A Computational Account of the Role of Cochlear Nucleus and Inferior Colliculus in Stabilizing Auditory Nerve Firing for Auditory Category Learning

Irina Higgins
irina.higgins@gmail.com
Simon Stringer
simon.stringer@psy.ox.ac.uk
Department of Experimental Psychology, University of Oxford,
Oxford, OX2 6GG, U.K.

Jan Schnupp
jan.schnupp@googlemail.com
Department of Physiology, Anatomy and Genetics, University of Oxford,
Oxford, OX1 3QX, U.K.

It is well known that auditory nerve (AN) fibers overcome bandwidth limitations through the volley principle, a form of multiplexing. What is less well known is that the volley principle introduces a degree of unpredictability into AN neural firing patterns that may be affecting even simple stimulus categorization learning. We use a physiologically grounded, unsupervised spiking neural network model of the auditory brain with spike time dependent plasticity learning to demonstrate that plastic auditory cortex is unable to learn even simple auditory object categories when exposed to the raw AN firing input without subcortical preprocessing. We then demonstrate the importance of nonplastic subcortical preprocessing within the cochlear nucleus and the inferior colliculus for stabilizing and denoising AN responses. Such preprocessing enables the plastic auditory cortex to learn efficient robust representations of the auditory object categories. The biological realism of our model makes it suitable for generating neurophysiologically testable hypotheses.

1 Introduction

The hierarchy of the auditory brain is complex, with numerous interconnected subcortical and cortical areas. While a wealth of neural response data has been collected from the auditory brain (Winter & Palmer, 1990; Recio & Rhode, 2000; Schnupp, Hall, Kokelaar, & Ahmed, 2006), the role of the computations performed within these areas and the mechanism by which the sensory features of auditory objects are transformed into higher-order representations of object category identities are yet unknown (Bizley & Cohen,

Neural Computation **30, 1801–1829** (2018)
doi:10.1162/neco_a_01085

2014). How does the auditory brain learn robust auditory categories, such as phoneme identities, despite the large acoustical variability exhibited by the raw auditory waves representing the different auditory object exemplars belonging to a single category? How does it cope once this variability is further amplified by the spike time stochasticity inherent in the auditory nerve (AN) when the sounds are encoded into neuronal discharge patterns within the inner ear?

One of the well-accepted theories explaining the information encoding operation of the AN is the so-called volley principle (Wever & Bray, 1930). It states that groups of AN fibers with a similar frequency preference tend to phase-lock to different randomly selected peaks of a simple sinusoidal sound wave when the frequency of the sinusoid is higher than the maximal frequency of firing of the AN cells. This allows the AN to overcome its bandwidth limitations and represent high frequencies of sound through the combined frequency of firing within groups of AN cells. It has not been considered before, however, that the information-encoding benefits of the volley principle may come at a cost. Here we suggest that this cost is the addition of the so-called spatial jitter to the AN firing.

It is useful to think of the variability in AN discharge patterns as a combination of temporal and spatial jitter. Temporal jitter arises when the AN fiber propensity to phase-lock to temporal features of the stimulus is degraded to a greater or lesser extent by Poisson-like noise in the nerve fibers and refractoriness (Eggermont, 2001). "Spatial jitter" refers to the fact that neighboring AN fibers have almost identical tuning properties so that an action potential that might be expected at a particular fiber at a particular time may be observed in one of the neighboring fibers (Wever & Bray, 1930). In this letter we hypothesize that space and time jitter obscure the similarities between the AN spike rasters in response to different presentations of auditory stimuli belonging to the same class, thus impeding auditory object category learning.

The reason we believe that excessive jitter in the AN can impair auditory object category learning in the auditory cortex is the following. Previous simulation work has demonstrated that one way category learning can arise in competitive feedforward neural architectures characteristic of the cortex is through the continuous transformation (CT) learning mechanism (Stringer, Perry, Rolls, & Proske, 2006; Evans & Stringer, 2012). CT learning is a biologically plausible mechanism based on Hebbian learning, that operates on the assumption that highly similar, overlapping input patterns are more likely to be different exemplars of the same stimulus class. CT learning then binds these similar input patterns together onto the same subset of higher-stage neurons that thereby learn to be selective and informative about their learned preferred stimulus class. The CT learning principle is a biologically plausible mechanism for learning object transformation orbits as described by Liao, Leibo, and Poggio (2013). CT learning breaks when the similarity between the nearest-neighbor exemplars within a stimulus class

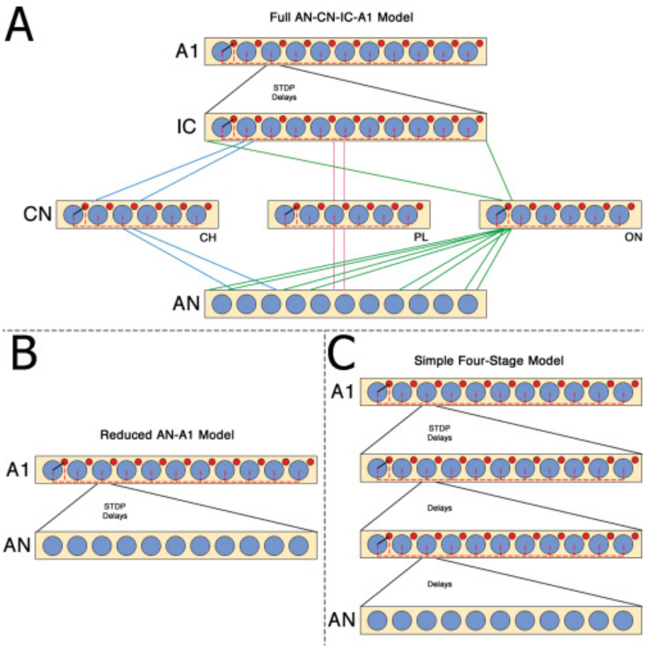

Figure 1: Schematic representation of the full AN-CN-IC-A1 (A), the reduced AN-A1 (B), and the simple four-stage (C) models of the auditory brain. Blue circles represent excitatory (E) and red circles represent inhibitory (I) neurons. The connectivity within each stage of the models is demonstrated using one excitatory cell as an example: E→I connection is shown in black, and I→E connections are shown in red. Feedforward connections between the last two stages of each model are modifiable through STDP learning. AN: auditory nerve. CN: cochlear nucleus with three subpopulations of cells: chopper (CH), primary-like (PL) and onset (ON), each exhibiting different response patterns by virtue of their distinct connectivity. IC: inferior colliculus. A1: primary auditory cortex.

become approximately equal to the similarity between the nearest-neighbor exemplars in different stimulus classes. A more detailed description of CT learning is provided in section 4.

In this letter, we hypothesize that the additional spike time variability introduced in the AN input representations of the different exemplars belonging to a single auditory object class break CT learning. We provide evidence for our hypothesis by training a biologically realistic feedforward spiking neural network model of the auditory cortex with spike-timing-dependent plasticity (STDP) learning (Bi & Poo, 1998) to perform simple categorization of two synthesized vowel classes using raw AN firing as input (see the AN-A1 model shown in Figure 1B). We show that such a model is unable to solve this easy categorization task. We suggest that this is because the reproducibility of AN firing patterns for similar stimuli necessary for CT learning to operate is disrupted by the multiplexing effects of the volley principle in the AN.

If our hypothesis about the disruptive effect of AN stochasticity on vowel categorisation learning is true, it would suggest that an extra preprocessing stage was necessary between the AN and the plastic A1 in order to reduce the jitter (noise) found in the temporal and spatial distribution of AN spikes in response to the different exemplars of the same auditory stimulus class. This reduction in jitter would be necessary to enable the plastic auditory cortex to learn representations of auditory categories through CT learning. We hypothesized that this preprocessing could happen in the intermediate subcortical stages of processing in the auditory brain, such as cochlear nucleus (CN) and inferior colliculus (IC), whereby the essential contribution of the precise microarchitecture and connectivity of the CN and IC would be able to help dejitter and stabilize the AN firing patterns, thereby enabling the plastic cortical area A1 to develop informative representations of vowel categories through CT learning. The hypothesized increase in the stability of firing responses across the AN and A1 is in line with the evidence reviewed in DeWeese, Hromadka, and Zador (2005), which suggests that many neurons in the auditory cortex tend to have transient, binary, and highly regular responses at the onset of simple auditory stimuli.

In other words, this letter proposes the minimal subset of subcortical auditory brain areas that allow the primary auditory cortex to learn "good representations" of speechlike auditory objects through STDP learning mechanisms as described by Bi and Poo (1998). A "good representation" is defined as that which is informative of the stimulus class regardless of the variability in the raw input, whether this variability is due to the speaker-dependent stochasticity and hence inherent to the input stimulus or due to the stochasticity introduced through the initial stages of the auditory processing in the cochlea and the AN. A good representation should be less redundant (or more compressed) than the representations within the initial stage of the auditory processing, such as the AN (Barlow, 1961). This can be measured by looking at the amount of mutual information between the stimulus class and the responses of single neurons within the AN and A1, with the expectation that individual A1 cells will be more informative than the individual AN cells.

We tested our hypothesis by comparing the performance of a biologically realistic four-stage hierarchical feedforward spiking neural network model of the auditory brain incorporating both subcortical (AN, CN, IC) and cortical (A1) stages (full AN-CN-IC-A1 model shown in Figure 1A) to the performance of two models that either omitted areas CN and IC (reduced AN-A1 model shown in Figure 1B) or had the same number of processing stages as the full AN-CN-IC-A1 model but lacked the precise CN and IC microarchitecture and connectivity (simple four-stage model shown in Figure 1C). Our simulations demonstrated that both the reduced AN-A1 and simple four-stage models significantly underperformed the full AN-CN-IC-A1 model on the two vowel classification task.

The contributions of this work are three-fold. First, we show how simple, local synaptic learning rules can support unsupervised auditory category learning in a biologically inspired spiking model that mimics the connectivity and microarchitecture of the auditory brainstem (CN and IC) feeding into the primary auditory cortex (A1). Second, we provide computational evidence for the hypothesis that the stochasticity introduced in the auditory nerve is detrimental to auditory category learning in the A1 unless it is reduced by the auditory brain stem processes (CN and IC). Third, we provide a quantitative theoretical framework that explains the diverse physiological response properties of identified cell classes in the ventral cochlear nucleus and generates neurophysiologically testable hypotheses for the essential role of the nonplastic CN and IC as the AN jitter removal stages of the auditory brain.

2 Results

2.1 Quantifying Spike Jitter in the Auditory Nerve. In this letter, we hypothesize that the reproducibility of AN firing patterns for similar stimuli necessary for CT learning to operate is disrupted by the multiplexing effects of the volley principle in the AN. We tested this hypothesis by generating examples of two vowel classes, /i:/ and /a/. Each example of a particular vowel class was generated with different formants, thus simulating the variability inherent in the stimulus. Each example was presented a number of times, hence simulating the stochasticity introduced in the AN. We measure these two types of stochasticity using a quantitative analysis of the similarity or dissimilarity between AN firing patterns in response to the different vowel stimuli (see section 4 for calculation details). We found that the AN spike rasters for repeat presentations of the same exemplar of a vowel or of different exemplars of the same vowel category were as dissimilar to each other as the AN responses to the vowels from different vowel categories (see the AN scores in Table 1). This highlights the high level of stochasticity introduced in the AN, which appears to be of similar magnitude to the intrinsic stimulus variability.

2.2 Reduced AN-A1 Auditory Brain Model. We begin by presenting simulation results from the reduced AN-A1 spiking neural network model of the auditory brain shown in Figure 1B, in which the intermediate CN and IC stages were omitted (see section 4 for model architecture details). The input stage of the AN-A1 model is a biologically realistic AN model by Zilany, Bruce, Nelson, and Carney (2009), and the output stage is a loose and simplified approximation of the A1 in the real brain.

Table 1: Similarity Measure Scores Between the AN and IC Spike Rasters in Response to Different Presentations of the Same Exemplar of a Stimulus (Same Exemplar Index), Different Exemplars of the Same Stimulus Class (Different Exemplars Index), and Different Stimulus Classes (Different Categories Index).

	/i:/		/a/		/i:/ and /a/	
	AN	IC	AN	IC	AN	IC
Same exemplar index	0.45	0.9	0.57	1	–	–
Different exemplars index	0.52	0.91	0.63	1	–	–
Different categories index	–	–	–	–	0.42	0.67

Note: Scores vary between 0 and 1, with higher scores indicating higher levels of similarity and consequently low levels of jitter.

We tested the ability of the AN-A1 model to learn robust representations of auditory categories using a controlled yet challenging task, whereby 12 different exemplars of each of two classes of vowels, /i:/ and /a/, were synthesized and presented to the network (see Figure 2 and section 4). The biologically plausible unsupervised CT learning mechanism implemented through STDP (Bi & Poo, 1998) within the AN→A1 afferent connections was expected to enable the model to learn the two vowel categories (see section 4 for an overview of CT learning). In particular, we investigated whether localist representations of auditory categories emerged, whereby individual neurons would learn to respond selectively to all exemplars of just one preferred stimulus class (DeWeese & Zador, 2003).

The ability of the AN-A1 model to learn robust vowel categories depends on how it is parameterized. A hyperparameter search using a grid heuristic was therefore conducted. Mutual information between the stimuli and the responses of singles cells within the output A1 stage of the model was used to evaluate the performance of the AN-A1 model on the vowel categorization task (see section 4). It was assumed that the performance of the network changed gradually and continuously as a function of its hyperparameters, since learning in the real brain has to be robust to mild variations in biological parameters. It was therefore expected that the best model performance found through the grid parameter search would be a good approximation of the true maximal model performance. The detailed description of the parameter search can be found in the supplemental materials. The following parameters were found to result in the best AN-A1 model performance: LTP constant (α_p) $= 0.05$; LTD constant (α_d) $= -0.02$; STDP time constants (τ_p/τ_d) $= 15/25$ ms; initialization magnitude of AN→ A1 connections (w_{ij}^{BL}) $\in [30, 35]$ nA; and level of inhibition in the A1 (w_{ij}^{IE}) $=$ -6 nA.

The performance of the best AN-A1 model found through the parameter search is shown in Figure 3 (solid dark blue line). The average information

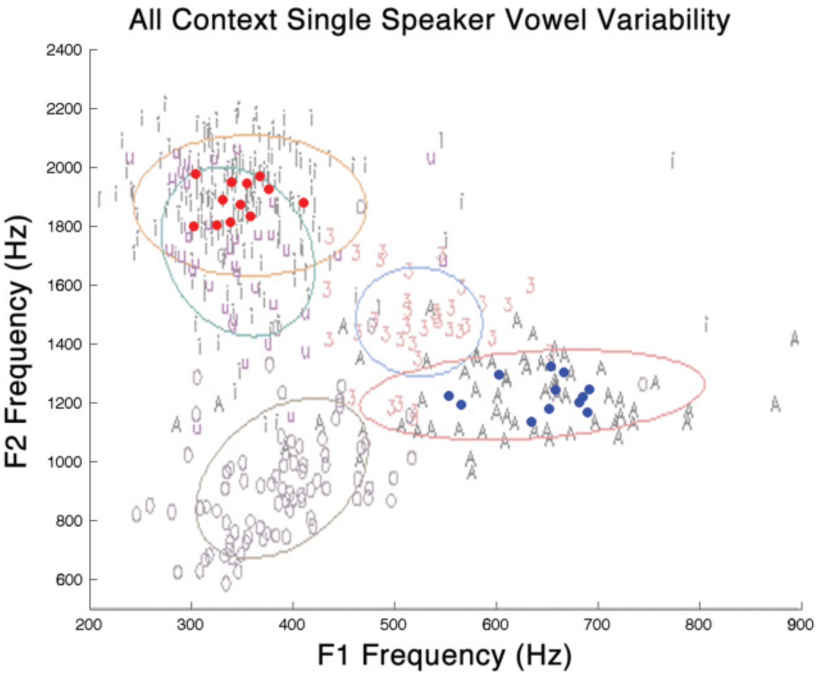

Figure 2: Schematic representation of 12 transforms of two synthesized vowels (/a/ - blue, /i:/ - red) projected onto the two-dimensional plane defined by the first two formants of the vowels. Each transform was generated by randomly sampling three formant frequencies from a uniform 200 Hz distribution centered around the respective average values reported by Peterson and Barney (1952) for male speakers. It can be seen that the generated vowel transforms are in line with the vowel distribution clouds produced from natural speech of a single speaker (Huckvale, 2004). All transforms were checked by human subjects to ensure that they were recognizable as either an /a/ or an /i:/. The ellipses approximate the 70% within-speaker variability boundary for a particular phoneme class.

about the vowel class identity among the top 10 most informative A1 cells was 0.21 bits, and the maximum A1 information was 0.57 bits out of the theoretical maximum of 1 bit. This is not enough to achieve good vowel recognition performance using a small population of cells. In fact, when we trained a nonlinear decoder (a multilayer perceptron with a single hidden layer and a two-class classification objective), we found that it was able to reach only 97% accuracy given the responses of 120 most informative A1 cells. Saying this, a certain amount of useful learning did occur in the reduced AN-A1 model as evidenced by more A1 information after training

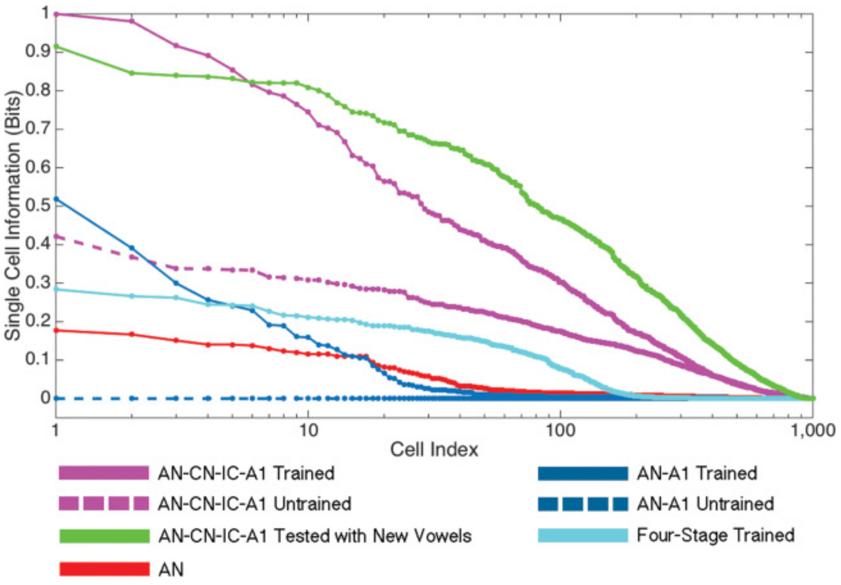

Figure 3: Single cell information carried by cells in a specified model neural area during the vowel classification task. The cells are ordered along the abscissa by their informativeness. Maximum theoretical entropy for the task is 1 bit. It can be seen that the output A1 neurons of the full AN-CN-IC-A1 spiking neural network model of the auditory brain after training carry more information about the two vowel classes than the input auditory nerve (AN) fibers, or the A1 cells of the reduced AN-A1 model, simple four-stage model, or any of the models before training.

than before training and more information in the A1 compared to the AN input (see Figure 3, dotted dark blue and solid red lines, respectively).

2.3 Removing Auditory Nerve Jitter. The reduced AN-A1 model was unable to learn the identities of the two vowel classes through unsupervised CT learning implemented through STDP within the plastic AN→A1 connections. Successful CT learning relies on the discovery of correlations, or overlap, in the neural representations of stimuli that belong to the same object or stimulus class. We attribute the failure of the A1 neurons in the reduced model to discover stimulus classes to the fact that the biologically realistic AN input to the model contains large amounts of physiological noise or space and time jitter in the spike times, which obscure the similarities between the AN spike rasters in response to different stimuli belonging to the same vowel class. Since such similarities are necessary for CT learning to operate, the output A1 stage of the reduced AN-A1 model was unable to learn robust representations of the two vowel classes directly from the AN input.

Reducing time and space jitter in AN response spike rasters should aid unsupervised learning in the auditory brain, and it can be achieved through the following mechanisms: (1) information from a number of AN fibers with similar characteristic frequencies (CFs) is integrated in order to remove space jitter, and (2) AN spike trains for different cells are synchronized, whereby spikes are realigned to occur at set points in time rather than anywhere in continuous time, thus removing time jitter.

We consider space and time jitter removal to be one of the key roles of the subcortical areas CN and IC, whereby jitter reduction is initiated in the CN and completed within the IC, as convergent inputs from different subpopulations of the CN are integrated in such a way that facilitates effective stimulus classification by CT-like learning mechanisms in subsequent stages, such as A1. We envisage the following processes: (1) chopper (CH) cells within the CN remove space jitter, (2) onset (ON) cells within the CN remove time jitter, and (3) the IC produces spike rasters with reduced jitter in both space and time by combining the afferent activity from the cochlear nucleus CH and ON cells.

2.3.1 Space Jitter Removal. CH neurons in the CN are suitable for the space jitter removal task due to their afferent connectivity patterns from the AN. Each CH cell receives a number of afferent connections from AN neurons with similar CFs. The incoming signals are integrated to produce regular spike trains. We use a small number of afferent connections to match the number of strongest afferent inputs provided by Ferragamo, Golding, and Oertel (1998) and Young and Sachs (2008); however, that is not to deny any higher estimates that may exist.

In the full AN-CN-IC-A1 model shown in Figure 1A, a CH subpopulation was simulated by adding 1000 class 1 neurons by Izhikevich (2003) with gaussian topological connectivity from the AN, whereby each CH cell received afferents from a tonotopic region of the AN. A hyperparameter search was conducted to maximize the space jitter removal ability of CH neurons (see the supplemental materials), and the following parameter values were found to be optimal: gaussian variance of the AN→CH afferent connectivity (σ) = 26 cells and magnitude of the AN→CH afferent connections (w_{ij}^{BL}) $\in [30, 35]$ nA. While it is suggested that CH neurons receive inhibitory inputs in the real brain, the hyperparameter search we conducted suggested that our model worked best with no inhibitory inputs to the CH layer (within-CH inhibition (w_{ij}^{IE}) $= 0$ nA). We believe that this is due to the particular implementation of the inhibitory feedback we have chosen; it should not affect the validity of our results, since the discharge

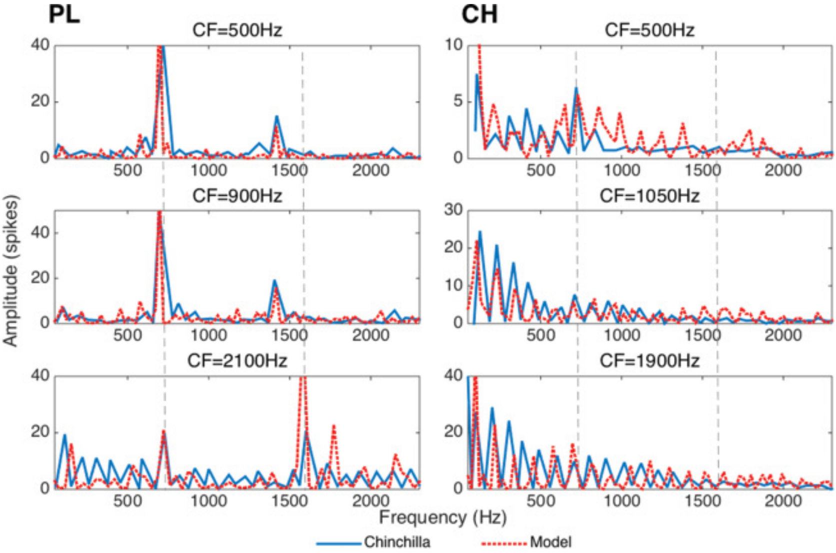

Figure 4: Spectra (computed as fast fourier transforms of period histograms) of primary-like (PL) (left column) and chopper (CH) (right column) cochlear nucleus neuron responses to a synthetic vowel /a/ generated using the Klatt synthesizer (Klatt, 1980). The ordinate represents the level of phase locking to the stimulus at frequencies shown along the abscissa. Dotted lines show the positions of the vowel formant frequencies F_1 and F_2. Data from chinchilla CN fibers reproduced from Recio and Rhode (2000) are shown in blue solid lines. Data collected from the corresponding model CN fibers are shown in red dashed lines. The similarity between the real and model fibers' response properties suggests that the model's performance is comparable to the neurophysiological data.

properties of the optimized CH cells in our model corresponded closely to those reported experimentally for biological CH neurons (see Figure 4, right column).

2.3.2 Time Jitter Removal. Time jitter removal is thought to be facilitated by ON neurons in the CN. ON cells are relatively rare, constituting approximately 10% of the ventral CN (Rhode, Roth, & Recio-Spinoso, 2010). They have been estimated to each receive connections from up to 65 AN fibers across a wide stretch of the cochlea, which results in broadly frequency tuned response properties (Rhode et al., 2010). These cells are characterized by fast membrane time constants, which makes them very leaky, with high spike thresholds. Consequently, ON cells require strong synchronization from many AN fibers with a wide range of CFs in order to produce a discharge (Oertel, Bal, Gardner, Smith, & Joris, 2000). The cross-frequency coincidence detection inherent in the ON cells makes them able to phase-lock to the fundamental frequency (F_0) of vowels, as supported by neurophysiological evidence (Winter, Palmer, Wiegrebe, & Patterson, 2003).

We propose that the interplay between the converging ON and CH cell inputs to the IC can reduce jitter in the neural representations of vocalization sounds. Since ON cells synchronize to the stimulus F_0, they can introduce regularly spaced afferent input to the IC. Such subthreshold afferent input would prime the postsynaptic IC cells to discharge at times corresponding to the cycles of stimulus F_0. If IC cells also receive input from CH cells, then ON afferents will help synchronize CH inputs within the IC by increasing the likelihood of the IC cells firing at the beginning of each F_0 cycle. This is similar to the encoding hypothesis described in Hopfield (1995).

In the full AN-CN-IC-A1 model, a population of ON cells was simulated using 100 class 1 neurons by Izhikevich (2003) sparsely connected to the AN. A hyperparameter search was conducted to maximize the ability of ON neurons to synchronize to the F_0 of the stimuli (see the supplemental materials), and the following parameter values were found to be optimal: AN→ON afferent connection weight magnitudes (w_{ij}^{BL}) = 21 nA, sparseness of AN-ON connectivity = 0.46 (54% of all possible AN-ON connections are non-zero), and within-ON inhibition magnitude (w_{ij}^{IE}) = −75 nA.

2.4 Full AN-CN-IC-A1 Auditory Brain Model. The full AN-CN-IC-A1 model of the auditory brain was constructed as shown in Figure 1A to test whether the addition of the subcortical stages corresponding to the CN and IC would remove space and time jitter contained within the input AN firing rasters as described and thus enable the output plastic cortical stage A1 to learn invariant representations of the two vowel categories, /i:/ and /a/ (see section 4 for details of the model architecture). Similar to the reduced AN-A1 model, the output stage of the full AN-CN-IC-A1 model is a loose and simplified approximation of the A1 in the real brain.

In the brain, subpopulations of the CN do not necessarily synapse on the IC directly. Instead, they pass through a number of nuclei within the superior olivary complex (SOC). The nature of processing done within the SOC in terms of auditory object recognition (rather than sound localization), however, is unclear. The information from the different CN subpopulations does converge in the IC eventually, and for the purposes of our argument, we model this convergence as direct. The same simplified connectivity pattern (direct CN-IC projections) was implemented by Meddis and O'Mard (2006) for their model of the subcortical auditory brain.

Apart from the CH and ON subpopulations described above, the CN of the full AN-CN-IC-A1 model also contained 1000 primary-like (PL) neurons. PL neurons make up approximately 47% of the ventral CN in the brain (Winter & Palmer, 1990), suggesting that they might play a significant role in auditory processing. Although their contribution to the preprocessing of

AN discharge patterns is perhaps less clear than that of the CH and ON subpopulations, PL cells were included in the model architecture to investigate their effect on auditory class learning. PL cells essentially transcribe AN firing in the brain (Winter & Palmer, 1990) and were therefore modeled using strong ($w_{ij}^{BL} = 1000$ nA) one-to-one AN→PL afferent connections and no inhibition ($w_{ij}^{IE} = 0$ nA) within the PL area. The discharge properties of the model PL neurons were found to correspond closely to those reported experimentally (see Figure 4, left column).

A grid search heuristic was applied to the full AN-CN-IC-A1 model to find the hyperparameters that produce the best model performance on the two-vowel category learning task (see the supplemental materials for details). Similar to the reduced AN-A1 model, mutual information was calculated to evaluate the performance of the full AN-CN-IC-A1 model. The following parameter values were found to result in the best model performance: CH→IC (w_{ij}^{BL}) = 400 nA, PL→IC (w_{ij}^{BL}) = 400 nA, and ON→IC (w_{ij}^{BL}) = 3 nA connection magnitudes; the magnitude of the within-IC inhibition (w_{ij}^{IE}) = 0 nA; and the LTD magnitude of the IC→A1 connections (α_d) = -0.015.

It was found that unlike the reduced AN-A1 network, a well-parameterized full AN-CN-IC-A1 model of the auditory brain was able to solve the two-vowel categorization task by developing many A1 neurons with high levels of vowel class identity information approaching the theoretical maximum of 1 bit (see Figure 3, pink). The vowel category information carried in the discharges of the A1 neurons of the full AN-CN-IC-A1 model increased substantially during training (see Figure 3, dotted pink versus continuous pink). Note, however, that this is not to say that the information about the stimulus category was not present in the input AN layer of the model. According to the data processing inequality (Cover & Thomas, 1991), postprocessing of a signal cannot increase information. Hence, all the information was already present in the AN. What can change, however, is the relative amount of information carried by single cells within the different stages of processing. We found that in order for a nonlinear classifier to categorize the two stimulus classes, it required a population of at least 80 AN cells compared to just 1 cell in the trained A1. Our results therefore suggest that the presence of the nonplastic CN microarchitecture converging on the IC indeed helped the plastic A1 learn to produce stimulus class-selective responses. Furthermore, unlike the AN-A1 model, the resulting firing properties of the A1 cells of the AN-CN-IC-A1 model were in line with the firing properties described in DeWeese et al. (2005), as shown in Figure 5.

2.4.1 Generalization of Learning. We have demonstrated that the trained full AN-CN-IC-A1 model was capable of correctly recognizing different exemplars of vowels belonging to either vowel class /i:/ or /a/, despite the high variability even between the input AN spike rasters in response to

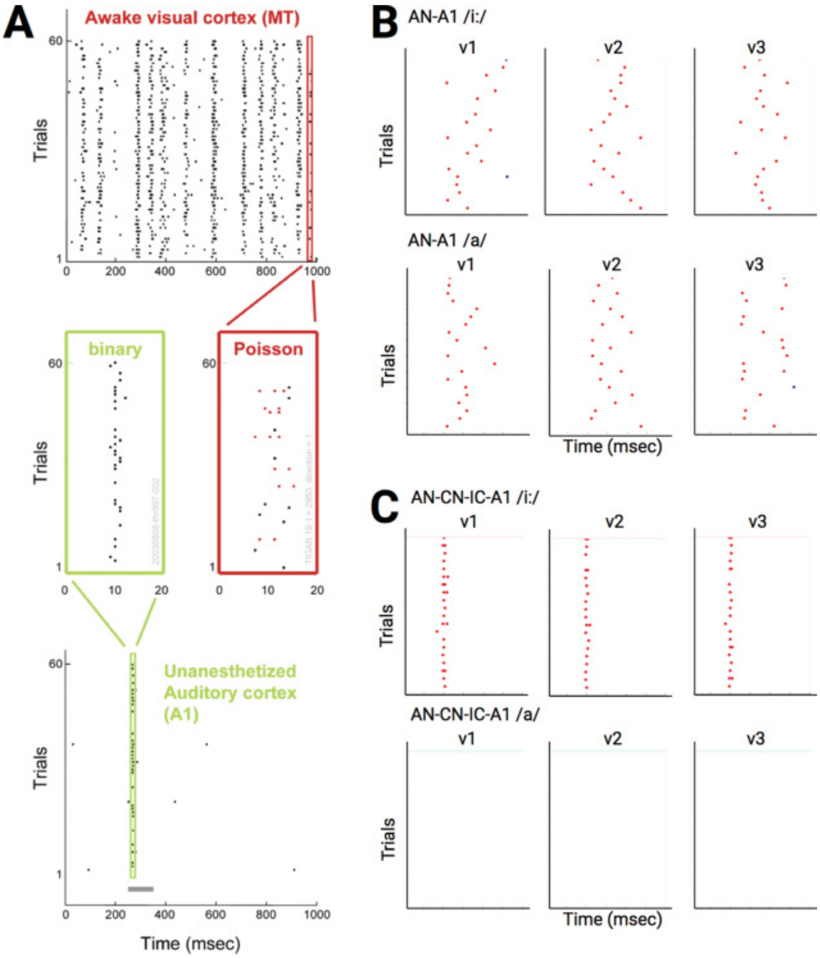

Figure 5: (A) Schematic demonstrating jittery Poisson-like firing of a motion-sensitive neuron in the monkey visual cortex and stable transient binary firing of a sound-sensitive neuron in the rat auditory cortex. (Figure adapted from DeWeese et al., 2005.) (B) Firing rasters of a single cell with the A1 of a trained AN-A1 model to 20 trials of three versions (v1–3) of vowels /i:/ and /a/. It can be seen that the model fires in a manner that resembles the Poisson neuron in panel A and is not differentiating between the two vowel classes well. (C) Firing rasters of a single cell with the A1 of a trained AN-CN-IC-A1 model to 20 trials of three versions (v1-3) of vowel /i:/ and /a/. It can be seen that the cell fires in a way that resembles the A1 responses recorded in an awake rat shown in panel A. It also shows high preference for vowel class /i:/ but not /a/.

the different presentations of the same vowel exemplar. It was possible, however, that the model overfit the data and learned only the particular vowel exemplars presented during training instead of exploiting the statistical regularities within the stimuli to develop generalized representations of the two vowel classes. To test whether this was the case, we synthesized 12 new exemplars for each of the two vowel classes /i:/ and /a/. The formants of the new vowel stimuli were different from those used in the original stimulus set. Each of the new vowels was presented to the network 20 times. It can be seen in Figure 3 (green) that many of the A1 cells of the full AN-CN-IC-A1 network trained on the original and tested on the new vowels reached high (up to 0.92 bits) levels of single cell information about the vowel class identity, approaching the theoretical maximum of 1 bit. This suggests that the network indeed learned general representations of vowel classes /i:/ and /a/ rather than overfitting by learning only the particular vowel exemplars presented during training.

2.4.2 The Importance of CN and IC Microarchitecture and Connectivity. Having shown that unlike the reduced AN-A1 model, the full AN-CN-IC-A1 model was capable of learning robust representations of vowel class identities, we investigated next whether the particular microarchitecture and connectivity of the subcortical stages CN and IC were important for the improved AN-CN-IC-A1 model performance.

An additional simulation was run to confirm that the particular microarchitecture of the CN and its subsequent convergence on the IC, rather than the pure addition of extra processing layers, improved the performance of the four-stage AN-CN-IC-A1 model compared to the two-stage AN-A1 model on the vowel class identity learning task. To this accord, a simple four-stage fully connected model lacking the detailed CN and IC microstructure and connectivity (see Figure 1C and section 4 for details) was constructed and evaluated using the original two-vowel-class learning paradigm. Figure 3 (teal) demonstrates that this simple four-stage network achieved very little information about the identity of the vowel stimuli (no more than 0.28 bits). This suggests that the pure addition of extra processing stages within a spiking neural network model does not help with auditory category learning. Instead, the preprocessing within the particular microarchitecture of the three subpopulations of the CN followed by their convergence on the IC is necessary for such learning to occur.

2.4.3 The Importance of CN Subpopulations. In order to verify that each of the three CN subpopulations—CH, ON, and PL—was important for enabling the full AN-CN-IC-A1 network to learn robust representations of vowel class identities, the performance of the model was evaluated when each of the CN subpopulations was ablated one by one. Every time one of the CN subpopulations was eliminated from the model, the network parameters were reoptimized to find the best possible classification

Table 2: Maximum Single Cell Information within the Output A1 Stage of the Best-Performing Reoptimized Full AN-CN-IC-A1 Model When Different CN Subpopulations of Neurons Were Selectively Knocked Out.

Chopper	Onset	Primary-Like	A1 Information (bits)
Yes	Yes	Yes	1
Yes	No	Yes	0.93
Yes	Yes	No	0.89
Yes	No	No	0.81
No	Yes	Yes	0.36
No	No	Yes	0.18
No	Yes	No	0

Notes: "Yes" and "no" indicate that the relevant subpopulation is either present or absent, respectively. The theoretical maximum for the single cell information measure for two auditory classes is 1 bit. The maximum information is achieved only when all three subpopulations are present.

performance by the new reduced model architecture. Table 2 demonstrates that the removal of any of the three subpopulations of the CN resulted in significantly reduced performance of the AN-CN-IC-A1 model on the vowel class identity recognition task, thus suggesting the importance of all three CN subpopulations in enabling auditory class learning. Note that one explanation for the increased informativeness within the A1 of the AN-CN-IC-A1 model compared to the AN-A1 model is the tuning to the amplitude modulation rate within the CN and IC rather than the reduction in AN jitter due to CH and ON processing. CH cells in the CN, and consequently the IC cells they connect to, do show amplitude modulation tuning due to the local connectivity of the CH cells. This, however, is not enough to account for the full increase in the informativeness within the output layer of the AN-CN-IC-A1 model. This is demonstrated by the lower single cell information achieved by the ablated versions of the AN-CN-IC-A1 model shown in Table 2, which included the CH cells but not the ON cells. ON cells do not carry any information about the amplitude modulation rate of the input due to their broad connectivity patterns. They do, however, play a large role in synchronizing the activity of the IC cells and hence removing jitter.

2.4.4 Quantifying Jitter Removal in CN and IC. So far we have demonstrated that the precise microarchitecture and connectivity of the CN and IC are important for enabling the full AN-CN-IC-A1 model to learn robust representations of vowel class identities. Here we test whether the subcortical stages CN and IC indeed remove AN jitter as originally hypothesized. To confirm this, we compared the firing pattern similarity scores between the AN and IC (see section 4 for details). The scores varied between 0 and

1, with high scores indicating high levels of similarity between the corresponding spike rasters and, consequently, low levels of jitter. The high same exemplar and different exemplars IC scores in Table 1 suggest that the IC firing rasters in response to the different presentations of the same vowel exemplar or in response to the different exemplars of the same vowel category are highly similar and hence are mostly jitter free. This is in contrast to the corresponding AN scores, which are all significantly lower due to the space and time jitter. (Also see Figure 11 in the supplementary materials for a visualization of jitter reduction in the spike rasters of the AN and the IC in response to the different presentations of the different exemplars of the two stimulus classes.)

3 Discussion

This work has hypothesised that spike time jitter inherent in the auditory nerve (AN) firing may prevent auditory category learning in the plastic cortical areas of the auditory brain as evidenced by the poor performance of the reduced AN-A1 or the simple four-stage spiking neural network models of the auditory brain on a controlled and very simple vowel categorization task. While past research has suggested that input spike jitter can be reduced by the intrinsic properties of spiking neural networks (Diesmann, Gewaltig, & Aertsen, 1999) and STDP learning (Bohte & Mozer, 2004), such jitter reduction works on the scale of a few milliseconds rather than tens of milliseconds characteristic of the AN jitter. A jitter removal preprocessing stage may therefore be important in order to enable the plastic auditory cortex to learn auditory categories. Here we have shown that the ventral cochlear nucleus (CN), followed by the inferior colliculus (IC), may be able to do just that. In particular, we have demonstrated that chopper (CH) and onset (ON) subpopulations of the CN and their subsequent convergence on the IC have the right connectivity and response properties to remove space and time jitter in the AN input respectively.

Our simulation results also demonstrated the importance of primary-like (PL) neurons in the CN for enabling the auditory cortex to learn auditory categories. The PL subpopulation simply transcribes AN firing and therefore is unlikely to play a role in AN jitter removal. We are therefore still unsure what its role in auditory category learning might be. It is possible that PL input is necessary to simply introduce a base level of activation within the IC. Our simulations nevertheless have demonstrated that the removal of any of the three subpopulations of the CN (CH, ON, or PL) resulted in a significant drop in maximum single cell information within the A1 stage of the trained AN-CN-IC-A1 model (see Table 2). This suggests that the AN-CN-IC-A1 model may have the minimal sufficient architecture for learning auditory categories.

In this letter, we hypothesized that the full AN-CN-IC-A1 model would use the continuous transformation (CT) learning mechanism to develop stimulus class-selective response properties in the A1. For CT learning to be able to drive the development of output neurons that respond selectively to particular vowel classes, the spike rasters in the preceding neuronal stage, in response to the different presentations of the same exemplar of a vowel or of different exemplars belonging to the same vowel class, must be similar to each other. The presence of spike jitter at any stage of processing will destroy these similarity relations needed for CT learning to operate. The firing pattern similarity scores shown in Table 1 demonstrated that the spike raster similarity or dissimilarity relations required for CT learning to operate were restored in the IC compared to the AN of the full AN-CN-IC-A1 model through the dejittering preprocessing within the CN and IC. We hypothesize that this, in turn, enabled the plastic A1 of the full AN-CN-IC-A1 model to learn vowel categorization through the CT learning mechanism. The structurally identical A1 layer of the reduced AN-A1 or the simple four-stage models failed to learn from the unprocessed input AN firing patterns due to the space and time jitter breaking the stable AN firing patterns that are necessary for CT learning by STDP to operate. While our simulations were able to provide certain evidence to support our hypothesis, more work needs to be done to establish the full causal relationship between the levels of jitter in the AN, CN, and IC and the ability of the auditory cortex to learn auditory categories.

We hypothesized that space jitter in the AN was removed by CH neurons in the CN because anatomical studies suggested that CH neurons had the appropriate connectivity from the AN for the task. Similar connectivity, however, is shared by the primary-like with notch (PLn) subpopulation of the CN, suggesting that they may also take part in AN space jitter removal. Neurophysiological evidence, however, suggests that the two cell types have different intrinsic properties (Joris & Smith, 2008; Winter & Palmer, 1990), and the response properties of the CH stage of the AN-CN-IC-A1 model optimized for space jitter removal were found to be more similar to those of the real CH rather than PLn cells (i.e., they do not phase-lock to the stimulus). This suggests that CH cells are more likely to be important for auditory category learning in the brain than PLn neurons.

The simplicity of the synthesized vowel stimuli and the small number of exemplars in each stimulus class are not representative of the rich auditory world that the brain is exposed to during its lifetime. The model therefore needs to be tested on higher numbers of stimuli, as well as on more complex and more realistic stimuli, such as naturally spoken whole words, in future simulation work (see Higgins, Stringer, & Schnupp, 2017, for a first attempt at extending this model to more realistic stimuli). The two-vowel classification problem nevertheless was suitable for the purposes of demonstrating the importance of subcortical preprocessing in the CN and IC for preparing the jittered AN input for auditory category learning in the cortex. The appropriateness of the task is demonstrated by the inability of the reduced AN-A1 and the simple four-stage models of the auditory brain to solve it.

We took inspiration from the known neurophysiology of the auditory brain in order to construct the spiking neural network models described in this section. As with any other model, however, a number of simplifying assumptions had to be made with regard to certain aspects that we believed were not crucial for testing our hypothesis. These simplifications included the lack of superior olivary complex or thalamus in our full AN-CN-IC-A1 model, the nature of implementation of within-layer inhibition in both the AN-A1 and AN-CN-IC-A1 models, and the lack of top-down or recurrent connectivity in either model. While we believe that all of these aspects do affect the learning of auditory object categories to some extent, we also believe that their role is not crucial for the task. Therefore, we leave the investigation of these effects for future work.

The full AN-CN-IC-A1 model described in this letter possesses a unique combination of components necessary to simulate the emergent neurodynamics of auditory categorization learning in the brain, such as biologically accurate spiking dynamics of individual neurons, STDP learning, neurophysiologically guided architecture, and exposure to realistic speech input. Due to its biological plausibility, the model can be used to make neurophysiologically testable predictions and thus lead to further insights into the nature of the neural processing of auditory stimuli. For example, one of the proposed future neurophysiological studies would compare the levels of jitter in the real AN and IC in response to the same auditory stimuli, with the expectation being that the level of jitter will be significantly reduced in the IC.

4 Materials and Methods

4.1 Stimuli. A stimulus set consisting of 12 exemplars of each of two vowels, /i:/ and /a/, was generated using the Klatt synthesizer (Klatt, 1980). Each 100 ms long sound was created by sampling each of the three vowel formants from a uniform 200 Hz distribution centered around the corresponding formant frequency as reported by Peterson and Barney (1952) for male speakers. The variability in formant frequencies among the 12 stimulus exemplars was consistent with the range of variation present in natural human speech as demonstrated in Figure 2. Furthermore, informal tests showed that greater variation in vowel formant frequencies resulted in vowel exemplars that sounded perceptually different from /i:/ or /a/. A fundamental frequency (F_0) of 100 Hz was used for all stimuli.

The vowel stimuli belonging to the two classes, /i:/ and /a/, were presented in an interleaved fashion and separated by 100 ms of silence. The silence encouraged the models to learn separate representations of each individual vowel class and to avoid learning any transitions between vowel classes. We used 200 training and 20 testing epochs, whereby each epoch consisted of the first exemplar of vowel /i:/, followed by the first exemplar of vowel /a/, followed by the second exemplar of vowel /i:/, and so on up to the last twelfth exemplar of vowel /a/. Twenty (rather than one) test epochs were used because, due to the stochasticity of AN responses, input AN spike patterns in response to repeated presentations of the same sound were not identical. Informal tests demonstrated that on average, the order in which the vowel exemplars were presented did not make a qualitative difference to the performance of the trained models. It did, however, introduce higher trial-to-trial variability. Hence, we fixed the presentation schedule for the simulations described in this letter for a fairer model comparison.

4.2 Continuous Transformation Learning. The CT learning mechanism was originally developed to account for geometric transform invariance learning in a rate-coded neural network model of visual object recognition in the ventral visual stream (Stringer et al., 2006), but has recently been shown to also work in a spiking neural network model of the ventral visual stream (Evans & Stringer, 2012). A more detailed description of CT learning for vision can be found in Tromans, Higgins, and Stringer (2012).

In vision, simple changes in the geometry of a scene, such as a shift in location or rotation, can generate a multitude of visual stimuli that are all different views, or transforms, of the same object. CT learning was at its origin an attempt to understand how the brain can form representations of visual objects that are not confused by such transformations, this is, they are transform invariant. At first glance, it may seem that there is no obvious analogue of such transformations in the auditory world. For many classes of natural auditory stimuli, however, their location in frequency space depends on the physical characteristics of the sound source. For example, the changes in physical dimensions of the resonators of the vocal tract would create transformations of vocalization sounds. Such changes would happen due to variations in the placement of the tongue or the jaw when the same or different speakers produce the same speech sound. Thus, many natural auditory objects are prone to shifts in frequency space that are not too unlike the shifts in retinotopic space observed when visual objects undergo geometric transformations. We therefore propose that CT learning may play a crucial role in auditory category learning.

The original CT learning mechanism relies on the presence of a significant overlap between input representations of temporally static stimulus transforms; in other words, neural representations of snapshots of the same object taken from somewhat different points of view often exhibit areas of high correlation that can be discovered and exploited by an associative learning mechanism (Evans & Stringer, 2012; Stringer et al., 2006). Unlike snapshots of visual objects, auditory stimuli have an essential temporal structure. In order for CT learning to associate similar temporal presynaptic patterns of firing onto the same output neuron by STDP, it is important that the volley of spikes from the presynaptic neurons arrive at the postsynaptic neuron almost simultaneously (Evans & Stringer, 2012). If this is not the case, connections corresponding to the presynaptic spikes that arrive after the postsynaptic neuron fires will be weakened due to the nature of STDP, whereby there is strengthening of connections through long-term potentiation (LTP) if the presynaptic spike arrives before the postsynaptic spike and weakening of connections through long-term depression (LTD) otherwise, thus preventing effective CT learning of the input patterns.

In order to allow CT learning to work for the temporal auditory stimuli, a distribution of heterogeneous axonal conduction delays needs to be added to the plastic afferent connections. These axonal delays would transform temporal input sequences into patterns of spikes arriving simultaneously at individual postsynaptic cells. The patterns of coincident spikes received by each postsynaptic cell would depend on the cell's transformation matrix of axonal delays. If an appropriate delay transformation matrix is applied to the input spike pattern, a subset of postsynaptic neurons will receive synchronized spikes from the subset of input neurons encoding similar exemplars of a particular stimulus class, such as a vowel, thus enabling CT learning. Neurophysiological data collected from different species suggest that cortical axonal connections, including those within the auditory brain, may have conduction delays associated with them on the order of milliseconds to tens of milliseconds (Salami, Itami, Tsumoto, & Kimura, 2003; Miller, 1996).

It is therefore suggested that the CT mechanism can enable a spiking neural network to learn class identities of temporal auditory stimuli if, over the whole space of different stimulus exemplars belonging to one class, stimuli that are similar to each other physically also evoke similar spatiotemporal firing patterns (i.e., have sufficient overlap). Spatial and temporal jitter, for example, in the input auditory nerve (AN), add noise to the spatiotemporal firing patterns and therefore make responses to similar stimuli more dissimilar, hence preventing effective CT learning without additional preprocessing to reduce such jitter.

4.3 Information Analysis. One common way to quantify learning success is to estimate the mutual information between stimulus category and neural response $I(S; R)$. It is calculated as $I(S; R) = \sum_{s\in S, r\in R} p(s, r) \log_2 \frac{p(s,r)}{p(s)p(r)}$, where S is the set of all stimuli and R is the set of all possible responses, $p(S, R)$ is the joint probability distribution of stimuli and responses, and $p(s) = \sum_{r\in R} p(s, r)$ and $p(r) = \sum_{s\in S} p(s, r)$ are the marginal distributions (Nelken & Chechik, 2007). The upper limit of $I(S; R)$ is given as $H(s) = \sum_{s} p(s) \log_2 \frac{1}{p(s)}$, which, given that we had two equiprobable stimulus classes, here equals 1 bit.

Stimulus-response confusion matrices were constructed using a simple binary encoding scheme (DeWeese & Zador, 2003) and used to calculate

$I(S; R)$. Binary encoding implies that a cell could either be on (if it fired at least once during stimulus presentation) or off (if it never fired during stimulus presentation).

We used observed frequencies as estimators for underlying probabilities $p(s)$, $p(r)$, and $p(s, r)$, which introduced a positive bias $Bias \approx \frac{\#bins}{2N \log_2 2}$, where $\#bins$ is the number of potentially nonzero cells in the joint probability table and N is the number of recording trials (Nelken & Chechik, 2007). Given the large value of $N = 960$ in our tests of model performance, the bias was negligible ($Bias = 0.004$ bits) and was ignored.

4.4 Quantifying Spike Raster Similarity. We hypothesize that a spiking neural network can learn auditory categories through the CT learning mechanism. CT learning relies on a high degree of similarity or overlap between spike rasters in response to different exemplars of one particular stimulus class, such as /i:/ or /a/. Here we describe three indices that quantify the degree of similarity or dissimilarity between spike rasters in response to different presentations of the same exemplar of the same stimulus class (same exemplar index), different exemplars of the same stimulus class (different exemplars index), or different stimulus classes (different category index). Each index varies between 0 and 1, with higher scores indicating a higher degree of similarity between the corresponding firing rasters. Lower scores suggest that the firing rasters being compared are dissimilar due to either the inherent differences between the input stimuli or the presence of spike time jitter that diminishes the otherwise high similarity between the firing rasters being compared.

4.4.1 Same Exemplar Index. The same exemplar (SE) index quantifies the degree of similarity between the firing rasters within a particular area (such as AN or IC) in response to different presentations of the same exemplar of a stimulus. Broadly, it calculates the average number of identical spikes across the different presentations of each exemplar $e_{k(s)} \in \{e_{1(s)}, \ldots, e_{12(s)}\}$ of a stimulus $s \in \{s_1, s_2\}$ in proportion to the total number of stimulus exemplar presentations ($n \in [1, N]$, where $N = 20$ testing epochs). For each presentation of each stimulus, we therefore constructed a $T \times J$ matrix $M^n_{e_{k(s)}}$ (where $T = 100$ ms is the number of 1 ms time bins spanned by the auditory input, and $J \in \{100, 1000\}$ neurons is the size of the chosen neural area of the model). Each element m_{tj} of matrix $M^n_{e_{k(s)}}$ contained the number of spikes produced by the particular neuron $j \in [1, J]$ within the time bin $t \in [1, T]$ in response to the stimulus exemplar $e_{k(s)}$. We chose 1 ms time bins to have a conservative estimate of raster similarity that matches the lowest amount of jitter reported in the auditory cortex (DeWeese & Zador, 2003; DeWeese et al., 2005; Heil, 2004; Chimoto, Kitama, Qin, Sakayori, & Sato, 2002; Barbour & Wang, 2003).

If the firing rasters of the chosen area of the model in response to the different presentations $n \in [1, N]$ of the same stimulus exemplar $e_{k(s)}$ are similar to each other, then the same slots of the firing pattern matrices $M^n_{e_{k(s)}}$ should be nonzero for different $n \in [1, N]$. Consequently, the following becomes more likely when the proportion of stimulus presentation epochs n for which elements of $M^n_{e_{k(s)}}$ are nonzero across the different presentations of the same stimulus exemplar becomes large: (1) the firing responses within the model area are more likely to be similar; (2) it is likely that less jitter is present in the chosen area of the model; and (3) CT learning is more likely to enable postsynaptic cells to learn that the similar, stable, jitterless responses within the model area belong to the same stimulus class.

We therefore computed the matrix $M_{e_{k(s)}} = \langle M^n_{e_{k(s)}} \rangle$, where $\langle \cdot \rangle$ signifies the mean over all the presentation epochs $n \in [1, N]$, and then identified the mean $\mu_{e_{k(s)}}$ of the 100 largest elements of $M_{e_{k(s)}}$. These were used to compute the final SE_s score for each stimulus $s \in \{s_1, s_2\}$ as $SE_s = \langle \mu_{e_{k(s)}} \rangle$, where $\langle \cdot \rangle$ signifies the mean over all exemplars $e_{k(s)}$ of stimulus s. A higher SE_s index points to more similarity between the chosen firing rasters in response to the different presentations of the same exemplar of stimulus s. Consequently, this also signifies lower levels of jitter present within the layer, since high levels of jitter would disrupt the similarity in firing patterns and result in a lower SE_s index.

4.4.2 Different Exemplars Index. The different exemplars (DE) index quantifies the similarity of the firing rasters within a chosen neural area of the model in response to the different exemplars of the same stimulus class. It is somewhat similar to the SE_s index described above; however, instead of comparing the firing matrices across the different presentations n of the same stimulus exemplar $e_{k(s)}$, the firing matrices are compared across the different exemplars $e_{k(s)}$ of each stimulus class $s \in \{s_1, s_2\}$. Consequently, firing raster matrices $M^n_{e_{k(s)}}$ were calculated once again, but this time, the average was taken over all the different exemplars $e_{k(s)} \in \{e_{1(s)}, \ldots, e_{12(s)}\}$ of stimulus $s \in \{s_1, s_2\}$. That is, we computed $M^n_s = \langle M^n_{e_{k(s)}} \rangle$, where $\langle \cdot \rangle$ signifies the mean over all the stimulus exemplars. We then identified the mean μ^n_s of the 100 largest elements of M^n_s and used them to compute the final DE_s score for each stimulus $s \in \{s_1, s_2\}$ as $DE_s = \langle \mu^n_s \rangle$, where $\langle \cdot \rangle$ signifies the mean over all $n \in [1, N]$ presentation epochs of each exemplar $e_{k(s)}$ of stimulus s. A higher DE_s index points to more similarity between the firing rasters within the chosen model neural area in response to the different exemplars $e_{k(s)}$ of stimulus s. Consequently, this also signifies lower levels of jitter present within the layer, since high levels of jitter would disrupt the similarity in firing patterns and result in a lower DE_s index.

4.4.3 Different Category Index. The Different Category (DC) index quantifies the similarity of the different firing rasters within a chosen neural area of the model in response to different stimulus classes. This score is somewhat similar to the SE_s and DE_s scores described above; however, here the rasters are compared across the different stimulus categories $s \in \{s_1, s_2\}$. To this accord, firing raster matrices $M^n_{e_{k(s)}}$ were calculated once again, but this time the average $M^n = \langle M^n_{e_{k(s)}} \rangle$ was taken over all the different exemplars $e_{k(s)} \in \{e_{1(s)}, \ldots, e_{12(s)}\}$ and over all the stimuli $s \in \{s_1, s_2\}$. We then identified the mean μ^n of the 100 largest elements of each matrix M^n and used them to compute the final DC score as $DC = \langle \mu^n \rangle$, where $\langle \cdot \rangle$ signifies the mean over all $n \in [1, N]$ presentation epochs. A lower DC index points to more differences between the chosen firing rasters in response to the different stimulus categories s.

4.5 Spiking Neural Network Models

4.5.1 Neuron Model. Apart from the AN, all other cells used in this letter were modeled according to the spiking neuron model by Izhikevich (2003). We chose this model because it combines much of the biological realism of the Hodgkin-Huxley model with the computational efficiency of integrate-and-fire neurons. We implemented our models using the Brian simulator with a 0.1 ms simulation time step (Goodman & Brette, 2008). A range of conduction delays between layers is a key feature of our models. In real brains, these delays might be axonal, dendritic, synaptic, or due to indirect connections, but in the model, for simplicity, all delays were implemented as axonal. The [0, 50] ms range was chosen to approximately match the range reported by Izhikevich (2006).

Excitatory cells. Neurophysiological evidence suggests that many neurons in the subcortical auditory brain have high spiking thresholds and short temporal integration windows, thus acting more like coincidence detectors than rate integrators (Sadagopan & Wang, 2009; Abeles, 1982). This is similar to the behavior of Izhikevich's class 1 neurons (Izhikevich, 2003). All subcortical (CN, IC) excitatory cells were therefore implemented as class 1. To take into account the tendency of neurons in the auditory cortex to show strong adaptation under continuous stimulation (Ulanovsky, Las, Farkas, & Nelken, 2004), we chose Izhikevich's spike frequency adaptation neurons to model the excitatory cells in the auditory cortex (A1).

Inhibitory cells. Since inhibitory interneurons are known to be common in most areas of the auditory brain (Frisina, 2001; Ulanovsky et al., 2004) except the AN, each stage of the models apart from the AN contained both excitatory and inhibitory neurons. Inhibitory cells were implemented as Izhikevich's phasic bursting neurons (Izhikevich, 2003). Sparse connectivity between excitatory to inhibitory cells within a model area was modeled using strong one-to-one connections from each excitatory cell to an inhibitory partner. Each inhibitory cell in turn was fully connected to all excitatory cells. Such inhibition implemented dynamic and tightly balanced inhibition as described in Deneve and Machens (2016), which resulted in competition between excitatory neurons and provided negative feedback to regulate the total level of firing within an area. Informal tests demonstrated that the exact implementation of within-layer inhibition did not have a significant impact on the results presented in this letter, as long as the implementation still achieved an appropriate level of within-layer competition and activity modulation.

4.5.2 Spike Time Dependent Plasticity (STDP) Learning. We used an implementation of STDP based on the work by Bi and Poo (1998). The following equations describe the implementation of STDP-based learning within the proposed neural network model of the auditory brain. The weight update is scaled by

$$f(s_{ij}) = \begin{cases} \alpha_p e^{-s_{ij}/\tau_p}, & if\ s_{ij} > 0 \quad LTP \\ \alpha_d e^{s_{ij}/\tau_d}, & if\ s_{ij} < 0 \quad LTD \end{cases}, \tag{4.1}$$

where τ_p and τ_d are STDP time constants, α_p and α_d are constant coefficients, and s_{ij} is the time difference between a post- and a presynaptic spike calculated according to

$$s_{ij} = t_i - (t_j + \Delta_{ij}),$$

where t_i is the time of the postsynaptic spike, t_j is the time of the presynaptic spike, and Δ_{ij} is the magnitude of the axonal conduction delay between the pre- and postsynaptic cells. All delays were treated as axonal, whereby each presynaptic spike time was set to be the time of spike arrival to the postsynaptic cell rather than the time of presynaptic spike discharge. Neurophysiological evidence suggests that STDP time constants are asymmetric and are equal to 17 ± 9 ms for LTP and 34 ± 13 ms for LTD (Bi & Poo, 1998). Therefore, the default values of τ_p and τ_d were set to 15 ms and 25 ms correspondingly, as suggested by Perrinet, Delorme, Samuelides, and Thorpe (2001).

The equations above calculate a scaling variable $f(s_{ij})$ that can be used to update the synaptic weights according to one of three paradigms: additive, multiplicative, or mixed. Neurophysiological data suggest that the LTD magnitude is independent of the instantaneous synaptic strength (w_{ij}), while the magnitude of LTP changes inversely proportionally to connection strength, with stronger synapses resulting in less LTP than weaker synapses (Debanne, Gahwiler, & Thompson, 1996, 1999; Bi & Poo, 1998). These findings are modeled using a mixed STDP paradigm as shown in the equations below, whereby additive learning is used for LTD and multiplicative learning is used for LTP. The minimum bound of 0 nA was set to ensure that the additive LTD did not run away to $-\infty$. The differential equations that follow lead to the bounding of weights in the interval $[0, w_{ij}^{\mathrm{max}}]$:

$$w_{ij}(t+1) = \begin{cases} w_{ij}(t) + (w_{ij}^{max} - w_{ij}(t))f(s_{ij}), & if\ s_{ij} > 0 \quad LTP \\ w_{ij}(t) + w_{ij}^{max} f(s_{ij}), & if\ s_{ij} < 0 \quad LTD \end{cases} . \tag{4.2}$$

For the purposes of the simulations described in this letter, whenever the same presynaptic cell fired more than once within the STDP time window, only the first spike was used in STDP calculations (van Rossum, Bi, & Turrigiano, 2000). It has been demonstrated that this "nearest-only" paradigm results in the same equilibrium state when used with mixed learning as the alternative "all-pairings" paradigm but if less computationally expensive (van Rossum et al., 2000).

4.5.3 Reduced AN-A1 Model Architecture. The reduced AN-A1 spiking neural network model of the auditory brain consisted of two fully connected stages of spiking neurons, the AN (input) and the A1 (output) (see Figure 1B). The AN consisted of 1000 medium spontaneous rate neurons modeled by Zilany et al. (2009) with CFs between 300 and 3500 Hz spaced logarithmically and with a 60 dB threshold. The firing characteristics of the model AN cells were tested and found to replicate reasonably accurately the responses of real AN neurons recorded in neurophysiology studies.

The AN and A1 stages were fully connected using feedforward connections modifiable through spike time dependent plasticity (STDP) learning. The connections were initialized with a uniform distribution of axonal delays (Δ_{ij}) between 0 and 50 ms. The randomly chosen axonal delay matrix was fixed for all simulations described in this letter to remove the confounding effect of different delay initialization values on learning. Informal testing demonstrated that the choice of the axonal delay matrix did not qualitatively affect the simulation results. The initial afferent connection strengths (w_{ij}^{BL}) were randomly initialized using values drawn from a uniform distribution. A grid search heuristic was used to find the optimal model hyperparameters (see Table S1 in the supplemental materials for full model parameters).

4.5.4 Full AN-CN-IC-A1 Model Architecture. The full AN-CN-IC-A1 spiking neural network model of the auditory brain consisted of four stages of spiking neurons as shown in Figure 1A. In contrast to the reduced AN-A1 network, the full AN-CN-IC-A1 model included two intermediate stages between the input AN and output A1 stages to remove time and space jitter present in the AN. These intermediate stages were the CN with CH, ON, and PL subpopulations and the convergent IC stage. The architecture of the three subpopulations of the CN and their corresponding connectivity from the AN is discussed in section 3. The CN→IC connectivity was the following: CH had gaussian topological connectivity, whereby each cell in the IC received afferents from a small tonotopic region of the CH subpopulation ($\sigma = 2$ cells); PL→IC connections were set as one-to-one; and ON→IC connections were set up using full connectivity. The AN and A1 stages of the full AN-CN-IC-A1 model were equivalent to those in the AN-A1 model. The IC→A1 connections in the full AN-CN-IC-A1 model were set up equivalent to the AN→A1 connections of the reduced AN-A1 model. Full model parameters can be found in Table S2 in the supplemental materials.

4.5.5 Simple Four-Stage Model Architecture. The simple fully connected feedforward four-stage model was initialized with randomly distributed synaptic weights (w_{ij}^{BL}) and axonal delays (Δ_{ij}) between each of the stages and with STDP learning for the stage 3→A1 connections (see Figure 1C). The magnitudes of the feedforward connections (w_{ij}^{BL}) were chosen to ensure that the rate of firing in stage 3 was similar to that of the equivalent IC stage of the AN-CN-IC-A1 model (approximately 9 Hz). The STDP parameters of the stage 3→A1 connections were set to the mean of the corresponding optimal values found through the respective parameter searches for models AN-A1 and AN-CN-IC-A1. Full model parameters can be found in Table S3 in the supplemental materials.

Acknowledgments

This research was funded by a BBSRC CASE studentship award (No. BB/H016287/1), and the Oxford Foundation for Theoretical Neuroscience and Artificial Intelligence.

References

Abeles, M. (1982). *Local cortical circuits: An electrophysiological study*. Berlin: Springer.

Barbour, D., & Wang, X. (2003). Auditory cortical responses elicited in awake primates by random spectrum stimuli. *J. Neurosci.*, *23*, 7194–7206.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*, *1*, 217–234.

Bi, G.-Q., & Poo, M.-M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.*, *18*, 10404–10472.

Bizley, J. K., & Cohen, Y. E. (2014). The what, where and how of auditory-object perception. *Nat. Rev. Neurosci.*, *14*(10), 693–707.

Bohte, S. M., & Mozer, M. C. (2004). Reducing spike train variability: A computational theory of spike-timing dependent plasticity. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *14*. Red Hook, NY: Curran.

Chimoto, S., Kitama, T., Qin, L., Sakayori, S., & Sato, Y. (2002). Tonal response patterns of primary auditory cortex neurons in alert cats. *Brain Research*, *934*, 34–42.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Debanne, D., Gahwiler, B. H., & Thompson, S. M. (1996). Cooperative interactions in the induction of long-term potentiation and depression of synaptic excitation between hippocampal CA3-CA1 cell pairs in vitro. *Proc. Natl. Acad. Sci. USA*, *93*, 11225–11230.

Debanne, D., Gahwiler, B. H., & Thompson, S. M. (1999). Heterogeneity of synaptic plasticity at unitary CA1-CA3 and CA3-CA3 connections in rat hippocampal slice cultures. *J. Neurosci.*, *19*, 10664–10671.

Deneve, S., & Machens, C. K. (2016). Efficient codes and balanced networks. *Nature Neuroscience*, *19*, 375–382.

DeWeese, M. R., Hromadka, T., & Zador, A. M. (2005). Reliability and representational review bandwidth in the auditory cortex. *Neuron*, *48*, 479–488.

DeWeese, M. R., & Zador, A. M. (2003). Binary coding in auditory cortex. In S. Thrün, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems*, *16*. Cambridge, MA: MIT Press.

Diesmann, M., Gewaltig, M. O., & Aertsen, A. (1999). Stable propagation of synchronous spiking in cortical neural networks. *Nature*, *402*, 529–533.

Eggermont, J. (2001). Between sound and perception: Reviewing the search for a neural code. *Hearing Res.*, *157*, 1–42.

Evans, B., & Stringer, S. (2012). Transformation-invariant visual representations in self-organizing spiking neural networks. *Front. Comput. Neurosci.*, *6*(46), 1–19.

Ferragamo, M., Golding, N., & Oertel, D. (1998). Synaptic input to stellate cells in the ventral cochlear nucleus. *J. Neurophysiol.*, *79*, 51–63.

Frisina, R. (2001). Subcortical neural coding mechanisms for auditory temporal processing. *Hearing Res.*, *158*, 1–27.

Goodman, D. F., & Brette, R. (2008). Brian: A simulator for spiking neural networks in Python. *Front. Neuroinform.*, *2*(5), 2–5.

Heil, P. (2004). First-spike latency of auditory neurons revisited. *Curr. Opin. Neurobiol.*, *14*, 461–467.

Higgins, I., Stringer, S., & Schnupp, J. (2017). Unsupervised learning of temporal features for word categorization in a spiking neural network model of the auditory brain. *PLoS ONE*, *12*.

Hopfield, J. (1995). Pattern recognition computation using action potential timing for stimulus representation. *Nature*, *376*, 33–36.

Huckvale, M. (2004). *How to: Phonetic analysis using formant measurements*. https://www.phon.ucl.ac.uk/resource/sfs/howto/formant.php

Izhikevich, E. (2003). Simple model of spiking neurons. *IEEE Trans. Neural. Netw.*, *14*, 1569–1572.

Izhikevich, E. (2006). Polychronization: Computation with spikes. *Neural. Comput.*, *18*, 245–282.

Joris, P., & Smith, P. (2008). The volley theory and the spherical cell puzzle. *Neuroscience*, *154*, 65–76.

Klatt, D. H. (1980). Perception and production of fluent speech. In R. A. Cole (Ed.), *Speech perception: A model of acoustic-phonetic analysis and lexical access*. Hillsdale, NJ: Erlbaum.

Liao, Q., Leibo, J. Z., & Poggio, T. (2013). Learning invariant representations and applications to face verification. In J. C. Burges, L. Bottou, M. Welling, K. Ghahramani, & K. Q. Wenberger (Eds.), *Advances in neural information processing systems, 26*. Red Hook, NY: Curran.

Meddis, R., & O'Mard, L. P. (2006). Virtual pitch in a computational physiological model. *J. Acoust. Soc. Am.*, *120*(6), 3861–3869.

Miller, R. (1996). *Axonal conduction times and human cerebral laterality. A psychobiological theory.* Reading, UK: Harwood Academic.

Nelken, I., & Chechik, G. (2007). Information theory in auditory research. *Hearing Res.*, *229*, 94–105.

Oertel, D., Bal, R., Gardner, S., Smith, P., & Joris, P. (2000). Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. *PNAS*, *97*(22), 11773–11779.

Perrinet, L., Delorme, A., Samuelides, M., & Thorpe, S. J. (2001). Networks of integrate-and-fire neuron using rank order coding a: How to implement spike time dependent hebbian plasticity. *Neurocomputing*, *38–40*, 817–822.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

Recio, A., & Rhode, W. (2000). Representation of vowel stimuli in the ventral cochlear nucleus of the chinchilla. *Hearing Res.*, *146*, 167–184.

Rhode, W. S., Roth, G. L., & Recio-Spinoso, A. (2010). Response properties of cochlear nucleus neurons in monkeys. *Hearing Research*, *259*, 1–15.

Sadagopan, S., & Wang, X. (2009). Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. *J. Neurosci.*, *29*(36), 11192–11202.

Salami, M., Itami, C., Tsumoto, T., & Kimura, F. (2003). Change of conduction velocity by regional myelination yields constant latency irrespective of distance between thalamus and cortex. *PNAS*, *100*, 6174–6179.

Schnupp, J. W. H., Hall, T. M., Kokelaar, R. F., & Ahmed, B. (2006). Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. *J. Neurosci.*, *26*(18), 4785–4795.

Stringer, S., Perry, G., Rolls, E., & Proske, J. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biol. Cybern.*, *94*, 128–142.

Tromans, J. M., Higgins, I. V., & Stringer, S. M. (2012). Learning view invariant recognition with partially occluded objects. *Frontiers in Computational Neuroscience*, *6*, 48.

Ulanovsky, N., Las, L., Farkas, D., & Nelken, I. (2004). Multiple time scales of adaptation in auditory cortex neurons. *J. Neurosci.*, *24*, 10440–10453.

van Rossum, M. C. W., Bi, G. Q., & Turrigiano, G. G. (2000). Stable Hebbian learning from spike timing-dependent plasticity. *Journal of Neuroscience*, *20*(23), 8812–8821.

Wever, E., & Bray, C. (1930). The nature of acoustical response: The relation between sound frequency and frequency of impulses in the auditory nerve. *J. Exper. Psychol.*, *13*, 373–387.

Winter, I., & Palmer, A. (1990). Responses of single units in the anteroventral cochlear nucleus of the guinea pig. *Hearing Res.*, *44*, 161–178.

Winter, I., Palmer, A., Wiegrebe, L., & Patterson, R. (2003). Temporal coding of the pitch of complex sounds by presumed multipolar cells in the ventral cochlear nucleus. *Speech Commun.*, *41*, 135–149.

Young, E. D., & Sachs, M. B. (2008). Auditory nerve inputs to cochlear nucleus neurons studied with cross-correlation. *Neuroscience*, *154*, 127–138.

Zilany, M., Bruce, I., Nelson, P., & Carney, L. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. *J. Acoust. Soc. Am.*, *126*(5), 2390–2412.

Received July 29, 2017; accepted January 23, 2018.