Tongue-in-Cheek: Using Wireless Signals to Enable Non-Intrusive and Flexible Facial Gestures Detection

Mayank Goel[1], Chen Zhao[2], Ruth Vinisha[2], Shwetak N. Patel[1,2]

[1]Computer Science & Engineering | DUB Group
University of Washington
Seattle, WA 98195 USA

[2]Electrical Engineering | DUB Group
University of Washington
Seattle, WA 98195 USA

{mayankg, chzhao, vinisha, shwetak} @uw.edu

ABSTRACT

Serious brain injuries, spinal injuries, and motor neuron diseases often lead to severe paralysis. Individuals with such disabilities can benefit from interaction techniques that enable them to interact with the devices and thereby the world around them. While a number of systems have proposed tongue-based gesture detection systems, most of these systems require intrusive instrumentation of the user's body (*e.g.,* tongue piercing, dental retainers, multiple electrodes on chin). In this paper, we propose a wireless, non-intrusive and non-contact facial gesture detection system using X-band Doppler. The system can accurately differentiate between 8 different facial gestures through non-contact sensing, with an average accuracy of 94.3%.

Author Keywords

Tongue-computer interface; Tongue gestures; Wireless signals; Accessibility.

ACM Classification Keywords

H.5.2. Information interfaces and presentation: User Interfaces – *input devices and strategies.*

INTRODUCTION

Motor neuron diseases such as Amyotrophic Lateral Sclerosis (ALS), as well as serious brain or spine injuries, often cause severe paralysis. Many of these individuals retain strong cognitive abilities, and half of them are between the ages of 16 and 30 years [1]. The debilitating effects of these injuries and diseases impede an otherwise able person from fully participating and engaging with certain aspects of their environment, such as computing technologies. Therefore, there is great value in empowering these individuals by helping them communicate with and control their environments.

Many brain injuries, spinal injuries, and motor neuron diseases that lead to paraplegia, do not however affect the cranial nerves [3]. These nerves control various facial organs such as eyes, jaws, tongue, and cheeks. While the eyes and tongue have been used extensively by researchers for building accessible interaction systems [1,2,3,4,7], many other facial organs also offer advantages for robust and non-intrusive gestural interaction. For example, due to their utility in chewing and vocalization, jaws offer low perceived exertion and, unlike the eye, produce no interference with the user's visual activity. Moreover, most of the tongue-based gesture recognition approaches require varied levels of on-body instrumentation; including magnetic piercings in the tongue [1], dental retainers [3], and an array of eight sEMG sensors attached on the user's face and chin [4,7]. The more invasive technologies, such as the tongue piercing, offer very fine-grained tracking of the user's tongue [1] and can be used for many advanced applications such as controlling wheelchairs or handwriting recognition. However, in cases where the user needs simpler input and does not require fine-grained tracking of tongue, sensing can be significantly less intrusive. Moreover, in many cases, these individuals can still use other parts of their body controlled by the cranial nerves, such as their jaws and cheeks [5]. A system that could non-intrusively track movement of all these body parts can provide additional flexibility for users. The users can then easily switch between the different ways of gesture input depending on their preferences and exertion over time.

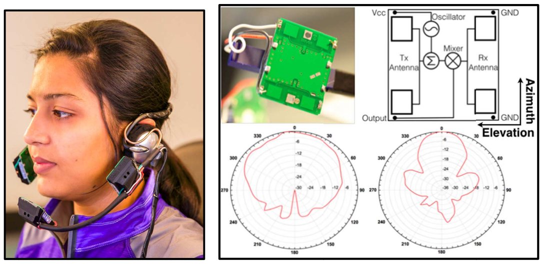

Figure 1. ***(Left)*** Tongue-in-Cheek is a non-contact facial gesture detection system. It is integrated into a pair of off-the-shelf headphones. ***(Right)*** The motion sensor module and Azimuth (bottom-left) and elevation (bottom-right) of the antenna beam pattern.

In this paper, we present *Tongue-in-Cheek* (Figure 1), a system that uses 10 GHz wireless signals to detect different facial gestures in four directions. It detects the movement of cheeks caused by moving different parts of the mouth:

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
Copyright 2015 ACM 978-1-4503-3145-6/15/04…$15.00
http://dx.doi.org/10.1145/2702123.2702539

touching of tongue against the inside of the cheeks, puffing the cheeks, and moving the jaws. Tongue-in-Cheek places three small, inexpensive 10 GHz Doppler radar units around a user's face (Figure 1) and measures the Doppler shifts caused by fine movements of the cheeks. Unlike earlier systems, these sensors do not need to be in contact of the user's skin and require no intricate calibration or placement. Tongue-in-Cheek easily integrates with existing headphones and treats motions from the tongue, cheeks, and jaws as the same so that the user can seamlessly switch between these three body parts. This feature also makes the system adaptable to individuals with different variations of paralysis. For example, the system used by an individual with facial paralysis can seamlessly be used by another individual with tongue paralysis.

We evaluated our system design and gesture set with 3 individuals with facial paralysis and 2 of the 3 participants felt Tongue-in-Cheek was perfect suited to their needs. The accuracy of Tongue-in-Cheek was evaluated in a controlled study of 8 participants. The system was tested for eight different gestures performed through three methods: tongue movement, cheek puffing, and jaw movement. Our findings show that Tongue-in-Cheek differentiates between different gestures in four directions: *up, down, left,* and *right,* with an average accuracy of 94.3%. The system requires minimal calibration from the user. It simply checks whether the user has worn the headset properly by asking them to perform one gesture in each of the four directions. On an average, participants took 10.2 seconds to adjust the headset. This adjustment is not significantly different than adjusting one's headphones. Lastly, we also tested the effectiveness of Tongue-in-Cheek for text entry using EdgeWrite [6] and for playing video games.

DESIGN OF TONGUE-IN-CHEEK

The Tongue-in-Cheek prototype consists of three small microwave motion sensor modules. These sensors are attached to a pair of off-the-shelf headphones such that that each sensor covers one of the three sides: *left, right,* and *bottom* (Figure 1, *Left*). The outputs of the sensors are connected to a National Instruments USB-6009 data acquisition unit (DAQ). The unit takes voltage samples at 48 kS/s and digitizes them with 14-bit resolution. The system samples each of the three sensors at 16 kS/s. The output of the DAQ was connected to a computer through USB, where it was processed in MATLAB.

Motion Sensors

The microwave motion sensor modules used in this prototype are X-band bistatic Doppler transceiver modules (Parallax 32213). The module's built-in dielectric resonator oscillator uses a pair of transmitting patch antennas (Figure 1, *Right*) to transmit a 10.5 GHz waveform with a peak power of 5 mW. Another pair of antenna receives the microwave energy reflected by the objects in front of the module. If the object moves, the received frequency is shifted away from the transmit frequency of 10.5 GHz due to the Doppler effect. This frequency-shifted received signal is mixed with the transmitted signal to obtain a low-frequency voltage representing the amount of frequency shift.

Apart from being inexpensive ($4 USD) and easily available, these sensors offer some unique advantages for our purposes. Considering that they operate in the microwave region, the sensing is immune to the effects of temperature (within normal bounds of temperature variation), acoustic noise, light, *etc.* Their low output power also prevents any harm to the human body. These modules are not FCC approved but comply with FCC Part 15 Rules and Regulations. Additionally, our system requires sensors that can be placed in close proximity of the face and can detect subtle facial movements The antennas on these modules are highly directional and have almost uniform beam patterns from -60° to 60°. If the sensors are placed in close proximity of the face, the 60°conical radiation pattern of the antenna in front of the sensors covers the majority of user's cheeks and chins. The system also needs to ensure that the modules do not 'jam" each other as they are running at same frequency. It presents a simple UI to guide the user to adjust the modules such that they are close to the user's cheeks and their face occludes the opposing modules in such a way that the side lobes of the modules do not interfere with one another.

Gestures

Tongue-in-Cheek supports gestures in four directions: *left, right, up,* and *down* and two modes: *tap* and *hold.* In case of *tap* gestures, the user moves the body part with which they want to perform the gesture and then retract to the rest position. For *hold* gestures, the user "holds" the gesture instead of retracting back to the resting position after a tap. All the (both *tap* and *hold*) gestures can be performed either by touching the tongue against the inner side of the cheeks or moving the lower jaw, or puffing the cheeks. In case of puffing of cheeks, the *up* and *down* gestures require the user to puff the upper lip and lower lip portion, respectively. All these gestures result in a, sometimes subtle, motion of cheeks.

We aimed to make our device flexible so that the user could

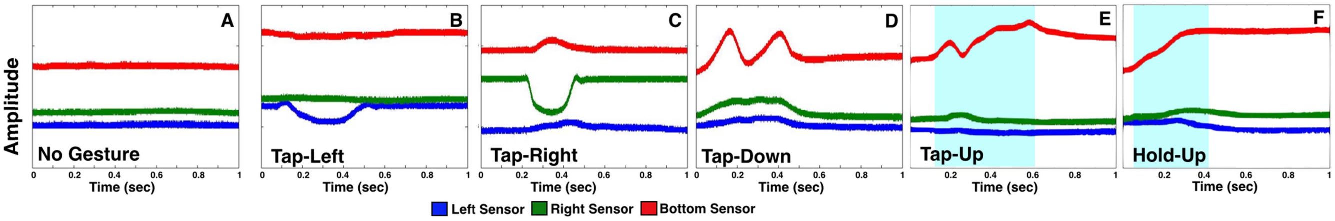

Figure 2. (A-E) Output signal from the each of three microwave Doppler sensors for no-gesture, 4-directional tap gesture. **(F)** *hold-down* gesture. The highlighted part in **E** and **F** shows the difference between a tap and a hold gesture.

switch between different interaction modes. This was motivated by two observations: (1) cranial muscles are not trained for gestures and have a tendency to get tired after prolonged use, and (2) different individuals have different conditions. For example, in our evaluation one participant could not use their tongue for interaction but could easily puff the cheeks. This goal was aided by our choice of sensing approach. The motion sensors used in our prototype sense the direction of the motion. The system is only aware of the direction in which the gesture was performed and not the way the user performed it. This method helps the user to seamlessly switch between the gesture modes without the need to retrain or reconfigure the system. We carefully selected our gesture set to make sure that they are simple and modular. The gestures can be performed in all four directions and the differentiation of *tap* and *hold* means that the gestures can be combined to perform more complicated interactions. The video figure shows one such example for cursor control. The modularity of the gestures helps in overcoming the limitation of discrete gesture detection.

The placement of three sensors around the face enables a simple detection of motion in three directions: left, right, and down. A gesture in one direction changes the signal in the corresponding sensor significantly more than the rest. We do not have a sensor in the *"up"* direction, which adds some complexity. Ideally, the upward motion of the skin would generate a reverse Doppler shift detected by the bottom sensor. However, this simple model does not accommodate the complexity of facial muscle structure; when a user touches their upper lips with their tongue, the entire area around the mouth moves, and all three sensors experience deviations in their low-frequency components. The bottom sensor experiences the most deviation, but not necessarily in the opposite direction as in the case of the *"down"* gesture (Figure 2). In the both the *"up"* and *"down"* gestures, the skin moves out (*i.e.,* away from the user's body) more than it moves up. However, our empirical experiments show that the signal is different for each of the four gestures and can be separately classified using standard machine learning techniques.

Figure 2F shows an example of the observed low frequency signal for a *hold-up* gesture. This signal for bottom sensor is very different from the *tap-up* gesture shown in Figure 2E. In case of *hold*, the magnitude does not oscillate because the muscles do not retract to resting position immediately.

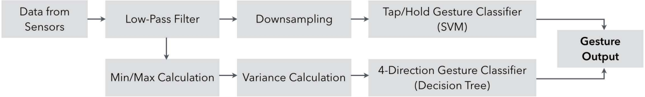

Figure 3. Tongue-in-Cheek's algorithm

Algorithm

The output of the microwave motion sensors is a baseband signal representing the amplitude of the Doppler shift. An analog to digital converter (ADC) first digitizes the output from the motion sensors. Then a third order Butterworth Filter low-pass filters the data up to 15 Hz (Figure 3). After this we use two different processing pipelines to differentiate between *tap* and *hold,* and the four directions. The outputs from both the classifiers are combined in the end to form eight different gestures: *tap-left, tap-right, tap-up, tap-down, hold-left, hold-right, hold-up,* and *hold-down.*

Tap vs. Hold Classification

Figure 2 shows that the signal looks substantially different for the *tap* and *hold* gestures. We use a support vector machine (SVM) classifier here and the features are calculated across 1 sec windows with a 0.8 sec overlap. The overlap of 0.8 sec assumes that the user would not perform two gestures within 0.2 sec. We first use mean filtering to downsample the low-pass filtered signal into 20 samples. These 20 data points from each of the three sensors gives us a total of 60 features for the SVM-based model. This particular processing pipeline does not do any segmentation. It simply calculates features across window. The second processing pipeline that classifies the gesture into one of the four directions performs the segmentation.

Direction Classification

Figure 3 shows that while differentiation between *left, right,* and *down* is very clear, the case of *down vs. up* is not that straightforward. Therefore, we cannot use a rule-based system and use decision trees to classify between the four directions, instead. We also added a class to represent when no gestures is being performed. This helped us in robust segmentation of real-time data. We calculate 3 features for each gesture: minimum, maximum, and variance of each 1 second window. The variance is the most important feature in this classifier because it is clear from Figure 3 that the amount of change in signal is a big differentiating factor. The minimum and maximum features help in overcoming the noise in the data and provide dynamic thresholding to the decision trees. All the features are calculated over 1 second window with a 0.8 seconds overlap.

QUALITATIVE EVALUATION

We evaluated our choice of sensing mechanism and gesture-set by demonstrating Tongue-in-cheek to three individuals with neuromuscular conditions. Two of the participants tried the system for almost 10 minutes each, and commented on the performance and the design. The third participant could not use the system personally and commented on the overall design and utility. P1 could not open his jaw far enough to instrument the inside of the mouth easily. Another side effect of his condition was tongue fasciculation – minor but constant involuntary muscle fluctuations across the tongue. These fluctuations were large enough to make tongue tracking hard, but the cheeks were relatively stable. Therefore, Tongue-in-Cheek was a relatively better approach for him as compared to [1,3]. P2 preferred switching between different modes and said, "the fact that you support several avenues of interaction is great." P3 actually preferred using a magnet attached to her tongue for more fine-grain control, but saw the appeal for people not wanting to place a magnet in their mouth [1]. Tongue-in-Cheek could be useful for a subset of the population due to its flexibility and non-intrusiveness.

TECHNICAL EVALUATION & RESULTS

To evaluate the gesture detection accuracy, the users had to perform repeated sets of gestures and our participants from qualitative evaluation had limited availability; hence we recruited eight participants (3 females) who did not have any neuromuscular condition. They all used the same pair of modified headphones. The initial adjustments were not very different from the kind a user makes when using a pair of headphones, *i.e.,* adjusting the size of the headband and adjusting the angle and reach of the microphone.

For each gesture class, participants had to perform 20 gestures. The order of the gesture classes was random. The participants were told that they could perform the gestures by either using their tongue, moving their lower jaw, or by puffing their cheeks.

The system performs leave-one-out cross validation and the system never uses the data for the same participant in training as well as testing. The models are therefore global and do not need to be adjusted for each user.

The system was able to correctly predict the direction of the gesture with 94.3% accuracy. The confusion was greatest between the *up* and *down* gestures. This was expected since both of these gestures depend heavily on the variation in output of the sensor placed under the user's chin. The accuracy for differentiating between *tap* and *hold* was 97.4%, and the segmentation accuracy for the system was 93.6%.

Tongue-in-Cheek is agnostic to the source of movements, *i.e.,* tongue, cheek, or jaw, and we did not formally evaluate how the system performed for these different movements. The participants were told that they could use the three motions independently.

APPLICATIONS

We applied Tongue-in-Cheek-based input to two different applications. We recruited three of our original eight participants to evaluate the applications (See video figure).

EdgeWrite

Inputting text on computing devices is an important capability. We integrated EdgeWrite [6] into our system to test if Tongue-in-Cheek could be used for text. Each participant was asked to enter three simple English language phrases. All three participants were able to correctly enter the three phrases with an average text entry rate of 9 wpm.

Video Games

In order to test the responsiveness of our gesture detection system, we asked the participants to play two different games: Pac-Man and Contra. All participants were able to play the games and expressed satisfaction with the system's responsiveness.

These are just two example applications for Tongue-in-Cheek. Our evaluation has shown that the system can reliably detect gestures in four directions and can be useful for many other general-purpose applications.

DISCUSSION AND LIMITATIONS

Systems like Tongue-in-Cheek that are designed for users with neuromuscular conditions face an important challenge. In most cases, such systems require the user to change their lifestyle and go through training. Hence the users should be able to use the device properly and actually enjoy using it. Tongue-in-Cheek detects modular, and directional gestures with no contact with the body. However, our interviews highlighted that different individuals have different needs and preferences. For example, Tongue-in-cheek is less intrusive, but it is more visible than [1,3].

Near infrared (NIR) sensors can also be used for proximity and motion sensing. Apart from being susceptible to dust and light, NIR is less immune to physical properties of the reflecting surface, hence might work differently for bearded or clean-shaven user. On the flip side, RF is more prone to ambient motion and headset movement. While this limitation can be largely countered with an Inertial Measurement Unit, we believe that our algorithms will work similarly for both the sensors and the final decision on sensors will depend on the user's preference.

CONCLUSION

Motor neuron diseases such as ALS as well as serious brain or spine injuries often result in severe paralysis. Patients can benefit from systems that allow them to control their environment non-intrusively. We present a facial gesture detection system that robustly detects various facial gestures in four directions. These gestures can be performed using the user's tongue, cheeks, or jaws. This capability allows the user to seamlessly switch between different gesture modalities and potentially feel less exerted.

REFERENCES

1. Huo, X., Wang, J., and Ghovanloo, M. A Magneto-Inductive Sensor Based Wireless Tongue-Computer Interface. *Neural Systems and Rehabilitation Engineering 16*, 5 (2008).
2. Lukaszewicz, K. The Ultrasound Image of the Tongue Surface as Input for Man/Machine Interface. *Proc. INTERACT '03.*
3. Saponas, T.S., Kelly, D., Parviz, B. a., and Tan, D.S. Optically sensing tongue gestures for computer input. *Proc. UIST'09.*
4. Sasaki, M., Arakawa, T., Nakayama, A., Obinata, G., and Yamaguchi, M. Estimation of tongue movement based on suprahyoid muscle activity. *Proc. IEEE Engineering in Medicine and Biology Society. 2013.*
5. Wilson-Pauwels, L., Akesson, E.J., and Stewart, P.A. *Cranial Nerves: Anatomy and Clinical Comments*.1988.
6. Wobbrock, J.O., Myers, B.A., and Kembel, J.A. EdgeWrite: A Stylus-Based Text Entry Method Designed for High Accuracy and Stability of Motion. *Proc. UIST 2003.*
7. Zhang, Q., Gollakota, S., Taskar, B., and Rao, R.P.N. Non-Intrusive Tongue Machine Interface. *Proc. CHI'14.*