Directional Silent Speech Recognition with Jaw Motion

Aditi Ravindra, Helen Liu, Rachel Jee

Introduction

The directional silent speech recognition project aims to create an ear-worn system for identifying unvoiced human commands that is an alternative to voice-based interactions. The specific voice commands that the project focuses on classifying correctly are directional based, using the four primary directions of “left”, “right”, “up”, and “down”, for data collection and testing. The core idea behind this is that with an IMU sensor placed on the lower jaw, the acceleration and gyroscopic signals are used to break down the articulation of a word and further used to determine the directional command the user is saying via jaw motion data without relying on any verbal or audio noise. The importance of this project comes from the drawbacks of voice-based interactions, which can be unreliable in noisy environments, disruptive to one’s surroundings, and potentially compromise one’s privacy. Silent speech recognition overcomes these issues while also maintaining the usability of voice-based interaction and communication, and additionally allowing people with speech disorders or others who cannot produce any sound but can still articulate words with their jaw to also take advantage of voice-based-like operations.

Our group decided to focus on recognizing the four directional commands specifically because the directions could be used for many different further applications if the project were to be expanded on, meaning it can be easily scalable and offers flexibility for future steps. For example, it could be used for anyone to play *Dance Dance Revolution*, open doors for commanding toy vehicles without a remote, moving a paint brush for digital painting, amongst many other potential uses.

After reading the *MuteIt* paper that gave its audience a basic understanding of how a user’s jaw motion can be broken down into articulation for each word to its constituent syllables and phonemes, our group was intrigued by the concept of recognizing speech solely with jaw motion sensors. We were inspired by the benefits of silent speech recognition compared to voice-based interactions and became drawn to the *MuteIt* system’s purpose for the greater good of users and positive impact it would have on society, which led us to work on a jaw motion project with a similar directive.

Project Overview

The project set-up involved using a single IMU sensor and attaching it to the user’s lower jaw. When the physical set-up was completed, data collection proceeded. With similar code from assignments 1, 2, and 3, the accelerometer and gyroscope data was collected with the user saying each direction multiple times and labeled with the respective direction. After data collection, the data was then normalized and processed, and used to train a Support Vector Machine classifier. As for evaluation, the classifier was evaluated with a train-test split, and our main performance metric for the classification was accuracy between actual and predicted labels. After making sure the accuracy of the classifier was sufficient enough, which is explained further in the report, we implemented real-time prediction by collecting data from the sensor and normalizing the data live while the user was speaking, and returning the predicted word for end-to-end integration of silent speech recognition of “up”, “down”, “left”, and “right”.

Project Implementation

To collect data from our hardware we have the same configuration from Assignments 1 and 2 from class. This included the use of Python to collect data, convert it to a csv file, and train our model. Some of the libraries we used were time, sys, threading, pickle, numpy, os, pandas, and serial.

For our physical setup, we attached an IMU sensor to the jawline to collect data from jaw movement (refer to the image below). From the sensor, we collected acceleration and gyroscopic data for 4 seconds per word. This was used to train and test our model.

Figure 1: Initial IMU Sensor placement

Figure 2: Final IMU Sensor placement

Training + Testing

Once data was collected (25 csv files for each direction), we normalized all the training and testing data. When needed, we sliced the data to have the same size and padded it with zeros. This was to increase the accuracy of our model.

To train our model, we used SVM (Support Vector Machines). Essentially, SVM is a supervised learning method that classifies data through regression and outlier detection. Using a kernel, this method transforms low-dimensional input to a higher one. The goal is to set a hyperplane between 2 classes such that there is a maximum margin between support vectors. Using the pickle library, we stored our SVM model to use for testing.

After training, we implemented an end-to-end system. Instead of a sliding window, we looked at the data every 3 seconds to determine the word. Like the training data, the testing data was normalized and sliced before being fed to our model. To prevent an infinite stream, we added an adjustable time limit for the duration of our system. Every four seconds the prediction is printed and will tell the user to begin their next word.

Challenges

The process itself for developing this ear-worn system for recognizing unvoiced human commands posed several challenges throughout the entirety of the process, from data collection all the way to testing. As mentioned above, the primary goal of this project was for our model to be able to accurately identify directional words, specifically "up", "down", "left", and "right" based on the movement of the jaw as captured by the IMU sensor. The initial testing that we did revealed that there were some issues that resulted from poor sensor placement on the jaw, unclear word articulation, and the model having difficulty distinguishing between certain pairs of commands due to similar jaw movements. To overcome these obstacles, we went through a pretty repetitive process of data collection, refining details with each iteration to improve the accuracy of our model.

Sensor Positioning

One of the most significant challenges we faced initially revolved around finding the optimal placement of the IMU sensor on the jaw to allow it to effectively capture the jaw motion data. In our initial round of data collection, we decided to place the sensor based on information and diagrams we found in the paper we read in class: directly under the ear, at the very top of the jaw. However, this placement of the sensor turned out to be inadequate as it captured very minimal jaw movement associated with the unvoiced commands. In fact, in our first testing iteration, our SVM classifier had an initial accuracy of 44.235%, which was significantly lower than our target.

To address this, we decided to go back to the data collection phase, and consider the placement of the sensor. Though through much deeper exploration it would be possible to complete the project by placing the sensor at our original position, after speaking with the professor, we decided to place the sensor a bit lower on the jaw. We experimented with different placements along the jawline, and settled on placing around the middle of the jaw, where the sensor would be able to reliably detect the directional changes in the jaw motion as we tested the different commands. This one adjustment alone actually led to a relatively significant improvement in our data quality and played a crucial role in the accuracy boost we had in our second iteration of our data collection phase.

Word Enunciation

In addition to the placement of the IMU sensor, another critical change we made that helped us substantially improve our SVM classifier's accuracy was the enunciation of the commands during our data collection phase. The first time we collected our data, we didn't create any emphasis on enunciating the word clearly in order to "exaggerate" the jaw movements. The words were spoken in a relatively minimalistic manner, with minimal articulation. This resulted in data that lacked distinct features between the different words we were testing, i.e. the CSVs containing the sensor data for all four words looked extremely similar. Without having the data look different, our model was unable to correctly distinguish between the different words in our dictionary; this issue contributed to our original accuracy of 44.235%.

To respond to this issue, we focused on consciously enunciating each of the four directional commands, emphasizing the movement of the mouth and jaw, as well as trying our best to maintain consistency with each word. To go into a bit more detail, the jaw/mouth motion we had for each word in this second iteration was as follows:

1. Down: Start with the mouth a little wide; the jaw moves down and then out
2. Right: The lips/mouth move out and then down; the mouth widens at the end
3. Up: The jaw moves up/down a lot; the mouth width stays relatively consistent
4. Left: Quick up/down motion with minimal jaw movement.

This allowed our model to be able to detect the distinctions between each of the 4 commands more clearly and distinguish between them more accurately. This change, in unison with our new sensor placement, resulted in a 90.265% accuracy, over double our first iteration, of our SVM classifier during our second data collection phase. The confusion matrix for this second data collection phase is pictured in Figure 3.

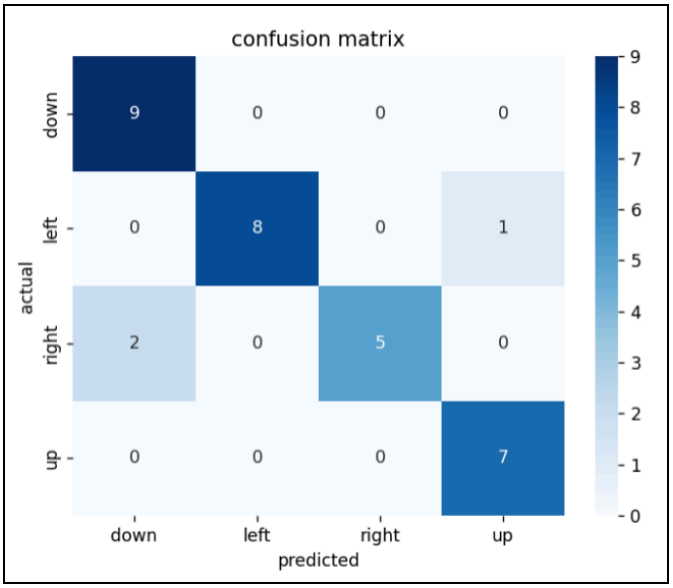

Figure 3: Confusion matrix for 2nd data collection phase: 90.265% accuracy.

Final Iteration: What Works and What Did Not

Though our model's accuracy dramatically improved, we noticed that the model still had some notable difficulty distinguishing between the "down" and "right" commands. We predicted that this was more than likely due to the similarity in the combination of the down and outwards movement in saying both of these words. To address this issue, we went back to our data collection phase and refined our enunciations of the word "right". We introduced a quicker, more deliberate outward motion, with a little less up/down movement. Our goal in doing this was to minimize the confusion the model had between "right" and "down". This adjustment, once again, yielded a noticeable improvement and raised our system's accuracy to 93.750%. The confusion matrix for our third and final data collection iteration is pictured in Figure 4.

Our final model was able to (relatively) accurately distinguish between the four commands (up, down, left, right) in our dictionary, with minimal errors. One last error that we were unable to resolve, though we tried many times, was the model's difficulty in differentiating between "left" and "right". We predict that this is because both commands involve relatively rapid and subtle jaw movements. Our model is able to correctly classify and predict these commands most of the time, however, there are a few occasional misclassifications that occur.

Evaluation and Demonstration

We evaluated the accuracy of our ear-worn system for recognizing the words in our dictionary (up, down, left, and right) on a combination of quantitative metrics and live testing. Our quantitative metrics relied on the confusion matrices and accuracy percentages that our SVM classifier outputted. Our live testing focused on testing various sequences of the four words. This blend of evaluation types allowed us an in-depth insight on the model's robustness and reliability under more "real-life" conditions.

SVM Accuracy

The confusion matrix essentially represents how well the model classified the different classes. In this situation, we have 4 classes, which correspond to the 4 movements of the sensor: up, down, left, and right.

The rows represent the actual class labels and the columns represent the predicted labels. The diagonal values represent the # of correct predictions the model made (i.e. a movement to the right by the sensor was correctly classified as moving to the right). The values that are not on the diagonal represent the # of misclassified predictions.

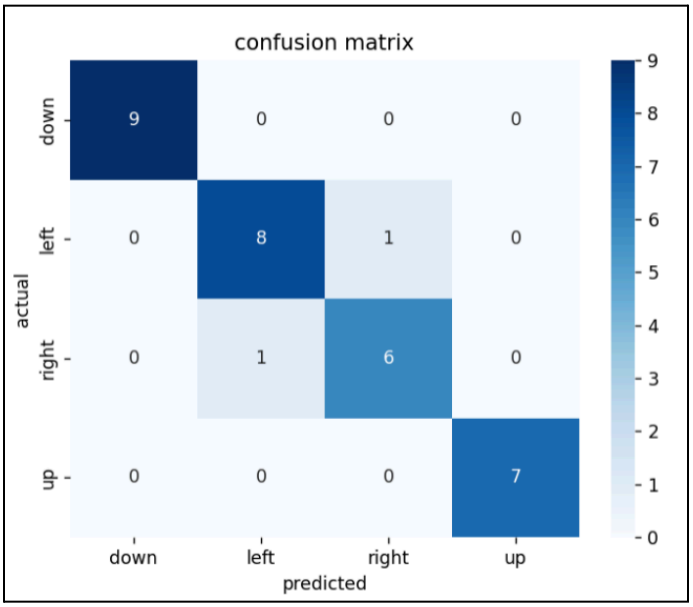

Figure 4: Final confusion matrix: 93.750% accuracy

In our iteration of testing, our SVM classifier performed with a 93.750% percent accuracy as mentioned above. Our data set consisted of 25 CSVs for each of the 4 commands and we used a 30/70 split for our train and test sets. The model incorrectly predicted one instance of our tester saying “right” as “left”, as well as one instance of our tester saying “left” as “right”, which highlighted that the model had some trouble differentiating between these two words.

Live Testing

To test our system's robustness and to evaluate its functionality in more practical scenarios, we tested it's live classification accuracy through using a variety of word sequences.

1. Repeating each individual command ("up", "down", "left", "right") to verify that the model can consistently recognize each word.
2. Testing alternate sequences of words to evaluate the model's ability to differentiate between commands that were frequently confused: up/down alternating sequence and right/left alternating sequence.
3. Testing more complex patterns using various sequences that included all four words.

These live tests illustrated the system's relative reliability. While there was some occasional confusion between "left" and "right", as demonstrated by the confusion matrix in Figure 4 as well, the system consistently had a high accuracy throughout the range of testing scenarios.

Demonstration

The following link (https://drive.google.com/file/d/1WZ5CPIfOSKI4jIV4Uth8mMBpRspeGYeQ/view?usp=sharing) directs

to the video demonstration of our project's real-time prediction in action. In the video, the speaker cycles through directions, saying "left", "right", "up", and "down" repeatedly. The real-time classification showing in the terminal returns the predicted direction based on jaw motion data, and signals the user to say their next direction. The classified labels returned in the video match what the speaker is saying, showing that our end-to-end system successfully reads the sensor data and uses the trained classification to accurately predict the direction the user is articulating.

GitHub Repository

The github repository of our code and data for implementation of this project can be found here: https://github.com/helenliuf/silent-speech.

Conclusion

Thus, the directional silent speech recognition project is a developed end-to-end integration of real-time prediction of unvoiced directional human commands ("left", "right", "up", "down") with jaw motion data collected with a jaw-worn IMU sensor, allowing them to be identified without relying on an audio input. By correctly classifying the directional commands, the project addresses the limitations of voice-based systems like decreased performance in noisy environments, the potential disruptiveness to others, privacy concerns, and accessibility for people who cannot produce sound, so that these commands can be recognized solely with jaw motion. After different trials of data collection and refinement in sensor placement and word enunciation, the project achieved the final accuracy of 93.75%. Live testing confirms that the end-to-end system is for the most part reliable, even though there are some occasional misclassifications.

Limitations

We recognize that there are some limitations to our project. We particularly found misclassifications occurring when words were pronounced off, meaning that there is not a lot of room for human error when saying the directions with where our project is at the moment. Thus, the user has to enunciate their directions very specifically, and the classification might be off if they are not enunciating clearly for the jaw motion to be profound enough to find differences in the acceleration and gyroscope data for classification. This could come from our enunciation in data collection, where we recollected data with more pronounced jaw movement. This error could be reduced if more data was collected, even with less clear enunciation. The lack of difference in jaw motion also brings about the model sometimes struggling to differentiate between "left" and "right" jaw movements because they are both rapid jaw movements. Adding timestamps to data collection and utilizing them for analysis could improve the model if it also takes time into account for the movement, reducing a bit of the reliance on speed at which words are said to identify them.

Future Applications

As mentioned in the introduction, this project can be easily scalable and applied to many other projects for next steps. The directional commands recognized by the system can be seen and used for various industries, like gaming, where a user can control movements without verbal or physical controllers or participate in games in a non-verbal and hands-free way. Creative tools can also make use of the directional commands, where users can navigate a paintbrush or pencil in digital art. There is also robotic control, allowing the user to command toy vehicles silently.

Hence, the end-to-end system can relatively accurately recognize the "up", "down", "left", and "right" commands with just jaw motion and without using any audio signal, though there are still some limitations and errors with enunciation and speed that can lead to misclassifications of commands.

However, the project successfully addresses the issues with voice-based interaction and opens the opportunity for silent speech recognition of directions that can be used in a wide variety of applications, enhancing the usability and inclusivity of speech-based technologies that are commonly used in society's everyday lives.