The Disadvantage of Proficiency:

A Neural Network Simulation of the Non-Native Phoneme Generalization Problem

Jared Jolton

University of Colorado at Boulder

Author Note

Jared Jolton, Department of Psychology, Department of Computer Science, Institute of Cognitive Science, University of Colorado at Boulder.

This research was completed as part of the requirements for the Fall 2015 PSYC 4175 course at the University of Colorado at Boulder.

Abstract

Speakers have a hard time distinguishing between non-native speech sounds. This study attempts to model and explain this phenomenon using a computational neural network. The model produces the same form of generalization errors as seen in human data. These errors are interpreted as direct results of the network's proficiency with the native language.

Introduction

Research Question

Many previous studies have documented the surprising inability for adults to recognize and differentiate between the sounds of nonnative languages (Francis & Nusbaum 2002; McLelland, 2002; Escudero, 2005). In a sense, this inability is easily explained - adults have little to no exposure to these foreign sounds, whereas native speakers likely use and hear them daily. What is more interesting is the fact that young children have no problem differentiating between these sounds, regardless of the language(s) they are exposed to. The human brain adapts efficiently and perhaps essentially to better suit our individual environments, as explicitly documented in phoneme discrimination studies (McLelland, 2002), face recognition studies (Walker, 2003), and studies of linguistic determinism (Boroditsky, 2001). This study intends to explore the relationship between the neural adaptation process and Hebbian learning principles - noting their influence on the structuring and restructuring of the brain's networks. Conclusions will focus on not only Hebbian learning's facilitation of the neural adaptation process, but also its enforcement of neural adaptation, an explanation for why our robust and adaptable neural networks have an inherent difficulty generalizing to novel or untrained inputs.

Background

There are many explanations for why the ability to discriminate between the universal set of phonetic sounds quickly deteriorates. Some studies (i.e. Francis & Nusbaum 2002; McLelland, 2002) attribute the difficulty to a limited perceptual space that develops over time. Studies like these suggest that selective attention allows us to specialize our perceptual space to afford efficiency and accuracy at recognizing and differentiating between the percepts that we most frequently encounter. Other research (i.e. Escudero, 2005) attributes the difficulty with foreign sounds to the feature geometry of the phonemes themselves. The similarity in native acoustic signals allows them to be categorized based on phonological structure, whereas non-native signals that are inherently less similar than familiar sounds are categorized based solely off of their acoustic features. Universal phonetics and native phonology must be, to some extent, separate within the brain.

Current Study

Having acknowledged that, "the difficulties adults may have in learning new speech contrasts in adulthood might reflect an undesirable characteristic of Hebbian synaptic modification," (McClelland, 2002) it follows that the phoneme discrimination problem can be modeled computationally by creating a neural network that is capable of learning to recognize and categorize individual phonemes, using Hebbian and error-driven learning. By visually representing phonemes, the model can learn to differentiate between phonemes based on their features, processing the sound by means of the visual representation. While the layers and connections of the network are not structured according to the biology of the speech perception system in the brain, the network's capacity for building complex featural representations of input stimuli makes it an excellent tool studying this particular artifact of speech perception.

After training the network to recognize and differentiate between native speech sounds, the network's ability to generalize to novel, non-native speech sounds was tested. In this experiment, the network was first trained on a large subset of Japanese speech sounds, and then on its ability to differentiate between English phonemes. All sounds were pronounced by a male and a female voice, protecting the model from overfitting and ensuring that some level of generalizability would be present before ever testing the network with the English sounds. The network was trained until it could perfectly distinguish between all Japanese speech sounds, regardless of how long this learning process took.

Previous research has confirmed Japanese speaker's difficulty distinguishing between the /r/ and /l/ sounds found commonly in English, as differentiating between them is not required in the Japanese language (McLelland, 2002). As such, it was hypothesized that the network trained on the Japanese speech sounds would not be able to differentiate between these particular English phonemes.

Methods

Materials

Dataset

To represent a complete Japanese phonetic inventory, 46 Japanese speech sounds were obtained from the companion website for Kim's "A Guide to Japanese Grammar" (Kim, 2014)[1] see references for a download link), each sound being pronounced by both a male and a female voice. Importantly, these sounds were not phonemes by definition. Rather, they were the individual components of *hiragana*, a Japanese syllabary that contains all possible sounds in the

[1] To download the speech sounds, visit http://www.guidetojapanese.org/audio/basic_sounds.zip

language. These sounds are known as *morae* (*mora* singular). Certain phonemes in the Japanese phonetic set, such as the vowel [i], are not annunciated unless they occur after a palatalized consonant, which makes *morae* a better tool for representing the speech sounds heard by native speakers.

After all 92 Japanese speech sounds - and the 2 English phonemes - were normalized in both length and amplitude, each was converted into a sonogram (Figure 1) using software created by Cristoph Lauer (Lauer, 2015). The images generated represented the frequency spectrum of the sounds across time, such that frequency was plotted on the y-axis and time was plotted on the x-axis. The frequencies were logarithmically scaled, because most of the sounds only consisted a small range of lower frequencies. Scaling them ensured that the full image was representative of the entire sound, and maximized the amount of information that was encoded in the image.

Network

The neural network used for this research[2] employed a simple three layer structure. All neurons in the input layer were bidirectionally connected to the neurons in the hidden layer, and all neurons in the hidden layer were bidirectionally connected to the neurons in the output layer. The input layer consisted of 1,000 neurons, each representing an individual frequency detector. During the presentation of a Japanese sound - the activation phase - the activation of the of each input unit was based on the corresponding pixel value of the image. The network can use up to 100,000 input units, but the computation times required for the network to compute activations and weight updates for this many units largely outweigh the benefits of more feature detectors.

[2] The network and all of the stimuli can be viewed and downloaded at https://github.com/2PacIsAlive/phoneme_classification_network

The network has converged on solutions for small training sets (less than 8 total speech sounds) using as few as 10 feature detector neurons in the input layer, but as the size of the training set increases, so does the amount of input units required to differentiate between the sounds - a trivial insight.

The hidden layer in the network can similarly be of varying size, but utilized 36 neurons in all relevant trials. This layer was connected only to the input and the output layers, allowing it to form distributed representations of the speech sounds used by the output layer to distinguish between the sounds. Neurons in the output layer represented speech sound detectors. Prior to training, each neuron was arbitrarily assigned a *mora* to identify. Whether or not this neuron was maximally activated a trial where its corresponding sound was presented to the input layer determined the direction of learning that occurred. Due to the structure of the neural network code, adding neurons to the output layer resulted in cubic runtime increases, drastically increasing in the time required for training. As such, it was infeasible to train the network on the entire Japanese phoneme set, and the size of the output layer was restricted to 34 units. However, there is no reason to believe that the network would be incapable of learning the entire set, and because the program can be paused, efforts to train the network on all Japanese speech sounds are ongoing.

In order to facilitate learning, the network employed feedforward activation and error backpropagation (O'Reilly, 2014). The activation of each hidden and output unit was computed by summing over the product of the activation of the units in the preceding layer and the weight for the connection between these units and the given hidden or output unit. The final activation

was passed through a sigmoid activation function. The sigmoid function is bounded by -1 and 1, ensuring that the final value would be within this range.

During the backpropagation phase, the network computed weight updates in cascading fashion from the output layer down to the input layer. The error values for the output units were computed by simply subtracting the actual activation for the neuron from the expected activation. Neurons were expected to have an activation of -1 if they were not associated with the presented sound, and and activation of 1 if they were indeed responsible for representing the sound. Computing the weight updates for the connections between these units and the hidden layer units is simple, the change in weights is simply the product of the activation of the hidden unit and the error value of the output unit. It is much more difficult to compute the weight updates for the connections between the hidden and input units, as these units are not intended to be maximally or minimally activated. Rather, their activations must be precisely defined in order for the network to learn. As such, the error must be propagated down the network, using the derivative of the sigmoid activation function. The network must also compute two sums for this phase - the sum over all output units of the product of the expected output unit activation and the corresponding weight, and the sum over all output units of the product of the actual output unit activation and the corresponding weight. To compute the final error value for the hidden units, this derivative is multiplied by the difference of these sums, and the resulting value is multiplied by the activation of an input unit to compute the weight change for the hidden/input connection.

Procedure

Before training, the network was tested on its ability to recognize all phonemes using initially generated random weights. During training, two output units were reserved for

non-native speech sounds, reflecting the separation of universal phonetics and native phonology in the brain (Escudero, 2005). These units did not undergo learning during the Japanese training, and were reserved for English speech sounds. During testing, the network was first tested on its ability to classify all Japanese sounds. Then, the English sounds were presented to the network. As the problem non-native speakers have is differentiating between non-native sounds, having different maximally activated output units for each sound was taken to mean that the network was capable of distinguishing between them, even if these output units were not the ones delegated for English. Because the extra English output units never undergo training, what they represent is random, and it is a better measure of performance to see what the network thinks the novel sounds are closest to. The /r/ and /l/ sounds are each similar to several different sounds in the Japanese training set, so having these similar output neurons be activated is as revealing as having the extra output neurons be activated.

After testing, the network was trained further, to see if exposure to the non-native sounds would allow the network to learn to classify all speech sounds. While adults often have problems when first exposed to novel speech sounds, they can learn to differentiate between them with repeated exposure (McLelland, 2002).

Results

Analysis

Network Performance

The network was able to quickly learn to classify the Japanese phonemes (Figure 2), and was easily able to identify male and female voices. Quickly, of course, is a relative term, as learning to distinguish between 32 *morae* took over 24 hours. However, this only required 9

epochs of learning within the network - meaning each sound was presented to the network only 9 times before the network was capable of differentiating between them. Interestingly, the number of epochs taken to learn a solution to the task seemed to go down as the input size increased (Figure 2). While the performance of the network on any given trial was largely due to the set of initial random weights used by the network, the general trend suggested that an increased input size led to a shorter training time. Even when the random weights were extremely well defined at the beginning of training, allowing the network to start out at around 83% error, more epochs were needed than with the largest input set, which began at 100% error (Figure 2). However, the stochastic effects of the initial random weights cannot be ignored as an incredibly important factor in the learning process. People often have varying levels of difficulty learning the speech sounds of other languages, and these difficulties could reflect the random weight patterns within their own neural networks.

As predicted, the network was incapable of differentiating between the two English phonemes in the generalization task. Without training, both phonemes maximally activated the "ko" *mora* pronounced by a male. While the English phonemes were also pronounced by a male, nothing about the /r/ and /l/ sounds or their sonograms suggests that they related in any way to the "ko" *mora*. It is difficult to say whether the /r/ and /l/ sonograms were at all similar to any of the Japanese sonograms. This failure directly reflects the human performance on similar speech sound recognition tasks. After 3 epochs of further training, the network successfully learned to represent the new sounds using the extra output units, just as human participants can learn to differentiate between previously unrecognizable sounds (McLelland, 2002).

Receptive Field Analysis

To better understand the kinds of representations that were being formed by the network, a receptive field analysis was conducted on the neurons in the hidden layer. An image was generated for each hidden unit that represented all the connections between that hidden unit and all the input units (Figures 3, 4, and 5). Blue squares represent connections with negative weights, and red squares represent connections with positive weights. More vibrant colors correspond to weights closer towards -1 and 1, respectively, and darker colors correspond to weights closer to 0. While these receptive fields were only compared qualitatively, there are several interesting insights they afford. Most of the hidden layer neurons appeared to have relatively structured receptive fields, consisting of patches or areas of neurons that had strong positive or negative weights (Figures 5, 6), and few neurons appeared to be representing the same areas. In addition, there are several neurons that formed mostly negative weights, only responding strongly to a very small selection of input layer neurons. These inhibitory neurons play an important role in allowing the network to learn. Throughout the early stages of training, the network often classifies all inputs in the training set as the same output. Inhibitory neurons must be essential for preventing inputs with the same feature as being classified as the same. Many of the sonograms are incredibly similar, and likely have hundreds of the same features. The receptive fields of the network showcase the ways in which these features are carefully parsed, allowing the network to form complex distributed representations of very similar inputs.

Figures

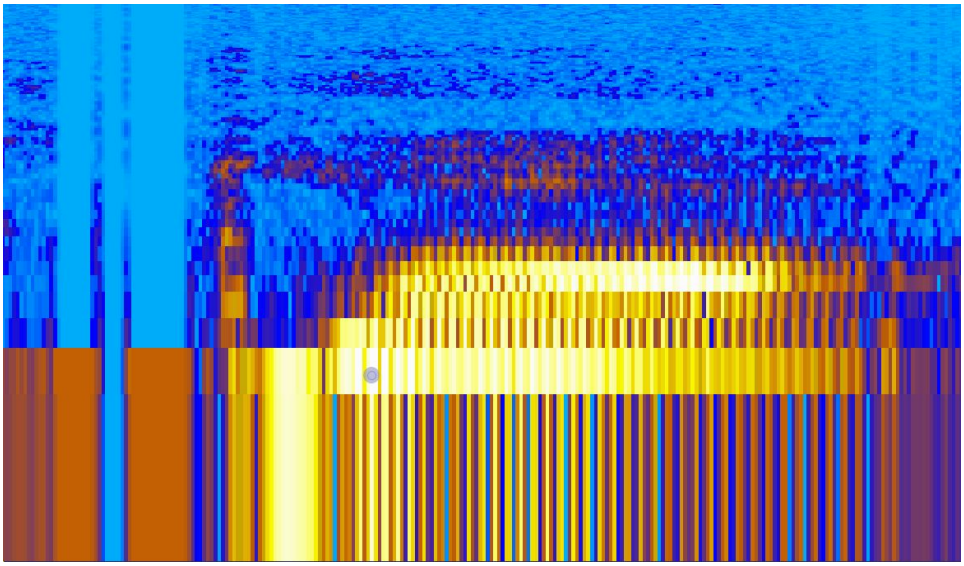

Figure 1: Sonogram, Logarithmically Scaled (ke mora pronounced by female voice)

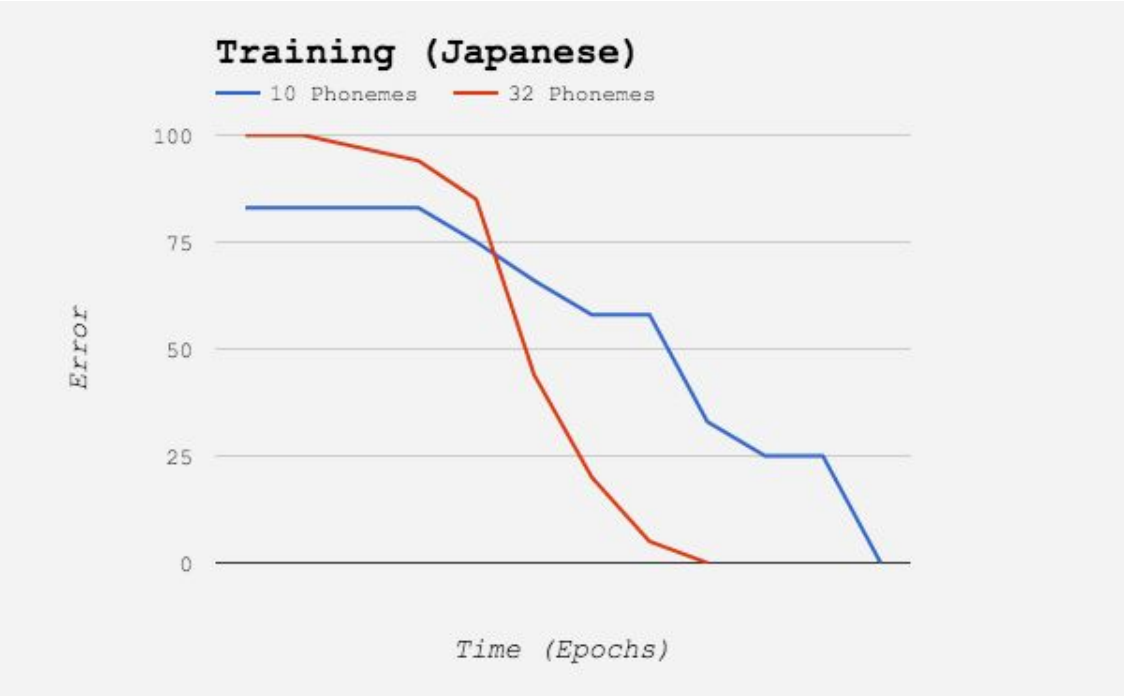

Figure 2: Network Performance

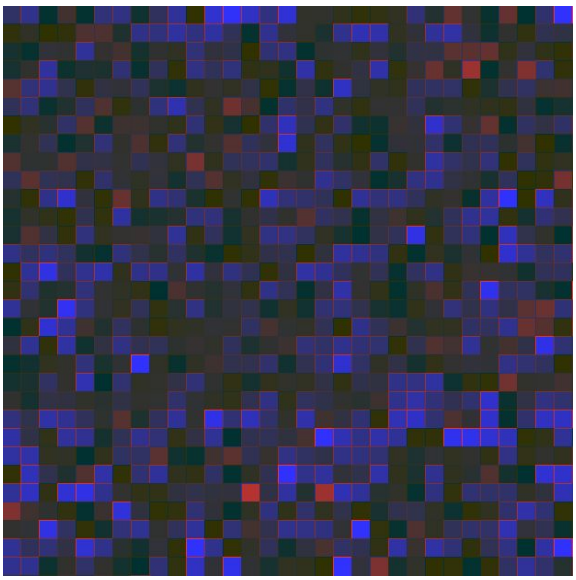

Figure 3: Receptive Field for hidden unit 20 (Inhibitory Neuron)

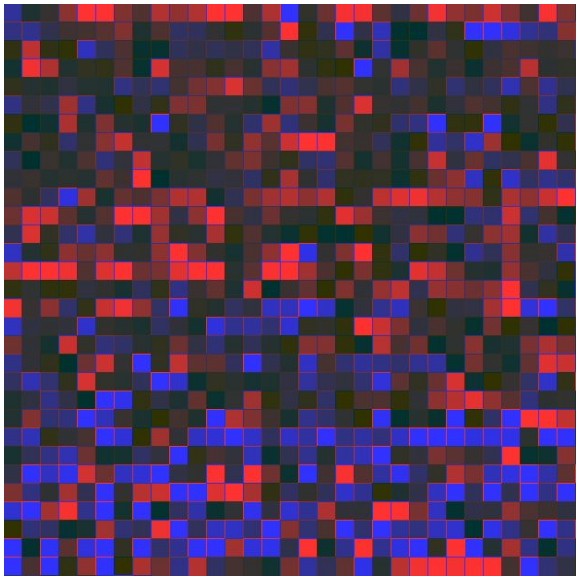

Figure 4: Receptive Field for hidden unit 31

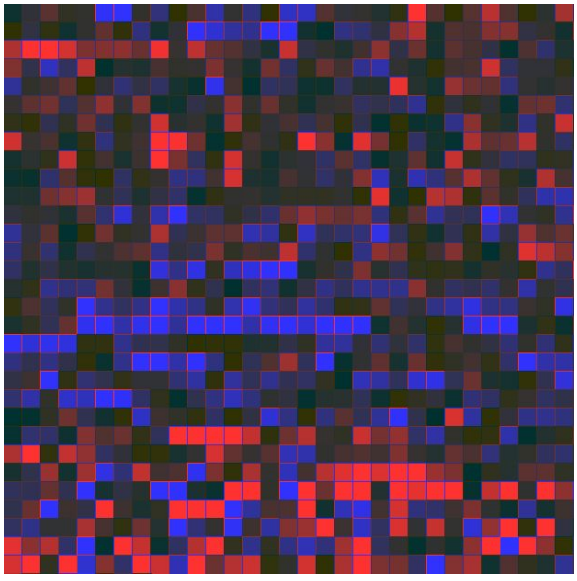

Figure 5: Receptive Field for hidden unit 2

Discussion

We are constantly bombarded with speech, incessantly faced with the difficult task of identifying infinitesimal sounds at blazing fast speeds. Yet, normally functioning brains are capable of doing this with little cognitive effort, and can easily generalize to many different pitches, accents, and tones. The reason we are so adept at this task is undoubtedly due to Hebbian synaptic modification. Each time the feature detectors required to identify a speech sound are activated simultaneously, the representation they foster becomes more and more specialized, allowing for even better recognition of the speech sound later on. As was depicted by the network, neurons require incredibly precise weight patterns to be able to differentiate between sounds, and without training or exposure, the Hebbian process required to tune these weights will never occur, making it unlikely that novel sounds will have their own representations downstream. The feature detectors have no trouble identifying and encoding the new sounds, but they lack the precise connections to their own output neurons needed for identification and differentiation.

It is intriguing that without further training, the Japanese trained network consistently identified the English speech sounds as the "ko" *mora*. It was predicted that the novel sounds would at the very least activate somewhat similar speech sounds, as neural networks are usually fairly proficient at generalizing to new inputs (O'Reilly, 2014). Perhaps a gaussian activation system which, in a graded fashion, activates the neighbors of a maximally activated neuron in the hidden layer, would allow for better generalization. Similarly, using dropout or inhibitory competition would foster sparser representations, giving the network a better chance to attribute the novel features of the new stimuli to similar, existing representations. Yet, it is important to

acknowledge that the pattern of weights required to differentiate between the original language's sounds must be very specifically defined. Without carefully adjusting these weights using Hebbian and error-driven learning, representations will likely never be formed. Further, the precise weight adjustments that do occur during training make it harder for new representations to be formed, as these weights must be readjusted during further training, perhaps destabilizing the existing representations of the native sounds. The success of the network in classifying the Japanese *hiragana* ends up being the reason it has difficulty distinguishing between the English sounds. We shape our minds based on our experiences, and our proficiency with our native languages causes our inability to differentiate between non-native speech sounds.

References

Boroditsky, L. (2003). Does Language Shape Thought?: Mandarin And English Speakers' Conceptions Of Time. *Cognitive Psychology,* 1-22.

Escudero, P. (2005). Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization. Netherlands Graduate School of Linguistics.

Francis, A., & Nusbaum, H. (2002). Selective attention and the acquisition of new phonetic categories. Journal of Experimental Psychology: Human Perception and Performance, 349-366.

Hebb, D. (1949). The organization of behavior; a neuropsychological theory. New York: Wiley.

Iverson, P. (2002). A Perceptual Interference Account Of Acquisition Difficulties For Non-native Phonemes. Cognition, B47-B57.

Kavanagh, B. (2007). The phonemes of Japanese and English: A contrastive analysis study. **青森県立保健大学雑誌**, 8(2), 283-292.

Kim, T. (2014). *A guide to Japanese grammar*. http://www.guidetojapanese.org/learn/grammar/hiragana

Lauer, C. (2015). Sonogram. http://www.christoph-lauer.de/sonogram

Mcclelland, J., Fiez, J., & Mccandliss, B. (2002). Teaching the /r/–/l/ discrimination to Japanese adults: Behavioral and neural aspects. Physiology & Behavior, 657-662.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., and Contributors (2014). Computational Cognitive Neuroscience. Wiki Book, 2nd Edition. URL: http://ccnbook.colorado.edu

Sharma, A., & Dorman, M. (2000). Neurophysiologic Correlates Of Cross-language Phonetic Perception. The Journal of the Acoustical Society of America, 2697-2697.

Walker, P., & Tanaka, J. (2003). An encoding advantage for own-race versus other-race faces. *Perception,* 1117-1125.