Non-Intrusive Tongue Machine Interface

Qiao Zhang, Shyamnath Gollakota, Ben Taskar, Rajesh P. N. Rao
University of Washington
{qiao, gshyam, taskar, rao}@cs.washington.edu

ABSTRACT
There has been recent interest in designing systems that use the tongue as an input interface. Prior work however either require surgical procedures or in-mouth sensor placements. In this paper, we introduce TongueSee, a non-intrusive tongue machine interface that can recognize a rich set of tongue gestures using electromyography (EMG) signals from the surface of the skin. We demonstrate the feasibility and robustness of TongueSee with experimental studies to classify six tongue gestures across eight participants. TongueSee achieves a classification accuracy of 94.17% and a false positive probability of 0.000358 per second using three-protrusion preamble design.

Author Keywords
Electromyography (EMG); Tongue Gesture Interface

ACM Classification Keywords
H.5.2 [User Interfaces]: Input devices and strategies

INTRODUCTION

Researchers have recently explored the feasibility of using the tongue as an assistive interface, to help millions of patients [2] suffering from conditions such as tetrapelgia and Amyotrophic Lateral Sclerosis. Prior approaches, however, require devices to be inserted within the mouth to directly sense the tongue motion. Solutions such as the Tongue Drive [1] extract rich tongue gestures by instrumenting the tongue with magnetic piercings; such solutions are however unlikely to extend to the general user base. Identifying this challenge, researchers have proposed detecting tongue gestures using optical sensors placed on a tongue retainer that is placed in the mouth [3].

We introduce TongueSee, a non-intrusive tongue gesture recognition system that requires neither surgical procedures nor inconvenient in-mouth sensor placements. TongueSee can achieve classification accuracies similar to, if not better than, prior sensor placement approaches [3]. TongueSee leverages the basic observation that different tongue movements often trigger slightly different electrical activity on the surface of the skin at the chin and the throat. TongueSee measures these signals non-intrusively using surface electromyography (sEMG) sensors that measure the gross electrical activity produced during muscular contractions.

The challenge, however, is that the muscles connecting the tongue to other structures in the mouth are hidden under facial muscles and blood vessels. Thus, sEMG sensors typically do not directly capture the electrical signals from the tongue muscles. Instead, they measure the EMG signals responsible for the minute jaw movements corresponding to the tongue gestures. As one would expect, a number of muscles at the lower face and throat are responsible for moving the jaw, and as a result, the sEMG sensors register the cumulative effect of all these muscles. To address this issue, we pick the sensor placement in a physiologically informed manner, to ensure that the sensors are in close proximity to certain muscles to maximize the signals from them. Table. 1 shows the tongue gestures and the corresponding muscle groups that are mainly involved, and the labels of the sEMG sensors covering them.

We are not the first to use EMG signals for gesture sensing. Saponas *et al.* [4] have demonstrated the feasibility of using EMG signals from the forearm to accurately classify finger gestures. Similarly, Sasaki *et al.* [5] use EMG signals to estimate tongue position and contact force, but they have to insert a mouthpiece at the roof of the mouth for training and feedback.

We make the following contributions:

- We introduce a non-intrusive tongue gesture system that classifies six tongue gestures – left, right, up, down, protrude, and rest, with an average accuracy of 94.17%.[1] To do this, we identify EMG sensor positions near the chin and throat, in a physiologically informed manner.
- We describe an anytime classification algorithm that provides a tradeoff between classification accuracy and the real-time responsiveness. We show how to achieve this without retraining the model for different response times.
- We build a TongueSee prototype using the the Biosemi ActiveTwo system, evaluate it across eight participants, and demonstrate its robustness with time. Further, we design a preamble mechanism that allows TongueSee to address the problem of false positive events.

Our work is also related to concurrent work [6] that has close parallels to TongueSee. In contrast, however, TongueSee takes a systematic approach that not only works with a larger gesture set but also addresses key challenges such as false

[1] TongueSee's classification accuracy for the four gestures in [3], is about 96%.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

http://dx.doi.org/10.1145/2556288.2556981

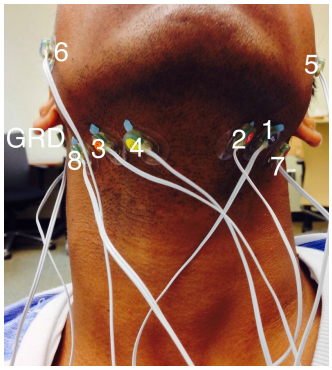

Figure 1. Sensor placements on an experimental participant

Gestures	*Muscle Group*	*Sensors*
Left	Masseter	5,6
Right	Masseter	5,6
Up	Mandibular, Digastric, Omohydroid	1,2,3,4,7,8
Down	Mandibular, Digastric, Omohydroid	1,2,3,4,7,8
Protrude	Digastric, Mylohyoid	1,3,5,6
Rest	-	-

Table 1. Tongue Gestures and Sensor Placements

positives and robustness across sessions, which are crucial in practice. Specifically, our system successfully classified a larger set of six instead of three voluntary gestures, which we achieved by carefully selecting our EMG sensors locations. Further, we design our system to work across sessions over long periods of time (our results show four hours of experiments). Finally, we introduce the preamble design and have determined the appropriate preamble gesture sequence that addresses the problem of false positives. We believe that the above contributions make the experiments, results and algorithmic designs in our paper a significant step towards making this line of work practical.

THE TongueSee SYSTEM

TongueSee leverages EMG signals to reliably classify tongue gestures without requiring any sensor placement in the mouth. We choose the left, right, up, down, protrude, and rest gestures since they are the basic building blocks in any computer input modality. As shown in Fig. 1, our system uses six EMG sensors along a thin strip below the jaw and two EMG sensors covering the Masseter muscles. TongueSee's anytime algorithm can return a classification result at any time (with a resolution of 0.0625 s); the algorithm achieves higher classification accuracies, the longer it runs.

Hardware Equipment and Setup

Our system uses the BioSemi ActiveTwo device to collect EMG signals from the sensor positions shown in Fig. 1. We sample at 2048 Hz from the eight sensors, and use two ground sensors placed behind the ear to reduce line noise and other types of noises. Experimental results are recorded using the BioSemi 2, DummySignalProcessing and StimulusPresentation modules in BCI2000 – BCI2000 is a data acquisition and processing system typically used with EMG signals. The signals are processed with a notch filter at 60 Hz to further remove the line noise and a high pass filter at a low frequency of 0.1 Hz to remove drifts seen in the captured signals.

Feature Extraction

TongueSee extracts two features from each sensor, for a total of 16 features across all the eight sensor locations. The first feature is the root-mean-square (RMS) value of a disjoint window of 128 samples from the sensor. The second is the RMS value of the signed difference between the signal value and mean signal value. Since the sampling rate is 2048 Hz, we get one training point every 0.0625 seconds. Note that our feature extraction mechanism requires minimal computation on the EMG signals, and hence has the advantage of being fast enough to enable real-time testing.

Classification

Our goal is to design an algorithm that can output classification results at any time, while the user performs the gesture. Such an algorithm can enable TongueSee to adapt with the interactivity requirements of the applications. Further, it would provide feedback to the users so that they can continue to perform the gesture until the system gets the classification right. Specifically, we train our multi-class SVM classifier with gaussian kernel using the features extracted from the EMG signals every 0.0625 seconds. However, since the tongue gestures are typically much longer than the above duration, TongueSee performs classification on the test data at coarser resolutions through majority voting. In particular, we divide the test data into non-overlapping intervals of 0.0625 seconds and use our SVM classifier to predict for each interval. To generate a tongue gesture prediction, we apply majority voting to combine the predictions at the fine time scale to coarser time resolutions. Our algorithm thus allows a tradeoff between the interactivity requirements of the application and classification accuracy. Further, our results show that majority voting can significantly increase the accuracy if we combine predictions over long enough durations.

EXPERIMENTS AND RESULTS

We run our experiments with eight volunteers from our research organization. The participants consist of 6 males and 2 females, with an average age of 24.4 ± 1.2, average height of 172.2 ± 9.9 cm, and an average BMI of 23.3 ± 2.8. They were previously inexperienced with EMG sensing technology. In our experiments, they are given a description for the sensor placements, and are instructed to put on the sensors themselves. The sensor placements are therefore approximate, yet effective as shown by our experimental results.

Feasibility of Tongue Gesture Recognition

First, we evaluate the feasibility of using TongueSee to detect and distinguish tongue gestures.

Experiments: We run experiments with all the participants performing six tongue gestures (left, right, up, down, protrude, and rest). Since the participants are unable to replicate sensor placements exactly, we collect the training and test data for each participant separately. In principle, one

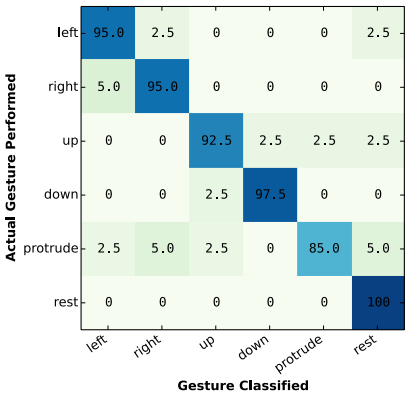

Figure 2. Confusion matrix for tongue gestures: The average accuracy across eight participants for all six tongue gestures is 94.17%. The matrix also shows that the rest action is unlikely to be misclassified.

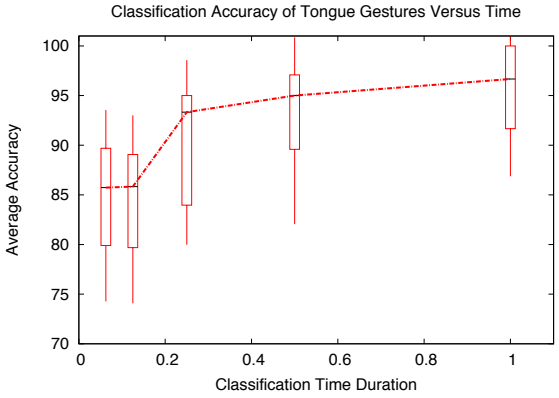

Figure 3. Boxplot of classification accuracy against time resolution for features. The horizontal axis shows the time duration over which the classification was performed. Each candlestick shows the accuracy averaged across all eight participants. The whisker of the box plot shows the error bar (one standard deviation), while the box itself extends from the first quartile to the third quartile of the accuracies. The median is marked in the box.

could integrate the EMG sensors with a chin support band to achieve a more accurate sensor placement and avoid per-participant training; this however is not in the scope of this paper. We run experiments in blocks where each block consists of a sequence of six tongue gesture stimuli that are presented to the participant. Each stimulus consists of a caption shown visually in a 600×480-resolution window on the top left corner of the computer screen. At the beginning of each block, there is a three-second pause for the participants to adjust comfortably for data collection. Each stimulus is then presented for four seconds, with a four-second inter-stimulus pause. During the stimulus presentation, the participants are instructed to perform the appropriate tongue gesture as shown in the captions. They are also asked to hold their tongues in position during the stimulus presentation, and perform the tongue gestures as consistently as possible. During the inter-stimulus pause, the caption screen is blank and the participants are told to relax. We collect 20 blocks of data for each participant; the first 15 blocks are used for training and the last 5 blocks for testing.

Results: We classify the test gestures over a one-second duration using our majority voting mechanism. Fig. 2 plots the confusion matrix computed across participants and the six gestures. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. This matrix shows the extent to which one tongue gesture is mislabeled as another by TongueSee. Our findings are as follows:

- The average classification accuracy across eight participants for all six tongue gestures is $94.17\% \pm 7.29\%$. This is in comparison to an accuracy of 17% with random guesses for six gestures. The average accuracy and standard deviation for these gestures are 94.17% (90.83%) and 7.29% (7.72%) when using the first 15 (10) blocks for training and the last 5 (5) blocks for testing.

- Of the misclassifications, the rest action was the least likely to be confused as tongue gestures. This is expected, since in the absence of tongue motion, the EMG signals do not undergo significant changes and have a low value. As a result, the rest action has a distinct EMG signal pattern that is the farthest from actual tongue motions.

Next, we demonstrate the trade-off between accuracy and real-time responsiveness of our anytime classification algorithm. Specifically, we compute the average accuracy across the eight participants for different classification time durations. In Fig. 3, we plot the accuracy for six tongue gestures against the time duration over which the classification is performed on the test data. The plot shows an increase from 83.91% to 94.17% as the classification time duration increases from 0.0625s to one second. The trade-off for the increase in accuracy is that instead of detecting a tongue gesture every 0.0625 of a second, the system now detects a gesture every second. Hence, depending on the application, we can trade off classification accuracy for the optimal time responsiveness on the system.

TongueSee's Robustness

We next evaluate the robustness of TongueSee in classifying tongue gestures across time.

Experiment: Since TongueSee requires training per participant, it is crucial to ensure that training model is robust over time. To evaluate this, the participant performs the six tongue gestures across five test sessions that are separated in time. Specifically, the training session consists of 15 blocks, each consisting of a sequence of the six tongue gestures. After the training session, the participant performs 5 blocks for the first testing session. The participant then waits for an hour before the second test session of 5 blocks; the third, fourth and fifth sessions are performed, each after an hour gap and consists of 5 blocks each.

Results: Fig. 4 shows the average accuracy over time; the average is computed across the six tongue gestures for a single participant. Each plot in the figure shows the accuracies for six tongue gestures for different time durations over which

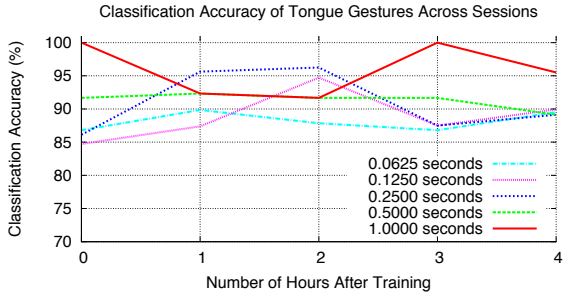

Figure 4. Classification accuracy on the test set is plotted against the number of hours after training for the left and the right tongue gestures. The accuracy is found to be fairly constant over a period of four hours.

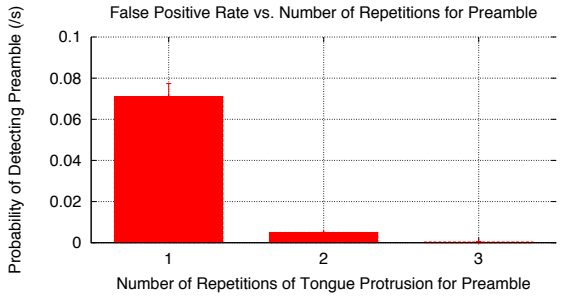

Figure 5. False positive rate. TongueSee uses a repetitive protrusion gesture sequence to gain access to the system. The false positive rate reduces as the number of repetitions in this sequence increases.

the majority voting for classification was performed. The figure shows that the classification accuracies are all above 85% for different time durations used for majority voting. The accuracies are above 90% when one second is used for majority voting. We also note that the one-second accuracies are as high as 100% for a smaller gesture subset, i.e., left and right gestures. The important trend however is that accuracy does not significantly change across time. This shows that TongueSee generalizes over time and is robust across sessions.

False Positives with TongueSee

In practice, users perform a number of involuntary tongue motions that can create false positives in our system. To address this problem, TongueSee uses a repetitive gesture sequence as a starting preamble. Specifically, in TongueSee, users perform a repetitive tongue gesture sequence to gain access to the system. Once TongueSee recognizes the repetitive sequence, the users can perform normal non-repetitive tongue gestures to interact with the device.

Experiments: We compute the false positive rate by running the experiment for a period of 90 minutes. Four participants perform 15 blocks of training, each consists of six tongue gestures. After the training session, the four participants wear the EMG sensors and perform normal activities such as eating, talking, moving his head, and working at the desk for the whole duration of the 90 minutes. Our preamble gesture consists of a repetition of the tongue protrusions. We classify the captured EMG signals for the whole duration to detect the repetitive preamble gesture. We compute the false positive rate as the probability of detecting the preamble gesture each second over the whole 90 minute duration.

Results: In Fig. 5, we plot the false positive rate as a function of the number of repetitions of the protrusion gesture in the preamble. The figure shows that the probability of detecting the preamble per second is 0.071 ± 0.0064, with a single protrusion gesture as the preamble. Using the multiplication rule in probability, the false positive rate decreases with the number of repetitions – it reduces to 0.000358 with three protrusion gestures. We note that if we use three up tongue gestures as the preamble, the false negative rate would be 0.540 ± 0.157 since users are far more likely to move their tongues up during normal activities. These results show that by using three tongue protrusions as the preamble gesture, TongueSee can significantly reduce the false positive rate.

CONCLUSION

We demonstrate the feasibility and robustness of a non-intrusive tongue machine interface that achieves comparable, if not better, classification accuracy for a rich set of six tongue gestures than prior work. Our proposed algorithm trains and classifies at a finer time scale but uses majority voting at a coarser scale, allowing us to maximizes the classification accuracy for different interactivity requirements. Our system can deliver high-fidelity tongue gesture recognition with the potential to help millions of patients.

ACKNOWLEDGEMENTS

We thank Sidhant Gupta, members of UW Networks and Wireless Lab, and the anonymous reviewers for their valuable feedback. We also thank Devapratim Sarma for helping with the experimental set-up. Finally, we thank the experiment participants from our department. This work was partially supported by NSF grant EEC-1028725 and ARO Award no. W911NF-11-1-0307.

REFERENCES

1. Krishnamurthy, G., and Ghovanloo, M. Tongue drive: A tongue operated magnetic sensor based wireless assistive technology for people with severe disabilities. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, IEEE (2006), 4–pp.
2. Lee, B., Cripps, R., Fitzharris, M., and Wing, P. The global map for traumatic spinal cord injury epidemiology: update 2011, global incidence rate. *Spinal cord* (2013).
3. Saponas, T. S., Kelly, D., Parviz, B. A., and Tan, D. S. Optically sensing tongue gestures for computer input. In *UIST '09* (2009), 177–180.
4. Saponas, T. S., Tan, D. S., Morris, D., and Balakrishnan, R. Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces (2008).
5. Sasaki, M., Arakawa, T., Nakayama, A., Obinata, G., and Yamaguchi, M. Estimation of tongue movement based on suprahyoid muscle activity (2011). 433–438.
6. Sasaki, M., Onishi, K., Arakawa, T., Nakayama, A., Stefanov, D., and Yamaguchi, M. Real-time estimation of tongue movement based on suprahyoid muscle activity. In *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE* (2013), 4605–4608.