Investigating Speech Adaptation with CNN

By: Esther Ahn, Shlok Goyal, Yuk Yeung Lam, Lesley Lyu

Advisor: Charles Wu
Supervisor: Peter Freeman

Introduction

This project investigates how the brain achieves speech adaptation to accents with EEG (electroencephalogram) data recorded from brain activities. During the recording, subjects hear typical English speech sounds “Beer” and “Pier” (canonical group), as well as variations of these two sounds (reversed group) generated by mismatching acoustics dimensions, VOT and F0. Previous behavioral data indicate a downweighting on the F0 dimension when people hear mismatched sounds, which leads to difficulty differentiating the sounds. Now we are interested in if such a pattern can be observed in brain data as well.

The goal of the study is to employ classification models to test whether it’s easier to differentiate brain activities in the canonical group or in the reversed group. Our hypothesis is that the model will perform higher accuracy on the canonical group than the reversed group.

Data

The dataset includes a two-hour brain recording with sampling rate of 512 Hz (512*60*60*2 data points) for a single electrode and a single subject. There are a total of 32 electrodes and 23 participants. We then applied a 0.1 - 32 Hz filter and noise removal.

Event-ids	Subjects	Channels	Signal Timepoints
• Stimuli - standard, deviant • can/standard • can/deviant • rev/standard • rev/deviant • can/test1 • can/test2 • rev/test1 • rev/test2	• Subject 001 - 011, 015 - 016, 021 - 025, 027 - 033 • 23 subjects participated total.	• An electrode capturing brainwave activity. • 32 channels total. • A1 - A32.	• Specific time-windows extracted from the continuous EEG signal. • These are also called Epochs. • 91 timepoints obtained total for each subject. • Interval : -0.20315 ~ 0.5

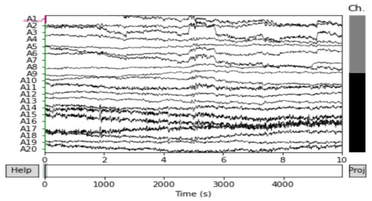

Figure 1 : Image of EEG activity of subject 001 for channels A1 - A20.

Methods

- We followed the general design of deep Convolutional Neural Network by Schirrmeister, which includes four convolution-max-pooling blocks and uses exponential linear unit as the activation function. We also use batch normalization and dropout.
- L1 and L2 regularizer are applied to the kernel of convolution layer
- Two models with exact same architecture and parameters are trained on each subject, one for the canonical case and the other for the reverse case
- 5-fold cross-validation accuracy of two models are compared for each subject

Data Input: (156 trials, 1 subject, 32 channels, 91 timepoints)
Analogous to Image Representation:
(Batch Size, 3 Color Channels, Height, Width)

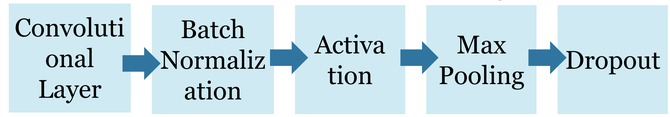

Figure 2: Convolutional Neural Network’s block. We repeat each of these blocks four times in the neural network.

Analysis

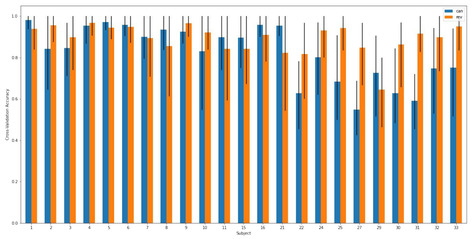

Figure 3: Individual Cross validation accuracy of two types of sounds for each subject

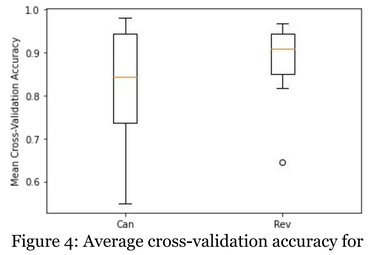

Figure 4: Average cross-validation accuracy for two types of sounds

The canonical and reverse models have fairly similar validation accuracies when compared for all subjects. We used a one-sided two-sample t-test where:

- Null hypothesis:
  Classification Accuracy (Rev) ≥ Classification accuracy (Can)
- Alternative hypothesis:
  Classification accuracy (Rev) < Classification accuracy (Can)

The t-test yielded a p-value of 0.99 when applied over all subjects. So, we fail to reject the null hypothesis.

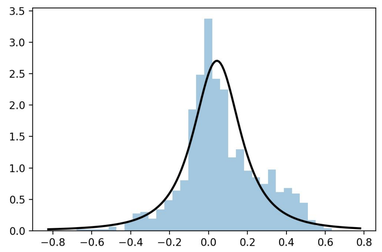

Figure 5: Empirical Distribution of the difference in validation accuracy between the canonical and reverse cases approximates a t-distribution.

There is a significant decline in the validation accuracy in the canonical case for the second half of subjects, suggesting changes in data collection methods.

In fact, a one-sided two-sample t-test does find that subjects 1, 8, 16, 21, and 29 do have p-values less than 0.05 when tested individually.

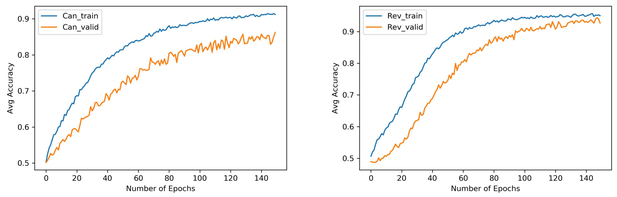

Figure 6: Train and validation accuracy for two cases (averaging all subjects). The neural network doesn’t overfit, suggesting that in general, brain activity is consistent across tests within both canonical and reverse cases.

Conclusions

The result contradicts behavioral data as subjects had a lower accuracy in distinguishing between “beer” and “pier” in the reverse case than in the standard case but our neural network did not. This suggests that even in the accented case, there are patterns in the brain data that can be used to distinguish the exact word being spoken. It must be these patterns that the neural network trained on the reverse case is picking up. Thus, even though subjects are less able to classify between beer and pier when listening to accented sounds, their brain has already started noticing patterns to make the classification, allowing for the rapid acclimatization of accented speech that we see in daily life.

We would need data from more subjects to conclude this hypothesis convincingly. Furthermore, we can try different neural network architectures and see if the result holds. Another extension would be to train the model with all subjects, for both the standard and canonical cases. We opted not do so here since people process speech differently, but doing this will increase the number of observations passed into the model

References

Schirrmeister, Robin Tibor et al. “Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization.” Human Brain Mapping 38.11 (2017): 5391–5420. Crossref. Web.