A New Model of Statistical Learning: Trajectories Through Perceptual Similarity Space

Elizabeth A. Hutton* (ehutton@usc.edu), Felix Hao Wang* (wang970@usc.edu), Jason D. Zevin (zevin@usc.edu)

University of Southern California Department of Psychology, Department of Linguistics
3620 S. McClintock Ave., SGM-501, Los Angeles, CA 90089
* Denotes co-first authorship.

Abstract

Existing models of statistical learning involve computation of conditional probabilities over discrete, categorical items in a sequence. We propose an alternative view that learning occurs through a process of tracking changes along physical dimensions from one stimulus to the next within a "perceptual similarity space." To test this alternative, we examined a situation where it is difficult or impossible to label stimuli in real time, and where the two assumptions lead to conflicting hypotheses. We conducted two experiments in which human participants passively listened to a familiarization sequence of frequency-modulated tones and were then asked to make familiarity judgments on a series of test bigrams. Behavioral results were broadly consistent with a conceptualization of learning as tracking trajectories through perceptual similarity space. We also trained a neural network that codes stimuli as values along two continuous dimensions to predict the next stimulus given the current stimulus, and show that it captured key features of the human data.

Keywords: statistical learning; similarity space; connectionist modeling

Introduction

In as little as two minutes of exposure to a stream of stimuli, humans are able to absorb an underlying pattern based on statistical regularities (Saffran, Aslin & Newport, 1996). This phenomenon of learning through passive observation is called statistical learning and it has been observed in humans of all ages including neonates (Gervain, Macagno, Cogoi, Pena, & Mehler, 2008), infants (Saffran et al., 1996; Aslin, Saffran & Newport, 1998), and adults (Saffran, Johnson, Aslin, & Newport, 1999, inter alia).

Statistical learning is generally understood by assuming that learners are able to extract information from the environment by subconsciously recording and computing statistical relationships in sequences. By predicting upcoming stimuli from prior stimuli, for example, learners track transitional or conditional probabilities—that is, the probability of "x given y" (Aslin et al., 1998). These models, such as PARSER (Perruchet & Vinter, 1998) or the simple recurrent network (Elman, 1990; 1991), therefore rely on discrete representations of stimuli to segment a stream using statistics. All of these models assume that participants are quickly and accurately categorizing stimuli according to labels intended by the experimenter.

It may be problematic to assume that learners are able to make these categorical judgments in real time, in particular if statistical learning is thought to extend to natural stimuli, which are often ambiguous and highly dependent on context to identify (Hockett 1960). Here we present a novel approach to understanding statistical learning that does not assume participants are categorizing stimuli in real time. We propose that participants rely on situating stimuli within a perceptual similarity space and learn by tracking the change from one stimulus to the next within this similarity space (Emberson et al., 2013; Wang & Zevin, submitted).

We propose that by continuously tracking the perceived change from one stimulus to the next in a sequence, the learner represents stimuli relative to one another along a number of perceptual dimensions (for example, two dimensions were used in our experiments and simulations). Thus, each stimulus can be situated in a feature space defined by these dimensions (Shepard, 1965), where transitions from one stimulus to the next can be understood as the trajectory between two locations in this space. Concretely, we can model this by coding stimuli in two or more continuous dimensions. Rather than predict a discrete, symbolic stimulus from the current stimulus, such a model would predict the next location in terms of continuous values on its dimensions. A simple connectionist model provides a logical approach to simulating the phenomenon.

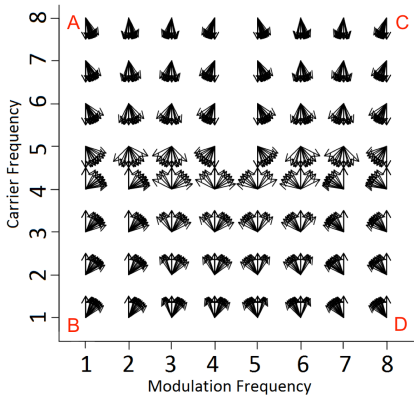

Figure 1: The angles of all possible trajectories from each point in the acoustic space following the grammar for Experiment 1 (ABCD).

For our experiments, we adapted stimuli from Holt and Lotto's studies of auditory categorization (2006). Stimuli were frequency-modulated tones uniformly distributed over a two-dimensional acoustic space that can be visualized as a grid with carrier frequency on the y-axis and modulation

frequency on the x-axis. Each stimulus was assigned a category based on its location in this space by dividing the grid into four quadrants, labeled A, B, C, and D (see Figure 1). During the familiarization phase of the experiments, stimulus tones were presented as a stream of bigrams organized by the four experimenter-defined categories (e.g. a tone from quadrant A was always followed by a tone from B). In this way, the sequences can be described alternatively as a sequence of category labels, or as a sequence of trajectories through similarity space, leading to different predictions about how participants should process the test stimuli. For example, some test stimuli violate predictions based on a sequence of category labels, but are broadly consistent with the direction of change in similarity space.

A neural network simulation provided a qualitative fit to the results from two experiments with different sequences. In fact, the model fit a difference between the two experiments that we did not predict when designing the stimuli.

Experiment 1: ABCD

Experiment 1 was motivated by a desire to replicate, with more power, an earlier study on the same topic (Wang & Zevin, submitted). Twice as many subjects were recruited and an extraneous test condition was excluded for the new version of the experiment. The experiment was designed to test the different predictions made by the two accounts discussed in the introduction: the categorization-based approach and the similarity space approach. Specifically, two different types of non-words were created: one for which the items violate the grammar but whose transition trajectory was similar to other transition trajectories in the training (Correct Trajectory Non-Word), and one for which the items never occurred in the training and the transitional trajectory was very dissimilar to other transitional trajectories in the training (Incorrect Trajectory Non-Word). If participants relied on identifying the incoming units as categories A, B, C or D, they would treat words better than non-words and treat both types of non-words as equally unfamiliar. If participants made use of the transition trajectories, Correct Trajectory Non-Words should not be as good as Words but Incorrect Trajectory Non-Words should be much worse than both Words and Correct Trajectory Non-Words.

Methods

Participants: 78 undergraduate students from The University of Southern California were recruited from the Psychology Department subject pool. They received either course credit or a payment of $5 for their participation. Due to technical errors, data was only collected for 72 of the 78 who participated.

Stimuli: The stimuli were frequency-modulated tones adapted from the studies of Holt and Lotto (2006). 64 tones were uniformly distributed over a two-dimensional acoustic space in perceptually equivalent steps (30 Hz in carrier frequency, 18 Hz in modulation frequency). The stimuli were divided into four even quadrants each containing 16 tones and labeled A, B, C, and D. Each stimulus comprised 300ms of sound and 300ms of silence.

Familiarization Phase: The entire experiment was controlled using Paradigm (Perception Research Systems, 2007) on a Windows desktop computer. Participants were allowed to read the material of their choice while passively listening through headphones to 10.5 minutes of a sound stream. The sound stream consisted of a total of 512 AB words and 512 CD words, such that all possible A-B transitions and C-D transitions were presented twice. Only half of all possible part-word transitions (from B to A or C and from D to A or C) occurred. The stimuli were chosen using a recursive algorithm to ensure even sampling from the distribution. Consequently, the transitional probability of a tone from B following one from A is 1, while the probability of a tone from A following one from B is 0.5.

Testing Phase: Immediately following the training phase, participants were instructed to make a series of familiarity judgments on 36 pairs of tones. During each trial, participants clicked anywhere on the screen to begin and a consecutive sequence of two tones was played. Following presentation of the sequence, participants were asked to indicate their familiarity with the pair of tones. A text prompt was displayed ("Do you think that you heard this sequence in the previous section?") and participants responded by clicking on one of five ratings ("Definitely", "Maybe", "Not Sure", "Maybe Not", "Definitely Not"), ending the trial. There were a total of 36 trials, 12 of each from 3 test conditions: Word, Correct Trajectory Non-Word, and Incorrect Trajectory Non-Word. Each test category had 4 unique test items that were repeated 3 times each, for a total of 12 trials per condition. To maintain consistency across conditions, all test items were novel (i.e. none of the bigrams were present in the familiarization sequence) and followed trajectories with a length of 3 arbitrary units from the first to second tone in the bigram. The Word condition contained two AB and two CD pairs, where bigrams that started in quadrant A followed the median angle for 3 units from the starting stimulus, terminating in quadrant B. In the Correct-Trajectory Non-Word condition, each pair of stimuli began and ended in the same quadrant (e.g. AA or BB) but followed a trajectory along the median angle established during the training phase (in general, towards the center of the acoustic space). The Incorrect Trajectory Non-Word condition contained the same pairs of sounds from the Correct Trajectory condition, but reversed the order in which they were played such that they followed the opposite, more unfamiliar trajectory (i.e. outwards from the center of the acoustic space). To reiterate, although the distance in feature space between each tone of a bigram remained at a constant 3 units, only items in the Word condition crossed a quadrant boundary.

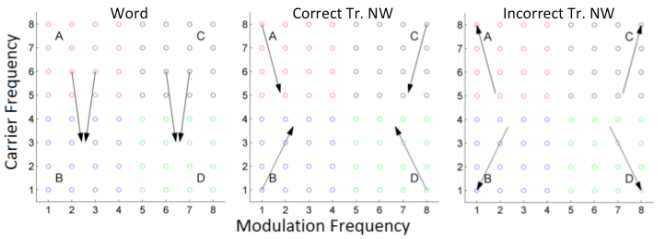

Figure 2: Visualization of the 4 test items from each of the 3 test conditions for Experiment 1 (ABCD).

Results and Discussion

Inferential tests for both experiments are based on linear mixed effects models created in Stata (StataCorp, 2013). Words were rated as significantly more familiar than items in both of the non-word conditions: Correct Trajectory ($\beta = 0.27$, $z = 5.01$, $p < 0.05$) and Incorrect Trajectory ($\beta = 0.37$, $z = 6.91$, $p < 0.05$). This result demonstrates that learning has occurred, as participants treated the grammatical bigrams as different and more familiar than the other sequences. The difference between ratings for Correct Trajectory Non-Words and Incorrect Trajectory Non-Words was marginally significant ($\beta = 0.10$, $z = 1.90$, $p = 0.057$). Although the increase from Correct to Incorrect Trajectory Non-Words was only marginally significant, it is important to note the overall trend of increasing unfamiliarity across the 3 conditions (see Figure 3) is consistent with data from Wang & Zevin (submitted).

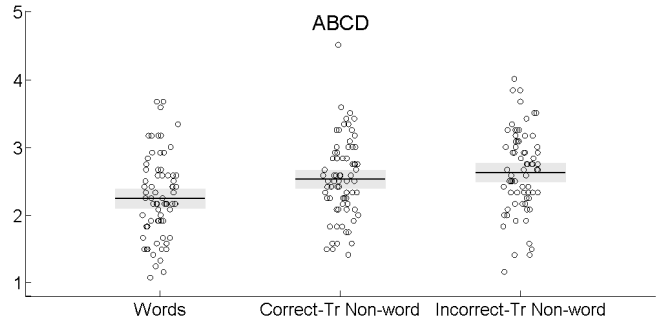

Figure 3: Ratings by test category for Experiment 1. Each dot in the scatter represents a subject's mean rating on a scale from 1 to 5 (where 1 is most familiar and 5 is most unfamiliar) for that category. The line and shadow indicate the mean rating and 95% confidence interval for all subjects in that category.

Experiment 2: ABDC

In Experiment 1 (ABCD), words were defined as transitions from a tone in quadrant A to one in B or from a tone in quadrant C to one in D, such that words could always be recognized as going down in carrier frequency. In other words, participants could have used a single dimension to learn the regularities in Experiment 1. However, we wanted to examine how participants would learn when the grammar was more complicated. So, in Experiment 2 (ABDC), words were defined as transitions from A to B or D to C, making it necessary to use both carrier frequency and modulation frequency to identify grammatical bigrams. This more complicated grammar should be harder for subjects to learn because it requires tracking two dimensions rather than one.

Methods

Participants: 84 undergraduate students from The University of Southern California were recruited from the Psychology Department subject pool. They received either course credit or a payment of $5 for their participation. Due to technical errors, data was only collected for 72 of the 84 who participated.

Stimuli: Stimuli were taken from the same acoustic space as Experiment 1. Each stimulus comprised 300ms of sound and 300ms of silence.

Familiarization Phase: In Experiment 2, words were defined as transitions A-B and D-C (rather than C-D as in Experiment 1). The sound stream contained a total of 512 AB words and 512 DC words, such that all possible A-B transitions and D-C transitions were presented twice. The procedure used was identical to Experiment 1, where participants listened to the familiarization stream passively.

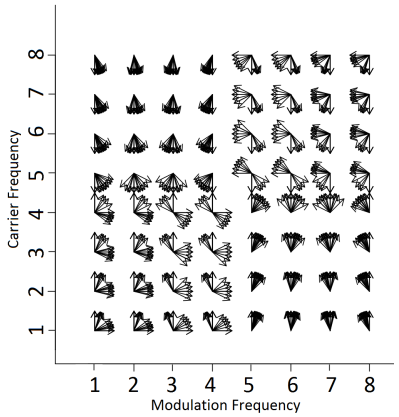

Figure 4: The angles of all possible trajectories from each point in the acoustic space following the grammar for Experiment 2 (ABDC).

Testing Phase: The testing procedure for Experiment 2 was consistent with Experiment 1, but used a different set of test items. There were a total of 36 trials, 12 of each from the same 3 test conditions: Word, Correct Trajectory Non-Word, and Incorrect Trajectory Non-Word. Each test category had 4 unique test items that were repeated 3 times each, for a total of 12 trials per condition. All test items were novel and followed trajectories with a length of 3 arbitrary units from the first to second tone in the bigram. As before, the Correct-Trajectory Non-Word pairs of stimuli began and ended in the same quadrant (e.g. AA or DD) but followed a trajectory along the median angle established

during the training phase (in general, towards the center of the acoustic space). The Incorrect Trajectory Non-Word condition contained the same pairs of sounds from the Correct Trajectory condition, but reversed the order in which they were played such that they follow the opposite, more unfamiliar trajectory (i.e. outwards from the center of the acoustic space).

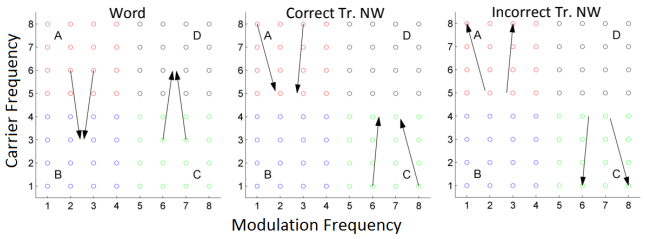

Figure 5: Visualization of the 4 test items from each of the 3 test conditions for Experiment 2 (ABDC).

Results and Discussion

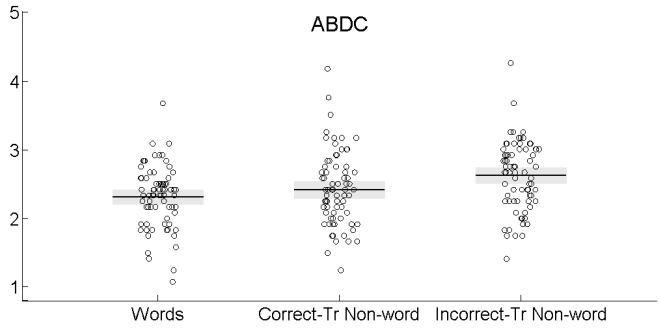

Figure 6: Ratings by test category for Experiment 2. Each dot in the scatter represents a subject's mean rating on a scale from 1 to 5 (where 1 is most familiar and 5 is most unfamiliar) for that category. The line and shadow indicate the mean rating and 95% confidence interval for all subjects in that category.

As in the previous experiment, there is robust evidence of statistical learning in Experiment 2. Unlike in Experiment 1, however, there was only a marginally significant difference between average ratings for Words and Correct Trajectory Non-Words ($\beta = 0.11$, $z = 1.96$, $p = 0.05$). As before, Words were rated as significantly more familiar than Incorrect Trajectory Non-Words ($\beta = 0.30$, $z = 5.62$, $p < 0.05$). Further, Correct Trajectory Non-Words were rated as significantly more familiar than Incorrect Trajectory Non-Words ($\beta = 0.20$, $z = 3.66$, $p < 0.05$), which indicates sensitivity to the direction of change.

Thus, results from both Experiment 1 (ABCD) and Experiment 2 (ABDC) follow the same general trend: words were rated as most familiar, followed by Correct Trajectory Non-Words, with Incorrect Trajectory Non-Words rated as most unfamiliar, although particular pairwise contrasts differ in significance across experiments. Curiously, and contrary to our initial predictions, the difference between ratings for Words and Correct Trajectory Non-Words is smaller in Experiment 1 than in Experiment 2, ($\beta = -0.17$, $z = -2.20$, $p < 0.05$).

Computational Modeling

Design and Procedure

In order to simulate learning in our experiments, we developed a simple feed-forward back-propagation, neural network using PDPTool (McClelland 1986; 2015). The neural network used a logistic activation function and had two input units, two output units, a two-unit hidden layer and a bias. Two versions of the model were trained ten times each: ABCD and ABDC, which were identically constructed but received different inputs corresponding to the 1024 stimulus sequences from Experiment 1 and 2, respectively. Each stimulus was coded as a pair of coordinates representing its location in the acoustic space. Inputs and outputs were scaled to fit within the valid range of input [-1,1] and output [0,1] values. The model was trained to predict the next stimulus from the current stimulus as bigram pairs, including both Words and Part-Words, so for example the sequence ABCD would be presented to the model in three discrete trails: AB, BC, and CD. As an initial measure of learning, we trained multiple runs for 100 epochs, collecting pattern sum of squares (pss) on the training items after each of the first ten epochs, and every fifth epoch thereafter. We observed that the model reached asymptote by this measure after ten epochs, to an error of 0.12 for ABCD and 0.14 for ABDC.

Error scores for the test items were generated by presenting the first stimulus in each test pair to the model and calculating the summed squared error (pss) for the model's output relative to the second item in the test bigram. This measure was taken for all 12 test items every 5 epochs from the 10th to the 50th, and the mean of these observations taken for each run. Means of all ten runs and standard deviations across runs are reported in Tables 1 and 2.

Results and Discussion

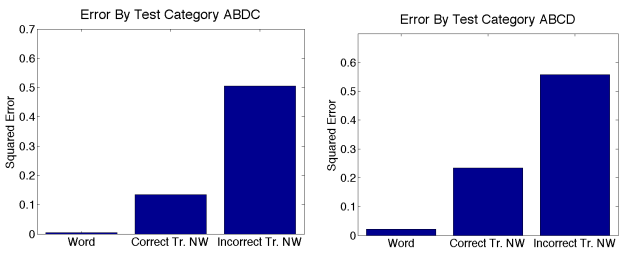

Figure 7: Average error by test category over 10 runs of the model for ABCD (right) and ABDC (left).

The computational models for ABCD and ABDC qualitatively replicated the human data. Figure 7 displays each model's error by test condition, measured as the squared distance between the model's prediction and the target (i.e. the second point in the test item). Thus, the higher the error for a test item, the further away in feature space the second tone in the pair was from the model's prediction. As such, the model's error on each test item parallels the measures of familiarity collected from the human data. As in the human data, both ABCD and ABDC display the overall increasing trend across the three test conditions, with the Incorrect Trajectory Non-Words treated as significantly different than the Words. Furthermore, there is a difference in the Correct Trajectory Non-Word condition between ABCD and ABDC. The ratio of the errors between the categories Word and Correct Trajectory Non-Word is larger in ABCD (0.0955) than in ABDC (0.0387), and it is clear from Figure 7 that this increase is larger for ABCD than ABDC. This difference qualitatively mimics the observed discrepancy between ratings for Correct Trajectory Non-Words in Experiments 1 and 2.

Table 1: Average error and standard deviation by test category over 10 runs of the simulation for ABCD.

Condition	Avg. Error	Std. Deviation
Word	0.0229	0.0094
Correct Tr. NW	0.2399	0.0165
Incorrect Tr. NW	0.5484	0.0333

Table 2: Average error and standard deviation by test category over 10 runs of the simulation for ABDC.

Condition	Avg. Error	Std. Deviation
Word	0.0052	0.0007
Correct Tr. NW	0.1342	0.0051
Incorrect Tr. NW	0.5060	0.0039

General Discussion

The behavioral results from the two experiments presented here are broadly consistent with our conceptualization of statistical learning as occurring by situating stimuli in a perceptual similarity space. Further, the computational model we designed according to this conceptualization fits the data quite well. The auditory stimuli were specifically designed to be difficult to categorize, yet participants were able to distinguish between words and non-words after brief, passive familiarization with a sequence of grammatical bigrams. Although results for the Correct Trajectory Non-Word condition differed between Experiments 1 and 2, the overall trend of increasing unfamiliarity across conditions indicates that learners are sensitive to the trajectory from one stimulus to the next in feature space.

Using the same stimuli – indeed, the same ABCD familiarization sequence used in the current Experiment 1 – Wang and Zevin (submitted) observed a small difference between Words and Correct Trajectory Non-Words, and a much larger difference between Correct and Incorrect Trajectory Non-Words. Across a number of experiments we are not reporting here due to space limitations, the general pattern of decreasing familiarity from Words to Correct to Incorrect Trajectory Non-Words is always present, although different contrasts are significant by inferential tests under different conditions. We therefore suggest that this overall pattern is the most critical feature of the data to simulate.

Interestingly, there are more subtle differences between Experiments 1 and 2 that are also captured by the simulation. Both the model and the human participants treated Correct Trajectory Non-Words as more similar to Incorrect Trajectory Non-Words in Experiment 1, but more similar to Words in Experiment 2. Until examining the simulation results, we failed to consider an idiosyncrasy with how the test items were chosen between experiments. The test items for ABCD and ABDC differed slightly in how they were sampled from throughout the feature space. As shown in Figures 2 and 5 above, the four non-word pairs for ABCD were taken from each of the four quadrants while in ABDC the four non-word items were drawn from only two quadrants (two from A and two from D). Therefore, half of the Correct Trajectory Non-Words in ABCD followed the correct trajectories for words and the other half for part-words while in ABDC they all followed trajectories for words. This could explain why the Correct Trajectory items were rated as more unfamiliar in ABCD than in ABDC for both the human experiments and the computational models.

Interestingly, the simulation's overall error, especially for Words, is lower in ABDC than ABCD. One possible explanation is that having two meaningful dimensions to define words provides the model with more information over which it can track probabilities, increasing its ability to learn the grammar. In contrast, the extra dimension introduces additional complexity that makes the sequences more difficult for humans to learn. This gets at one of the problems with the model: it is almost too good at learning the pattern. While humans must approximate each stimulus's location in similarity space, the model receives exact coordinates so naturally the model will produce more accurate and precise predictions. A further problem with the simulation lies in the fact that connectionist models like the SRN (Elman, 1990) and the one presented here all learn with supervision. While the model receives feedback on its predictions for every stimulus, human learners are thought to be dependent on unsupervised mechanisms under similar conditions (McClelland, 2006).

Furthermore, because the model was designed for a very specific experimental setting, it has limited applications. We have proposed elsewhere (Wang & Zevin, submitted) that the trajectory-tracking approach may provide an explanation for statistical learning phenomena hitherto unaccounted for by existing models. For example, word segmentation during initial language acquisition is a real-life situation in which category labels are not readily available and the sequence

signal may be ambiguous due to natural variation in human speech (Shannon, 1948; Hockett, 1960).

However, there is no reason to believe that the trajectory-tracking model tells the whole story. It is more likely that learners utilize different mechanisms, either simultaneously or individually, depending on the situation and the information that is readily available in the stimuli sequence. Relying on perceptual similarities is useful when stimuli are defined on the same dimensions and low-level physical features are readily extracted. When it is easy to abstract and divide stimuli into categories, however, there may be situations in which stimuli are readily recognizable, and it is simpler (i.e. involves lower computational load) to compute transitional probabilities over labels.

In conclusion, the results of this series of experiments and their remarkably close fit to the simulations provide overwhelming support for our theory that learning occurs by tracking changes in perceptual features from one stimulus to the next in a sequence. Although we observed a difference in one of the test conditions between the two experiments, the simulations reproduced the phenomenon, leading us to believe that it was a result of an idiosyncrasy in our test stimuli. Results from both experiments were otherwise consistent with our assertion that participants are situating stimuli within a perceptual similarity space and learn the pattern by tracking their trajectories through this space.

Acknowledgments

We would like to thank Guo Dong for help with the recursive algorithm for generating balanced stimulus sequences, and Jay McClelland for suggestions on coding the two-dimensional space for the simulations.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, 9(4), 321-324.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*(2-3), 195-225.

Emberson, L. L., Liu, R., & Zevin, J. D. (2013). Is statistical learning constrained by lower level perceptual organization?. *Cognition*, *128*(1), 82-102.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458.

Gervain J, Macagno F, Cogoi S, Pena M, Mehler J. The neonate brain detects speech structure. *Proc Natl Acad Sci U S A* 2008, 105:14222–14227.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-436.

Hockett, C. D. (1960). The origin of speech. *Scientific American*, 203, 88–111.

Holt, L. L. & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, 119, 3059-3071.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148.

McClelland, J. L., Rumelhart, D. E., and the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*. MIT Press, Cambridge, MA.

McClelland, J. L. (2006). How far can you go with Hebbian learning, and when does it lead you astray. *Processes of Change in Brain and Cognitive Development: Attention and Performance XXI*, *21*, 33-69.

McClelland, J.L. (2015). PDPTool [Computer Software]. Retrieved from http://web.stanford.edu/group/pdplab/pdphandbook

Perception Research Systems. 2007. *Paradigm Stimulus Presentation*. Retrieved from http://www.paradigmexperiments.com

Perruchet P, Vinter A. (1998). PARSER: a model for word segmentation. *J Mem Lang*, **39**:246–263.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423.

Shepard, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. *Stimulus Generalization*, 94-110.

StataCorp. 2013. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP. Retrieved from http://www.stata.com/

Wang, F. H., & Zevin, J.D. (submitted) Statistical learning of unfamiliar sounds as trajectories through a perceptual similarity space.