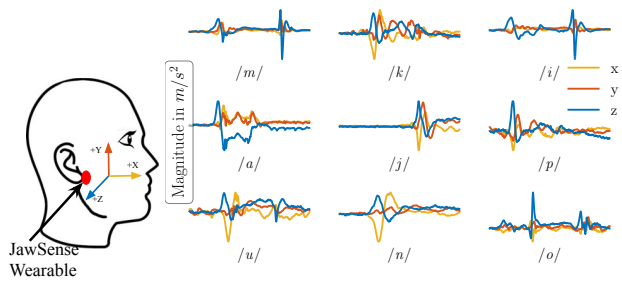

Figure 2: Time domain representation signals when user speaks different phonemes.

Ph	/m/	/k/	/i/	/a/	/j/	/p/	/u/	/n/	/o/
Wd	man	king	bit	cat	jug	pay	put	net	pot

Table 1: Nine most used phonemes of occurrence across 451 languages. (Ph - Phoneme, Wd - Word)

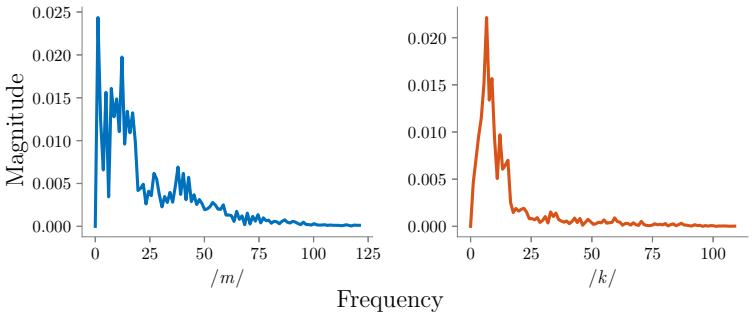

Figure 3: */m/* has a wider frequency profile compared to the plosive phoneme */k/.

Placing it lower on the TMJ, allows *JawSense* to be retrofitted on any off-the-shelf headphones.

Our experiments are designed to answer the following questions: (1) *Can a single accelerometer sensor be used to capture human jaw movement during phoneme articulation?* (2) *What noise sources affect the system performance?*

Figure 2 shows time domain accelerometer signals for different phonemes, which demonstrates distinguishable characteristics. Figure 3 further shows an example of the frequency domain representation of the dominant axis (y-axis) for two phonemes */m/* and */k/*. The plosive phoneme */k/*, in which airflow is blocked and then released in a burst, is more contracted in bandwidth compared to the other nasal phoneme */m/*. This set of experiment establishes that jaw movement signal exhibits characteristics both in time and frequency domain for a particular phoneme.

■ **Challenges.** While the preliminary results are very promising, realizing *JawSense* is difficult due to the following challenges. Accelerometer, along with jaw movement, can be affected primarily by two other sources: **body movements** and **mechanical waves** (music) from external sources vibrating the system. Figure 4 shows the spectrogram when a user is performing different activities (nodding, head rotation, yawning, etc.) and in different conditions (noisy acoustic environment) while wearing the *JawSense* prototype. As

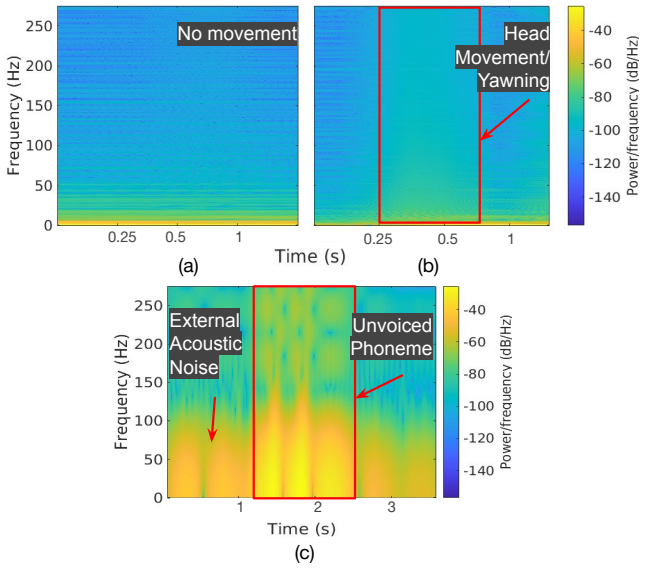

Figure 4: Spectrogram of signals in different conditions. (a) Quiet environment. (b) Body movements (c) External acoustic noise.

seen, body movements (nodding, yawning) have far lower energy and frequency than other activities. Previous studies show that accelerometer's performance is affected by acoustic signals [23]. Unsurprisingly, our prototype's accelerometer is also affected in an external acoustic environment.

3 SYSTEM DESIGN

JawSense enables unvoiced sound recognition via a ear-worn system. Figure 5 shows the *JawSense*'s system overview. *JawSense* includes (1) an ear-worn device that non-obtrusively tracks the human jaw movement in real-time, and (2) an algorithm to remove the noise from human artifacts (e.g., nodding, yawning), extract unvoiced signal, and classify the detected unvoiced phoneme.

■ **Preprocessing**. First step is to preprocess the raw accelerometer data. We take samples equivalent to 2 seconds of data and call it as a *window*. We then divide the window into *frames* of half a second. We take this frame size as it is long enough to capture an entire jaw movement. From the 3 axis accelerometer data, we remove the effect of gravity (offset removal). A mean average filter is used to smooth the signal and remove jitters.

■ **Motion artifacts mitigation**. Human jaw motion creates millimeter motion captured by the accelerometer at the skin surface. With a low-cost sensor, it results in only few mV of voltage captured by the Analog-to-Digital Converter (ADC) on the microcontroller. At the same time, motion artefacts create similar or often stronger signals in amplitude that are the mixed with the jaw movement signal readings. To remove the noise caused by human artefacts, including but not limited to nodding, head movements, and yawning, *JawSense* applies a high-pass filter on the pre-processed signals. Spectrogram in Fig. 4 shows that these body movements have a low-frequency range – less than 1 Hz – while jaw movements associated with phonemes have a frequency profile in [5,100] Hz range. As a result, *JawSense* applies a 1 Hz high-pass filter on the pre-processed signals to remove low frequency motion artifacts.

■ **Unvoiced signal extraction from acoustic noise**. As discussed in Section 2 performace of accelerometers is affected by interference of acoustic signals. While human artifacts induce noise levels in a lower frequency range, audible sounds from external sources causes frequency disturbances over a wide range of frequency as shown in Figure 4 (c). The frequency profile of phoneme-induced jaw movement is from 5 to 100 Hz. Assuming only one phoneme in a two second window, no two consecutive frames are expected to have [5,100] Hz frequency range in the absence of external noise. If there are more than two such frames, we identify the window as acoustic noise event. The sound waves interacting with the sensor have lower energy compared to those produced by jaw movements. This allows for energy based thresholding. We find the ratio of energy for each frame with the window. The frame with highest ratio is identified as a frame with phoneme. Wiener filter is used to remove acoustic noise from unvoiced phoneme.

■ **Voiced and unvoiced sound detection**. After reducing effect of noise components we intend to detect if the phoneme is said in voiced or unvoiced manner. This assists users when they want to interact with the *JawSense* system and system outside it. E.g. interacting with VR using unvoiced commands and voiced commands for communicating with people. Figure 6 shows the time domain representation and frequency spectrum (y-axis) for the same phoneme, once in an audible manner and then just jaw movement. In Figure 6, (1) is associated with jaw opening, (2) deformation of skin around jaw, (3) with jaw closing. (1) and (3) are not affected by voiced phoneme, but as vibrations travel through skin, there is a high frequency component associated with (2) that is marked in (4). The higher frequency component associated with audible phoneme is in the base frequency of human speech [80,255] Hz. This is from the vibrations caused by vocal cords while making a sound that travel up-to prototype through cheek muscles. *JawSense* uses this property for detecting audible phonemes. Specifically, if the energy of spectrum in [80,220] Hz is greater than thrice the noise floor for more than 0.2 sec, jaw movement is considered voiced.

■ **Unvoiced phonemes classification**. Once we remove human artifacts and the effect of external noise components, we classify non-audible phonemes. We assume that if phoneme is voiced, users must interact with components outside our system and need not be classified. As observed in Figure 2 and 3 both time and frequency domain representation have characteristics based on manner of jaw movement. Hence, we extract time and frequency domain features. In time domain, we select features that are independent of magnitude as different users may have different extent to which they can open their jaw. We use skewness, area under the curve, and kurtosis as time domain features and first eight Discrete Fourier Transform (DFT) coefficients for getting frequency domain characteristics. This set of nine features is used to train a Support Vector Machine (SVM) classifier.

4 EVALUATION

We conducted a series of real-world experiments to validate the performance of *JawSense*. In this section, we first describe our experimental setup. We then analyze the performance of voiced vs unvoiced phoneme detection. Next, we evaluate the overall phoneme

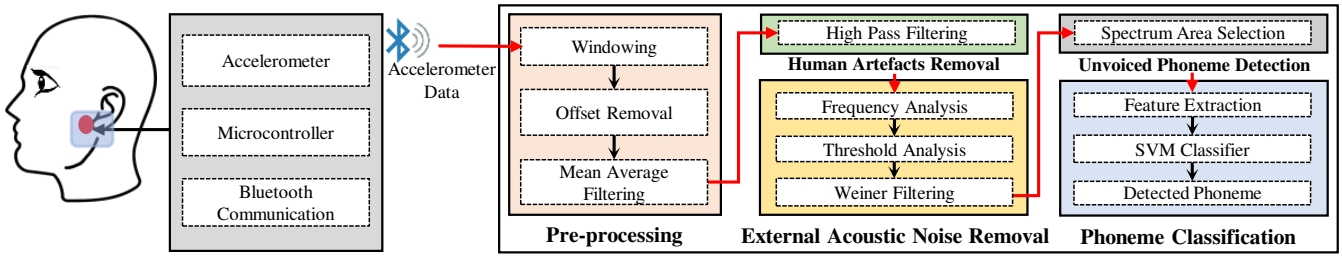

Figure 5: *JawSense* system overview

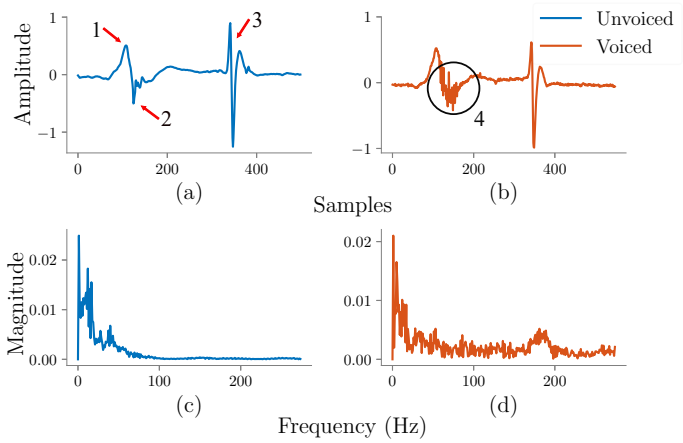

Figure 6: (a), (b) and (c), (d) are time and frequency representations of unvoiced and voiced phoneme /*m*/ respectively. The figure shows: jaw opening (1), skin contraction (2), jaw closing (3), and voiced phoneme introducing high frequency component (4).

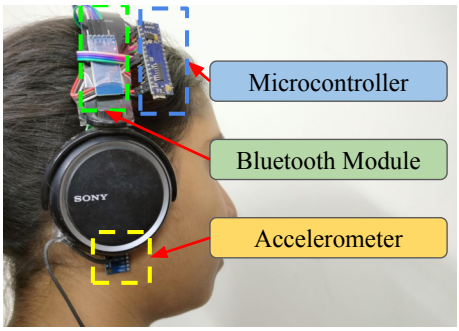

Figure 7: *JawSense* prototype

classification model. Lastly, we evaluate the overall system performance in a real-world scenario.

■ **Experimental setup.** We build the *JawSense* prototype with two key design goals: (1) To have a small form-factor wearable that does not interfere with the jaw functioning, and (2) To have a socially acceptable design. The *JawSense* prototype consists of an Arduino Nano 33 BLE Sense and an off-the-shelf IMU MPU 9250 retrofitted on a headphone. Surgical tape was used so the wires do not hurt the user. The prototype did not cause discomfort to the user. Our prototype collects the three axes accelerometer data at 550 Hz. The system communicates to an HP Notebook laptop via Bluetooth. Fig. 7 shows our prototype worn by a participant.

We conduct experiments with six volunteers. The age group of participants is 17-55 years, with three males and three females. They were asked to record each of the nine phonemes 20 times. Two of them were asked to conduct additional experiments under different scenarios to assess the robustness of the system. They were asked to perform the unvoiced phonemes 20 times while performing the following movements: head nodding, yawing, voiced phoneme articulation, and music playing in the background.

■ **Detecting unvoiced phoneme.** *JawSense* distinguishes skin surface vibrations and deformation of unvoiced phoneme from voiced phonemes by analysing the energy of the frequency spectrum. Fig. 8 shows the precision and recall of the unvoiced phoneme detection.We reduce false negatives for unvoiced phonemes to prevent the system from omitting any phonemes and skipping classification. We observe that the unvoiced phoneme shows a precision rate of 0.91 and a recall rate of 1, which implies a lower false negative rate. Voiced phonemes show a precision rate of 1 and a recall rate of 0.9, which implies a lower false positive rate.

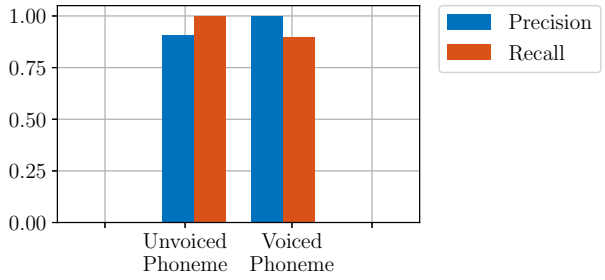

Figure 8: Unvoiced-Voiced phoneme detection rate.

■ **Phoneme classification.** *JawSense* achieves the unvoiced phoneme recognition accuracy of 92% across six subjects and nine phonemes. Figure 9 shows the confusion matrix of the recognition accuracy over the investigated nine phonemes. We observe that /*a*/ and /*i*/ are most erroneous phonemes. Despite their similarity, the system still achieves 94% and 75% recognition rate for/*a*/ and /*i*/. We believe that by training personalised models for each user we can achieve a higher accuracy since it would account for individual variations. We leave this avenue for further research.

Subject	S1	S2	S3	S4	S5	S6
Accuracy	0.87	0.86	0.82	0.81	0.80	0.88

Table 2: Leave one out accuracy for each subject. Standard deviation is 0.31.

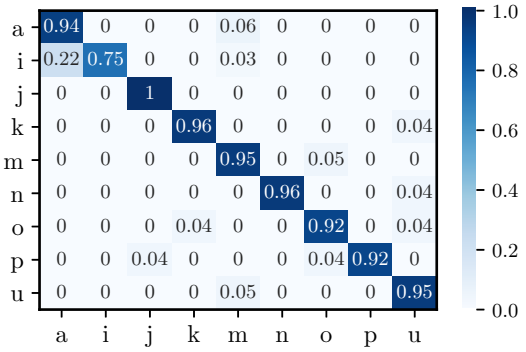

Figure 9: Classification accuracy for nine phonemes.

We observe that the acceleration of the jaw movement varies among users, which could be caused by differences in jaw structure across individuals. To assess the impact of this variation, we conduct leave-one-user-out validations and obtain a mean recognition accuracy of 84%. Table 2 shows accuracy for when each subject is left out from training set. The accuracy for each user is comparable with the others showing that the model is generalizable for all subjects.

■ **Sensitivity analysis.** To assess the robustness of the system, we evaluate *JawSense* with data containing human motion artifacts like head nodding and yawning, and in the presence of external acoustic noise like music playing in the background. Fig. 10 shows the precision and recall for the two subjects who performed the experiments in these conditions. *JawSense* achieved an accuracy of 96% under controlled (no external noise) conditions for these two subjects. Even after adding human motion artifacts and external acoustic noise *JawSense* attained an accuracy of 94% and 90% respectively, demonstrating the robustness in real-world environment.

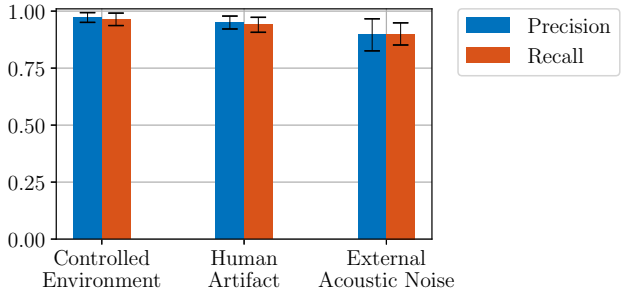

Figure 10: Phoneme classification accuracy in different conditions. 0.96 in controlled conditions, 0.94 with motion artifacts, and 0.90 in external acoustic noise.

5 DISCUSSION AND LIMITATIONS

We present our initial efforts towards building a low-cost unvoiced sound detection system. We further discuss limitations, directions for future research, and potential applications:

JawSense's limitations. The current prototype has following limitations: (1) Body movements that create high frequency noise may induce false-positives for audible phonemes. This can be avoided by using a microphone to distinguish between high frequency components induced from body movements and audible phoneme articulation. (2) We can not reliably understand the confounding nature of phoneme classification with the current system. In the future, we plan to map the motion signal to jaw movement. This would help better understand the phoneme identification. (3) In our current prototype, the accelerometer has to be placed in a fixed orientation; with change in the orientation the model would have to be retrained. We plan to integrate gyroscope measurements into the system, which will give information about current orientation hence making it possible to transform current orientation to a reference orientation. (4) Presently, we have not evaluated the system's efficiency such as energy consumption, latency, and change in temperature. We plan to address this in the next iteration of the system.

Generalizing JawSense. The current prototype of *JawSense* uses off the shelf accelerometer along with a battery and microcontroller. In the future, we plan to make the system robust to various noise sources and evaluate on a larger sample size. This can be achieved by the use of multi modal sensing systems like outward facing microphone and gyroscope. Also, a system with a smaller form factor can be retrofitted with existing behind the ear earphones and headphones. Apple's second-generation AirPods are enabled with accelerometers. Also, Apple is working towards building earphones capable of supporting on-device inference [13].

Recognizing words and sentences. We recognize nine phonemes at a sampling rate of 550 Hz. We envision more phonemes can be distinguished with 1) higher sampling rate and higher frequency details and 2) combination of accelerometer and gyroscope signals. Previous studies [11, 12] have shown that higher sampling rate can give information about fine grained bio vibrations. Also, with a gyroscope we get an estimate of change in orientation. We envision to use sensor fusion techniques to measure displacement and change in angle of jaw. These can serve as important parameters for phoneme classification. These further improvements would pave the way for word and sentence recognition with multi-modal and more granular data.

Speech disorder detection. Speech disorders limit an individual's ability of articulation. This includes slurred speech, stuttering, mumbling, etc. As these disorders are associated with TMJ (Figure 1), we believe *JawSense* can be used for diagnosis and evaluation of speech disorders [16].

Speech verification. Many IoT devices use speech commands to authenticate a user and for machine-user interaction. This leaves user privacy at stake. *JawSense* can be used to interact with IoT devices by using unvoiced commands, keeping user privacy intact.

6 RELATED WORK

Prior research has explored lip and mouth motion tracking via contactless and contact-based approaches. Non-contact systems utilize either vision-based methods or wireless signals, such as ultrasound, radio waves, etc. Vision-based methods [4, 15] use images captured from a camera to track the mouth's motion for interpreting silent talk. However, these systems are susceptible to variations in camera placement and lighting conditions. WiFi signals have been used to track motion of articulators. One such system is WiHear [22], which leverages multipath effects and wavelet packet transformation. Though the system overcomes the line of sight limitations found in vision based systems, it is easily affected by body movements. Ultrasound methods [5, 6] leverage Doppler Shift for motion detection. SilentTalk [20] uses ultrasound generated from mobile phone to analyze the frequency shift induced due to lip movements. These systems are not entirely *hands-free*, and do not distinguish between voiced and unvoiced sounds.

Contact based approaches use sensors like microphone, EEG, EMG, and systems combining signals from brain and muscles. Acoustic systems [2] place a microphone very close to the mouth and detecting non-audible murmur (NAM). SottoVoce [9] utilizes high frequency (3.5 MHz) ultrasonic sensor placed under jaw. Though accurate, the system requires human organs exposed to high frequency ultrasound, effects of which are still unknown. EMG and EEG have been widely studied for silent lip movement tracking [18]. Jorgensen et al. [7] recorded surface signals from the sublingual areas associated with vocal tract's muscle activity. AlterEgo[8] and TYTH [14] are examples of systems that use neuromuscular and combination of brain muscle signals for silent communication. These systems are robust in movement recognition but require use of multiple sensors setup. Prakash et. al [17] *re-task* earphones, using them as input to detect teeth tapping and sliding. The system focuses on tracking teeth grinding gestures instead of recognizing speaking phoneme.

7 CONCLUSION

In this paper, we explored unvoiced sound recognition leveraging the motion of articulators involved in human speech generation. *JawSense* enables recognition of unvoiced phonemes via motion signals from jaw and cheek muscles. We evaluate the system for six subjects across nine phonemes achieving an accuracy of 92%. We aim at continuous speech recognition with a smaller form factor in future research.

ACKNOWLEDGEMENTS

We thank the shepherd and the anonymous reviewers for their insightful comments. This material is based in part upon work supported by the National Science Foundation under Award No. 1932296.

REFERENCES

[1] Aarthi Easwara Moorthy and Kim-Phuong L Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction* 31, 4 (2015).

[2] Masaaki Fukumoto. 2018. SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In *Proceedings of UIST '18* (Berlin, Germany). Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/3242587.3242603

[3] Google. 2020. About Switch Access for Android. https://support.google.com/accessibility/android/answer/6122836?hl=en

[4] J. Han, L. Shao, D. Xu, and J. Shotton. 2013. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. *IEEE Transactions on Cybernetics* 43, 5 (2013), 1318–1334. https://doi.org/10.1109/TCYB.2013.2265378

[5] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52, 4 (2010), 288–300.

[6] Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. of ISSP* (2008), 365–369.

[7] Chuck Jorgensen and Kim Binsted. 2005. Web browser control using EMG based sub vocal speech recognition. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE.

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

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

[10] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of DIS 2018* (Hong Kong, China). Association for Computing Machinery, New York, NY, USA, 881–894. https://doi.org/10.1145/3196709.3196784

[11] Gierad Laput and Chris Harrison. 2019. Sensing fine-grained hand activity with smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[12] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. Viband: High-fidelity bioacoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 321–333.

[13] MacObserver. 2020. Future Apple Headphones Could Tell Which Ear They're In. https://www.macobserver.com/link/future-apple-headphones-orientation/

[14] 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 MobiSys 2018* (Munich, Germany). Association for Computing Machinery, New York, NY, USA, 269–282. https://doi.org/10.1145/3210240.3210322

[15] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha. 2013. Unsupervised random forest manifold alignment for lipreading. In *Proceedings of the IEEE International Conference on Computer Vision*. 129–136.

[16] Raquel Aparecida Pizolato, Frederico Silva de Freitas Fernandes, and Maria Beatriz Duarte Gavião. 2011. Speech evaluation in children with temporomandibular disorders. *Journal of Applied Oral Science* 19 (2011).

[17] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: Earphones as a Teeth Activity Sensor. In *Proceedings of MobiCom 2020* (London, United Kingdom). Association for Computing Machinery, New York, NY, USA, Article 40, 13 pages. https://doi.org/10.1145/3372224.3419197

[18] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of CHI EA 2017* (Denver, Colorado, USA). Association for Computing Machinery, New York, NY, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[19] C. M. Reed, W. M. Rabinowitz, N. I. Durlach, L. D. Braida, S. Conway-Fithian, and M. C. Schultz. 1985. Research on the Tadoma method of speech communication. *The Journal of the Acoustical Society of America* 77, 1 (1985), 247–257. https://doi.org/10.1121/1.392266

[20] J. Tan, C. Nguyen, and X. Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 1–9. https://doi.org/10.1109/INFOCOM.2017.8057099

[21] UCLA. 2020. UCLA Phonological Segment Inventory Database. http://web.phonetik.uni-frankfurt.de/upsid.html

[22] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni. 2016. We Can Hear You with Wi-Fi! *IEEE Transactions on Mobile Computing* 15, 11 (2016), 2907–2920. https://doi.org/10.1109/TMC.2016.2517630

[23] Yunfan Zhang, Hui Li, Shengnan Shen, Guohao Zhang, Yun Yang, Zefeng Liu, Qisen Xie, Chaofu Gao, Pengfei Zhang, and Wu Zhao. 2019. Investigation of Acoustic Injection on the MPU6050 Accelerometer. *Sensors* 19, 14 (2019), 3083.