Psychonomic Bulletin & Review (2025) 32:139–155
https://doi.org/10.3758/s13423-024-02541-0

THEORETICAL/REVIEW

Exploring the effectiveness of reward-based learning strategies for second-language speech sounds

Craig A. Thorburn[1] · Ellen Lau[2] · Naomi H. Feldman[2,3]

Accepted: 24 May 2024 / Published online: 7 August 2024

Abstract

Adults struggle to learn non-native speech categories in many experimental settings (Goto, *Neuropsychologia, 9*(3), 317–323 1971), but learn efficiently in a video game paradigm where non-native speech sounds have functional significance (Lim & Holt, *Cognitive Science, 35*(7), 1390–1405 2011). Behavioral and neural evidence from this and other paradigms point toward the involvement of reinforcement learning mechanisms in speech category learning (Harmon, Idemaru, & Kapatsinski, *Cognition, 189*, 76–88 2019; Lim, Fiez, & Holt, *Proceedings of the National Academy of Sciences, 116*, 201811992 2019). We formalize this hypothesis computationally and implement a deep reinforcement learning network to map between environmental input and actions. Comparing to a supervised model of learning, we show that the reinforcement network closely matches aspects of human behavior in two experiments – learning of synthesized auditory noise tokens and improvement in speech sound discrimination. Both models perform comparably and the similarity in the output of each model leads us to believe that there is little inherent computational benefit to a reward-based learning mechanism. We suggest that the specific neural circuitry engaged by the paradigm and links between striatum and superior temporal areas play a critical role in effective learning.

Keywords Speech perception · Category learning · Computational modeling · Reinforcement learning

Introduction

Language learners need to map a continuous, multidimensional acoustic signal to discrete abstract speech categories. The complexity of this mapping poses a difficult learning problem, particularly for second-language learners who struggle to acquire the speech sounds of a non-native language, and almost never reach native-like ability. A common example used to illustrate this phenomenon is the distinction between /r/ and /l/ (Goto, 1971). While these sounds are distinct in English, and native English speakers easily distinguish the two sounds, native Japanese speakers find this difficult, as the sounds are not contrastive in their language. Even with much explicit training, Japanese speakers do not seem to be able to reach native-like ability (Logan, Lively, & Pisoni, 1991; Lively, Logan, & Pisoni, 1993).

A particularly effective strategy for learning speech categories as an adult, however, is through implicit learning within a video game paradigm (Wade & Holt, 2005). In these experiments, participants control a spaceship at the center of a screen with aliens entering from various locations. Participants must shoot or capture these aliens and increase their score every time they do so. Auditory tokens are presented before each alien enters and are sampled from a category that corresponds to the aliens' location and color, providing an early cue for identifying which direction the participant should face. As the experiment progresses, aliens begin to move faster and it becomes increasingly difficult to turn and shoot or capture without attending to the auditory information. Improved performance for the recognition of categories between the start and end of gameplay is observed for various types of stimuli – both synthesized noise (Lim et al., 2019) and speech tokens (Lim & Holt, 2011). In the latter experiment, adults show as much improvement in the discrimination of non-native speech sounds after 2 h of video game training as they do during 10 h of explicit training over a period of 2–4 weeks (Logan et al., 1991). Various follow-up experiments show that implicit learning paradigms of this

✉ Craig A. Thorburn
craig.thorburn@austin.utexas.edu

[1] Department of Psychology, University of Texas at Austin, Sarah M. & Charles E. Seay Bldg 108 E Dean Keeton St, Austin, TX 78712, USA

[2] Department of Linguistics, University of Maryland, College Park, MD, USA

[3] Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA

sort can be effective in auditory category learning generally (Gabay, Dick, Zevin, & Holt, 2015; Roark & Holt, 2018; Roark, Lehet, Dick, & Holt, 2022), and that implicit learning might play a unique role in the learning of categories such as speech sounds that require integration of information across dimensions (Chandrasekaran, Yi, & Maddox, 2014; Chandrasekaran, Koslov, & Maddox, 2014).

A clue to the mechanisms behind this paradigm could lie in its engagement of reward-based learning. When undergoing learning in this environment, neural activity is observed in the dorsal striatum (Lim, Fiez, & Holt, 2014; Lim et al., 2019). This region is considered to play an important role in reward prediction through the activation of dopaminergic circuits. Activation of these circuits during a category learning task is not restricted to the domain of speech: in vision, corticostriatal connections are one of many neural circuits not only involved in perception but specifically category learning (Seger & Miller, 2010). While studies on the striatum often discuss reward prediction in this region in general terms, recent studies have proposed that the basal ganglia could specifically implement reinforcement learning – one particular reward-learning algorithm (Kawagoe, Takikawa, & Hikosaka, 1998; Joel, Niv, & Ruppin, 2002; Cohen & Frank, 2009; Dabney et al., 2020).

What is special about this reward-based circuit that potentially leads to particularly effective category learning in some situations? Why does a paradigm in which category information is presented implicitly result in at least as good performance as one where categories are explicitly given to the participants? The inherent computational properties of a reward-based algorithm that is implemented by the dorsal striatal implicit learning system could distinguish it from other learning systems, resulting in differences in the rate and quality of category learning when each system is activated. Alternatively, the algorithm deployed by reward learning circuits may not in itself provide a privileged role in category learning, and any differences in learning may come from the neural properties of circuits engaged by this learning paradigm.

In this paper, we provide new computational evidence that bears on this question. We show explicitly that a reinforcement implementation of reward-based learning provides a good model of human behavior in the video game paradigm proposed by Wade and Holt (2005) across two experiments – learning noise categories (Lim et al., 2019) and speech sound categories (Lim & Holt, 2011). We compare our models with a simulation of supervised learning and show that in the noise category experiment, reinforcement learning captures one specific aspect of human data which a supervised learning model does not. In the speech sound learning experiment, supervised learning and reinforcement learning provide a similar match to human behavior.

Our work provides the first algorithmic account of video game training for category learning and provides the first computational evidence that a reinforcement mechanism may be used during video game training. Our supervised algorithm is intended to provide a baseline to which we can compare reinforcement learning, however it also bears some resemblance to explicit category learning paradigms. While the reinforcement model shows a marginally closer model fit to human data in some contexts, the benefit of the algorithm is minimal, suggesting that any differences in effectiveness between the two learning paradigms may not lie in the algorithm but in the neural circuitry involved. We argue that since the learning signal in the video game paradigm comes in the form of a reward, rather than simple category information, specific feedback loops are activated which better enables humans to update category representations. This raises new questions about the role of reinforcement learning in phonetic learning more generally. Our models also provide one of the first formal frameworks for further studying computational properties of implicit learning that could lead to particular learning outcomes in this paradigm.

The remainder of the paper is organized as follows. We first provide background on reinforcement learning and its applications to this paradigm and describe the models we use in our work. In our first simulation, we implement our reinforcement learning algorithm in a model of the Wade and Holt (2005) video game paradigm where synthesized noise sounds are presented and demonstrate that it reflects a specific pattern in human data better than a supervised algorithm. In this simulation, we do not build in prior representations to the model, allowing it to form representations solely from the input during training. Our second simulation shows that the algorithm can overcome native language knowledge and improves in discrimination of non-native speech sounds. Here, listeners possess pre-existing perceptual category mapping from their native language, and we build this prior knowledge into our model, before observing how it can change after training. We discuss the implications of these results and what may give rise to the effectiveness of the video game paradigm, suggesting that the specific neural circuitry engaged during the task may contribute to learning in this paradigm.

Background

Reinforcement learning vs. supervised learning

Reinforcement learning (Sutton & Barto, 1998) is a specific implementation of reward-based learning. In reinforcement learning, an agent takes actions within an environment and receives a reward upon reaching a favorable outcome. At each timestep, the agent receives information about the state of the environment and predicts the value of each action available to it. The agent then takes an action, observes the reward received, and updates its parameters according to the mismatch between the predicted and observed reward value. This process is iterative, enabling the agent to take actions that lead

to reward in the future. Even if the agent does not receive a reward immediately, it will still take actions that put it in a better position to receive a reward at another timestep.

While other reward-based learning algorithms exist (Rescorla & Wagner, 1972), the construction of the video game paradigm has components which lend itself clearly to an implementation of reinforcement learning. A participant receives information about the state of the environment (visual and auditory cues within the video game), takes an action within that environment (turning and shooting aliens), and receives a reward for correct actions (increases in score). The paradigm itself has been used in many experiments and results in robust learning across different variations of stimuli.

We can contrast reinforcement learning with the much more commonly used algorithm of supervised learning. In supervised learning, a function is created which maps between input and a corresponding output – in this setting, a function which maps auditory input directly to abstract perceptual categories. For any input, the model generates a corresponding output and receives the ground truth value of the output. The model can directly compare the predicted output from the model with the ground truth output and update its parameters so that its predicted output will be closer to the ground truth in the future. We make a clear distinction between these two algorithms, where reinforcement learning is given a reward signal, and the supervised algorithm is given specific category information. These are different learning signals – we consider that a reinforcement learning agent is only told *when* it performs correctly, but is not told the correct action when it does not. In contrast, the supervised network receives information about the correct decision on each trial. We introduce supervised learning, not because we believe it is a particularly strong model of a specific task, but because it provides a baseline to consider the computational properties of reinforcement learning. However, since the supervised model receives direct feedback about correct categorizations, it most closely resembles an explicit learning paradigm, where participants are asked to categorize sounds and receive feedback on the correct category after every trial.

Reinforcement learning in speech

Previous findings have indicated that reinforcement learning might be utilized in human sound category learning, motivating our choice of algorithm in this paper. When participants take part in the video game paradigm, neural regions typically associated with reinforcement and reward prediction are engaged [ie. (Lim et al., 2014; Lim et al., 2019)]. In Lim et al. (2019), participants play a video game where non-speech auditory categories are presented while activity within subcortical areas is measured using fMRI. When these categories are correlated with rewarding actions in the game, participants show increased activity in the striatum – specifically the caudate body and putamen – a subarea of the basal ganglia. This region of the brain has long been thought to play an important role in reward, through the activation of dopaminergic circuits and recent studies have suggested that the striatum could implement reinforcement learning specifically (Kawagoe et al., 1998; Joel et al., 2002; Cohen & Frank, 2009; Dabney et al., 2020).

The striatum, however, is not a monolith, and the Dual Learning Systems (DLS) Model of auditory category learning specifically proposes a split between two different systems used during learning (Chandrasekaran, Yi, & Maddox, 2014; Chandrasekaran, Koslov, & Maddox, 2014), taking inspiration from models of category learning in vision (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Ashby & Maddox, 2011). This model posits that humans have available two distinct learning systems, which perform better at learning different category types. The *reflective* system is most similar to supervised learning and is recruited where a category boundary is clearly defined over one dimension. These are considered to be 'rule-based' categories and this system engages primarily the anterior regions of the striatum. Alternatively, categories that require the integration of cues from various categories are proposed to use the *reflexive system*. These categories are classified as 'information integration' categories and are proposed to engage the caudate body and tail, and putamen, which are active during implicit learning tasks.

Given evidence from implicit learning tasks such as the video game which appear to give rise to particularly effective category learning, as well as evidence from dual systems experiments, where experimental manipulations push participants towards one learning system or the other (Yi & Chandrasekaran, 2016), it has been proposed that the learning of speech categories is *reflexive-optimal*. Because speech categories tend to require the integration of a variety of different acoustic cues, learning phonetic categories as an adult potentially benefits from the involvement of the reflexive learning system. The DLS model of category learning proposes that the posterior caudate and putamen – and therefore the reward-based algorithms that are thought to be implemented by these regions – are particularly important for second-language sound category learning.

Implicit learning has also been expanded to use the SMART task, a paradigm designed to engage the same implicit systems as the video game paradigm, but with less complexity (Gabay et al. 2015). Experiments with this paradigm indicate that learning only occurs when the actions that a participant must make are aligned directly with the auditory categories in the input, suggesting that motor actions are important and an alignment to statistical input is required for learning to occur (Roark et al., 2022). Listeners cannot be simply mapping visual and auditory cues by seeing a specific

alien appear after a specific sound, but instead are relying on a motor response to make this connection. This implies that participants are not using an unsupervised learning mechanism, where discrimination increases by passively listening to the stimuli. This fact is important for the simulations we present in this paper, as we simulate learning by modeling the actions that an agent takes when presented with auditory input.

While there are differences between implicit and explicit learning, it may not be the case that implicit learning is always more effective than explicit learning (Roark & Holt, 2018; Barrett et al., 2022). The closest comparison in the auditory domain between the two learning paradigms comes from Roark and Holt (2018), where explicit learning leads to better learning outcomes and implicit learning. However, the effectiveness seen in some papers (Lim & Holt, 2011; Lim et al., 2014, 2019), alongside the evidence posited by the DLS model outlined above, suggest that there are differences between implicit and explicit learning and we aim to shed light on any computational differences in this paper.

Computational models

We implement two computational models that take inspiration from a theoretical framework of auditory category learning in dorsal striatal regions during the video game paradigm, originally proposed by Lim et al. (2014). First, we implement a reward-based algorithm where an agent maps between state and action, and a learning signal is generated as the difference between expected and received reward. This kind of learning is described by Lim et al. (2014) as an 'indirect reward prediction error' signal, where feedback only updates auditory space indirectly through the visual domain. Our model differs in that the reward signal is simultaneously backpropagated to auditory and visual networks as described below. However, the feedback is still 'indirect', as network outcomes that the learning signal operates over are actions within the game environment and not directly mapped to auditory categories.

The second model we implement is a supervised learning algorithm, which can be considered a network that implements 'direct reward prediction error' where a feedback signal directly alters the category output of the model. Our model is a computational implementation aligned closely with the direct model discussed in Lim et al. (2014), where the outcome of the network is the auditory categories themselves.

Reinforcement learning

We define the reinforcement framework formally as a Markov decision process where the environment is given by a set of states $S = \{s_1, s_2, s_3, ...\}$, one of which is presented to the agent at each time step t. The agent then has a set of actions $A = \{a_1, a_2, a_3, ...\}$ that it can take. At each timestep, the agent receives information about the state, chooses an action, and receives a reward r_t before the environment transitions to the next state. The transition between states can depend on both the actions of the agent and external factors of the environment. The goal of the agent is to find a policy mapping between states and actions (s, a) that will maximize the total reward for this and future steps, by updating its parameters depending on the reward it receives.

In this paper, we use model-free reinforcement learning, where the policy of the agent is determined by the Q-function, defined as follows.

$$Q(s_t, a_t) = r_t + \gamma * \max Q(s_{t+1}, a_{t+1}) \quad (1)$$

This function outputs the maximum possible reward for a specific action, by summing the reward at the current timestep, r_t, plus the maximum possible reward from the next timestep, $\max Q(s_{t+1}, a_{t+1})$, multiplied by a discount parameter γ. This function is iterative since the maximum value at the next timestep will apply the same function, allowing the model to account for future reward. The result of Eq. 1 (i.e., the 'value' of a state/action pairing including current and future reward) is not directly observable for the agent, and so it must estimate the value of each action at each state by approximating this function. While there are various methods for performing this approximation, we use a specific implementation of deep Q learning (Mnih et al., 2015) to build a deep neural network that represents the Q-function (a DQN).

This network takes state information as input and outputs a distribution of values over all actions the agent can take. During training, transitions between states, the action taken, and reward received are stored in memory. At each timestep, we sample a batch of transitions, calculate the error across this batch using a Smooth L1 loss function to calculate the difference between expected and actual reward, and update parameters using stochastic gradient descent. Our simulations implement a model-free reinforcement algorithm, where the agent does not have explicit knowledge of transition probabilities, and instead must learn the action-reward mapping without this knowledge. The selection of model-free over model-based reinforcement learning is unlikely to have a large impact in our specific case, where the relevant state transitions during each trial are deterministic and the number of learning trials is large relative to the number of states and possible actions.

The agent's decisions in this model follow an ϵ-greedy algorithm to allow for a trade-off between exploring the environment to discover patterns and choosing the most optimal action to receive a reward. Throughout training, the agent will choose a random action with probability ϵ and choose

the optimal outcome with probability $1-\epsilon$, where the optimal outcome is the action with the highest value, as determined by the network. The parameter decays during training to ensure that the agent has enough time to explore the full spectrum of state-action pairings at the start of the simulation, before starting to behave more optimally.

In the experimental paradigm (Lim & Holt, 2011; Lim et al., 2019), the player receives both auditory and visual cues as they can see the location/color of the alien and hear the corresponding sound. Over time, the game speed increases and in order to have a chance at shooting the alien, the participant must rely increasingly on the auditory cues as these are presented before the alien enters the screen. In our simulations, we present the agent with one repetition of an auditory token on each timestep, as well as the current direction the spaceship is facing. We do not present visual information identifying where the alien is situated, meaning that the agent does not have grounding for the correspondence between alien locations and the auditory category. This would be like a human playing a version of the game where they can see the direction that the spaceship is facing, but cannot see the location of any aliens. In theory, this should make the task more difficult, as the agent cannot bootstrap the learning of auditory categories from any visual information present. This ensures that any success from our model is due to the exploration of the reinforcement algorithm and the reward received, rather than simple correlation between visual and auditory signals.

We discretize the continuous nature of the game, where the participant can take actions at any time, into timesteps where at each step the participant receives information about their location and environment and selects one action out of those available to them. Actions consist of turning left or right, or shooting an alien. The reward is derived from the increase in score by correctly performing these actions.

The environment for the video game in our simulations is constructed with eight different directions in which the agent can face (N, NE, E, SE, S, SW, W, NW), where a left or right action turns the agent by 45°. Aliens can enter from four directions (N, E, S, and W), and each of these is associated with a different category. Each timestep consists of the presentation of one auditory token and one action by the agent. If the agent is facing the opposite direction to the location of a newly presented alien, it will need a minimum of five actions to turn toward and successfully shoot it. To allow enough time for this to happen, each stimulus is presented eight times. If the agent does not shoot correctly within this time, then the alien disappears without the agent receiving any reward and the next stimulus is presented. During training, there are four tokens within each category presented during training and the agent has three actions available to it: TURN LEFT, TURN RIGHT, or SHOOT. The agent receives one point of reward for every alien it shoots (unlike the original study, our simulations do not have a 'capture' mechanic). Location information is presented to the model as a one-hot vector of length 8 – i.e., the value for the corresponding direction where the agent is facing is equal to 1 with all other values set to 0. This is not visual information per se, as it simply instills knowledge of the direction that the agent is facing into the model and does not encode any visual information about the location of the alien, making our model slightly different to the theoretical framework proposed by Lim et al. (2014). However, we conceptualize this similarly to the 'visual' component in that framework.

Supervised learning

For this algorithm, we use identical structure for initial network layers, however we do not include location information and instead of outputting an action, the network outputs a category directly. Throughout training, the network is presented with a batch of inputs and predicts the corresponding category for each. These predicted categories are compared to the ground truth categories for each input, and the model updates its parameters using this difference. We use a Smooth L1 loss function, which provides a squared loss at small error values and an L1 loss elsewhere, meaning it is less sensitive to outlier data points than a squared loss, to avoid the issue of exploding gradients (Girshick, 2015). This signal is backpropagated and parameters are updated using stochastic gradient descent.

For each of our comparisons, we aim to make the amount of data presented to the reinforcement and supervised learning models as close as possible. We keep the batch size and number of presentations of each stimulus constant between the two simulations, although the exact number of presentations for each token in the reinforcement learning model will vary as it will take the agent different numbers of timesteps to face and shoot the alien. As the reinforcement algorithm outputs actions and not categories, the architecture of the two models cannot be identical, however we keep the architecture of the initial layers of each model constant.

Simulation 1 - synthesized sounds

Our first simulation models the results of Lim et al. (2019) and demonstrates that a reinforcement learning algorithm is able to reflect human behavior in discrimination of auditory noise categories with simple boundaries. We simulate human experiments by training a reinforcement learning network to map between states and actions within a video game environment and a supervised model which is directly given category labels. We evaluate the models by observing how much reward the reinforcement agent receives for various categories of stimuli and how many trials the supervised network

can correctly identify. A comparison between the outcomes of the behavioral results across various token types shows that reinforcement learning better models human generalization to novel tokens and the impact of specific variability in the input over supervised learning.

Methods

We use stimuli constructed following the parameters from Lim et al. (2019), with offset and onset categories created across an experimental and control condition (Fig. 1). Offset categories are consistent across the two conditions and can be identified by just one dimension – the trajectory of change after 150 ms – where onset categories are constructed differently between the two conditions. In the original experiment, onset and offset categories also differ in the type of wave a noise carrier versus a sawtooth carrier. In the experimental condition, one must attend to both the trajectory of change during the first 150 ms of the stimulus and the starting frequency in order to identify the correct token. In the control condition, onset tokens are randomized and have no identifiable category boundary.

The construction of stimuli in this way yields three types of category boundaries, which we refer to as SIMPLE, COMPLEX, and INCOHERENT. Offset categories in both conditions have a SIMPLE boundary, as these can be distinguished along only one dimension. Onset categories have a COMPLEX boundary in the experimental condition due to the two-dimensional nature of the category boundary, whereas onset categories in the control condition are INCOHERENT as there is no category boundary available.

We simulate the synthesized sounds in this experiment by using a Gaussian distribution centered around each frequency peak at each timestep. This means that the simulation does not differentiate the offset and onset stimuli by wave type, requiring it to discover specific patterns in the pseudo-acoustic signal. Each Gaussian distribution has a peak of 10 and a variance of 150 Hz. For each of the four categories, 11 tokens are synthesized with four presented during training and the remaining seven withheld for testing. Results for both learning types are averaged over ten model runs. To simplify the presentation of tokens for the INCOHERENT boundary condition in this simulation, four specific tokens are presented, rather than a range of randomized tokens as in the original experiment.

Our reinforcement network (Fig. 2) combines the auditory representation with the location information and outputs a value assigned to each action using a deep Q network, trained with the algorithm outlined above. The network is trained over a total of 1000 episodes, each consisting of 128 tokens and tested on an additional 500 episodes of 176 tokens with weights frozen. We define performance by looking at the

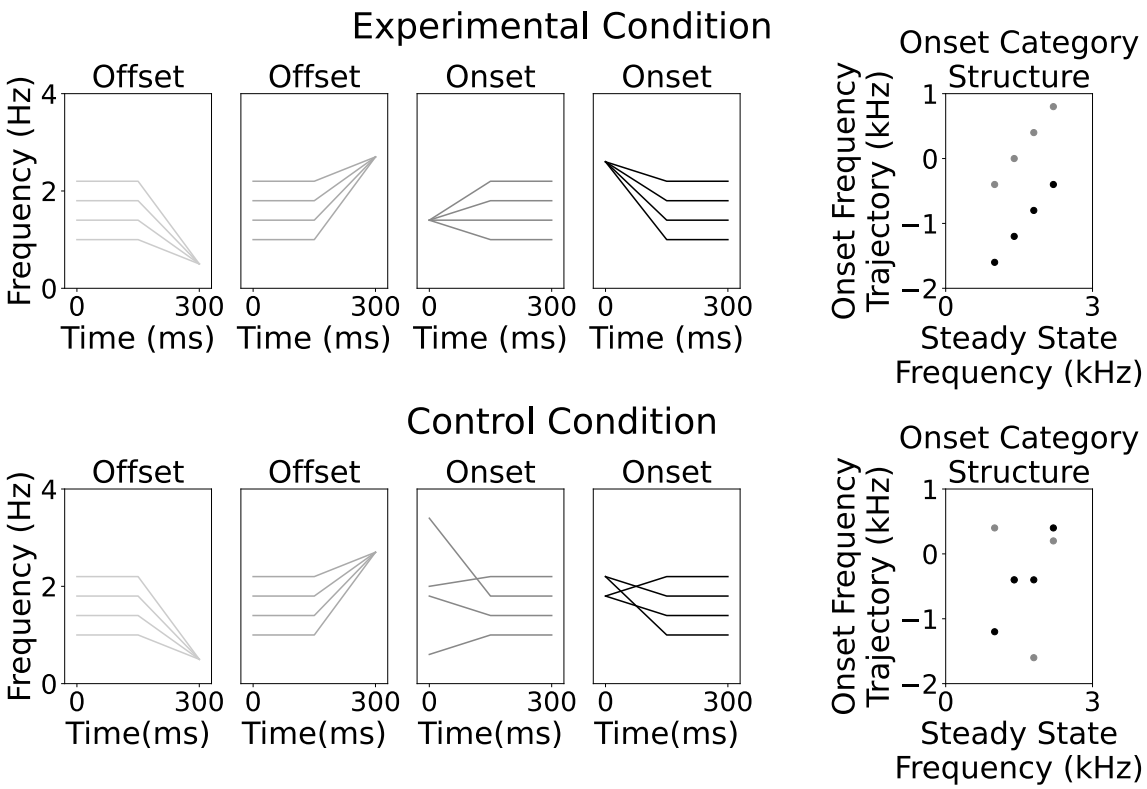
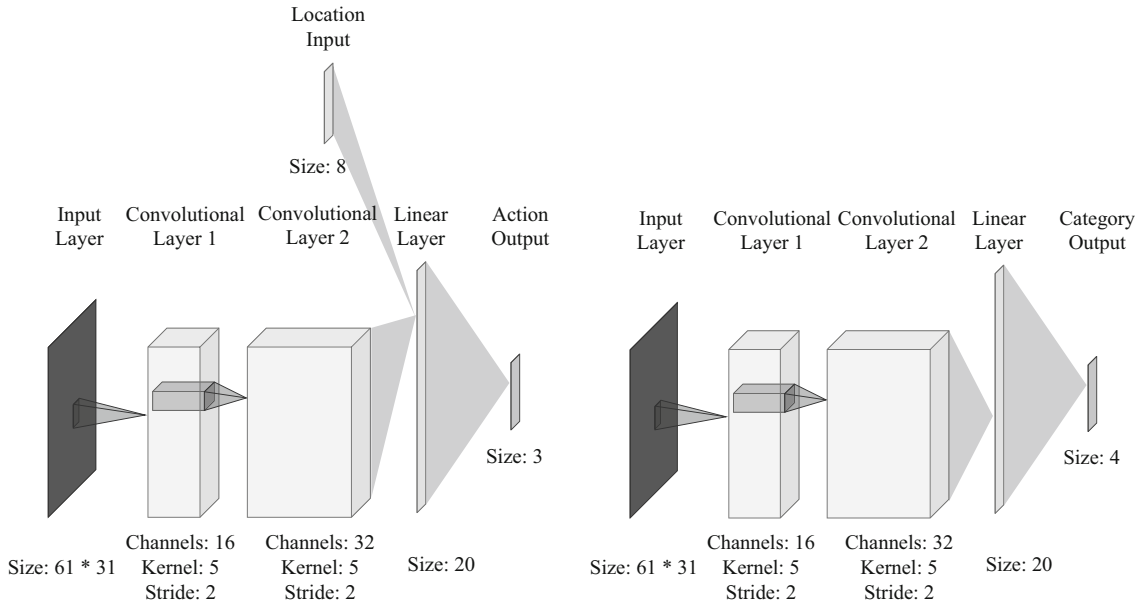


Fig. 1 Frequency profiles (*left*) and onset category structure (*right*) for simulation 1 noise categories in the experimental (*top*) and control (*bottom*) condition. Profiles from Lim et al. (2019)



(a) Reinforcement Learning Network (b) Supervised Learning Network

Fig. 2 Network diagrams for Simulation 1

proportion of total possible reward the agent receives for each category type during this test phase.

The supervised network (Fig. 2) uses solely the auditory representation and is trained to predict the category associated with each input, where each node in the output vector corresponds to one category. It is trained similarly over 128,000 tokens (equaling the 128 tokens per 1000 episodes for reinforcement learning) and evaluated by taking the node with the highest value as the predicted category for a specific stimulus. This is taken for each of the 11 tokens and averaged over different model runs.

Results

As expected, both models perform successfully in the paradigm, receiving a high proportion of the reward that is available (results are shown in Fig. 3). Like the humans, both models exhibit better performance for tokens presented during training than novel tokens, but this effect is much more pronounced for the models than the humans. The networks' generalization to new tokens allows us to observe its patterns of performance without these ceiling effects, which likely occur due to over-fitting to the data presented during training, a common occurrence with neural networks. Results show the performance of the simulated model on novel data during the testing phase closely aligns with the post-test categorization responses of humans from the original experiment.

As with humans, higher performance is observed for offset tokens than onset tokens for both models. This is because the boundary between categories has lower dimensionality in the offset condition than in the onset condition: onset tokens have a more complex boundary, being defined by both the frequency trajectory and the steady-state frequency. As we present four specific tokens for the INCOHERENT boundary, rather than a range of randomized tokens as in the original experiment, our model has a higher success rate for these tokens than the original experiments. As there is no boundary to generalize, however, we see that performance for novel tokens is low across both models – mirroring the human results.

For each model, we treat the average values for each token type across models as the mean of a distribution reflecting performance and measure how close this is to the mean of the performance distribution of human experiments using cross-entropy.[1] A high number (i.e., a negative number closer to zero) indicates that the models' distribution is closer to that of the original results. These values (Table 1) demonstrate that a reinforcement learning model captures human data better

[1] We use cross entropy to calculate model fit here — we cannot calculate log-likelihoods as we do not have access to the original experimental data and the exact number of data points for each token. Furthermore, the parameters in our models are not fit to the experimental data but rather are learned during a simulated learning process. However, if there were an equal number of trials between the experimental data and our models, then the cross entropy values shown here would be proportional to the log-likelihoods. We calculate cross entropy with the formula: $C = \sum_i x_i * log(\hat{x_i})$, where x_i is a probability distribution representing the experimental data (the mean value presented in the original paper), $\hat{x}_i$ is the estimate of this value produced by our model, and $\sum_i$ indicates the sum over both successes and failures. We subtract 0.01 from the performance of each model to avoid undefined values by calculating $log(0)$.

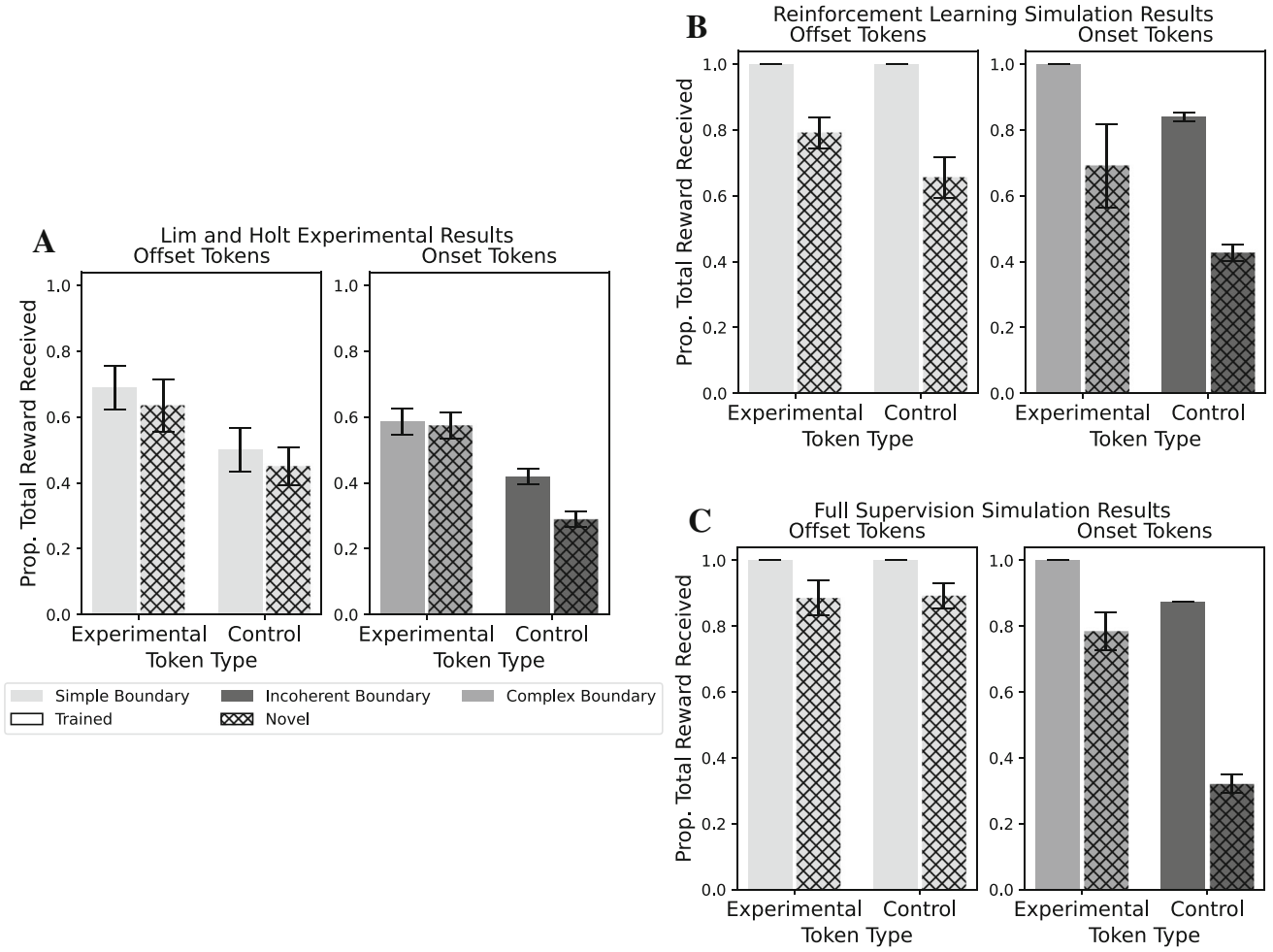


Fig. 3 Simulation 1 Results: **A** Experimental results from Lim et al. (2019). **B** Reinforcement learning results. **C** Supervised learning results with 95% confidence intervals

across multiple conditions and token types, including novel offset tokens across both conditions. The supervised model better captures data for only onset tokens in the control condition. Summing over these values we get an overall cross entropy of −9.73 for the reinforcement learning model and −10.51 for the supervised model, showing that that reinforcement learning overall reflects human results better than supervised learning.

Qualitatively, where supervised and reinforcement learning models diverge is the difference in generalization for SIMPLE stimuli between the two conditions for novel tokens. Despite the fact that these tokens are the same, humans are worse at discriminating between SIMPLE (offset) stimuli in the control condition than they are at discriminating between them in the experimental condition. This suggests that the noise provided by the incoherent tokens is enough to decrease generalization across the board, and not just for the onset tokens themselves. To test whether each of our models gives rise to this specific pattern, we compare the difference between experimental and control conditions in generalization to offset tokens across the two models, performing an ANOVA with condition and model as inde-

Table 1 Cross entropy values comparing our two models to experimental data for each type of token

	Offset tokens				Onset tokens			
	Experimental		Control		Experimental		Control	
	Trained	Novel	Trained	Novel	Trained	Novel	Trained	Novel
Reinforcement learning	−1.47	−0.72	−2.35	−0.77	−1.95	−0.71	−1.10	−0.64
Supervised learning	−1.47	−0.86	−2.35	−1.25	−1.95	−0.79	−1.24	−0.60

Reinforcement learning captures the results better for novel off set tokens in both conditions and trained onset tokens in the control condition

pendent variables and proportion of successful trials as the dependent variable. We found a main effect of both condition ($F(1, 36) = 11.55, p < 0.001$) and training ($F(1, 36) = 62.48, p < 0.001$) along with a significant interaction ($F(1, 36) = 9.36, p = 0.004$). Tests of simple effects revealed a significant difference between successfully identifying these categories in the experimental condition versus the control condition for the reinforcement learning network ($t(18) = 4.22, p < 0.001$), but not the supervised learning network ($t(18) = 0.26, p = 0.79$). This suggests that for this specific behavioral finding, a reinforcement learning algorithm better captures human data than supervised learning.

We do not perform a hyperparameter sweep, however we do test the model across a range of parameters. For the reinforcement learning model, we test learning rates between 0.01 and 0.09, and a γ value of 0.9 – values typical to Deep Q Networks. We also test a variety of game types, including where the agent receives many attempts to shoot the alien, and where it receives only one attempt. The supervised model does not have these parameters, but we match learning rates to those used in the reinforcement learning model. Data presented here use a learning rate of 0.05, however, results are consistent across these simulations.

This simulation demonstrates that it is possible to gather implicit knowledge about auditory categories through a reinforcement signal, giving rise to human-like behavior, and one specific pattern in the experimental data – namely the difference between learning of offset categories between the two conditions – is mirrored by a reinforcement model, but not a supervised model. The reinforcement learning algorithm successfully discovers the structure of input present through the reward function and actions within the environment. Having shown that a reinforcement learning algorithm is able to use perceptual information within this task, the next step is to show that the same algorithm can overcome native language knowledge and will simulate human performance in its improvement in discrimination of non-native speech categories.

Simulation 2 - Speech sounds

Our second simulation models a learning experiment where Japanese and English speakers participate in the video game with English speech tokens presented throughout play (Lim & Holt, 2011). For native Japanese speakers, discrimination between English [r] and [l] sounds can be challenging, as these sounds are distinguished primary by the F3 cue, where Japanese speakers only have one sound along this acoustic dimension. Participants in the Wade and Holt (2005) video game paradigm hear tokens of [ra], [la], [da], and [ga] before each alien enters and Japanese speakers show improvement on [r]/[l] discrimination after 5 days of training without reaching native-level performance, as measured on an identification task presented before and after the training period. We once again compare a supervised learning algorithm to a reinforcement learning algorithm in this paradigm to investigate how each algorithm overcomes native knowledge to improve on discrimination between the auditory categories presented. This simulation consists of two parts: native language training where we model the knowledge of a native Japanese speaker, and non-native training, where we expose this native model to either the video game environment or supervised learning.

Methods

We model native language knowledge by training a supervised phoneme classifier on Japanese speech. This is a simple and efficient way to instill native language knowledge into the model. The network takes acoustic parameters as input and outputs a distribution over phonemes, indicating the identity of the phone at the center of the input window. Training data consist of auditory tokens sampled from labeled corpora as described below. The model is presented with a window of speech and a corresponding vector indicating the phone at the center of the window and is trained using stochastic gradient descent over the network's parameters. We do not mean this supervised training to be an accurate model of first language acquisition, but rather, a way to quickly create networks that have knowledge of a "native" language. We similarly train an English model to use as a native language control. In this experiment, we use real speech tokens sampled from spoken corpora – stimuli which are inherently noisier than controlled lab-prepared stimuli. Using this more naturalistic data for both training and testing demonstrates that the model can deal with variability and is not sensitive only to stimuli created specifically for this experiment.

The native network (shaded in Fig. 4) consists of 3 convolutional and batch normalization layers followed by 1 linear layer. Data for native language training is derived from the Wall Street Journal Corpus for English (Paul & Baker, 1992) and the GlobalPhone Japanese Corpus (Schultz, Vu, & Schlippe, 2013) for Japanese. The model is presented with 19.5 hours of speech and training occurs for 25 epochs.

For subsequent video game training we sample sounds in each category ([r], [l], [d], and [g]) from the Wall Street Journal Corpus with each category reflecting a different location of the alien. We expose our native language model to training in the video game paradigm, where phonetic information is combined with the location vector into a linear layer as in Simulation 1 (unshaded in Fig. 4) and the model once again outputs its expected value of each action. We add two new nodes to the phoneme layer to mirror the supervised model and allow the model to map new information at this layer. The

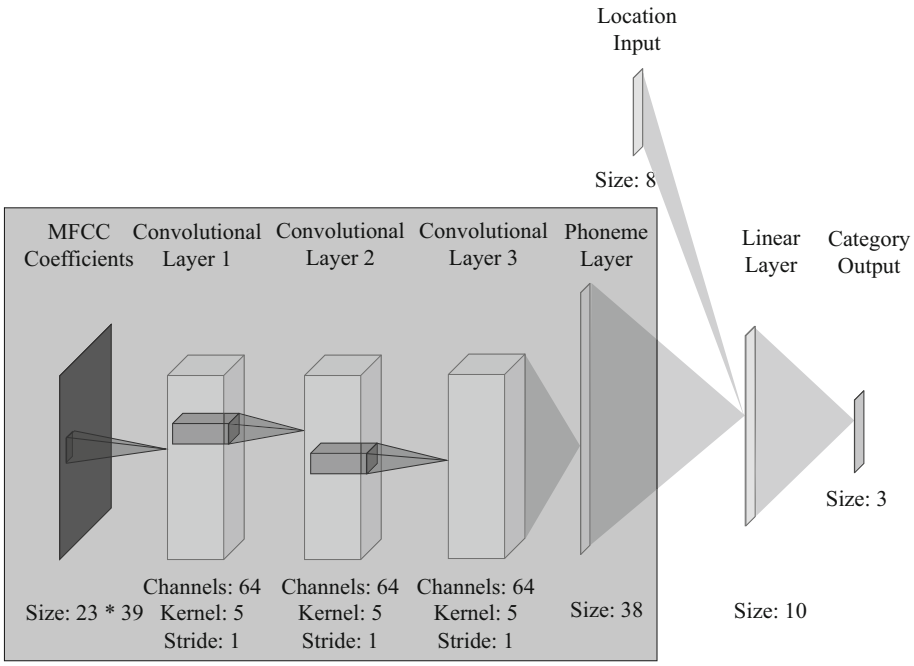


Fig. 4 Example network diagram for Simulation 2. The *shaded area* represents layers used for native language training and supervised learning. The entire network is used for reinforcement learning. The phoneme layer size is 36 for native Japanese training, 39 for native English training, and 38 for non-native training (both reinforcement and supervised)

training signal is allowed to backpropagate through all layers of the network and can change the initial parameters relating to the phone classification native language training. The environment and presentation of stimuli in the video game occurs as described in experiment 1, over 1500 episodes consisting of 128 tokens (32 for each category). As our primary goal in this experiment is seeing the outcome of video game training when there is some native knowledge already in the system, we simply attach the reinforcement learning to the output of this model at the phoneme layer.

The supervised learning models are presented with only examples of /r/ and /l/ sounds - the same tokens used in the video game training. Since the native model does not include nodes representing /r/ and /l/, we add two nodes to the phoneme layer corresponding to these two phonemes. We then continue training the model as a phoneme classifier with the same /r/ and /l/ input as in the reinforcement learning model. The model updates its parameters using the difference between its prediction of the phoneme category and the ground truth category (either /r/ or /l/). This is presented as a one-hot vector, except its length is two larger than in the native training, where the new /r/ or /l/ value is set to 1 and the other values, including those corresponding to categories learned during training, is set to 0, hence other than this increase in vector size by two, the architecture of the supervised model is virtually identical to that of the native pretraining. We aim to make the amount of training data equivalent between the reinforcement and supervised learning models. The reinforcement learning model makes a parameter update over a batch of 32 tokens every timestep. We present the supervised model with 1500 epochs of 64 datapoints – 32 from each of the /r/ and /l/ categories. We do not present /d/ and /g/ tokens as part of the supervised model as, even if they are acoustically different, these phonemes already exist in Japanese and are presented during the native training.

Neural networks often require large sets of data to train, so we train two different versions of each model. The first includes four tokens of each speech category as in the original experiment. These stimuli are those that the English native language training successfully categorizes and are kept constant throughout the simulation. In the second version of each model, we sample a large number of tokens during non-native training. Instead of sampling only four tokens of each category, we instead present a new token from the training portion of the Wall Street Journal corpus at each training step.

Acoustic input during all training is given as Mel frequency cepstral coefficients (MFCCs) (Mermelstein, 1976), which are designed to describe the overall shape of the acoustic at any one point in time. We take 13 coefficients and first- and second-order derivatives, giving 39 total dimensions. A 200-ms speech window is segmented into frames that are 25 ms wide at 10-ms intervals and zero-padded, yielding a total of 21 frames.

Neural networks are known to suffer from an issue known as catastrophic forgetting, where performance on one task is greatly reduced after training on a second task. This issue is relevant for our paradigm as we first train on a native language classification task before presenting the network with the video game training – two tasks which output quite different outputs. We are specifically focused on the discrimination of L2 contrasts in this work and use the native language model

only as a starting point for second-language learning. While we are not concerned with the continued perception of the native language, allowing the network to update all parameters in an unconstrained way would make the native language knowledge we instill irrelevant. For this reason, we add a regularization term during training which aims to find a solution to the second task (the non-native sound learning) which also performs well on the native language training (Kirkpatrick et al., 2017; Aljundi, Babiloni, Elhoseiny, Rohrbach, & Tuytelaars, 2018). This term penalizes parameters as the move away from the parameters learned during native language training and is implemented for both supervised and reinforcement networks.

We simulate the [r]/[l] pre- and post-test discrimination tasks from the original experiment as well as measure performance on the native language using a machine ABX test, which is a parameter-free method of measuring the distance between model representations (Schatz et al., 2013; Schatz, 2016). We take a vector of activations for a presented token at the 'phone' layer of the model. Two tokens, A and X, are taken from one category and a third token, B, is taken from a different category. We take the Euclidean distance between vectors A and X, and B and X and determine which distance is shorter. If X is determined to be closer to A, then the trial is a success, if X is closer to B, then the trial has failed. The ABX success rate is defined as the probability of success for two tokens from these categories selected at random from the corpus. An ABX success of 1 indicates perfect discrimination, with 0.5 being chance performance. We run the ABX task over approximately 9 million samples of [r] and [l] drawn from a portion of the WSJ and GPJ corpora withheld for testing, including 143 different speakers. Training and evaluation sets are completely independent and no token or speaker used during training is present during evaluation. For the ABX task used in our experiments, A, B, and X are all sampled from the same speaker, where X is sampled from a different 'context' (i.e., surrounding phonemes) than A and B. An ABX score is calculated by averaging the correct trials across all pairs within a specific context, before averaging performance across all contexts. All results are averaged over 6 model runs.

Results

Pre- and post-training ABX success rates are shown in Fig. 5[2] for both versions of our training – presentation of four exemplars during training, and presentation of a wide array of tokens sampled from the corpus. For models presented with only four tokens per category, the Japanese native model improves in [r]/[l] discrimination ability after non-native training across both supervised ($t(5) = 5.37$, $p = 0.002$) and reinforcement learning ($t(5) = 8.72$, $p < 0.001$) models when presented with only four tokens. This improvement is small but significant for both models and neither performance reaches that of the native English model. This is consistent with results from Japanese native speakers who participated in the experiment who show increased [r]/[l] discrimination performance but do not reach native-level performance. We note a statistically significant ($t(5) = 3.07$, $p = 0.014$), but very small difference between the performance of the supervised and reinforcement learning models after training (two tailed t-test), with reinforcement learning exhibiting slightly higher performance than supervised learning.

For the alternate version of our training with models presented with a large number of tokens sampled from the English corpus, the size of the improvement before and after video game training is much larger for both supervised ($t(5) = 24.01$, $p < 0.001$), and reinforcement training ($t(5) = 70.83$, $p < 0.001$). This model does not have an experimental counterpart, but the similar performance between supervised and reinforcement learning models is consistent over this different input. There is not a significant difference between the supervised and reinforcement trained models for this input ($t(5) = 1.93$, $p = 0.111$). We calculate cross entropy to compare our results to experimental data (Table 2). Presented results are for a learning rate of 0.1 over training, and a γ of 0.9, however, once again the results observed are consistent over various parameter settings of the network including varying learning rates learning rate ($0.01 - 0.9$) and game types as in simulation 1.

These results show that in principle, a reinforcement signal is enough to alter speech representations and enable an agent to improve its discrimination of 'non-native' speech sounds, providing additional evidence that that video game paradigm results arise due to the use of reinforcement learning. For training with limited tokens which mirrors the original experiment of Lim and Holt (2011) there is a statistically significant, but small difference in discrimination after training between the two models, where reinforcement learning shows slightly better performance than supervised learning. For training with a large number of tokens sampled from the corpus, supervised and reinforcement learning models perform equally well. Our results overall demonstrate that a reinforcement learning agent set within a video game paradigm can overcome native language knowledge. The reinforcement algorithm can use the category information implicitly provided through action and reward pairings to uncover structure in the stimuli to at least the same extent as an algorithm that takes explicit category information. This

[2] These figures show ABX scores computed at the 'phoneme' layer. In response to a reviewer's comment, we also computed ABX scores at earlier layers, but these did not change substantially during training in either the reinforcement learning or the supervised learning model, suggesting that most learning occurred at the phoneme layer.

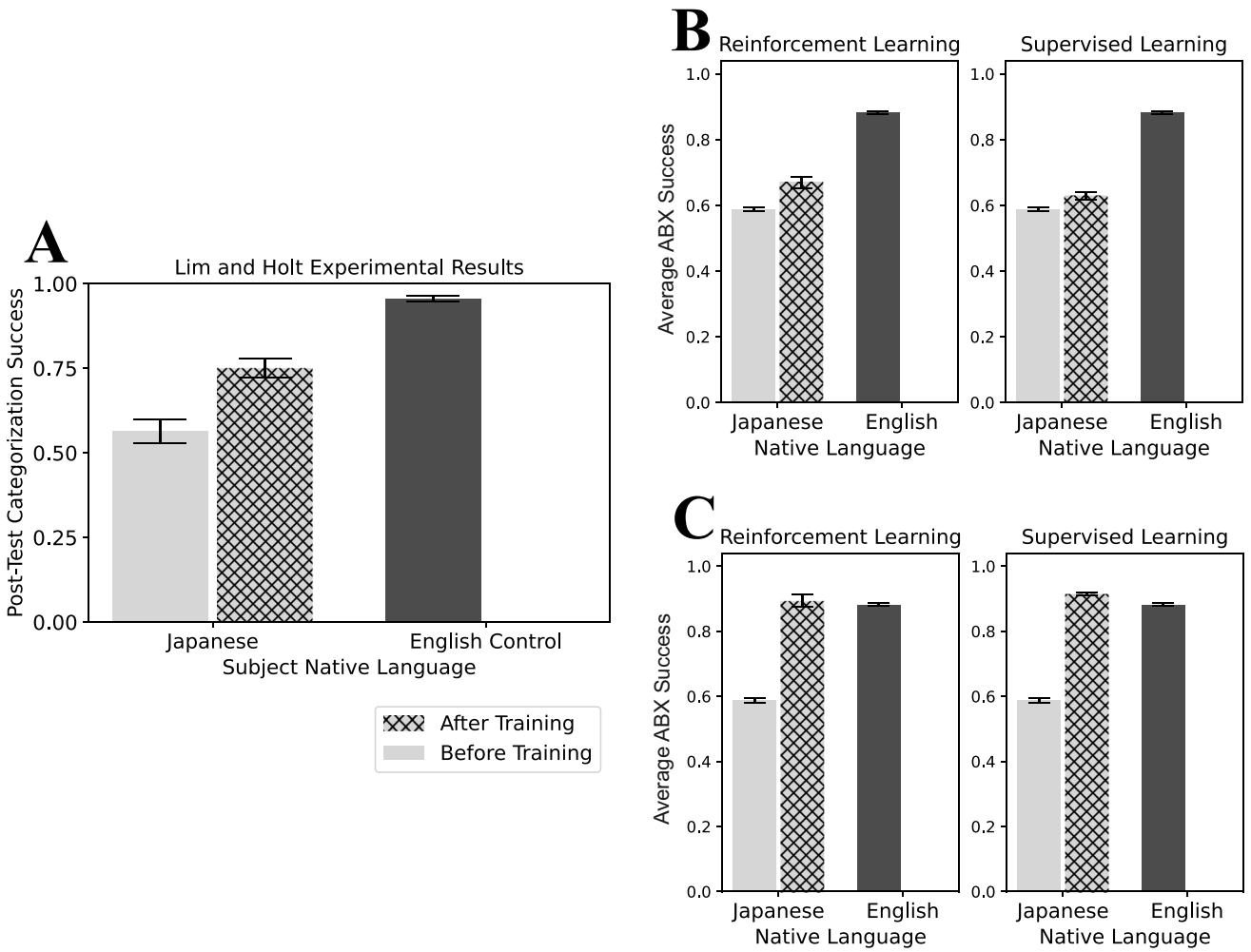


Fig. 5 Simulation 2 Results: **A** Experimental results from Lim and Holt (2011). **B** Simulation results run with only four tokens of each category presented during training. Reinforcement Learning Results (*left*) and Supervised Learning Results (*right*) with 95% confidence intervals. **C** Simulation results from run with large variety of token from each category presented during training

provides a specific account that can explain how humans are able to learn sound categories in the video game environment.

General discussion

Simulations in this paper show that a reinforcement learning network can model human behavior in the Wade and Holt (2005) video game paradigm and capture specific facets of human data better than a supervised learning algorithm. Our proposed model of the video game environment consists of a deep Q network – a reinforcement learning algorithm which maps an input of locations and auditory tokens to actions, learning at each timestep through a reward signal. In our first simulation, we demonstrate that while both algorithms are able to learn synthesized sound categories, reinforcement learning provides a better model of one specific aspect of experimental data than supervised learning. This shows that reward-based learning could be what is driving learning in the video game paradigm. In our second simulation, we build a model of a native Japanese and English speaker before exposing the Japanese-trained model to the video game paradigm

Table 2 Cross entropy values comparing our two models trained on limited tokens (as in original experiments) to experimental data

	Native language training		Limited token video game training	All tokens video game training
	English	Japanese	Japanese	Japanese
Reinforcement learning	−0.23	−0.68	−0.58	−0.64
Supervised learning			−0.60	−0.68

– again comparing this with supervised learning. We show that both the reinforcement and supervised networks perform similarly and reflect the improvement that humans show when learning /r/ and /l/ sounds.

In prior experiments, it has been shown that reward-learning systems in the striatum are active during incidental learning tasks and a high degree of functional connectivity is observed between these regions and the speech perceptual system in superior temporal gyrus (Lim et al., 2014, 2019). Taken alongside this evidence, the fact that in our simulations, reinforcement learning – a specific implementation of reward-based learning – successfully captures some qualitative aspects of the data better than a supervised algorithm supports the idea that usage of these circuits is important in adult speech category acquisition. It also demonstrates that corticostriatal loops may provide a particular role in updating perceptual representations.

Our model is similar to the framework proposed by Lim et al. (2014), and our reinforcement learning algorithm similarly generalizes across modes (vision and audition). One key difference is that both our reinforcement learning and supervised learning models in Simulation 2 include a layer that is intended to correspond to category information, whereas in the previous framework, this layer is only included in supervised learning. We choose to formulate our model in this way so we can instill phoneme category knowledge during native language training via a phoneme classification task. This also allows a fair comparison with supervised learning as both models are closely aligned in architecture and can be tested at the same layer.

Implications for second-language phonetic learning

While the model we propose in this paper shows a statistically significant difference compared to the performance of a supervised algorithm in some scenarios, the effect size is generally minimal. Behaviorally, the supervised learning simulation could be considered similar to a situation where an L2 learner recognizes there are two additional categories to learn and immediately matches any new information to one of those categories before receiving explicit feedback at the end of each trial. Previous work has given conflicting evidence on whether implicit learning is better than explicit learning. According to Lim and Holt (2011), 5 days of training in the video game paradigm appears to be more effective than 2–4 weeks of explicit training of this sort (Lively et al., 1993), although Roark and Holt (2018) show better learning outcomes for explicit learning for non-speech categories. The implicit learning paradigm is clearly effective though, and there appear to be differences in the category types that can be learned through training that engaged dorsal striatal circuits rather than frontal and anterior striatal circuits, as proposed by the Dual Learning Systems model (Chandrasekaran, Yi, & Maddox, 2014). What causes this difference? If we consider that supervised learning could model explicit paradigms, then why do our models show very similar results and what could give rise to the differences seen in prior literature?

We suggest that differences between these paradigms must arise from differences in the neural circuitry recruited in each task. Neural imaging of the video game paradigm not only shows activation of the striatum during category learning, but participants who exhibit learning show strong functional connectivity between the striatum and the superior temporal sulcus (STS) – the primary cortical center of speech processing (Liebenthal, Binder, Spitzer, Possing, & Medler, 2005). This suggests that the basal ganglia are using reward information to update auditory perceptual networks via corticostriatal loops. When the noise stimuli are randomized in the experimental counterpart to our Simulation 1 (such that the aliens do not correspond to distinct auditory categories), functional connectivity between STS and striatum is not predictive of category learning performance in the control condition where category identity is not discernable (Lim et al., 2019). This means that the connectivity is directly related to learning information about category membership of the stimuli. If this functional connectivity plays an important role in category learning, then the effectiveness of reinforcement paradigms could lie in the engagement of these specific reward circuits. When the basal ganglia are active, connections to the STS allow for effective learning; when the basal ganglia are less active, or not engaged at all, updates to the perceptual system in STS are reduced.

The neural signatures observed in initial video game experiments are also seen in other category learning paradigms (Golestani & Zatorre, 2004). These show that when learning is effective in explicit learning paradigms, greater activation is also seen in the basal ganglia. Basal ganglia activation may be particularly important for effective speech category learning in a wide range of situations. While explicit learning paradigms are able to activate these neural circuits required for effective learning, the video game by its nature and setup as a reward-based learning paradigm, may engage basal ganglia circuits more specifically and effectively. The fact that feedback in the video game paradigm comes as a reward signal, rather than simple category information, activates the reward-based circuits in this region.

However, this still leaves open the question of why updates induced by the basal ganglia are particularly effective. What is special about activating these circuits? We suggest the possibility that a frontal system may perform separate updates to the striatal system activated through reward-based learning. These circuits may have particular properties that modulate their effectiveness for speech sound learning.

One difference between the systems could arise from more rapid plasticity induced by basal ganglia feedback circuits. Theories in visual category learning propose that

different feedback loops may result in plasticity on different timescales. As one example, Seger and Miller (2010) suggest that the striatum forms rapid representations of reward predictions that can gate updates to cortex. Under this theory, striatal areas are responsible for forming the connection between specific cues and actions and frontal areas are responsible for generalization and identifying the properties that are consistent across categories. Within this framework, the effectiveness of the implicit learning paradigm would arise because – even though both striatal and frontal neural circuitry update the perceptual system in the same way – the striatal circuitry causes faster plasticity in perceptual regions.

Alternatively, our work conforms with the Dual Learning Systems (DLS) model for speech category learning recently proposed by Chandrasekaran, Koslov, and Maddox (2014). This theory builds on two system models in vision literature and posits that a *reflective* system generates category boundaries by forming explicit rule-based updates and a *reflexive* system generates implicit associations between tokens. A neurobiological split between the systems is similar to that proposed in vision research is suggested (Ashby et al., 1998; Ashby & Maddox, 2011), where the reflective system uses frontal circuits and and the anterior striatum, and the reflexive system using circuitry in the dorsal striatum. The framework is supported by both behavioral (Chandrasekaran, Yi, & Maddox, 2014), and neural studies (Feng et al., 2021).

The models presented in this paper could be considered to approximately equate with the two systems in this framework. Our reinforcement algorithm closely aligns with the *reflexive* system which engages reward circuits and our supervised algorithm could be considered to be more akin to the *reflective* system. Under the initial two systems models it is assumed that the more successful system will dominate at a given time (Ashby et al., 1998). If explicit rules are best suited to describing the task, the *reflective* system will dominate learning and if the boundary between categories is better described as an integration problem over multiple cues the *reflexive* system will dominate. Chandrasekaran, Yi, and Maddox (2014) propose that speech category learning is *reflexive-optimal*, due to the resources needed to generate rules to capture categories in such a highly multidimensional space. Under this assumption, in the video game paradigm the *reflexive* system gives rise to better learning outcomes and dominates over the *reflective* system. The results in our study are consistent with this idea as the reinforcement learning model, which closely aligns with the expected computations of the *reflexive* system, aligns more closely with human outcomes in specific situations. In our paper, we do not model the interaction between these systems, only the outputs that might be given by each. Future work should look more directly at the competition between the two systems.

Interestingly, the optimality of the *reflexive* system for speech could coincide with recent work in infant category learning which indicates that learning dimensions of an auditory perceptual space could be a separate process from the formation of specific categories during infancy (Feldman, Goldwater, Dupoux, & Schatz, 2021). We speculate one possibility that the reason the reflexive system is particularly effective in humans, could be because it targets an earlier stage of the mapping to phonetic categories – i.e., it targets the warping of perceptual space, from which phonetic categories can be built, as opposed to the stage which forms phonetic categories.

Our modeling results add to the growing body of evidence that reinforcement learning plays a role in speech category learning in adults. Previous modeling work outside of implicit learning, indicates that reinforcement learning provides a better model of human behavior than other algorithms. Many speech sound learning algorithms posit that learning occurs through gathering statistical information about one's environment, however Nixon (2020) shows that these algorithms cannot predict particular cue-outcome relationships. In humans, a learned informative cue which is associated with a specific outcome will block the learning of any other cues to the same outcome, and while statistical learning cannot account for this effect, error-prediction algorithms related to reinforcement learning can. Reward-based algorithms also predict ordering effects of cue and outcome (Nixon, 2020) as well as other cue-weighting effects (Harmon et al., 2019). The results we present in this paper, alongside these previous studies, point towards reinforcement learning as a mechanistic account of speech category learning in adults.

Infant phonetic learning

In the future, we hope to expand this algorithm to model native language learning in infants more generally, as they appear to acquire speech sound categories in a passive environment without direct feedback. Studies on auditory category learning show that adult listeners can learn one-dimensional stimuli through passive exposure, however they usually require explicit feedback to learn auditory stimuli that vary on more than one dimension (Goudbeek, Swingley, & Smits, 2009; Yi & Chandrasekaran, 2016). When infants learn the sounds of their native language, however, they are not told what specific sounds belong in what phoneme categories. While they receive little to no direct cstill becomes attuned to the speech sounds of their native language at a young age (Werker, Polka, & Pegg, 1997; Kuhl et al., 2006). Many theories posit that they may accumulate statistical information about their acoustic environment (Vallabha & McClelland, 2007), although there are some effects that purely statistical learning cannot account for (Nixon & Tomaschek, 2021). The idea of 'intrinsic reward' within reinforcement learning literature posits algorithms where

rewards come from within the system, rather than from the external environment (Singh, Lewis, Barto, & Sorg, 2010). For example, an algorithm that explores a wide area of the solution space may receive a reward for doing just that, even if it does not take rewarding actions (Eysenbach, Gupta, Ibarz, & Levine, 2018). These ideas could perhaps be applied to infant phonetic learning, where rewards are provided from higher levels of cognition, such as being able to form categories that allow for successful word and pattern recognition. This would contribute towards an explanation of how they are able to learn speech sound categories effectively at a young age without explicit feedback.

Reinforcement learning may also provide a good model of experimental paradigms that are typically used with infants. A common infant testing procedure is the conditioned head-turn (CHT) paradigm, where an infant on a caregiver's lap is trained to turn their head to a toy whenever they notice a change in a stimulus. When they do this correctly, they are rewarded with the toy lighting up and making a sound. This very easily fits as a reinforcement learning algorithm where the learning signal during the experiment that the infant receives comes in the form of a reward for good actions. While these paradigms are typically thought to reflect pre-existing knowledge that the infant has, the effectiveness of a reward-based paradigm in adults may lead us to consider that infants learn during the experiment as well.

Conclusions and future directions

Next steps require identifying the specific range of tasks that appear to activate the corticostriatal pathways recruited in this paradigm. Harmon et al. (2019) demonstrate that reinforcement learning reflects human behavior over other algorithms in a paradigm where the down-weighting of a phonetic cue will occur only if there is a more informative cue present. This task may be another which selectively activates basal ganglia circuits. There are, however, few additional examples in which we know that the basal ganglia are recruited or that specific predictions differ between reinforcement and supervised learning.

As discussed above, we have focused on learning to perceive L2 contrasts in this work, and have assumed that learning starts from native language speech perception. Since we do not model the continued perception of speech in the first language, we are agnostic with regard to the extent to which the first and second-language speech systems are shared or separate. For this reason, we do not discuss the ability of the system to discriminate native language contrasts after [r]/[l] training. Our regularization term ensures that the network uses native knowledge during non-native training, however both models that we study in this paper would show catastrophic forgetting for the native language tokens. Thus, for the present paper, we assume we are modeling an L2 perceptual system that is simply initialized to be identical to the L1.

Our framework opens avenues to investigating the flexibility of representations of speech and future work can manipulate which layers of the network can be influenced by the reinforcement learning signal, and how those weight updates occur. Our model could provide a basic framework to test whether there are particular parts of the processing system that are less flexibly updated, or not updated at all, during second-language learning. Further exploration of the regularization term as a way to constrain the plasticity of the system is also an interesting direction for future modeling work.

In the future our model could also be augmented to include varying levels of motivation as part of the simulation. Motivation and engagement could be manipulated by changing the amount of reward that the agent receives for taking correct actions, the precise actions for which the agent receives reward, or the probability by which the agent receives a reward for taking a correct action (i.e., the agent only receives a reward for the correct action 80% of the time). In the potential extension to infant phonetic learning, intrinsic motivation can be incorporated into the model, where the reward function accounts for actions that explore the entire action or reward space (an example of such a computational model is Eysenbach et al., 2018).

Finally, the simulations in this paper have implications for second-language teaching pedagogy. Learning the sounds of a second language has always been challenging, particularly when those sounds do not match up with one's native language phonology. Our work reinforces the idea that implicit learning can provide an effective strategy for learning non-native speech sounds. Understanding why this paradigm is effective will allow us to consider other tasks and activities that could lead to successful learning. Many people use mobile game-like applications to learn foreign languages, which often include a level system where learners receive rewards such as cosmetic upgrades, items, or points for a leaderboard system. This work raises questions about the exact role that these rewards play within second-language learning. Reinforcement learning – where improvement is motivated through rewards – is a particularly relevant topic and further simulations on the impact of rewards could potentially benefit second-language education.

Acknowledgements We thank William Idsardi, Saahiti Potluri, the UMD Language Acquisition Lab, and the UMD Phonology Circle group for helpful comments and discussion. This research was supported by NSF grant BCS-2120834.

Author Contributions Craig Thorburn: Conceptualization, Methodology, Software, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. Ellen Lau: Conceptualization, Writing - Review & Editing, Supervision. Naomi Feldman: Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Funding This research was supported by NSF grant BCS-2120834.

Availability of data and materials Not Applicable

Code Availability The code used in this project is available at: www.github.com/CraigThorburn/VidGameRL

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Not Applicable

Consent to participate Not Applicable

Consent for publication The authors give consent for publication

Open Practices Statement The code used to generate the data in Simulation 1 and Stimulation 2 are available, and neither simulation was preregistered.

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory aware synapses: Learning what (not) to forget. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), (Vol. 11207, pp. 144–161). https://doi.org/10.1007/978-3-030-01219-9_9

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442–481. https://doi.org/10.1037/0033-295x.105.3.442

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences, 1224*(1), 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Barrett, R. C. A., Poe, R., O'Camb, J. W., Woodruff, C., Harrison, S. M., Dolguikh, K., . . . Blair, M. R. (2022). Comparing virtual reality, desktop-based 3D, and 2D versions of a category learning experiment. *PLOS ONE, 17*(10), e0275119. https://doi.org/10.1371/journal.pone.0275119

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology, 5*

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review, 21*(2), 488–495. https://doi.org/10.3758/s13423-013-0501-5

Cohen, M. X., & Frank, M. J. (2009). Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural Brain Research, 199*(1), 141–156. https://doi.org/10.1016/j.bbr.2008.09.029

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature, 577*(7792), 671–675. https://doi.org/10.1038/s41586-019-1924-6

Eysenbach, B., Gupta, A., Ibarz, J., & Levine, S. (2018). Diversity is All You Need: Learning Skills without a Reward Function. arXiv:1802.06070 [cs].

Feldman, N. H., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do Infants Really Learn Phonetic Categories? *Open Mind, 5*, 113–131. https://doi.org/10.1162/opmi_a_00046

Feng, G., Gan, Z., Yi, H. G., Ell, S. W., Roark, C. L., Wang, S., & Chandrasekaran, B. (2021). Neural dynamics underlying the acquisition of distinct auditory category structures. *NeuroImage, 244*, 118565. https://doi.org/10.1016/j.neuroimage.2021.118565

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental auditory category learning. *Journal of experimental psychology. Human perception and performance, 41*(4), 1124–1138. https://doi.org/10.1037/xhp0000073

Girshick, R. (2015). Fast r-cnn.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: reallocation of neural substrates. *NeuroImage, 21*(2), 494–506. https://doi.org/10.1016/j.neuroimage.2003.09.071

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds L and R. *Neuropsychologia, 9*(3), 317–323. https://doi.org/10.1016/0028-3932(71)90027-3

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of experimental psychology. Human perception and performance, 35*(6), 1913–1933. https://doi.org/10.1037/a0015781

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition, 189*, 76–88. https://doi.org/10.1016/j.cognition.2019.03.011

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks: The Official Journal of the International Neural Network Society, 15*(4–6), 535–547. https://doi.org/10.1016/s0893-6080(02)00047-3

Kawagoe, R., Takikawa, Y., & Hikosaka, O. (1998). Expectation of reward modulates cognitive signals in the basal ganglia. *Nature Neuroscience, 1*(5), 411–416. https://doi.org/10.1038/1625

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., . . . Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13–F21. https://doi.org/10.1111/j.1467-7687.2006.00468.x

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15*(10), 1621–1631. https://doi.org/10.1093/cercor/bhi040

Lim, S.-J., Fiez, J., & Holt, L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences, 116*, 201811992. https://doi.org/10.1073/pnas.1811992116

Lim, S.-J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception? *Frontiers in Neuroscience, 8*. https://doi.org/10.3389/fnins.2014.00230

Lim, S.-J., & Holt, L. L. (2011). Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. *Cognitive Science, 35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic

environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America, 94*(3 Pt 1), 1242–1255. https://doi.org/10.1121/1.408177

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America, 89*(2), 874–886.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, 374–388.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition, 197*, 104081. https://doi.org/10.1016/j.cognition.2019.104081

Nixon, J. S., & Tomaschek, F. (2021). Prediction and error in early infant speech learning: A speech acquisition model. *Cognition, 212*, 104697. https://doi.org/10.1016/j.cognition.2021.104697

Paul, D. B., & Baker, J. M. (1992). The Design for the Wall Street Journal-based CSR Corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman*, New York, February 23–26, 1992

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: The effectiveness of reinforcement and non-reinforcement. *Classical Conditioning: Current Research and Theory*

Roark, C. L., & Holt, L. L. (2018). Task and distribution sampling affect auditory category learning. *Attention, Perception, & Psychophysics, 80*(7), 1804–1822. https://doi.org/10.3758/s13414-018-1552-5

Roark, C. L., Lehet, M. I., Dick, F., & Holt, L. L. (2022). The representational glue for incidental category learning is alignment with task-relevant behavior. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(6), 769–784. https://doi.org/10.1037/xlm0001078

Schatz, T. (2016). *ABX-disciminability measures and applications* (Doctoral Dissertation). Universite Paris 6 (UPMC)

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., Dupoux, E. (2013). Evaluating speech features with the Minimal-Pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013 14th Annual Conference of the International Speech Communication Association* (pp. 1–5). Lyon, France.

Schultz, T., Vu, T., & Schlippe, T. (2013). GlobalPhone: A Multilingual Text & Speech Database in 20 Languages.. https://doi.org/10.1109/ICASSP.2013.6639248

Seger, C. A., & Miller, E. K. (2010). Category Learning in the Brain. *Annual review of neuroscience, 33*, 203–219. https://doi.org/10.1146/annurev.neuro.051508.135546

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development, 2*(2), 70–82. https://doi.org/10.1109/TAMD.2010.2051031

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press.

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. *Cognitive, Affective, & Behavioral Neuroscience, 7*(1), 53–73. https://doi.org/10.3758/CABN.7.1.53

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics, 67*(6), 939–950. https://doi.org/10.3758/BF03193621

Werker, J. F., Polka, L., & Pegg, J. E. (1997). The conditioned head turn procedure as a method for testing infant speech perception. *Early Development and Parenting, 6*(3–4), 171–178. https://doi.org/10.1002/(SICI)1099-0917(199709/12)6:3/4<171::AID-EDP156>3.0.CO;2-H

Yi, H. G., & Chandrasekaran, B. (2016). Auditory categories with separable decision boundaries are learned faster with full feedback than with minimal feedback. *The Journal of the Acoustical Society of America, 140*(2), 1332–1335. https://doi.org/10.1121/1.4961163