Simulation of Human Speech Adaptation

David Xu, Madhuri Raman, Yitian Hu, Woo June Cha
Project Supervisor : Peter Freeman
Project Advisor : Charles Wu

Introduction & Background

This work simulates an acoustically driven adaptation phenomenon called dimension-based statistical learning. We conceptualize human speech adaptation as having two different pathways: one path learns slowly and stores long-term representations of how acoustic input can map to linguistic representation, while the other path learns quickly and constantly make dynamic adjustments. Throughout the experiment, we simplify the learning process by using two acoustic dimensions. With the error-driven learning idea, we modeled a dual-pathway design through a 2 hidden layer neural network with these two separate "paths". We successfully simulated a specific phenomenon, namely the down-weighting of the secondary dimension.

Data Generation

We created a simplified representation of acoustic data for canonical and accented speech using simulated data. The data was generated from two bivariate gaussian distributions. Each represents one of two acoustic dimensions of human speech. A third gaussian distribution is centered at each sample to represent excitement of different neurons to the sound. We specifically took 15 samples from each of these three distributions.

Model Description

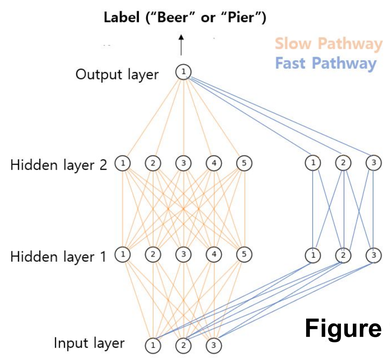

Figure 1.

- We add the fast pathway in a conventional NN to emulate the rapid changes of neurons during speech adaptations.
- We used tanh activation for the first hidden layer and linear activation for the second hidden layer. Sigmoid activation was used for the output layer to ultimately classify "Beer" or "Pier".
- We set a low learning rate of 0.01 for the slow layers and 0.18 for the fast layers.

Methods

General Procedure:

- **Pretrain phase:** We first train the slow pathway on canonical data.
- **Exposure phase**: We add the fast pathway to the model and fit the model in the order of reversed - canonical - reversed data set.
- During the exposure phase, we perform Time-Course Analysis on edge cases as test stimuli.

Time-Course Analysis:

To find out when the down-weighting of the secondary dimension happens, during exposure phase, we need to:

- Expose the model to only one single instance at a time.
- Test the model immediately on the two test stimuli.

Edge Cases as Test Stimuli:

- Vertical and horizontal axes are primary and secondary dimension.
- Orange (High) and purple (Low) circles are the edge case, or test stimuli.
- Note that test the two stimuli are indistinguishable in the secondary dimension.

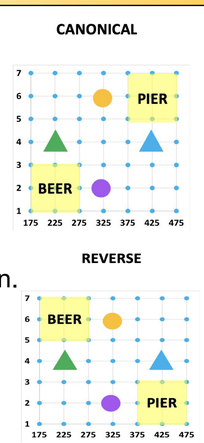

Figure 2.

Analysis & Results

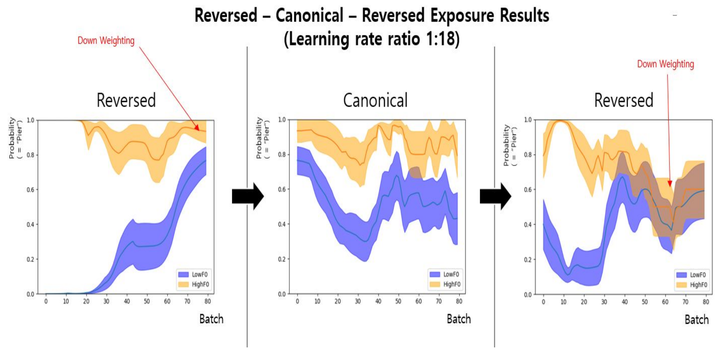

Figure 3.

In Figure 3, we show the proportion of instances classified as "Pier" after running the model on Reversed - Canonical - Reversed data sets in that order after the pretrain phase. The reverse data flips the dimensions of the stimuli from the canonical data, and are pictured in Figure 2 to the left. We hypothesized initially that when the model sees reversed data after pretraining, it will exhibit **down-weighting**, meaning that the model is clearly adjusting itself to the different, accented data by classifying a smaller proportion of High stimuli as "Pier" and higher proportion of Low stimuli as "Pier". When we continue with exposure, this time on Canonical data (the "regular" stimuli/accent), we see the consistent separation between proportions for Low and High stimuli. Finally, when we expose on Reverse data one last time, we again see this down-weighting effect occur with the model trying to distinguish between the two stimuli given one of the same dimensions.

In Figure 4, we visualize the weights for the slow and fast pathways after exposure to the Reversed data. We can observe that the slow pathway weights barely change, while the fast pathway weights change rapidly throughout the exposure. This implies that our model prediction is indeed more reliant on the fast path when the model is exposed to "weird" accent data, which is exactly how we wanted the model to behave.

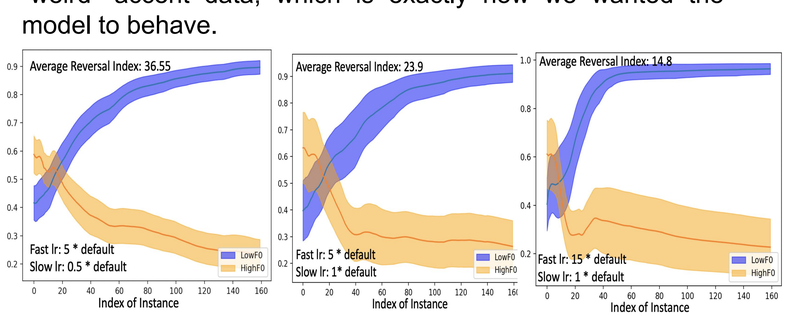

Figure 5.

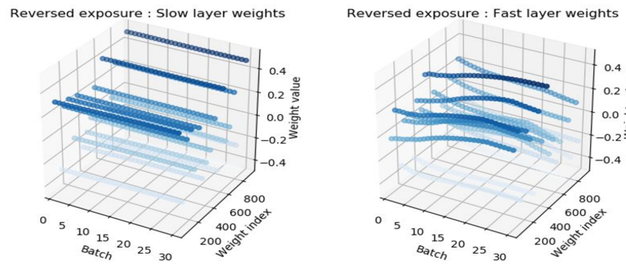

Figure 4.

We manipulated the ratio of learning rates for the two layers, and observed that with higher learning rates in the fast layer, we see faster downweighting during reversed exposure.

Conclusion

The goal of this project was to model human speech adaptation and accent learning abstractly in order to explain how the brain could potentially work in this task. We simplified the actual learning procedure of humans down to only two major dimensions in our modeling of the dimension-based statistical learning. We implemented a simple neural network with two hidden layers and two separate but ultimately concatenated pathways to simulate slow, long-term learning and fast, short-term adaptations to speech stimuli. After pretraining the model, we exposed it to reverse stimuli, canonical stimuli, and reverse stimuli again (different).

Our hypothesis was that the model would have more difficulty distinguishing between the reverse stimuli and be more reliant on the fast pathway for the classification decision. This behavior is exactly what resulted. The next steps we recommend our project advisor take based on our findings are to further run this simulation to see if the model can exhibit a full reversal in the proportion "P" prediction, where the Low stimuli get predicted a higher probability and the High stimuli a lower probability than the other. There is also room to explore the impact of differing weight decay in each path's learning. Moreover, real acoustic sounds could be ued instead of synthetic data. Finally, there is potential for self-supervised learning to be incorporated into the model to help truly model the human brain.

References: Wu, C. (2020). Neural Speech Adaptation [PowerPoint slides].