A Computational Model for Unsupervised Vowel Acquisition

Language Acquisition Research Report

Department of Linguistics and Philosophy, Massachusetts Institute of Technology

Tristan Thrush

Supervisor: Martin Hackl

May 26, 2018

Abstract

In this paper, I provide a computational model for how infants may acquire vowels. Unlike previous models where researchers have manually selected features of speech for vowel categorization, my model uses an optimization procedure to learn the most relevant features from a dataset of vowel recordings from human speech.

1 Introduction

Infants can acquire vowels (more generally, phonemes), without supervision. This statement must be true because infants (who are initially born with the ability to distinguish between vowel categories in every language) eventually become unable to distinguish many vowel categories in other languages that are within the same vowel category in their own language. This "learning to forget" becomes obvious around 6 months (Kuhl, 2004). This ability must be unsupervised because infants cannot be told how many vowels there are in their language, and cannot understand positive and negative examples, as they need to understand vowel categories to understand words in the first place. Because humans

can learn vowel categories without supervision, they must perform some sort of clustering algorithm, and that clustering algorithm must be done in a feature space that lends itself to the adequate separation of clusters that represent different vowels.

This paper focuses on a computational model for how humans may transform the sound of a human saying a vowel into a feature space that enables the vowel to be clustered with different utterances of the same vowel, and separated from utterances of different vowels. Previous work in this area has focused on manual selection of sound features with varying success (Molis, 2005). This paper attempts to explain the human vowel feature-mapping capacity as the result of a connectionist model. Particularly, I provide evidence for two hypotheses.

1. The human feature space representation for vowels could be the result of a connectionist model that is fine tuned by evolution to map sounds into a feature space where vowels can be clustered according to a clustering algorithm; and

2. The human feature space representation of vowels could provide a global optimal way to separate vowels for clustering, and the evidence for this is that an optimized way from any connectionist model that can learn a robust feature mapping provides a similar confusion matrix as the human confusion matrix. I use the definition of confusion matrix given by Zahorian and Jagharghi (1993), where each row represents a vowel, and each position in that row also represents a vowel. Each position in a row is assigned a percentage, which represents the percentage of human responses that categorized the row's vowel utterance as the vowel corresponding to that position. By "similar confusion matrix", I mean that most of the vowels that are misclassified by humans are misclassified by the model as the same wrong vowels, in essentially the same proportions.

To provide evidence for my hypotheses, I discuss implementations and tests of two neural network architectures that serve as examples of connectionist models. One of the architectures has too many parameters and overfits on the training set, which leaves me inconclusive about whether it can robustly learn a feature space that can be clustered in such a way to achieve human classification performance. The other model does not substantially overfit

on the training set and does end up yielding human-like performance. The model that yields human-like results provides evidence in favor of my hypotheses, and both models are consistent with my hypotheses.

2 Related Work

To the best of my knowledge, related work that focuses on vowel separation by category from speech data has never explored an optimization procedure that learns features. It has always explored manual selection related to formants, amplitude, spectral shape, and related features as shown by Molis (2005) and Zahorian and Jagharghi (1993). Molis (2005) also explores nonlinear kernels of such features, which offer more human-like performance according to Akaike's Information Criterion (AIC), but the authors only test their models on three different vowels, which is not even close to the number of categories in most languages (English has 12). In this paper, the model can handle all 12 vowels of English.

There has been other related work that was able to achieve essentially complete separation of all vowels, by category, in English, as shown by Coen (2006), but it relied on both formant features from the sound of the vowel and a video of a person saying the vowel. It was found that formant features from the sound of the vowel utterance paired with lip rounding features from the video of the utterance was enough to find a good feature space for vowels. However, the author did not address the fact that humans do not need visual information to extract information about lip rounding. Lip rounding cannot be obviously captured by simple audio features such as formants, but the model that I present in this paper can still identify lip rounding solely from the audio.

3 Methods

Here, I present an overview of my model, which takes in an audio recording of a human saying a vowel and outputs a category for that vowel. As it is a model for human vowel acquisition, it must learn vowel categories in an unsupervised way. The first component is a connectionist model that produces a feature space mapping from audio of a vowel utterance,

which I hypothesize humans have a priori, due to the effects of an optimization performed by evolution. The second component is a clustering algorithm that finds groupings of vowels within this feature space. The clustering algorithm is provided to prove that the connectionist model can learn a feature space that lends itself to a good clustering of vowels, and is not the focus of this paper.

3.1 Encoder

The connectionist encoder model takes in a time-varying Melfrequency Spectrogram (Xu et al., 2005) of a human saying a vowel as an input and converts it into a feature space. The spectrogram has been discretized such that it has 128 bins that span 0 Hz to 8192 Hz and the average energy that occurs over a bin's range is that bin's value. The bins change along the time axis of the spectrogram every 0.02 seconds. The encoder is trained using supervised learning with a second neural network that takes the encoding and outputs a vowel classification. The classifier network is a single fully connected layer with no nonlinearities, so it amounts to a linear classifier. The encoder training system can be seen in Figure 1.

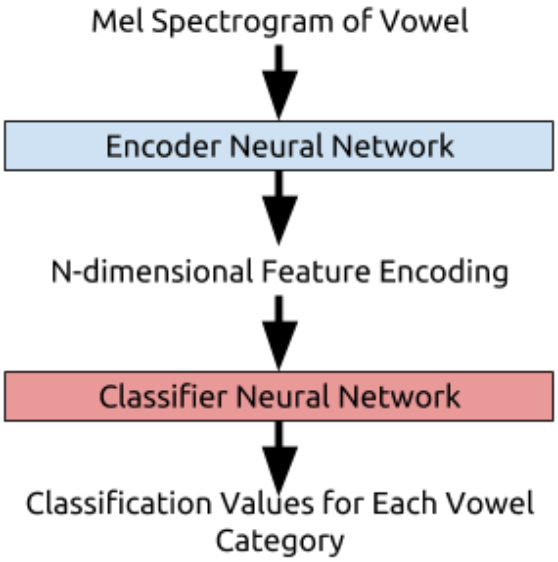

Figure 1: The apparatus that trains the encoder to learn the best N-dimensional feature encoding to separate vowels for linear classification. This encoding will presumably also be useful at separating vowels for a clustering algorithm.

The two networks are optimized together, such that the encoder learns how to create a feature space that can be given to a linear classifier to predict vowel labels. The output of the classifier is sent into the cross-entropy loss function (Murphy, 2012). The result of applying an optimizer such as Stochastic Gradient Descent (Bottou, 1998) to this system is that the category with the maximum value in the classifier net's output will be the predicted vowel label.

This paper examines two types of encoders: a convolutional encoder and a recurrent encoder. The architectures of both are discussed below.

3.1.1 Convolutional Encoder

The convolutional encoder handles the time-varying input with a 1-dimensional convolutional layer (Krizhevsky et al., 2012) that has 128 channels (1 for each spectrogram bin) and slides across the time dimension. The number of output channels are also 128. The outputs of this convolution are then all max-pooled together into one 128-dimensional vector, and sent into a sequence of fully connected layers, which are parametrized to return

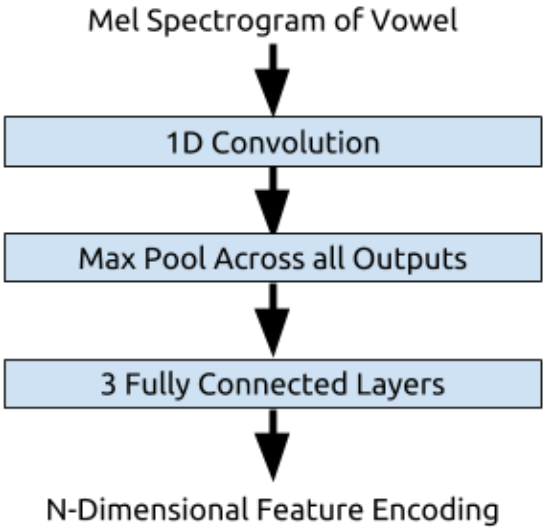

Figure 2: The neural architecture for the convolutional encoder.

an N-dimensional vector that represents the feature encoding of the input spoken vowel. The architecture can be seen in Figure 2.

3.1.2 Recurrent Encoder

The recurrent encoder handles the time varying input with an LSTM cell (Hochreiter and Schmidhuber, 1997). The LSTM takes in the sequence of 128-dimensional bin vectors across time and the final hidden state is taken as the overall LSTM cell's output. This hidden state is then passed into 2 fully connected layers, which are parametrized to return the N-dimensional feature vector. The architecture can be seen in figure 3.

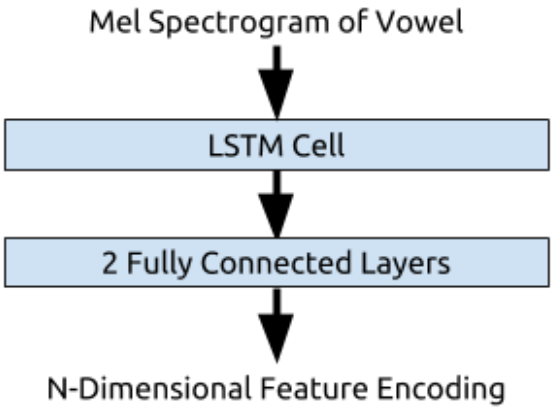

Figure 3: The neural architecture for the recurrent encoder.

3.2 Clustering Algorithm

After the encoder has been trained, the classifier neural network is discarded and replaced by a clustering algorithm. The clustering algorithm is designed to take in a set of encoded vowels and learn clusters that could represent distinct vowel categories. After it has completed its unsupervised learning procedure, it can then be used to predict the cluster that a test vowel belongs to. While the classifier neural network can accomplish this task, it cannot model how humans categorize vowels, because it learns with supervision, whereas a clustering algorithm does not.

The clustering algorithm that I use is Expectation Maximization on a Gaussian Mixture Model (Hastie et al., 2001). This algorithm is told that there are n clusters; it then randomly initializes n different Multivariate Gaussians, and iteratively fits the Gaussians such that it maximizes the probability that the data points were generated from those Gaussians. The result is that the algorithm converges to a local optimum clustering, so it should be run a few times to find the clustering with the highest probability. The only obvious inconsistency between this algorithm and human behavior is that humans can find vowel clusters without being told the number of distinct vowel categories. However, the focus of this paper is not on the specific clustering algorithm that I use; it is on the connectionist encoder model. The encoder to clustering algorithm system can be seen in

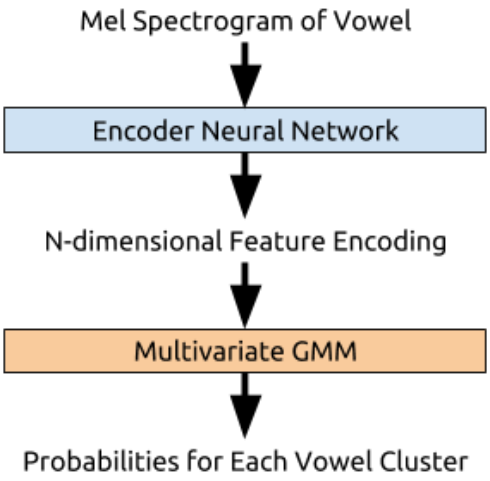

Figure 4: The apparatus that I present as a model for the part of human cognition that categorizes vowels. The trained encoder neural network represents a feature space mapping that I suppose humans have a priori, and the Multivariate Gaussian Mixture Model provides a mostly reasonable model for how humans can categorize vowels in this space by an unsupervised learning of clusters.

Figure 4.

3.3 Data and Analysis

The training data I use is a set of 1667 recordings of men, women, and children saying every English vowel surrounded by the consonants "h" and "d" , compiled by Hillenbrand (2018). I train the encoder training system on a random selection of 1000 of these vowels, until the classifier neural network is able to classify vowels with peak accuracy. I then perform the clustering algorithm on the other 667 vowels, after they have been encoded by the trained encoder. I manually annotate the learned clusters to correspond to what I perceive to be the correct vowel label, so that I can compare the clustering algorithm's cluster predictions with human predictions.

To verify that my Encoder + Gaussian Mixture Model produces a similar confusion matrix to that of humans, I ask it to predict what clusters the set of 667 test vowels belong to. This procedure generates a confusion matrix for my model, which I can use

to compare to a confusion matrix from Zahorian and Jagharghi (1993), gathered from experiments where humans were asked to do the same task (classify vowels when given consonant-vowel-consonant utterances).

4 Results

Here, I present the results of my Encoder + Gaussian Mixture Model system, for both the convolutional encoder and recurrent encoder. The convolutional encoder does not have as many nonlinearities and parameters, so it does not overfit substantially on the training dataset; the encoding provided is robust enough for the linear classifier network and Gaussian Mixture Model to work well on test vowels. The recurrent encoder overfits substantially, leaving me inconclusive about whether it would support my hypotheses if it was given more training data; it is still consistent with my hypotheses, as it cannot be adequately tested due to its inability to learn robust features, even for supervised classification.

4.1 Convolutional Encoder

After training the convolutional encoder (with a 3-dimensional encoding) and the classifier neural net with a batch size of 4 on 100 epochs through the training data, there was evidence that the encoder had learned a robust feature space representation. The classifier is able to achieve 97.5% train accuracy and 86.1% test accuracy, which is far above $1/12 = 8.3\%$ (chance). The Multivariate GMM is able to take the encodings for the 667 vowels that the encoder did not train on, and learn clusters that correctly categorize 70% of the vowels (the learned clusters were labeled by hand, as the vowel cluster that they appeared to align with).

The confusion matrix generated by the convolutional encoder + GMM is shown in Table 1. And a human confusion matrix for the same task, taken from Zahorian and Jagharghi (1993), is shown in Table 2. There are some poor clusterings; for example, my model misclassifies /ow/ as /ah/ 70% of the time. But most of the vowel clusterings and errors correspond across the two matrices, as evidenced by most of the high values being along

	iy	**ih**	**eh**	**ae**	**ah**	**aa**	**ao**	**ow**	**uh**	**uw**	**er**	**ei**
iy	100											
ih	5	87	2									7
eh			91	7				2				
ae			87	13								
ah				54	35			10				
aa		2				66			29	3		
ao							98				2	
ow					70			25	5			
uh								96	4			
uw										98		2
er			2				2			2	94	
ei	2	2									4	93

Table 1: The model's confusion matrix, generated from experiments where it was asked to classify vowels from the test set of 667 human recordings (which are consonant-vowel-consonant utterances)

	/iy/	/ih/	/eh/	/ae/	/ah/	/aa/	/ao/	/ow/	/uh/	/uw/	/er/
/iy/	100.0										
/ih/		99.1	0.9								
/eh/		1.5	96.2	2.3							
/ae/			3.3	82.3		14.5					
/ah/				0.3	93.4	0.3	0.8		4.8	0.5	
/aa/					5.2	73.6	19.5	1.6			
/ao/					0.9	15.8	82.9	0.2		0.2	
/ow/								99.6		0.4	
/uh/					11.5			0.7	83.3	4.4	
/uw/					0.7			0.7	4.5	94.0	
/er/					0.3				0.3		99.4

Table 2: A human confusion matrix, generated from experiments where humans were asked to classify vowels from consonant-vowel-consonant utterances.

the diagonals of the matrices.

Because the encoder was trained to learn a robust 3-dimensional encoding, the vowel space can be visualized. Figures 5 and 6 show different planes of the same feature space. Each point represents an audio recording of one of the vowel utterance test data points that has been mapped into the 3-dimensional space. The points are labelled according to their true classifications, although this is just for visualization purposes and the encoder is not given these labels. Interestingly, the encoder has learned 3 roughly orthogonal dimensions that all have obvious interpretations. Figure 5 shows the "V" or "U" shape that comes from graphing vowels according the first and second formants, although it is turned on its side in the image. Figure 6 shows the same plot, but rotated such that the red line shows the plane that Figure 5 was observed from. This 3rd dimension roughly corresponds to the

level of lip rounding that a vowel has.

The result of this encoding is a nearly complete separation of all of the distinct vowel clusters, except for /ae/ and /eh/, which many English speakers cannot tell apart. There is nothing that forced the neural net to learn these 3 orthogonal and interpretable features, except the drive to find an optimal way to separate vowels by category. This means that these three features are actually an optimal way to separate vowels.

An interesting finding is that when I run the encoder trainer with a 2-dimensional encoding, the classifier neural network is only able to achieve a training score of 72.1% and a test score of 61.3%; this is significantly worse than the 3-dimensional case. When I run the encoder trainer with a 4-dimensional encoding, the classifier neural network is able to achieve a training score of 99.1% and a test score of 86.3%, which is about the same as the 3-dimensional case. This is an indication that the first formant, second formant, and lip rounding are all important features to separate vowels, but there may not be another feature that can be added to significantly improve the performance. This finding makes intuitive sense because these three features already result in almost complete separation of vowel categories.

Finally, Figures 7 and 8 show the same mapping as Figures 5 and 6, except the labels identify clusters that the EM GMM clustering algorithm has learned. As you can see, these clusters roughly correspond to the true clusters of distinct vowels in Figures 5 and 6 (the colors do not stand for the same vowels).

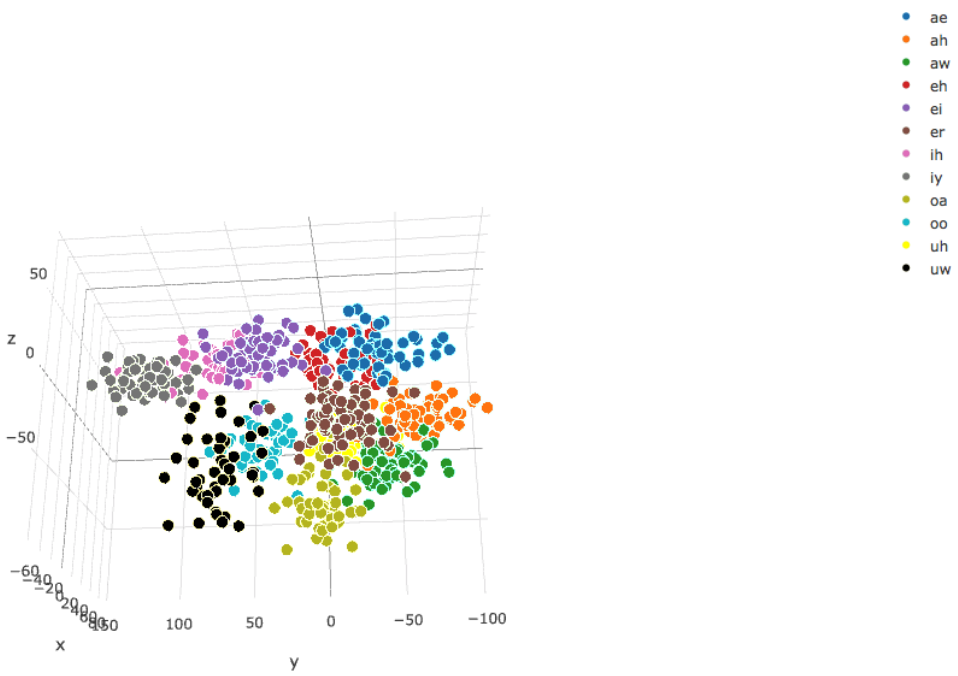

Figure 5: A view of the convolutional encoder's 3D encoding of vowel utterances, on a plane that corresponds the the first and second formants.

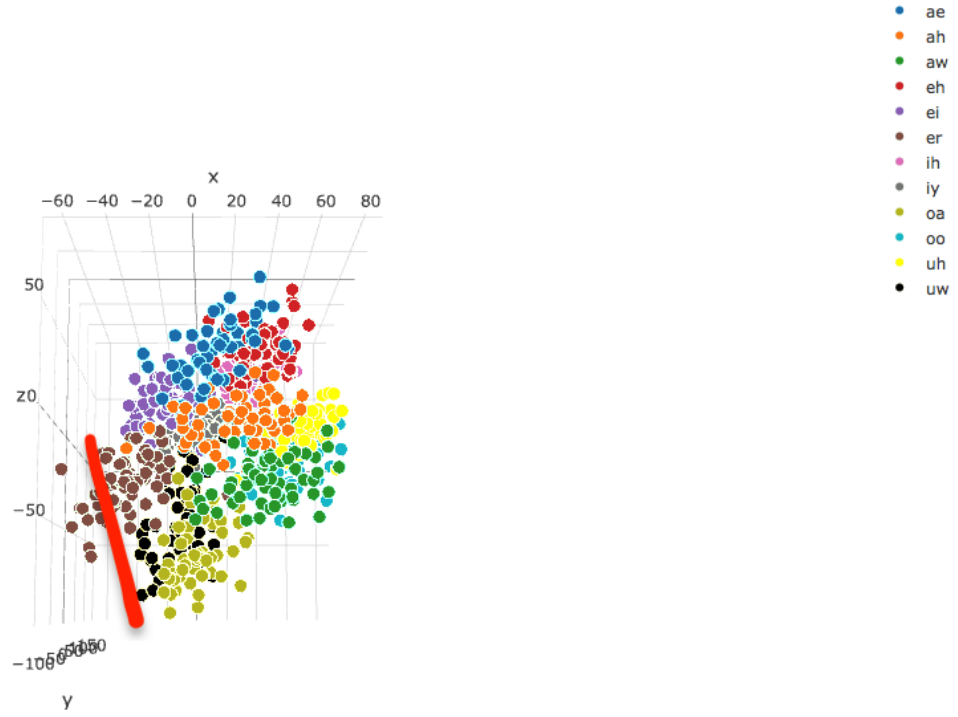

Figure 6: A view of the convolutional encoder's 3D encoding of vowel utterances, on a plane that is orthogonal to the plane that corresponds the the first and second formants. This plane has a rough correspondence to the amount of lip rounding that a vowel has.

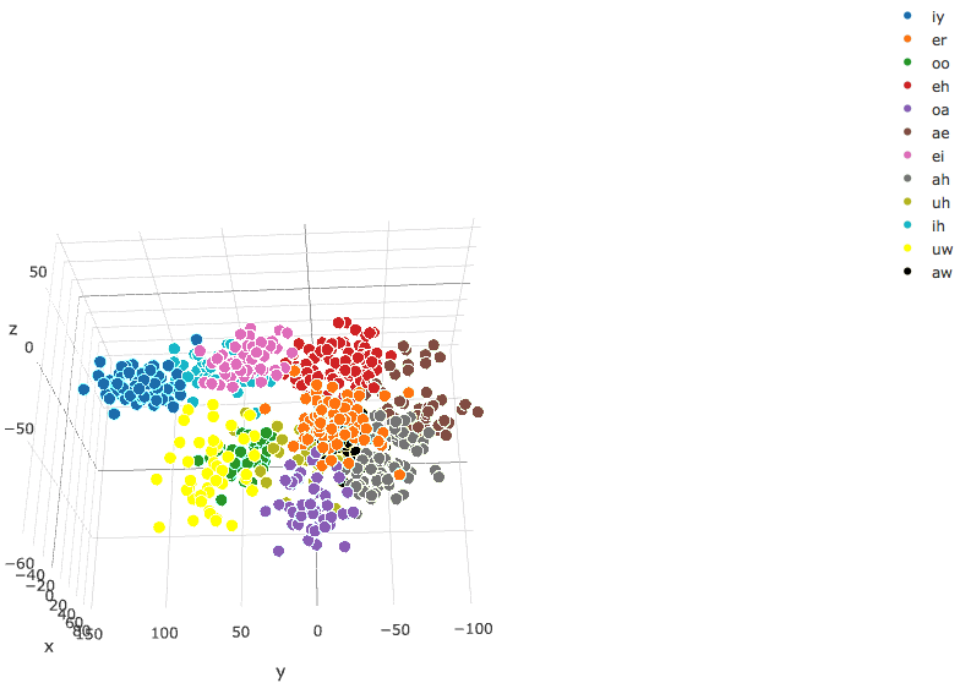

Figure 7: The same data as in Figure 5, except the colors are from learned clusters and are not from the true labels.

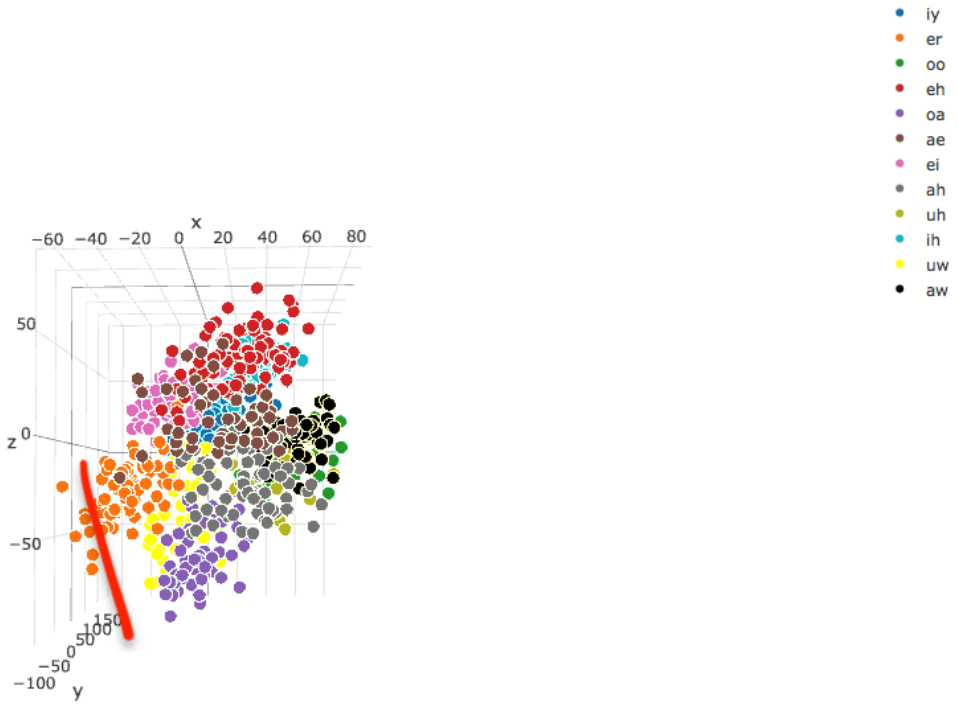

Figure 8: The same data as in Figure 6, except the colors are from learned clusters and are not from the true labels.

4.2 Recurrent Encoder

The results of the recurrent encoder are inconclusive. Given that this encoder has access to an LSTM cell, it has many more nonlinearities and parameters than its convolutional counterpart and is more prone to overfitting on a small dataset. The classifier net was able to achieve 100% training accuracy after just 20 epochs through the training data, although it could not reach a test accuracy above 33.0%, even when I increased the 3-dimensional encoding to 32 dimensions. Unsurprisingly, the accuracy of the clustering algorithm was 13.9%.

The result with the recurrent encoder is consistent with my hypotheses (it does not robustly learn features even for supervised classification, so it cannot be tested). But this encoder offers no support for my hypotheses. Further work can explore using such a model, if there is more training data.

5 Conclusion

This paper provides evidence for the hypothesis that a connectionist model that has been optimized by evolution can represent the human capacity to map vowel utterances to a space where they can be clustered into categories. My results are also consistent with the stronger assertion that humans optimally separate vowels for categorization and that a connectionist model that can sufficiently generalize to unseen vowels will end up producing a feature space that leads to a similar optimal seperation. A clustering algorithm trained on this feature space can yield a confusion matrix that is similar to the human confusion matrix. My recurrent encoder tests are at least consistent with these assertions. My convolutional encoder tests are supportive of my hypotheses, and even point to three potentially globally optimal features for vowel categorization: first formant, second formant, and rounding.

References

Bottou, L. (1998). *On-line Learning in Neural Networks*. Cambridge University Press, New York, NY, USA.

Coen, M. H. (2006). Self-supervised acquisition of vowels in american english. In *AAAI*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Hillenbrand, J. M. (Accessed May 10, 2018). *James M. Hillenbrand Homepage*. https://homepages.wmich.edu/~hillenbr/voweldata.html.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843.

Molis, M. R. (2005). Evaluating models of vowel perception. *The Journal of the Acoustical Society of America*, 118(2):1062–1071.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Xu, M., Duan, L.-Y., Cai, J., Chia, L.-T., Xu, C., and Tian, Q. (2005). Hmm-based audio keyword generation. In Aizawa, K., Nakamura, Y., and Satoh, S., editors, *Advances in Multimedia Information Processing - PCM 2004*, pages 566–574, Berlin, Heidelberg. Springer Berlin Heidelberg.

Zahorian, S. A. and Jagharghi, A. J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *The Journal of the Acoustical Society of America*, 94 4:1966–82.