PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Neural modelling of the encoding of fast frequency modulation

Alejandro Tabas[1,2]*, **Katharina von Kriegstein**[1,2]

1 Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, Technische Universität Dresden, Dresden, Saxony, Germany, **2** Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Saxony, Germany

* alejandro.tabas@tu-dresden.de

Abstract

Frequency modulation (FM) is a basic constituent of vocalisation in many animals as well as in humans. In human speech, short rising and falling FM-sweeps of around 50 ms duration, called formant transitions, characterise individual speech sounds. There are two representations of FM in the ascending auditory pathway: a spectral representation, holding the instantaneous frequency of the stimuli; and a sweep representation, consisting of neurons that respond selectively to FM direction. To-date computational models use feedforward mechanisms to explain FM encoding. However, from neuroanatomy we know that there are massive feedback projections in the auditory pathway. Here, we found that a classical FM-sweep perceptual effect, the sweep pitch shift, cannot be explained by standard feedforward processing models. We hypothesised that the sweep pitch shift is caused by a predictive feedback mechanism. To test this hypothesis, we developed a novel model of FM encoding incorporating a predictive interaction between the sweep and the spectral representation. The model was designed to encode sweeps of the duration, modulation rate, and modulation shape of formant transitions. It fully accounted for experimental data that we acquired in a perceptual experiment with human participants as well as previously published experimental results. We also designed a new class of stimuli for a second perceptual experiment to further validate the model. Combined, our results indicate that predictive interaction between the frequency encoding and direction encoding neural representations plays an important role in the neural processing of FM. In the brain, this mechanism is likely to occur at early stages of the processing hierarchy.

Author summary

Humans' ability to understand and produce speech is one of the most fascinating developments of evolution. It is critical for smooth daily routines at the individual to the societal level. The computational mechanisms that the human brain uses for excelling at speech recognition are far from understood. One of the fundamental building blocks of speech are so-called formant transitions that characterise different speech sounds. To date, formant transitions are assumed to be processed according to a representational framework.

OPEN ACCESS

Citation: Tabas A, von Kriegstein K (2021) Neural modelling of the encoding of fast frequency modulation. PLoS Comput Biol 17(3): e1008787. https://doi.org/10.1371/journal.pcbi.1008787

Editor: Frédéric E. Theunissen, University of California at Berkeley, UNITED STATES

Received: July 12, 2020

Accepted: February 12, 2021

Published: March 3, 2021

Data Availability Statement: All data, experiment scripts, code for the model, and scripts producing the figures in the manuscript are publicly available in a github repository: https://github.com/qtabs/fmPitch.

Funding: AT and KvK were funded by the ERC (erc.europa.eu) Consolidator Grant SENSOCOM 647051. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

In this view, the brain processes auditory signals in a hierarchical constructive way, where the higher levels of the hierarchy, that represent the formant transition directions, are informed by the neural representations of individual frequencies at the lower levels, but not vice versa. Here, we show that the representational framework does not fully explain human behaviour. Instead we develop a novel computational model in which the neural representations of formant transitions influence lower-level representations. This mechanism effectively increased the speed and efficiency of the recognition of formant transitions. The model explained previously unaccounted phenomena in human perceptual behaviour. These neural principles can be extended to other auditory processing networks and sensory modalities, and can be incorporated to neurobiologically-inspired automatic speech recognition algorithms.

Introduction

Frequency modulation (FM) is a basic acoustic feature of animal vocalisation, human speech, and music. In human speech, consonants preceding and following a vowel can be acoustically characterised by formant transitions: a series of simultaneous fast FM sinusoids of around 50 ms duration that start or finish in the frequencies characterising the vowel [1]. At all stages of the ascending auditory pathway, FM is represented along the tonotopic axis in a *spectral representation* that encodes the instantaneous frequency of the stimuli [2]. Individual neurons at higher levels of the processing hierarchy (inferior colliculus [3–5], medial geniculate body [6, 7], and auditory cortex [8–11]) also encode FM direction and FM rate, by responding selectively to certain rates and direction. We call this latter, more abstract representation, the *sweep representation*.

Despite the massive feedback projections that characterise the auditory pathway [12, 13], computational models to date use only feedforward mechanisms to explain FM encoding [14]. Given the importance of high-order predictive elements in the optimisation of speech (e.g., [15]) and FM [16] recognition, descending projections are likely to play an important role for the encoding of fast FM-sweeps in the auditory system. However, to date there is no comprehensive model of such fast FM encoding that incorporates both, the sweep and the spectral representations, and describes the potential sweep-to-spectral feedback mechanisms active during the processing of FM sounds. The aim of the present study was to develop such a model, with a focus on FM-sweeps of the duration and frequency span of formant transitions.

To do that we harnessed a classical behavioural effect from psychoacoustics first reported around 60 years ago [17], which we will refer to as the sweep pitch shift. In the original experiment, participants listened to fast rising and falling FM-sweeps. The authors discovered that the participants judged up sweeps as eliciting a higher pitch than down sweeps with the same average fundamental frequency. These findings were later replicated [18, 19], although a reliable quantitative assessment of the phenomenon using stimuli with a controlled spectrum is lacking. To explain the effect d'Alessandro and colleagues proposed a phenomenological model assuming that the pitch of a sweep is integrated using a fixed-size window from the instantaneous frequency of the stimulus across time [20, 21]. Due to the leaking memory of the integration, this process naturally favours the latest frequencies of the sweep, explaining the sweep pitch shift. However, the authors found that different integration weights were necessary to explain different partitions of their data, indicating that the phenomenological model is not a parsimonious explanation of the sweep pitch shift.

Using the classical behavioural effect we approached the development of a comprehensive FM-encoding model in three steps. First, we re-examined and quantified the sweep pitch shift in a behavioural experiment, and tested whether the experimental data could be explained by existing computational models [22–24]. We found that mechanistic models of pitch processing that attempt to describe the circuitry underlying perception rather than perceptual phenomena [25] could not explain the sweep pitch shift. Current models of FM encoding [14] consider a static representation of spectral information, and thus they predict that the sweep pitch shift would not occur. Thus neither existing models of pitch processing nor FM-encoding could explain the sweep pitch shift. In a second step we built a hierarchical model motivated by the hypothesis that the sweep pitch shift results from the modulation exerted by feedback projections between the sweep and spectral representations. The feedforward components of the model were based on results of previous studies on FM direction selectivity and included processing of instantaneous frequency and processing of FM direction [10, 14, 26]. The feedback architecture was based on generative hierarchical models and predictive coding [27, 28] and informed by the human psychophysics results from the first part of the study. In the third and last step, we used a new set of stimuli termed *sweep trains* to further validate the model. These stimuli, consisting of a concatenation of five FM-sweeps, preserve the same acoustical features of the original FM-sweeps but elicited different dynamics in the feedback system of the model than their single-sweep counterpart. The ability of the model to predict the pitch elicited by these new stimuli illustrated that the feedback mechanisms proposed in this work, and not bottom-up acoustical features of the stimuli, are the driver of the sweep pitch shift.

Results

The sweep pitch shift revisited

First, we re-examined and quantified the sweep pitch shift, measured as the difference between the perceived pitch and the average frequency of the sweep: $\Delta p = f_{\text{perceived}} - \bar{f}$. Eight participants matched a total of 30 fast FM sweeps with frequency spans $\Delta f \in [-600, 600]$ Hz and 3 average frequencies $\bar{f} \in \{900, 1200, 1500\}$ Hz (see Methods). Each sweep had a duration of 40 ms and was preceded and followed by 5 ms of constant frequency. Participants' task was to match each sweep to a pure tone, which frequency was used to determine the elicited pitch of the sweep.

The pitch shift Δp depended on the sweep's span Δf (Fig 1A and S1 Table). The exact dependence was consistent across participants for sweeps with $\Delta f \leq 333$ Hz lying in the vicinity of the linear fit $f_{\text{perceived}} \simeq \bar{f} + m\,\Delta f$. There was an average deviance from the fit of 46 Hz. Sweeps with larger frequency spans resulted in wider distributions of $f_{\text{perceived}}$ (Pearson's $r = 0.48$, $p < 10^{-14}$; Fig 1B); all subjects showed the same sweep pitch shift direction and comparable orders of magnitude on the dependence of Δp with Δf (S1 Fig). Presenting the sweep before or after the probe tone did not systematically affect the perceived pitch (S2 Fig). Raw data is available in an external repository (github.com/qtabs/fmPitch).

In their classical study, Brady and colleagues [17] showed that the absolute value of the sweep pitch shift $|\Delta p|$ is larger for down than for up sweeps. In a later study, Nabelek and colleagues [18] showed the reversed effect. To test if our data replicates any of these previous findings, or shows no up/down asymmetry at all, we drew, for each absolute frequency span $|\Delta f|$, the distribution of the differences between the pitch shift in up and down sweeps:

$$\text{asymm}^{\uparrow\downarrow}_{|\Delta f|} = |\Delta p(\Delta f)| - |\Delta p(-\Delta f)| \tag{1}$$

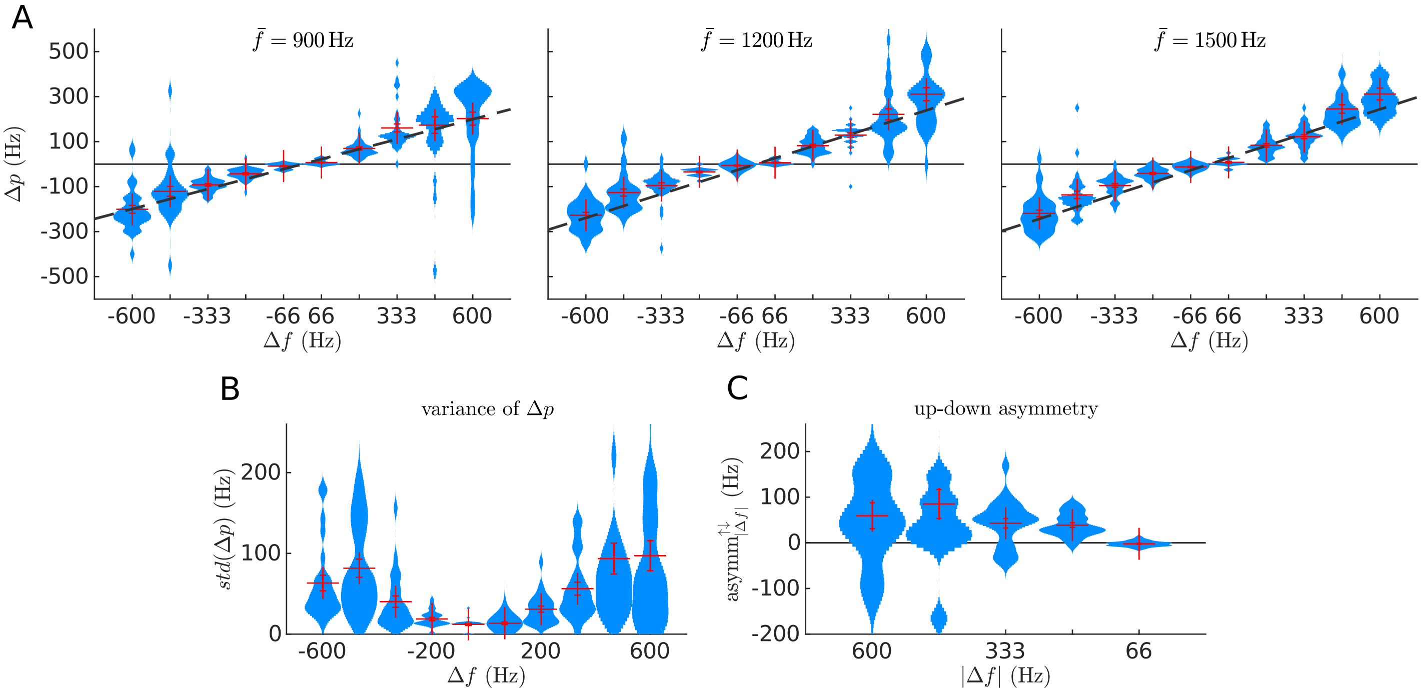

Fig 1. Sweep pitch shift. A) Kernel density estimations on the perceived pitch are plotted separately for each of the 30 FM-sweeps used in experiment 1; each of the three panels displays the data of each of the three average frequencies $\bar{f}$. Data points correspond to the average sweep pitch shift $\Delta = f_{\text{perceived}} - \bar{f}$ (y-axis) of a target sweep with a specific frequency span Δf (x-axis) reported by one participant. Red crosses show the mean and standard error of the average Δp across participants. Dark dashed lines show the linear fit of the average Δp across participants. B) Kernel density estimations of the subjects' standard deviation of the sweep pitch shift Δp, plotted separately for the different frequency spans Δf. Each sample in the distributions corresponds to the standard deviation of the perceived pitch of a sweep in one subject (i.e., in each distribution there are 8 × 3 points, one for each subject and $\bar{f}$). The standard deviation is monotonically correlated to the absolute span $|\Delta f|$ (Pearson's $r = 0.48$, $p < 10^{-14}$). C) Kernel density estimations of the up/down asymmetry asymm$^{\uparrow\downarrow}$ distributions as defined in Eq (1). Each sample of the distributions corresponds to the difference of the average absolute deviation from centre frequency between up and down sweeps of the same $|\Delta f|$ for a given subject and centre frequency ($N = 8 \times 3 = 24$). Red crosses show the mean and the standard error of the data.

https://doi.org/10.1371/journal.pcbi.1008787.g001

Our results robustly replicated the observations from Nabelek and colleagues (Fig 1C). The sweep pitch shift was significantly larger for up than down sweeps for $|\Delta f| \geq 200$Hz ($p < 2 \times 10^{-5}$) but not for $|\Delta f| = 66$Hz ($p = 0.77$), according to two-tailed rank-sum tests (number of samples $N = 96$).

Last, we tested if the dependence of the sweep pitch with Δf was robustly replicated across subjects. The slopes of the linear fits between $f_{\text{perceived}}$ and Δf were of similar magnitude in all participants (mean slope $m = 0.38$, standard deviation across subjects $\sigma_m = 0.07$, a 18.5% of the nominal value, corresponding to an effect size of $d = 5.4$; see S1 Fig).

Bottom-up models of pitch do not explain the sweep pitch shift

Current theories of pitch suggest that two complimentary codes of pitch coexist in the auditory system: the spectral or place code, produced by the spectral decomposition of the stimuli performed by the basilar membrane; and the temporal code, comprised in the spike timings of the neurons across the auditory pathway that are phase locked to the stimulus waveform (see [29] for a review). If the sweep pitch shift was a consequence of bottom-up pitch processing, we would expect the effect to be explainable by computational models that

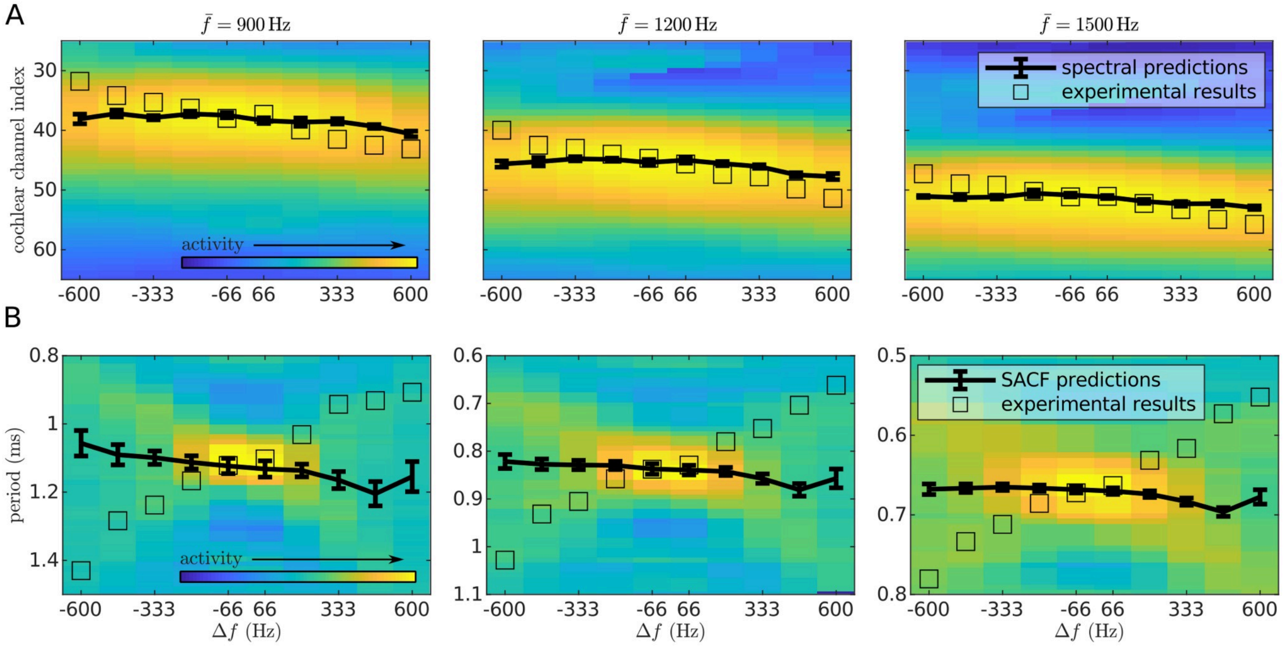

Fig 2. Predictions of the bottom-up models of pitch. A) Predictions of the spectral model. Heatmaps show the mean activation across the duration of the stimuli at different cochlear channels, as simulated by a model of the auditory periphery in response to each sweep; units are arbitrary. Error bars point to the expected value and standard deviation of the distribution across frequencies for each sweep. Each empty square denotes the expected channel elicited by a pure tone with the frequency of the average experimental data of the corresponding sweep. B) Predictions of the temporal model. Heatmaps show the distribution across periods elicited in the summary autocorrelation factor (SACF) for each sweep. The value corresponding to each period was computed as the average activation of the SACF across four harmonics (see Methods); heatmap units are arbitrary. Error bars point to the expected value and standard deviation of the distribution across periods for each sweep. Each empty square denotes the expected period elicited in the SACF by a pure tone with the frequency of the average experimental data of the corresponding sweep.

https://doi.org/10.1371/journal.pcbi.1008787.g002

use either of the two representations to infer pitch. To test this we computed the pitch predicted by one representative model of each family; i.e., one model using the spectral and one model using the temporal codes. Although the possibility that the auditory system integrates information from both codes has been theoretised (e.g., [25]) and implemented [30] before, the combination of both codes has so far proposed to be purely additive. Thus, a combined spectro-temporal approach could only explain the sweep pitch shift if at least one of the two codes shows a positive pitch shift for up sweeps ($\Delta f > 0$) and a negative pitch shift for down sweeps ($\Delta f < 0$).

In the spectral model, pitch can be directly inferred by computing the expected value of the activity across cochlear channels in the auditory nerve [22, 25]. Predictions of the spectral model approximate the empirical data for $\Delta f \sim 0$. Unlike the empirical data, however, predictions of the spectral model show no systematic dependence of $f_{\text{perceived}}$ on Δf (Fig 2A). More sophisticated spectral models designed to explain how the pitch of harmonic complex tones is encoded [25] would yield identical results because the sinusoidal FM-sweeps used in the present experiment evoke a single peak in the spectral distribution.

The temporal model was based on the principles of the summary autocorrelation function (SACF), that measures pitch according to the phase-locked response in the auditory pathway [23, 31]. We chose this model because it performs a relatively straightforward analysis of phase-locked activity. As in the spectral model, predictions of the SACF approximate the empirical data for $\Delta f \sim 0$, but do not show the dependence with Δf observed in the data

(Fig 2B). This is most likely a consequence of the 2.5 ms integration time window for phase-locked activity in the auditory system [32] being too large to integrate the rapidly changing frequencies of our stimuli (up to 15 Hz/ms).

We selected these two representative models of pitch processing because they keep the largest possible amount of information from the peripheral system. Our reasoning is that, if the sweep pitch shift cannot be derived from a minimally-processed code extracted from the peripheral output, it is very unlikely that it can be derived from any further bottom-up processing of this information.

The FM-feedback spectral model

In this section we introduce a hierarchical model of FM-encoding in the auditory system, termed *FM-feedback spectral model*, with two levels (Fig 3). In the first level, the *spectral* layer holds a spectral representation of the sound. In the second level, the *sweep* layer encodes FM-sweep direction. The spectral layer uses the spectral rather than the temporal code to represent

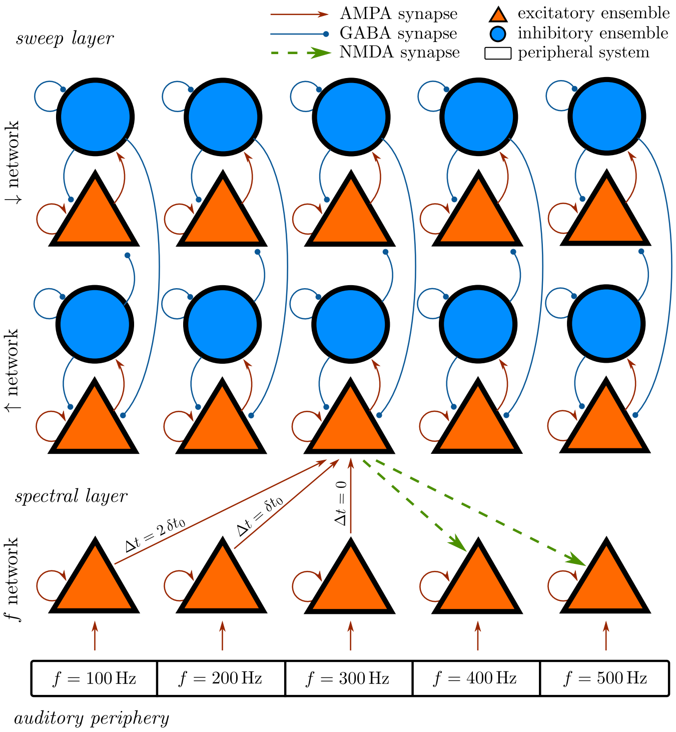

Fig 3. Diagram of the FM-feedback spectral model. The model consists of two levels: first, the *spectral layer*, with a network integrating the spectral information of the sound (*f* network); and second, the *sweep layer*, with one network specialised in detecting up (↑ network) sweeps and another network specialised in detecting down (↓ network) sweeps. The spectral layer integrates afferent inputs from the auditory periphery and encodes a representation of the stimulus that can be used to infer pitch. The sweep layer receives afferent inputs from the spectral layer that are used to decode the direction of the sweeps. Feedback connections from the sweep layer to the spectral layer modulate the time constants of the populations that are expected to be activated once the direction of the sweep has been decoded. The inhibitory ensembles in the up and down network enforce competition between up and down ensembles in a winner-take-all fashion. Note that the diagram is schematic and shows only 5 of the $N = 100$ populations and a single example of the connections between the sweep and the spectral layers. The labels of the boxes of the peripheral system are also schematic: the spectral resolution of the peripheral system is much higher. See Methods for specific details on the mathematical formulation of the model.

https://doi.org/10.1371/journal.pcbi.1008787.g003

the instantaneous frequency of the stimuli because the integration window of the auditory system is too large to integrate the rapidly changing frequencies of the sweeps used in our experiment [32]. The animal electrophysiology literature also converges in the notion that sweep direction and rate are decoded from the spectral, and not the temporal representation of the sounds [4, 6, 7, 11, 14, 33].

The main hypothesis introduced in the FM-feedback spectral model is that, once the direction of the sweep is encoded in the sweep layer, a feedback mechanism modulates the effective time constant of the populations encoding the frequencies that are expected to be activated in the next instant in the spectral layer. We expect this mechanism to qualitatively explain why the posterior parts of the sweep are given a higher weight during perceptual integration and to quantitatively reproduce the exact dependence of pitch with Δ*f* observed in our data. An implementation of the FM-feedback spectral model written in python is freely available at github.com/qtabs/fmPitch.

Example responses of the excitatory populations of the model to up and down sweeps are shown in Fig 4.

Modelling FM direction selectivity. We modelled FM direction selectivity using the principles of delayed excitation, a mechanism where neurons with different best frequencies output to the direction selective neuron with different delays [6, 10, 14, 26]. This mechanism introduced consistent delays between the populations in the spectral and the sweep layers. A sweep population receiving direct input from the spectral population encoding f_0 and responding

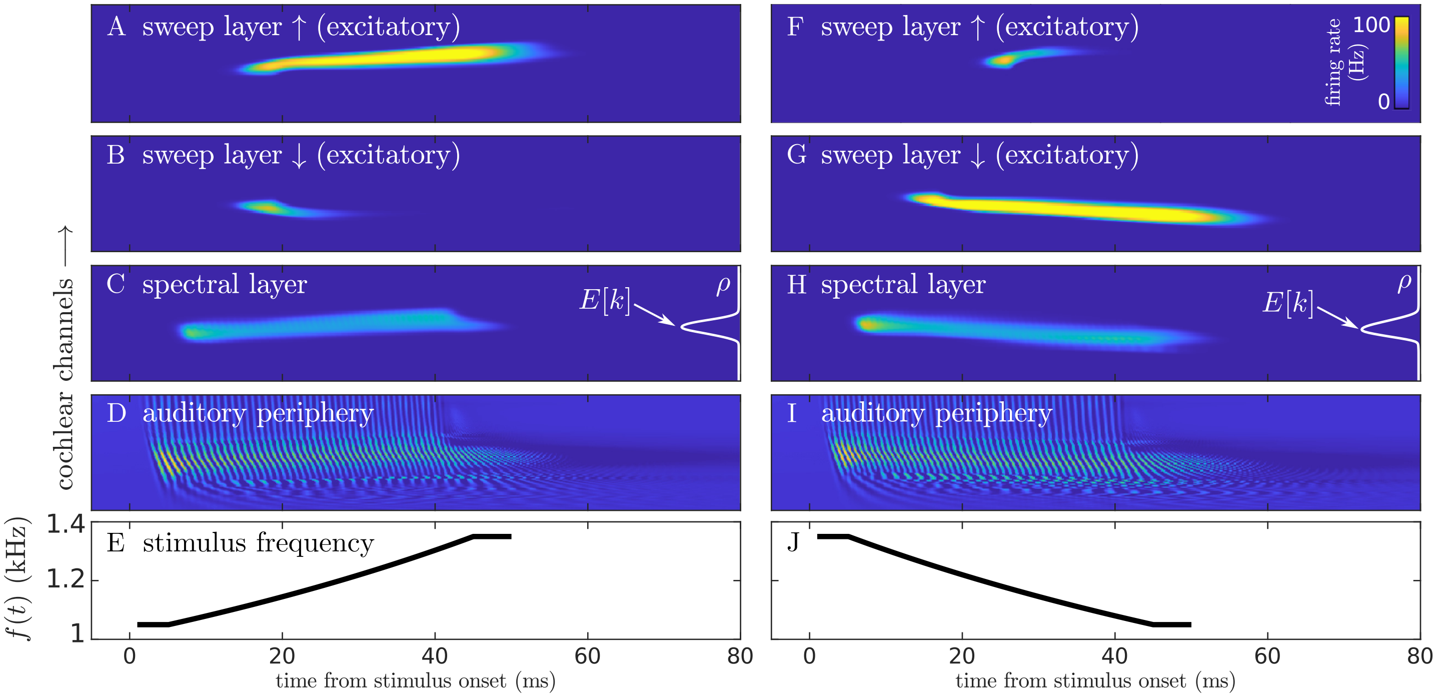

Fig 4. Model responses to an up and a down sweep. A-E show the responses to an up sweep, and F-J to a down sweep. From top to bottom: (A/F) the instantaneous predicted firing rate of the up-selective excitatory populations in the sweep layer; (B/G) the instantaneous firing rate of the down-selective excitatory populations in the sweep layer; (C/H) the instantaneous firing rate of the populations in the spectral layer and, in the right of the panels, a schematic view of the probability distribution of pitch derived from this representation; (D/I) the output of the model of the auditory periphery; (E/F) the instantaneous frequency of the sweeps along time. In all panels except for E/J, *y*-axis represents the cochlear channel *n*, ordered from bottom (lowest best frequency) to top (highest best frequency). The stimuli were the up and a down sweeps with Δ = ±300Hz and $\bar{f} = 1200$ Hz used in the experiment.

https://doi.org/10.1371/journal.pcbi.1008787.g004

selectively to up sweeps will receive increasingly delayed inputs from the spectral populations centred at $f < f_0$ (Fig 3). The relative delay in the connection between a spectral population m and a target sweep population n depends linearly on the spectral distance between the two ensembles: $\delta t_{nm} = |n - m|\delta t_0$. Although this configuration is optimal for linear sweeps with slopes $\simeq \delta t_0/(f_n - f_{n-1})$, adding parallel replicas of these populations with varying δt_0 would suffice to generalise the mechanism to a wider range of speeds and to non-linear sweeps. Populations selectively responding to specific rates have been reported in bats [34–37] and rodents [3, 8, 9].

The sweep layer consists of two networks, each encoding one of the FM directions and responding selectively to *up* (↑) and *down* (↓) sweeps. Each of the networks consists of N columns, each comprising an excitatory and an inhibitory population (Fig 3). Note that populations in the sweep network also have a best-frequency and they are thus arranged according to their corresponding cochlear channel: an up population in the sweep network responds selectively to up sweeps when these span through a certain frequency range (see Fig 4B).

To quantify direction selectivity, we used the standard direction selectivity index (DSI; e.g., [11]), defined as the proportion of the activity elicited in a network by an up sweep minus the activity elicited in the same network by a down sweep with the same duration and frequency span. An ideal network responding selectively to up sweeps will have a DSI = + 1 and an ideal network responding selectively to down sweeps will have a DSI = −1. Similar DSI magnitudes are measured in the down and the up network (Fig 5). Network selectivity to FM direction was robust to variations of around 20% of the fitted value of the main parameters of the model pertaining direction selectivity (δt_0, and the conductivities and dispersion in the connectivity matrices of the bottom-up connections). Deactivation of the feedback connections, however, resulted in a (16 ± 1.4)% average decrease in absolute DSI, indicating that the feedback connections sharpened direction selectivity.

Although we did not attempt to model FM rate selectivity, the DSIs monotonically increased with $|\Delta f|$ (Fig 5), a property that could be exploited in further developments of the model to encode modulation rate [3, 7, 9].

Predictive mechanisms. Once neurons in the sweep layer encoded the sweep direction, feedback connections targeting the spectral layer applied currents that facilitated the encoding of expected frequencies. We will call them facilitation currents in the following. Let i be the population in the up-sweep network receiving inputs from a population in the spectral layer encoding a certain frequency f_0. Due to delayed excitation, the population i becomes active when it detects an up sweep occurring in the neighbourhood of frequencies $f \leq f_0$. Although

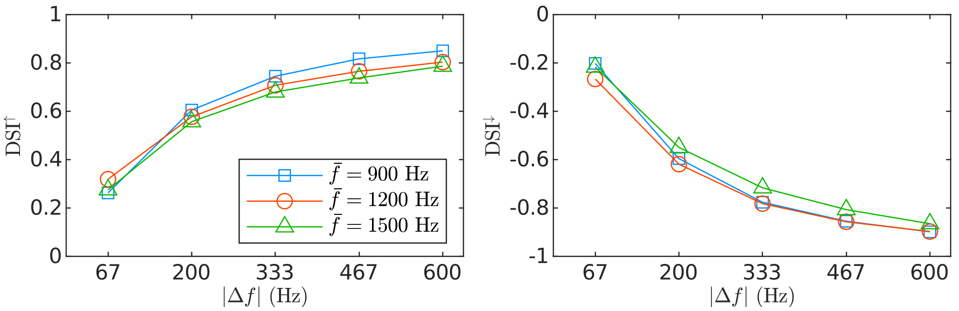

Fig 5. Direction selectivity indices for the FM-sweeps of the experiment. DSI$^{\uparrow}$ and DSI$^{\downarrow}$ to sweeps with different $\bar{f}$ and $|\Delta f|$. DSI is defined as the proportional activity to up in comparison to down sweeps in a given network.

https://doi.org/10.1371/journal.pcbi.1008787.g005

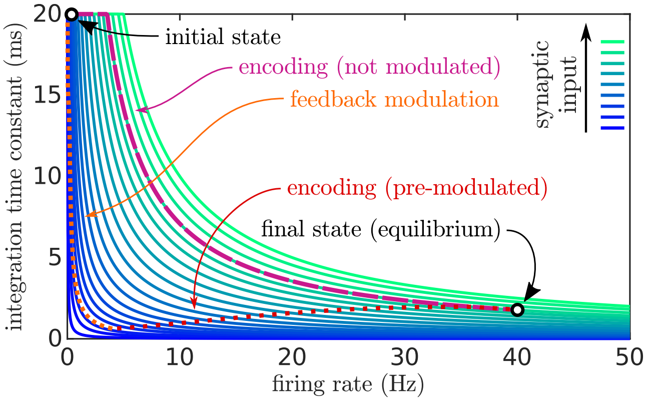

Fig 6. Effect of the predictive feedback mechanism on the population time constants in the spectral layer. $\tau^{\mathrm{pop}}(h, I)$ (green-blue solid lines) depends on the firing rate h and the synaptic input I (see Methods). The figure portraits two different trajectories of the variable $\tau^{\mathrm{pop}}(h, I)$ in the (h, I) space, both starting at an initial state ($h \sim 0$ in the regime of spontaneous activity with no inputs) and finishing at an equilibrium state (with $h \sim 40$ Hz). The dashed purple line shows a trajectory followed by the population when the forward synaptic input from the peripheral layer is plugged in without feedback modulation. In this case, the population reacts slowly to the strong synaptic input, and eventually converges to equilibrium. The dotted lines (orange and red) show the trajectory of the same population in the presence of feedback modulation (i.e., the facilitation currents). The low-current feedback excitatory signals drive the population to a state with a low effective time constant without substantially increasing its firing rate (orange section of the trajectory). When the strong synaptic input from the auditory periphery is switched on (red section of the trajectory) the population reacts quickly to the synaptic input, reaching equilibrium much faster than in the non-modulated case.

https://doi.org/10.1371/journal.pcbi.1008787.g006

when f_0 is the ending frequency of the sweep the following frequencies will not activate next, most often f_0 will be just an intermediate frequency within the sweep. Thus, activation of i would imply that populations in the spectral layer with best frequencies immediately higher than f_0 are likely to activate next. The facilitation currents, encoded in the feedback projections stemming from the sweep layer and targeting the spectral layer, reduce the reaction time of the populations in the spectral layer that are expected to activate next using low-current feedback excitatory signals. Similarly, feedback connections stemming from a population j in the down-network that received timely inputs from a spectral population with best frequency f will target populations in the spectral network with best frequencies immediately lower than f_0.

NMDA receptors are typically responsible for conveying feedback excitatory information in the cerebral cortex [38, 39]; specifically, NMDA-deactivation results in a reduced feedback control in the auditory pathway [40]. Thus, while bottom-up drive was modelled using AMPA dynamics, feedback connections were modelled according to NMDA-like synaptic gating dynamics with a finite rising time constant [41]. Feedback current intensity was kept low in comparison to the bottom-up driver by enforcing NMDA conductivity to be much smaller than the AMPA conductivity (i.e., $J^{\mathrm{NMDA}} \ll J^{\mathrm{AMPA}}$).

The facilitation currents modulated the spectral population that is expected to fire next so that it subtly increased its firing rate with respect to a not modulated population. Due to network effects captured in the mean-field model [42], this subtle activation driven by a low-current effectively reduces the neural population's decaying time constant τ^{pop} (Fig 6), equivalent to a smaller integration time window of a leaky integrator. Endowed with a smaller effective integration time constant, the population integrates the sensory input faster and spends more time in the high-firing-rate regime than a population that has not been facilitated. Since facilitated populations spend more time in the high-firing-rate regime, frequencies expressed in the last part of the sweep have stronger contributions to the probability distribution of pitch.

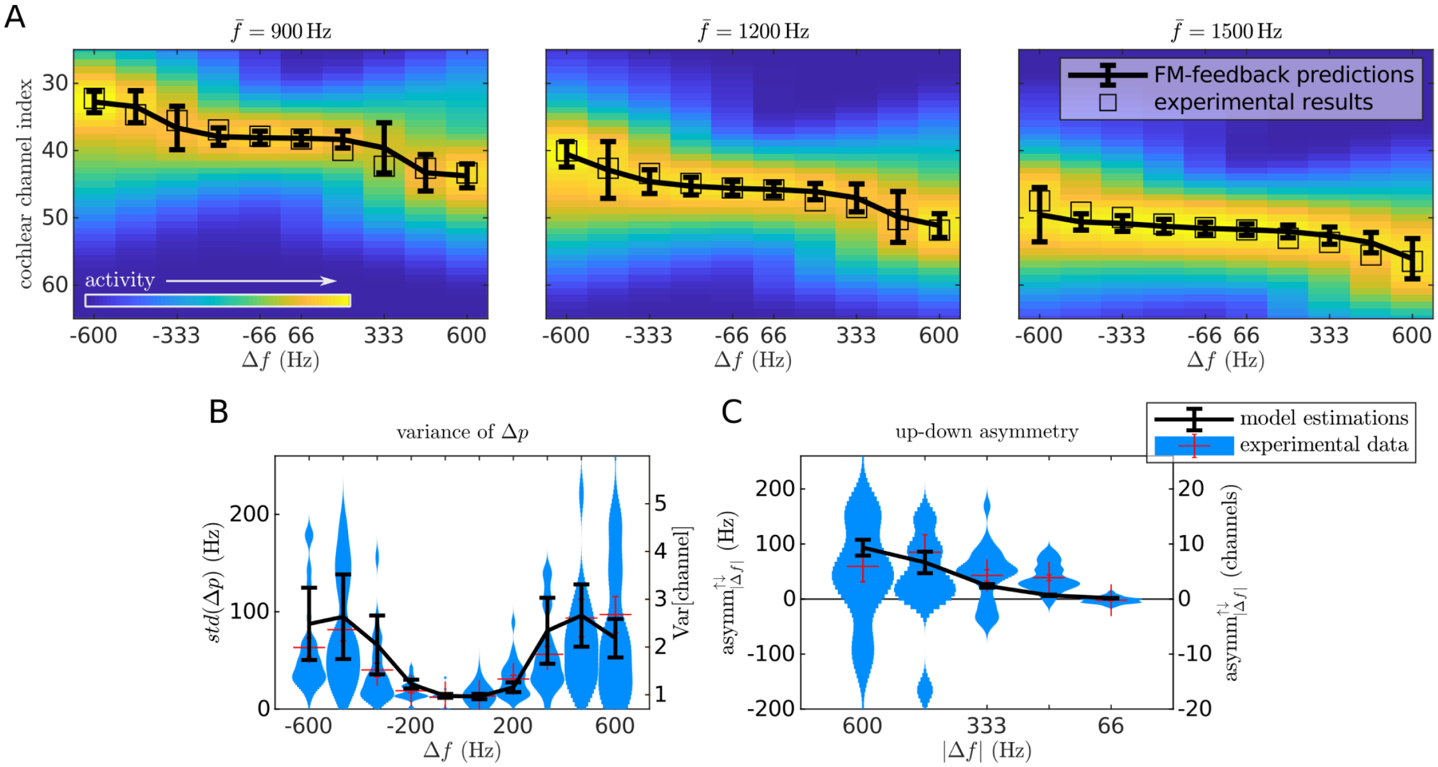

Fig 7. Predictions of the FM-feedback spectral model for FM-sweeps. A) Heatmaps show the mean activation at different cochlear channels as simulated by a model of the auditory periphery in response to each sweep; units are arbitrary. Error bars point to the expected value and standard deviation of the distribution across frequencies for each sweep. Each empty square denotes the expected channel elicited by a pure tone with the frequency of the average experimental data of the corresponding sweep. B) Predictions of the standard deviation of the perceived pitch. Error bars are estimations of standard error of the model calculated based on the dispersion of the centroids for different $\bar{f}$ and the standard deviations of the spectral distribution ρ of each condition. Kernel density estimations are the same experimental data as in Fig 1B. C) Predictions of the standard deviation and up/down asymmetry. Error bars show the model predictions of the up/down asymmetry coefficient asymm$^{\uparrow\downarrow}$ (see Eq (1)) in the channel space. Kernel density estimations are the same experimental data as in Fig 1C.

https://doi.org/10.1371/journal.pcbi.1008787.g007

Stimuli with constant frequency (e.g., pure tones) do not drive any of the sweep populations and thus do not activate any feedback mechanisms. Therefore, in the absence of FM, the model reduces to a purely bottom-up spectral model.

Reproduction of the sweep pitch shift by the FM-feedback spectral model. The FM-feedback spectral model explains $R^2 = 0.97$ of the variance of the experimental data (Fig 7A). Moreover, there was a significant correlation between the variance of the model responses and the standard error of the experimental data ($r_p = 0.63, p < 10^{-10}$), indicating that the larger variability in the sweep pitch shift observed for the larger Δf can be understood as a consequence of a wider spread activation across the spectral populations.

Up sweeps partially compensate for the differential delay in the basilar membrane responses to low frequencies with respect to high frequencies, provoking higher synchronisation in the auditory nerve [43]. Stronger peak activities result on slightly higher facilitation currents for up than for down sweeps, causing a noticeable stronger absolute mean pitch shift for up than for down sweeps, as observed in the experimental data (Fig 7B). Note that this is not the result of the model overfitting the data, since the average error of the model fit (E[error] $\simeq$ 1 channel $\simeq$ 50 Hz) is of the same order of magnitude as the up-down asymmetry $E[\mathrm{asymm}^{\uparrow\downarrow}] \simeq 100$ Hz.

To study the dependence of the model fit with the model's parameters we recomputed the explained variance R^2 across the parameter space of the model (Fig 8). The model explained the experimental data in a wide section of the parameter space, with an average R^2 across a

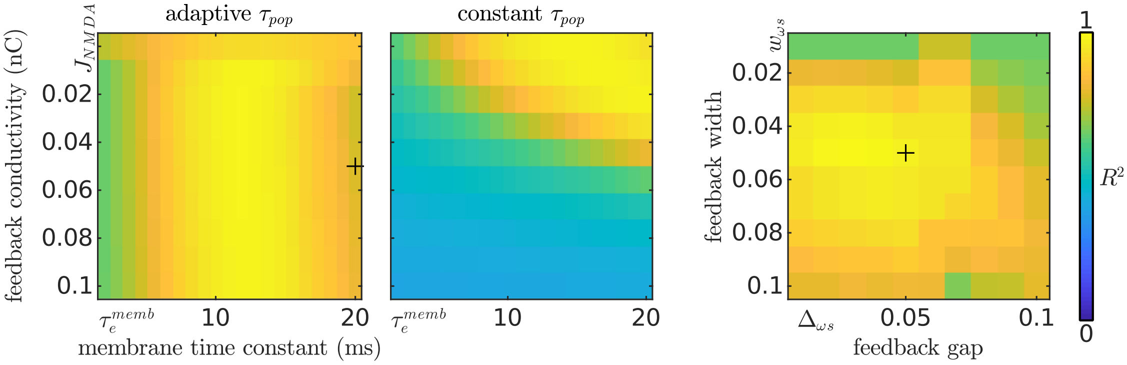

Fig 8. Experimental fit in relation to the FM-feedback spectral model parametrisation. Heatmaps show the explained variance of the experimental data R^2. Unless stated otherwise, parameters not varied in the matrices correspond to the values listed in Table 1. The two leftmost plots show the dependence of R^2 with the conductivity of the feedback connections and the dynamics of the excitatory population time constants. Different values of the nominal population's time constant were used to illustrate that the dynamic effect (rather than the resulting shorter time constant) is crucial to explain the experimental results; however, during the parameter tuning the temporal constant was constrained to $\tau^{\text{memb}} = 20$ ms based on physiological observations [44]. The rightmost plot shows the dependence of R^2 on the width and reach ($w_{\omega s}$ and $\Delta_{\omega s}$, respectively; see Methods) of the feedback connections. Black crosses in the parameter space signal the final parametrisation.

https://doi.org/10.1371/journal.pcbi.1008787.g008

5-point diameter sphere around the final parameters of $E[R^2] = 0.88 \pm 0.01$. To show that the fit of the model was not simply caused by an overall stronger activation provoked by the facilitation currents, but by a decrease in the effective time constant of the populations, we also computed the dependence of R^2 with the conductivity of the feedback J^{NMDA} while keeping the population time constant τ fixed to $\tau = \tau^{\text{memb}}$ (see Methods). Even considering lower τ^{memb} than the physiologically valid nominal value $\tau^{\text{memb}} = 20$ ms, without an adaptive τ, much stronger NMDA currents ($J^{\text{NMDA}} \sim J^{\text{AMPA}}$) are necessary to shift the peak of the distribution of the responses across the spectral layer towards the experimental results.

Reproduction of previous experimental results. We tested whether the FM-feedback spectral model was able to predict the pitch shift of additional data. For this, we chose the stimuli of Brady and colleagues [17], because this was the only study that investigated the dependence of the sweep pitch shift with properties different than Δf. Specifically, in the *experiment II* they considered FM-sweeps with a fixed 20 ms transition between 1000 Hz and 1500 Hz that was located at six different positions within a 90 ms stimulus (see schematics in Fig 9A, left). In the *experiment III*, they used FM-sweeps in the same Δf but with transitions of six different durations (see schematics in Fig 9A, right). All stimuli had the same duration (90 ms) and frequency span (1000-1500 Hz); in each of the two experiments there was a total of 12 stimuli (six up, six down).

We compared the predictions of the FM-feedback spectral model with the experimental results reported in the original paper (Fig 9B). The experimental trend is well reproduced by the model ($R^2 = 0.49$).

Testing the FM-feedback spectral model with a new class of stimuli

The results described so far are in favour of the hypothesis that there is a feedback system between populations of the spectral and sweep representations that has strong repercussions on perceptual behaviour. Next, to validate these findings, we introduced a new set of stimuli specifically designed to contest the main hypothesis of the model. The new stimuli consist of concatenations of several single sweeps with the same properties as the stimuli used in the first experiment. We call them *sweep trains* in the following. Sweep trains present the same

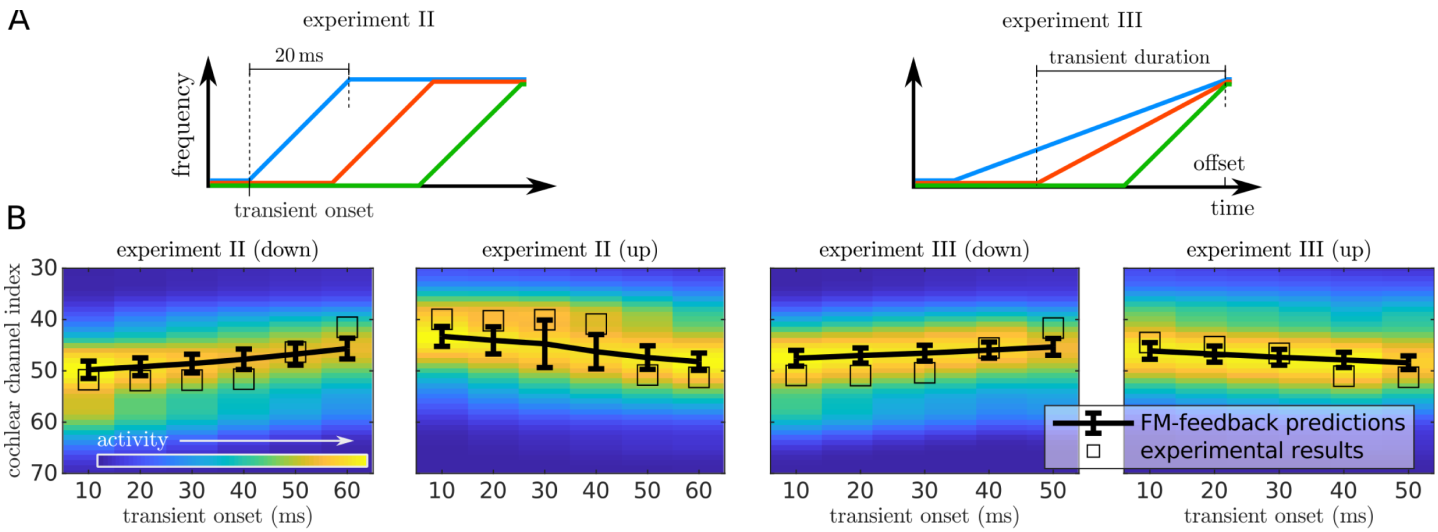

Fig 9. Predictions of the FM-feedback spectral model for Brady's stimuli. A) Schematic view of the stimuli from [17]. In experiment II (left) the transient was fixed to a 20 ms duration and its onset was systematically varied so that the transition falls at different segments of the stimulus. In experiment III (right) the stimulus offset was fixed at 90 ms and the transient's onset varied between 10 and 50 ms, resulting in transients of different durations. We extended the duration of these last stimuli to 95 ms to prevent the ramping at the end of the stimulus from overlapping with the FM transient. B) Model predictions. Heatmaps show the distribution of the activation across channels (y-axis) for different transient onsets (x-axis). Squares printed over the distributions mark the estimations of the experimental results in the channel space.

https://doi.org/10.1371/journal.pcbi.1008787.g009

acoustical properties as the single sweeps used in the first behavioural experiment and should nominally elicit the same pitch percept as their single-sweep subcomponents. However, the FM-feedback spectral model predicts that the feedback system will only reduce the time constant of the spectral populations during the processing of the first sweep in the train, because they will already have an elevated firing rate (and thus a low effective time constant) during the processing of the subsequent sweeps in the train. Consequently, the model predicts that the sweep trains will elicit a much more subtle pitch shift than their single sweep counterparts. We tested this prediction in a perceptual experiment analogous to the first experiment.

Sweep trains show minimal sweep pitch shift

Sweep trains were constructed using the sweeps from experiment 1. To ensure that each train was perceived as a single auditory object, we only used sweeps with $|\Delta f| \leq 333$ Hz, resulting in a total of $3 \times 6 = 18$ stimuli. As in the results from Experiment 1 (Fig 1), the magnitude of the pitch shift in sweep trains depended on Δf (Fig 10 and S1 Table, bottom). However, as qualitatively predicted by the FM-feedback spectral model, the effect sizes of the correlation were lower than in the single-sweep experiment (cf., S1 Table, top). Data also showed higher variability than in experiment 1 (S1 Fig). After completing the experiment, some participants reported in informal conversation that the sweep train stimuli were harder to match than the single-sweep counterparts. Although trains with small Δf were generally perceived as continuous tones, subjects reported that a few trains (putatively those with the largest Δf) elicited a ringing-phone-like percept. Stimuli are available in the supporting information (S1 Sounds).

Sweep-train stimuli show only a subtle up/down asymmetry that did not reach statistical significance even for the larger Δf ($p = 0.67$, $p = 0.96$, $p = 0.36$ for $|\Delta f| = 333$, $|\Delta f| = 200$, $|\Delta f| = 66$, respectively; according to two-sided Wilcoxon signed rank tests with 24 samples per condition).

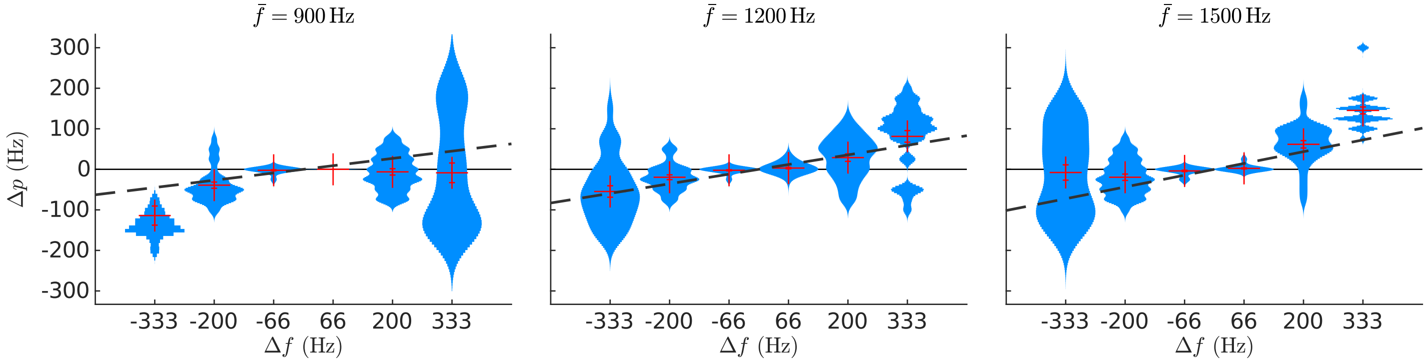

Fig 10. Sweep pitch shift for sweep trains. Kernel density estimations on the perceived pitch are plotted separately for each of the 18 sweep trains used in the experiment. The *y*-axis of each plot shows the magnitude of the sweep pitch shift Δ*p*. The x-axis list the spans of each of the sweeps. Red crosses show the mean and standard error of the data. Dark dashed lines show the group linear fit of the data.

https://doi.org/10.1371/journal.pcbi.1008787.g010

The FM-feedback spectral model explains the diminished pitch shift in the sweep trains

Next, we assessed the ability of the FM-feedback spectral model to quantitatively explain the effect size of the pitch shift observed in the sweep trains. The fit with the experimental data was comparable to that of the single sweep stimuli: the model explained $R^2 = 0.99$ of the variance of the data (Fig 11A). As in the first experiment, the standard deviations of the experimental data was strongly correlated to the width of the model responses ($r_p = 0.75$, $p < 10^{-10}$; Fig 11B).

Last, we tested whether the different up/down asymmetry ($\text{asymm}^{\uparrow\downarrow}$) observed in the single sweeps and sweep train data could be quantitatively explained by the FM-feedback spectral model. In the single-sweep data, the model predicts a stronger sweep pitch shift magnitude $|\Delta p|$ for up sweeps (Fig 7C) because these elicit a stronger peak activation in the auditory nerve [45], resulting in stronger facilitation currents. Qualitatively, a much weaker asymmetry was expected in the sweep-train data, since the spectral populations have already high firing rates (and thus low effective integration time constants) during the processing of the ending four fifths of the stimuli. In sweep trains, then, only the first sweep contributes to the sweep pitch shift, whereas the remaining sweeps provide equal contributions to the range of spanned frequencies, diluting the shift magnitude by four fifths. Modelling results on the up/down asymmetry closely reproduced the empirical data (Fig 11C), fully explaining the observed differences between the two classes of stimuli.

Discussion

In this work we have built a novel model that describes how feedback projections between the two different known representations of FM (i.e., spectral and sweep) could be used in the brain to facilitate encoding. This contrasts with the classical view of FM encoding as a bottom-up process [14]. The feedback mechanism proposed in this work uses predictions generated by populations encoding FM direction to aid encoding in populations encoding instantaneous frequency, enhancing direction selectivity and shortening FM processing time. Since this predictive facilitation is not intrinsically restricted to the fast-FM characteristic of formant transitions in speech, similar facilitation mechanisms could also boost encoding efficiency for the slower FM underlying the perception of prosody and melody.

In this work we have used the model to encode sinusoidal (pure-tone) FM stimuli that are far from the complexity of natural speech sounds. In speech, phonemes are characterised by concurrent formant transitions that span complementary frequency ranges. Moreover, each of

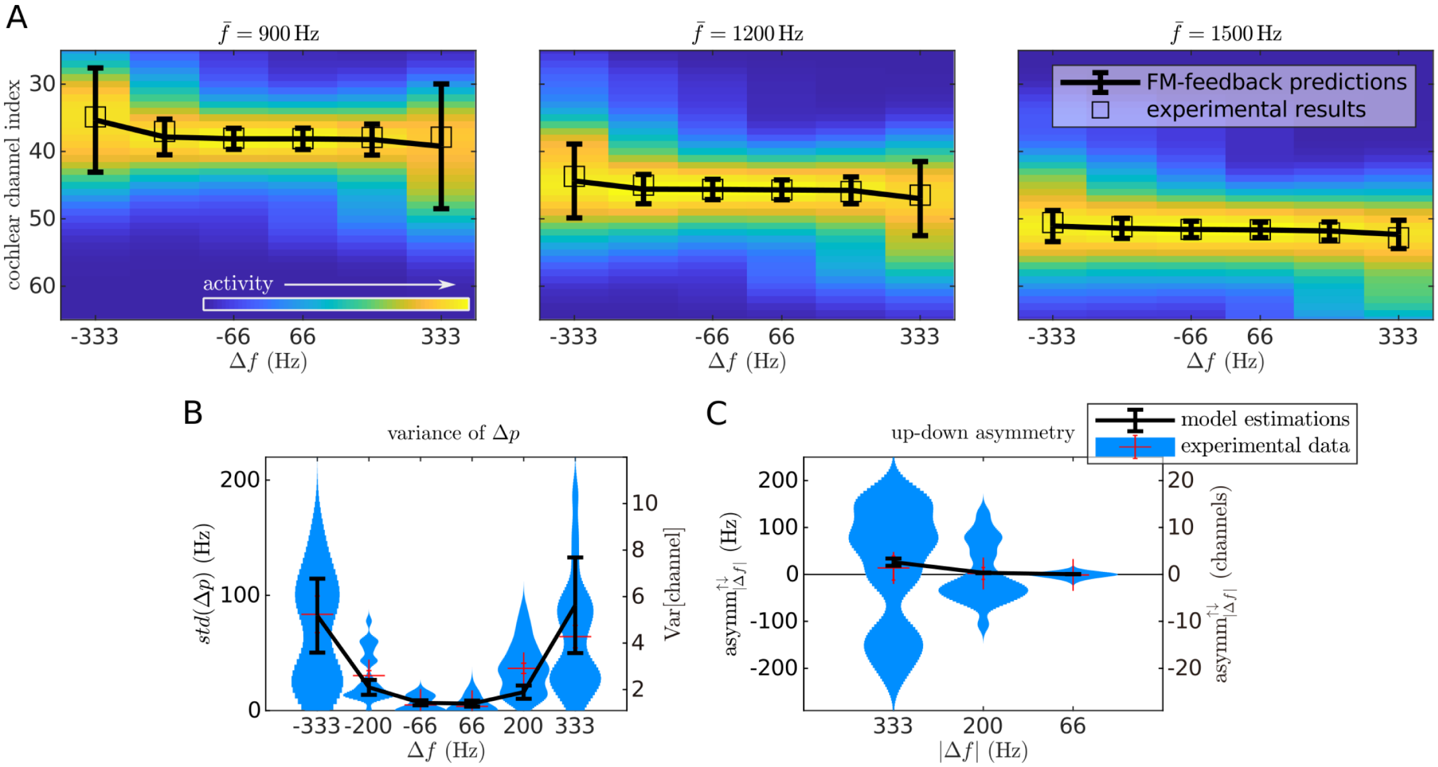

Fig 11. Predictions of the FM-feedback spectral model for sweep trains. A) Heatmaps show the distribution of the activation across the spectral representation (*y*-axis are cochlear channels) for different sweep frequency spans Δ*f* (*x*-axis). Error bars point to the expected value and standard deviation of the distribution across frequencies for each sweep. Each empty square denotes the expected channel elicited by a pure tone with the frequency of the average experimental data of the corresponding sweep. B) Predictions of the standard deviation of the perceived pitch. Error bars are estimations of standard error of the model calculated based on the dispersion of the centroids for different $\bar{f}$ and the standard deviations of the spectral distribution *ρ* of each condition. Kernel density estimations show the empirical subjects' standard deviation of the sweep pitch shift Δ*p* in sweep trains, plotted separately for the different frequency spans Δ*f*. Each sample in the distributions corresponds to the standard deviation of the perceived pitch of a sweep in one subject. The standard deviation is monotonically correlated to the absolute span |Δ*f*| ($r_s = 0.63$, $p < 10^{-27}$). C) Predictions of the standard deviation and up/down asymmetry. Error bars show the model predictions of the up/down asymmetry coefficient $\mathrm{asymm}^{\uparrow\downarrow}$ (see Eq (1)) in the channel space. Kernel density estimations show the empirical up/down asymmetry distributions in sweep trains, as defined in Eq (1). Each sample of the distributions corresponds to the difference of the average absolute deviation from centre frequency between up and down sweeps of the same |Δ*f*| for a given subject and centre frequency. Red crosses show the mean and the standard error of the data.

https://doi.org/10.1371/journal.pcbi.1008787.g011

these formant transitions are carried by harmonic complex tones rather than pure tones. It is currently unclear how the mechanisms introduced here could be used in natural settings to encode phonemes. One possibility is that sinusoidal sweeps are first decoded in the sweep layer and integrated in a later step of the ascending hierarchy. Since the populations encoding direction selectivity in the FM-feedback spectral model are tuned to specific best frequencies, the sweep layer is potentially able to encode simultaneous sweeps spanning complementary frequency ranges and represented by parallel harmonic series. Neural populations in the sweep layer could output to a third level of abstraction where combinations of concurrent sweeps across harmonic series are mapped into phonemes. Therefore, the FM-feedback spectral model could be the first basic building block towards a more comprehensive understanding of speech processing.

Bottom-up pitch models and pitch codes

The bottom-up integration of the spectral representation, cornerstone of the classical spectral or place theories of pitch [46], predicted a null sweep pitch shift. Other attempts of bottom-up models have also failed to parsimoniously explain the pitch shift: A previous phenomenological model suggested that a leaky integration of the instantaneous frequency could result in the ending segments of the sweeps having a stronger weight in the pitch decision, which would qualitatively explain the direction of the pitch sweep shift [21]. However, this model predicts the same pitch shift in single sweeps than sweep trains and the same absolute pitch shift in up and down sweeps, in direct contradiction with the empirical data.

Our simulations showed that current bottom-up modelling approaches based on a temporal code cannot even extract robust pitch information from most FM-sweeps used in the experiments. This is most likely a consequence of the fast change rate in the periodicities of fast FM stimuli. Typically, pitch decisions based on the auditory nerve temporal code are made after integrating over four cycles of the period of the stimuli [32, 47], coinciding with the duration threshold for accurate pitch discrimination [48]. However, our stimuli presented an average change of ~ 25 Hz across four repetitions of their average frequency, making this integration virtually impossible.

Another possibility is that a combination of the temporal and spectral codes is used to process the pitch of FM-sweeps, and that the sweep pitch shift emerges from this integration. Both, spectral and temporal representations of pitch play different roles in pitch processing, and it has been previously suggested that both codes could be added to perform pitch decisions (see [25] for a review). However, the temporal code had no usable pitch information at $|\Delta p| > 200$ Hz that could be integrated with the spectral code.

Although it is methodologically impossible to explore and anticipate the space of all possible models of pitch processing and more sophisticated bottom-up mechanisms might theoretically suffice to explain the sweep pitch shift in sweeps and sweep trains, the feedback modulation mechanism introduced in this study is, to date, the only available account of the experimental data.

Relation to predictive coding and hierarchical processing strategies

The presence of predictive feedback modulation in the subcortical sensory pathway has been shown before in humans [49, 50] and non-human mammals [51]. Previous studies often interpreted it in the context of the predictive coding framework [27, 28, 52], a theory of sensory processing that postulates that sensory information is encoded as prediction error; i.e., that neural activity at a given level of the processing hierarchy encodes the residuals of the sensory input with respect to a generative model encoded higher in the hierarchy.

The FM-feedback spectral model can also be understood in the light of this formalism: it presents three hierarchical layers of abstraction: the inputs from the peripheral system, the spectral layer, and the sweep layer. The top layer performs predictions on the sensory input incoming at the immediately lower representation of the hierarchy. However, unlike the classical predictive coding microcircuit where the generative model used to perform predictions and prediction error are kept in separate neural ensembles [53], the sweep network simultaneously holds a representation that is both, descriptive for its own representation and predictive for the immediately lower representation in the hierarchy.

Combining the generative model and stimulus representations in the same neural code solves some of the open questions of classical predictive coding architectures recently summarised by Denham and Winkler [54]: i) "what precisely is meant by prediction?", ii) "which generative models [within the hierarchy] make the predictions?", and iii) "what within the predictive framework is proposed to correlate with perceptual experience?". In the FM-feedback spectral model, the predictions can be summarised as the probability distribution of patterns of activation expected to come next in the lower level given what has been encoded so far in the higher level. These conditional probability distributions are encoded in the feedback connections stemming from the neurons holding the high-level representation and targeting the neurons holding the lower level representations. Such connectivity patterns would represent the statistics between the representations in the two levels, and could have arisen naturally during development after sufficient exposure to the stimuli. The perceptual experience in the FM-feedback spectral model is encoded in the activation along the two hierarchical stages, which encode different aspects of the stimuli.

Another key difference between the FM-feedback spectral model's architecture and the classical predictive coding microcircuit is that, rather than encoding the residuals of the spectral representation with respect to the FM-sweep representation, neurons in the spectral layer simply encode the spectral content of the stimulus. However, since the decoding of the predictable parts of the stimuli is faster, predictability potentially ensues a significant decrease on the amount of signal produced during the encoding. Such mechanism would explain why even expected stimuli, for which the residual should theoretically be zero, do still evoke measurable responses (e.g., [51]).

Comparison with previous measurements of the sweep pitch shift

Our experimental findings qualitatively replicated the sweep pitch shift found in previous studies; namely, we found that the pitch elicited by FM-sweeps was biased towards the frequencies spanned in the ending part of the sweeps [17], and that the perceptual bias is monotonically related to the frequency span Δf [18, 19]. On average, we estimated a putative linear relation between the pitch shift Δp and Δf of around $m \simeq 0.38$, slightly higher than Brady's [17] ($m \simeq 0.34$ with transitions of 50 ms) and Nabelek's [18] ($m \simeq 0.32$ with transitions of 40 ms) reports, and significantly higher than Rossi's [19] ($m = 1/6 \simeq 0.17$ with transitions of 200 ms) estimation. Since Rossi's transitions were 5 times longer than ours, the estimations are difficult to compare. However, the disagreement seems to indicate that the pitch shift is stronger with shorter durations. This observation would be fully compatible with the mechanism of predictive facilitation described in the FM-feedback spectral model: since the time to decode FM direction is independent of sweep duration, whilst only the most posterior part of the stimulus is facilitated in the short sweeps, in a long sweep the facilitation currents would affect a much larger portion of the sound, potentially including frequencies occurring even before $\bar{f}$.

In their original study Brady and colleagues [17] found that some sweeps of 20 ms duration elicited pitch values that coincided and even exceed the frequency span of the stimuli, especially in up sweeps. This perceptual extrapolation or *overshoot* has been replicated in two later studies [55, 56] and seems to further confirm the idea that the sweep pitch shift is driven by feedback between the sweep and spectral layer. However, the facilitation currents present in our model could not provoke activation of frequencies that are not initially present in the stimulus. One possibility is that the facilitation currents induce activation in neurons encoding frequencies that, although beyond the spectral range spanned by the sweep, are present in the stimulus spectrum. This scenario is unlikely in our data because we used frequency modulated sinusoids that elicit a unimodal distribution of responses in the auditory periphery. However, the three studies that showed the overshoot effect used spectrally rich stimuli. This would also explain why our model predicts smaller effect sizes than Brady's data in Fig 9. To clarify this point further work could either test if, as predicted by our model, sinusoidal sweeps do not produce an overshoot, and if an extension of our model able to handle spectrally rich stimuli does reproduce the overshoot effect shown in previous experimental data.

The FM-feedback spectral model also provides for a mechanistic interpretation of the two additional experiments reported in Brady's original study [17]. In Brady's experiment II, the transient duration of the sweep is kept constant but its onset is varied across the stimulus duration. When the transient is located near the beginning of the stimulus, the greatest part of the sounds excites neurons encoding frequencies close to the posterior parts of the transient pushing the distribution of the responses away from the average frequency towards the ending frequencies of the sweep f_1. This shift is larger than expected for a sound without a transient because of the feedback modulation of the later frequencies exerted by the sweep network. When the transient is located at the very end of the stimulus, the longer portion of the stimulus exciting f_0 compensates for the shift in the frequency distribution, bringing the perceived pitch closer to the starting frequencies of the stimulus.

In Brady's experiment III, the transient's onset is kept constant and it is the duration of the transient that is varied. The decreased sweep pitch shift observed for shorter in comparison to longer transient durations can be explained by the FM-feedback spectral model as a consequence of the stimuli presenting a larger segment with the initial frequency, thus shifting the distribution of the responses towards f_0.

FM encoding and physiological location of the sweep and spectral layers

The earliest neural centre within the auditory pathway showing FM direction selectivity in mammals is the inferior colliculus [3–6], although thalamic nuclei (medial geniculate body) [6, 7] and auditory cortex [4, 8–11] show generally stronger DSIs. Thus, the sweep layer postulated in the FM-feedback spectral model could be implemented even at early stages of the auditory hierarchy. Similarly, since all the nodes in the ascending auditory pathway contain tonotopically arranged nuclei, the spectral layer could be putatively located as early as in the cochlear nucleus. The physiological location of the mechanisms described here remains an open question.

Conclusion

In this work we have harnessed a well-established perceptual phenomenon to inform a model of FM direction encoding. We have shown that representative bottom-up models of pitch processing do not explain the pitch elicited by fast FM sweeps. Based on neurobiological considerations, we hypothesised that FM direction-selective neurons alter the way that spectral information is encoded via a feedback mechanism. We used the hypothesis to develop a model that proposes how this feedback modulation might be exerted and how it might affect the pitch percept elicited by FM sweeps. Although we cannot logically exclude other potential explanations for the effect, we provide evidence that our hypothesis is a likely and plausible mechanism underlying the encoding of formant transitions. These mechanisms could be part of a larger hierarchical network that transforms formant transitions into phonemes, phonemes into syllables, and syllables into words. Unravelling the fundamental building blocks of this hierarchy is a necessary prerequisite for a comprehensive understanding of the computational mechanisms underlying speech perception in the human brain.

Materials and methods

Ethics statement

The study was approved by the ethics committee of the University of Leipzig (ethics approval number 273/14-ff). All participants provided informed verbal consent.

Measuring the sweep pitch shift in single sweeps

Participants. Eight participants (4 female), aged 22 to 31 (average 26.9) years old, were included in the study. They all had normal hearing thresholds between 250 Hz and 8 kHz

(<25dB HL) according to pure tone audiometry (Micromate 304, Madsen Electronics). All reported at least five years of musical experience, but none of them was a professional musician. See the section Inclusion criteria bellow.

We considered a sample of $N = 8$ as sufficient for two reasons. First, the sweep pitch shift has been independently demonstrated in several previous studies [17–19]. The first experiment in our study is a replication of these previous studies that allowed us to quantify the sweep pitch shift. Second, low-level psychoacoustic phenomena are typically characterised by small inter-subject variabilities, so not many participants are necessary to demonstrate their generalisability [57]. We have selected 8 participants to ensure that that was the case, but the literature is populated with highly reproducible results that are inferred from experiments performed on populations as small as $N = 4$ [58]. Ours and previous data confirms that indeed the sweep pitch shift is present and shows the same direction and order of magnitude at the single subject level (S1 Fig; see also Tableau IV in [19] showing the effect in 18 subjects). Both experiments carried out in our study were taxing and extremely long, lasting for up to three hours per subject. Thus, while increasing the sample size would have not resulted in a stronger demonstration of the effect, it would have incurred an unjustified waste of resources.

Stimuli. Stimuli were 50 ms long frequency-modulated sweeps. Frequency was kept constant during the first and final 5 ms of the sweeps. The modulation was asymptotic (i.e., linear in the period $T = 1/f$ space) and carried out in 40 ms. Stimuli were ramped-in and ramped-out with 5 ms Hanning windows overlapping the sections with constant frequency.

There were 30 single sweeps with 10 linearly distributed frequency spans $\Delta f \in [-600, 600]$ Hz and 3 average frequencies $\bar{f} \in \{900, 1200, 1500\}$ Hz. For each sweep with a given Δf and $\bar{f}$, the initial and final frequencies were $f_0 = \bar{f} - \Delta f/2$ and $f_1 = \bar{f} + \Delta f/2$.

Sounds were delivered by over-ear headphones Sennheiser (Sennheiser electronic GmbH & Co. KG; Germany) HD201 connected to a Realtek (Realtek Semiconductor Corp.; Taiwan) ALC1150 soundcard. Participants were required to adjust the loudness of the stimuli to a comfortable level during the pure-tone-test (see Inclusion Criteria), so that they had a wide range of pure tones to use as reference. The experiment was carried out in a quiet room. Stimuli were produced and delivered by a custom-made MATLAB (MathWorks, Natik, USA) script. Scripts running the experiment and generating the sounds are freely available in github.com/qtabs/fmPitch/experiment.

Experimental design. Each trial consisted of a sequential presentation of a target sweep and a probe pure tone. After the presentation, the participant was asked whether the second sound evokes a higher, equal, or lower pitch percept than the first sound. Participants were allowed to replay the sounds as many times as needed in case of doubt. After the response, the software adjusted the frequency of the probe tone by steps of $\pm\epsilon = \pm 25$ Hz, bringing the pitch of the sound closer to the participant's percept (e.g., if the participant judged the target sweep as having a lower pitch than the probe tone, the frequency of the probe tone was reduced by 25 Hz). This procedure was repeated until the participant reported that the two sounds evoked the same pitch percept. Then, the frequency of the matched pure tone was stored as the perceived pitch of the sweep reported in that trial, and a new trial with a new target sweep began. The initial frequency of the probe tone was sampled from a Gaussian distribution centred on the average frequency $\bar{f}$ of the target sweep.

Each of the 30 sweeps was matched four times, so that there were 120 trials in total in the experiment. The relative order of the probe tone and the target sweep was reversed in half of the trials to assess if presentation order affects the sweep pitch shift. Thus, the experiment can be described as a 10 (10 different frequency spans) × 3 (3 average frequencies) × 2 (probe played first or last) factorial design.

Inclusion criteria. We initially recruited 22 participant candidates, which were screened by a first behavioural test assessing their capacity to match pure tones against pure tones (*pure-tone-test*), and then by a second behavioural test measuring their consistency when matching sweeps against pure tones (*sweep-test*). 14 of those 22 participants did not comply with the inclusion criteria: one was unable to match pure tones of the same frequency, 13 were unable to match sweeps against pure tones consistently. Consistency was assessed independently for each subject: i.e., the test did not evaluate whether the subject conformed to the results of other participants or if the subject showed a sweep pitch shift in any direction. The test only served to evaluate whether the subject approximately adjudicated the same pure tones to the same sweep, and was meant to exclude subjects that, due to poor pitch discrimination abilities or lack of motivation, were unable to perform the task.

The pure-tone-test was designed to ensure that participants had understood the task. We used the same experimental design as in the main experiment, including the same frequency step of 25 Hz, but both probe and target consisted of pure tones. During the pure-tone-test, the software provided feedback after each trial informing the participant whether the response was correct or incorrect. The pure-tone-test was divided in batches of six trials, and it concluded when the participant correctly matched the pitch of every trial in one batch. Most participants completed the pure-tone-test in the first batch; the participant excluded during the pure-tone-test failed to provide correct estimates for as many as six batches.

The *sweep-test* was used to evaluate whether participants could perform self-consistent judgements on the pitch of FM-sweeps. During the sweep-test, participants undertook a block of 12 trials consisting in 4 repetitions of the same 3 sweeps: $\{\Delta f = 67\,\mathrm{Hz}, \bar{f} = 900\,\mathrm{Hz}\}$, $\{\Delta f = -200\,\mathrm{Hz}, \bar{f} = 1200\,\mathrm{Hz}\}$, and $\{\Delta f = -67\,\mathrm{Hz}, \bar{f} = 1500\,\mathrm{Hz}\}$. We chose these sweeps to ensure diversity of the sweep properties while keeping $|\Delta f|$ small enough to ensure that the sweeps would elicit an unequivocal pitch percept according to Hart's law [59]. After the completion of this block, we scored the participant's pitch matching consistency as the inverse of the average of the absolute differences between the reported pitch in each sweep. Participants with an average standard deviation larger than twice the frequency increment step $2\epsilon = 50$ Hz were excluded from the experiment.

Since Hart's law [59] establishes that the sweeps used during the *sweep-test* elicit an unequivocal pitch percept, excluded participants were either unable or unwilling to perform consistent pitch judgements on sweeps. The inclusion of participant with inconsistent judgements would have contaminated the data with random guesses that could bias our estimations of the sweep pitch shift towards $\Delta p = 0$. Six of the excluded participants reported no previous musical experience; the remaining 8 had at least five years of musical training.

Experimental procedure. The 8 included participants matched the remaining 27 sweeps in four additional blocks. No sweep type was repeated within a single block, and all sweeps were presented 4 times across the entire experiment, resulting in 27 trials per block. The order of the sweeps within each block was randomised and the relative position of the probe tone with respect to the target stimulus was pseudorandomised so that in half of the trials in each block the probe tone was presented before the target sweep. Participants were instructed to take rests between blocks and were allowed to take as many shorter rests between trials as needed. To encourage precision, a 5€ award was offered to participants that kept their self-consistency along the main experiment with the same criterium as in the evaluation: a smaller standard error than $2\epsilon = 50$ Hz within each sweep type. Only sweeps expected to yield the most unequivocal pitch sensation according to Hart's law [59] (i.e., $|\Delta f| \leq 200$ Hz) were used to compute the overall self-consistency; participants were however unaware of this. Participants typically completed the experiment within 3 hours.

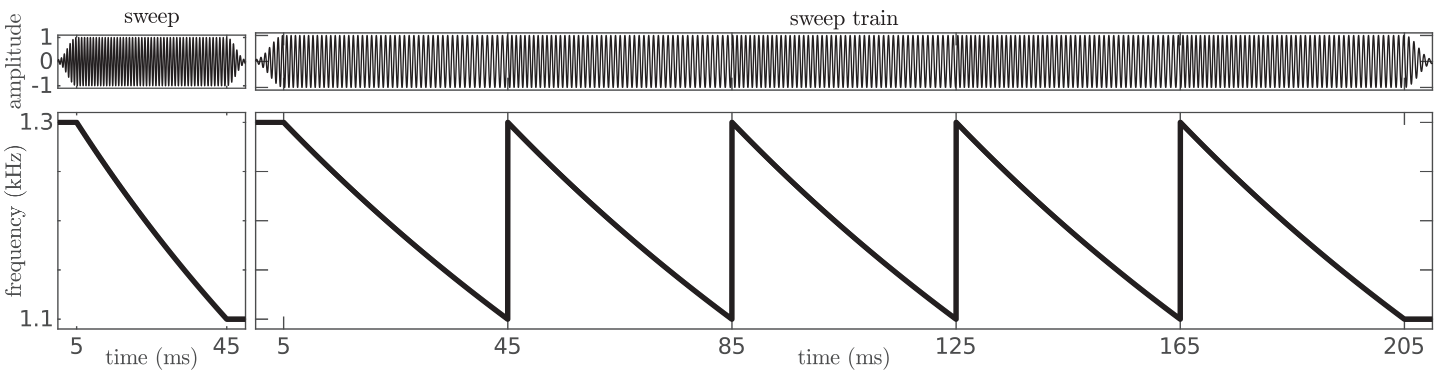

Fig 12. Examples of the stimuli. Waveform $s(t)$ and instantaneous frequency $f(t)$ of the sweep with $\bar{f} = 1200$ Hz and $\Delta f = -200$ Hz and its corresponding sweep train.

https://doi.org/10.1371/journal.pcbi.1008787.g012

Measuring the sweep pitch shift in sweep trains

Participants. The same 8 participants who completed the first experiment were invited to repeat the measurements with the new stimuli.

Stimuli. Stimuli were concatenations of 5 sweeps adding up to a total of 250 ms (sweep trains; see Fig 12). The sweeps were taken from a subset of 18 elements from the first experiment with 6 different frequency spans $\Delta f \in [-333, 333]$ Hz. To ensure continuity of the stimulus waveform, the sweeps were concatenated in the frequency domain (i.e., we computed the waveform of the stimuli by performing a Fourier transform over the concatenation of the time courses of the instantaneous frequencies). 5 ms Hanning windows were applied only at the very beginning and very end of the sweep trains.

Experimental design. The matching procedure was the same as in the first experiment: the participants matched the pitch of the sweep trains to probe pure tones whose frequency they could adjust with the aid of a computer software. To ensure that there were no effects of stimulus duration, the probe tones had the same duration as the sweep trains (i.e., 250 ms). As in the first experiment, each of the 18 sweep trains was matched four times, so that there were 72 trials in the second experiment. The relative order of the probe tone and the target sweep train was also reversed in half of the trials. Thus, the second experiment can be described as a 6 (different frequency spans) × 3 (average frequencies) × 2 (probe played first or last) factorial design.

Experimental procedure. Since the participants were already familiar with the task and proved to be able to match the pitch of FM-sweeps consistently, the experiment contained no pure-tone- or sweep- test. Four repetitions of the 18 sweep-trains were distributed across 5 blocks following the same principles as described in the first experiment. Participants typically completed the second experiment within 2 hours.

Bottom-up models of pitch

Spectral models of pitch processing. The responses at the auditory nerve were computed with a model of the peripheral auditory system [22, 60]. The model's output represents the expected firing rate $p_n(t)$ in a fibre of the auditory nerve associated with the nth cochlear channel ($n = 1, 2, \ldots, N$) at an instant t. The frequency range of the cochlear model was discretised in $N = 100$ channels, spanning frequencies from $f_{\text{min}} = 125$ Hz to $f_{\text{max}} = 10$ kHz.

The perceived pitch corresponded to the expected cochlear channel k, $E[k]$, according to a probability distribution ρ derived from the integral of $p_n(t)$ over the duration of the stimulus L:

$$E[k] = \sum_n n\rho_n \quad \text{with} \quad \rho_n = \frac{\int_0^L dt\, p_n(t)}{\sum_n \int_0^L dt\, p_n(t)} \tag{2}$$

To compare the predictions of the model with the experimental data, we also computed the expected channels $E[k]$ associated to pure tones with the frequency of the average perceived pitch of each sweep.

Temporal models of pitch processing. The SACF used in this work follows the original formulation by Meddis and O'Mard [23, 24]. Essentially, this model poses the existence of an array of M periodicity detectors responding more saliently to a preferred period δt_m. The instantaneous firing rate $A_m(t)$ of the mth periodicity detector ($m = 1, 2, \ldots, M$) follows:

$$\tau_m^{\mathrm{SACF}} \dot{A}_m(t) = -A_m(t) + \sum_n p_n(t)\, p_n(t - \delta t_m) \tag{3}$$

where the auditory nerve activity $p_n(t)$ in the cochlear channel n at an instant t is computed as in the previous section. The characteristic periods δt_m are uniformly distributed between $\delta t_m = 0.5$ ms and $\delta t_m = 30$ ms, which allows the model to capture periodicities corresponding to frequencies between 2 kHz and 135 Hz up to four lower harmonics. We kept a fixed integration constant $\tau_m^{\mathrm{SACF}} = 2.5$ ms; using variable τ_m^{SACF} that depend linearly on δt_m (see details in [31, 61]) did not result in substantial changes in our results.

Stimuli presenting periodicities at a certain frequency f typically elicit peaks of activation in the detectors tuned to the preferred period $\delta t_m = 1/f = T_0$ and to the periods corresponding to all subsequent lower harmonics $\delta t_m = 2T_0 = T_1$, $\delta t_m = 3T_0 = T_2$, etc. Thus, evidence for the period T at an instant t, $B(t)_T$ can be represented as the $B(t)_T = \sum_{m\in\{\mathcal{M}_T\}} A_m(t)$, where $\mathcal{M}_T$ are the indices of the periodicity detectors tuned to T, $2T$, $3T$, etc. (i.e., $\mathcal{M}_T = \{m : \delta t_m = nT \quad \forall\, n \in [1, 2, 3, \ldots]\}$). We estimated $B(t)_T$ using four harmonics; extending or reducing the number of harmonics used to estimate $B(t)_T$ did not significantly alter our results.

The perceived pitch corresponded to the expected period T, $E[T]$, according to a probability distribution ρ derived from the integral of $B_T(t)$ over the duration of the stimulus L:

$$E[T] = \sum_T T\rho_T \quad \text{with} \quad \rho_T = \frac{\int_0^L dt\, B_T(t)}{\sum_n \int_0^L dt\, B_T(t)} \tag{4}$$

To compare the predictions of the model with the experimental data, we computed the expected period $E[T]$ associated to pure tones with the frequency of the average perceived pitch of each sweep.

Details on the predictive model of FM encoding

Spectral layer and pitch estimations. The spectral layer consists on an array of $N = 100$ neural populations that integrate the output of the peripheral model. Neural populations are modelled according to a mean-field derivation [62] on linear integrate-and-fire neurons that, although first formulated to describe dynamics in cortical regions dedicated to visual decision making, has shown a great versatility approximating the dynamics of many different cortical areas (e.g., [63]). We decided to use a simple point model without physiological detail because we do not know exactly the location in the brain of the system we are modelling.

The firing rate $h_n(t)$ of the nth ensemble follows the dynamics of a leaky integrator:

$$\tau^{\text{pop}}\,\dot{h}^f_n(t) = -h^f_n(t) + \phi(I^f_n(t)) \tag{5}$$

where $\phi(x) = (cx - I_0)/(1 - e^{-g(cx - I_0)})$ is the transfer function of the mean-field model and τ^{pop} are adaptive time constants:

$$\tau^{\text{pop}}_{e,i}(h, I) = \tau^{\text{memb}}_{e,i}\,\Delta_T\,\frac{\partial_x\phi(x)|_{x=I}}{h} \tag{6}$$

Δ_T = 1mV is the size of the spike initialisation of the neural model and $\tau^{\text{memb}}_e = 20$ ms and $\tau^{\text{memb}}_i = 10$ ms [44] are the neural membrane time constants for excitatory and inhibitory populations, respectively. Using adaptive integration time constants makes the populations to react faster to changes when they are marginally active and have weak synaptic inputs, a behaviour often reported in tightly connected populations of neurons [42]. This component is the key of the feedback mechanism used to increase the responsiveness of the populations encoding the expected parts of the sweeps (Fig 6). The analytic formulation of $\tau^{\text{pop}}(h, I)$ stems from a theoretical study of networks of exponential-integrate-and-fire neurons [42].

Inputs $I^f_n(t)$ were modelled with AMPA synaptic dynamics [41]. AMPA synapses present short time constants that are able to preserve the fine temporal structure of auditory input, and thus are the major receptor type conveying bottom-up communication in the auditory pathway (e.g., [64]).

$$I^f_n(t) = J^{\text{AMPA}}_{\text{in}} \sum_k \omega^{\text{in}}_{nk} S^{\text{AMPA}}_{\text{in},k}(t) \tag{7}$$

where $J^{\text{AMPA}}_{\text{in}}$ is the effective synaptic efficacy of the peripheral input. We allowed some dispersion in the propagation from the peripheral model to the spectral layer by using a Gaussian-shaped connectivity matrix. This ensured that the bandwidth of the self-excitation in the spectral representation is independent of the number of cochlear channels:

$$\omega^{\text{in}}_{nm} = \frac{1}{\sqrt{\sigma_{\text{in}}}} e^{-\frac{(m-n)^2}{2\sigma^{\text{in}}}} \tag{8}$$

where the normalisation factor $\sqrt{\sigma_{\text{in}}}$ ensures that the total input to a population under a uniform peripheral input remains the same regardless of the dispersion σ_{in}. The synaptic gating variable $S^{\text{AMPA}}_{\text{in},n}(t)$ follows [41]:

$$\tau^{\text{AMPA}}\dot{S}^{\text{AMPA}}_{\text{in},n}(t) = -S^{\text{AMPA}}_{\text{in},n}(t) + p_n(t) \tag{9}$$

Note that we used the index ^{f} to denote variables in the spectral layer. The perceived pitch corresponded to the expected cochlear channel k, $E[k]$, according to a probability distribution ρ derived from the integral of $h^f_n(t)$ over the duration of the stimulus L (cf. Eq (2)):

$$E[k] = \sum_n n\rho_n \quad \text{with} \quad \rho_n = \frac{\int_0^L dt\, h^f_n(t)}{\sum_n \int_0^L dt\, h^f_n(t)} \tag{10}$$

The time constant τ^{AMPA} = 2 ms was taken from the literature [41]. The effective conductivity $J^{\text{AMPA}}_{\text{in}} = 0.38$ nA was manually tuned within the realistic range such that the peripheral system would elicit firing rates on the range 5Hz $\leq h_n(t) \leq$ 100 Hz in the integrator ensembles. The transfer function and its parameters, empirically derived for networks of integrate-and-fire neurons, were taken from [62].

Sweep layer and direction selectivity. We used delayed excitation [10, 26, 35] to model FM-direction selectivity. Two additional mechanisms for FM direction selectivity have been identified in IC, MGB and auditory cortex in the animal electrophysiology literature: asymmetric sideband inhibition [3, 35, 36], and duration sensitivity [6, 36, 37]. Although both delayed excitation and sideband inhibition contribute to direction selectivity in the mammal auditory pathway [3, 35, 36], the two mechanisms are often redundant and yield equivalent results when embedded in a neuronal model [14]. We chose to use delayed excitation alone for simplicity but, given that all models show similar direction and rate selectivity to FM-sweeps, replacing it by or adding any extra mechanism is unlikely to affect the model predictions.

The sweep layer consists of four arrays of $N = 100$ neural populations following the same dynamics described in the previous section (i.e., Eq (5)). From the four arrays, two (one excitatory, one inhibitory) are tuned to up sweeps, and two (again, one excitatory and one inhibitory) are tuned to down sweeps (Fig 3). The neural populations are characterised by the instantaneous firing rates $h_n^{\uparrow e}(t), h_n^{\uparrow i}(t)$ (up) and $h_n^{\downarrow e}(t), h_n^{\downarrow i}(t)$ $(down)$, and receive synaptic inputs $I_n^{\uparrow e}(t), I_n^{\uparrow i}(t)$ (up) and $I_n^{\downarrow e}(t), I_n^{\downarrow i}(t)$ $(down)$, respectively for excitatory and inhibitory populations. Although the transfer functions $\phi(x)$ are the same for all the ensembles, the parameters c, I_0, and g are different for excitatory and inhibitory populations [62].

Excitatory and inhibitory inputs to populations in the *sweep layer* are modelled according to AMPA-like and GABA-like synaptic gating dynamics [41]:

$$\begin{aligned}
\dot{S}_{\alpha,n}^{\mathrm{AMPA}}(t) &= -\frac{S_{\alpha,n}^{\mathrm{AMPA}}(t)}{\tau^{\mathrm{AMPA}}} + h_n^{\alpha e}(t) + \sigma\xi, \quad \alpha = \uparrow, \downarrow, f \\
\dot{S}_{\alpha,n}^{\mathrm{GABA}}(t) &= -\frac{S_{\alpha,n}^{\mathrm{GABA}}(t)}{\tau^{\mathrm{GABA}}} + h_n^{\alpha i}(t) + \sigma\xi, \quad \alpha = \uparrow, \downarrow
\end{aligned}$$

where ξ is an uncorrelated Gaussian noise sampled independently for each synapse and instant t, and $\sigma = 0.0007$ nA is the amplitude of the noise [62]. The total synaptic input for each population is then:

$$\begin{aligned}
I_n^{\uparrow e}(t) &= J_f^{\mathrm{AMPA}} \sum_m \omega_{nm}^{f\uparrow} S_{f,m}^{\mathrm{AMPA}}(t - \delta t_{nm}) - \\
&\quad J^{\mathrm{GABA}} \left(\sum_m \omega_{nm}^{ie} S_{\downarrow,m}^{\mathrm{GABA}}(t) + S_{\uparrow,n}^{\mathrm{GABA}}(t) \right) + I_{\mathrm{bkg}}^{E} \\
I_n^{\uparrow i}(t) &= J_s^{\mathrm{AMPA}} \sum_m \omega_{nm}^{ei} S_{\uparrow,m}^{\mathrm{AMPA}}(t) + I_{\mathrm{bkg}}^{I} \\
I_n^{\downarrow e}(t) &= J_f^{\mathrm{AMPA}} \sum_m \omega_{nm}^{f\downarrow} S_{f,m}^{\mathrm{AMPA}}(t - \delta t_{nm}) - \\
&\quad J^{\mathrm{GABA}} \left(\sum_m \omega_{nm}^{ie} S_{\uparrow,m}^{\mathrm{GABA}}(t) + S_{\downarrow,n}^{\mathrm{GABA}}(t) \right) + I_{\mathrm{bkg}}^{E} \\
I_n^{\downarrow i}(t) &= J_s^{\mathrm{AMPA}} \sum_m \omega_{nm}^{ei} S_{\downarrow,m}^{\mathrm{AMPA}}(t) + I_{\mathrm{bkg}}^{I}
\end{aligned}$$

where I_{bkg}^{E} and I_{bkg}^{I} are constant background inputs putatively sourced in external neural populations [62].

The excitatory-to-inhibitory and inhibitory-to-excitatory connectivity matrices ω^{ei} and ω^{ie} are Gaussian shaped and centred in the identity matrix:

$$\omega^{\alpha}_{nm} = e^{-\frac{(n-m)^2}{2\sigma_\alpha}}, \quad \alpha = ei, ie \tag{11}$$

The remaining connectivity matrices $\omega^{f\uparrow}$ and $\omega^{f\downarrow}$ are defined to constraint the up (down) feed to inputs from lower (higher) frequencies and to limit the range of the connection to a small number of populations $\Delta_{\omega f}$ of the spectral representation:

$$\begin{aligned}\omega^{f\uparrow}_{nm} &= \begin{cases} 1 & \text{if} \quad 0 \leq n - m \leq \Delta_{\omega f} \\ 0 & \text{otherwise} \end{cases} \\ \omega^{f\downarrow}_{nm} &= \begin{cases} 1 & \text{if} \quad 0 \leq m - n \leq \Delta_{\omega f} \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

The free parameters were initialised to standard values (the effective conductivities J_f^{AMPA}, J^{GABA}, and J_s^{AMPA}, according to [62]; the baseline delay δt_0 to 2 ms/channel; and the dispersion constants σ_{in}, σ_{ei}, σ_{ei}, and $\Delta_{\omega f}$, to $0.1N$) and manually tuned so that the networks showed direction selectivity for the FM-sweep characteristics (duration, rates, Δf) of the stimuli used in the first part of the study. Unless stated otherwise, all simulations listed in this work correspond to the parameters listed in Table 1.

The direction selectivity index (DSI; e.g., [11]) described in the Results section was computed as the proportion of the activity elicited in a network by an up sweep minus the activity elicited in the same network by a down sweep with the same duration and frequency span:

$$\text{DSI}^{\alpha} = \frac{\sum_n \int dt([h^{\alpha e}_n(t)]_{+\Delta f} - [h^{\alpha e}_n(t)]_{-\Delta f})}{\sum_n \int dt([h^{\alpha e}_n(t)]_{+\Delta f} + [h^{\alpha e}_n(t)]_{-\Delta f})} \quad \alpha = \uparrow, \downarrow \tag{12}$$

where $[h^{\alpha e}_n(t)]_{\Delta f}$ is the firing rate $h^{\alpha e}_n(t)$ elicited in the network by a sweep with a frequency span Δf.

Feedback connections. Feedback connections from the sweep layers to the spectral layer were modelled according to NMDA-like synaptic gating dynamics with a finite rising time constant [41].

$$\begin{aligned}\dot{S}^{\text{NMDA}}_{\alpha,n}(t) &= -\frac{S^{\text{NMDA}}_{\alpha,n}(t)}{\tau^{\text{NMDA}}} + \sigma\xi \\ &\quad + (1 - S^{\text{NMDA}}_{\alpha,n}(t))\gamma\, h^{\alpha e}_n(t), \quad \alpha = \uparrow, \downarrow\end{aligned}$$

with $\gamma = 0.641$. NMDA currents are added to the total synaptic input of the neurons in the spectral layer as an additional term in (7):

$$I^f_n(t) \rightarrow \hat{I}^f_n(t) = I^f_n(t) + J^{\text{NMDA}} \sum_{\alpha=\uparrow,\downarrow} \sum_m \omega^{\alpha f}_{nm} S^{\text{NMDA}}_{\alpha,m}(t)$$

The connectivity matrices $\omega^{\alpha\uparrow}_{nm}$, $\omega^{\alpha\downarrow}_{nm}$ were chosen such that the target of the NMDA-driven activation was limited to a number of $\Delta_{\omega s}$ and leave a gap of $w_{\omega s}$ populations between the

Table 1. Model parameters. Most parameters were taken from the original studies that derived the mean field approximations used in the model and are cited accordingly. Other free parameters, like the number of bins of the tonotopic axis *N*, were fixed to reasonable but arbitrary values at the beginning of the model construction and were not adjusted during the analyses (*ad-hoc*). Free parameters that were manually tuned are labelled as *tuned (x)*, where *x* is: 1, for parameters tuned so that the spectral layer integrates the peripheral representation correctly; 2, for parameters tuned to achieve FM-direction selectivity; and 3, for parameters tuned so that the feedback signalling resulted in a fair fit between the model's pitch predictions and the experimental observations. Short description of the parameters in the last column are further explained along the Methods section. Connectivity parameters are more eloquently described in Fig 13.

parameter	value (unit)	source	description
N	100 channels	ad-hoc	number of cochlear channels
dt	0.1 ms	ad-hoc	time step for the numerical integration
periph dt	0.01 ms	ad-hoc	*dt* in the peripheral model
periph $f_{\min}$	125 Hz	[60]	best frequency of the first cochlear channel
periph $f_{\max}$	10000 Hz	[60]	best frequency of the last cochlear channel
τ_e^{memb}	20 ms	[44]	membrane time constant of excitatory neurons
τ_i^{memb}	10 ms	[44]	membrane time constant of inhibitory neurons
Δ	1 mV	[42]	action potential initiation sharpness
$c^{\text{excitatory}}$	310 $(\text{V nC})^{-1}$	[62]	transfer function parameters (excitatory populations)
$I_0^{\text{excitatory}}$	125 Hz	[62]	transfer function parameters (excitatory populations)
$g^{\text{excitatory}}$	0.16 s	[62]	transfer function parameters (excitatory populations)
$c^{\text{inhibitory}}$	615 $(\text{V nC})^{-1}$	[62]	transfer function parameters (inhibitory populations)
$I_0^{\text{inhibitory}}$	177 Hz	[62]	transfer function parameters (inhibitory populations)
$g^{\text{inhibitory}}$	0.087 s	[62]	transfer function parameters (inhibitory populations)
I_{bkg}^{E}	0.23 nA	[62]	baseline input current (excitatory populations)
I_{bkg}^{I}	0.10 nA	[62]	baseline input current (inhibitory populations)
σ	0.0007 nA	[62]	synaptic noise amplitude
γ	0.641	[41]	NMDA coupling factor
τ^{AMPA}	2 ms	[41]	decay time constant of AMPA synapses
τ^{GABA}	5 ms	[41]	decay time constant of GABA synapses
τ^{NMDA}	100 ms	[41]	decay time constant of NMDA synapses
$J_{\text{in}}^{\text{AMPA}}$	0.38 nC	tuned (1)	AMPA conductivity of the peripheral input
J_f^{AMPA}	0.55 nC	tuned (2)	AMPA conductivity of the spectral populations
J_s^{AMPA}	0.67 nC	tuned (2)	AMPA conductivity of the sweep populations
J^{GABA}	0.30 nC	tuned (2)	GABA conductivity (sweep populations only)
J^{NMDA}	0.05 nC	tuned (3)	NMDA conductivity (feedback connections only)
σ_{in}	0.1*N* channels	tuned (2)	dispersion in peripheral input to spectral populations
σ_{ie}	0.5*N* channels	tuned (2)	dispersion in inhibitory-to-excitatory connections
σ_{ei}	0.03*N* channels	tuned (2)	dispersion in excitatory-to-inhibitory connections
Δt_0	1 ms/channel	tuned (2)	baseline delay of the delayed-excitation mechanism
$\Delta_{\omega f}$	0.05*N* channels	tuned (2)	connection width (spectral to sweep populations)
$\Delta_{\omega s}$	0.05*N* channels	tuned (3)	connection width of the feedback (sweep to spectral)
$w_{\omega s}$	0.05*N* channels	tuned (3)	connection gap of the feedback (sweep to spectral)

https://doi.org/10.1371/journal.pcbi.1008787.t001

centre frequency of the source and target ensembles (see Fig 13, right):

$$\omega_{nm}^{\uparrow f} = \begin{cases} 1 & \text{if} \quad w_{\omega s} \leq m - n \leq \Delta_{\omega s} + w_{\omega s} \\ 0 & \text{otherwise} \end{cases}$$

$$\omega_{nm}^{\downarrow f} = \begin{cases} 1 & \text{if} \quad w_{\omega s} \leq n - m \leq \Delta_{\omega s} + w_{\omega s} \\ 0 & \text{otherwise} \end{cases}$$

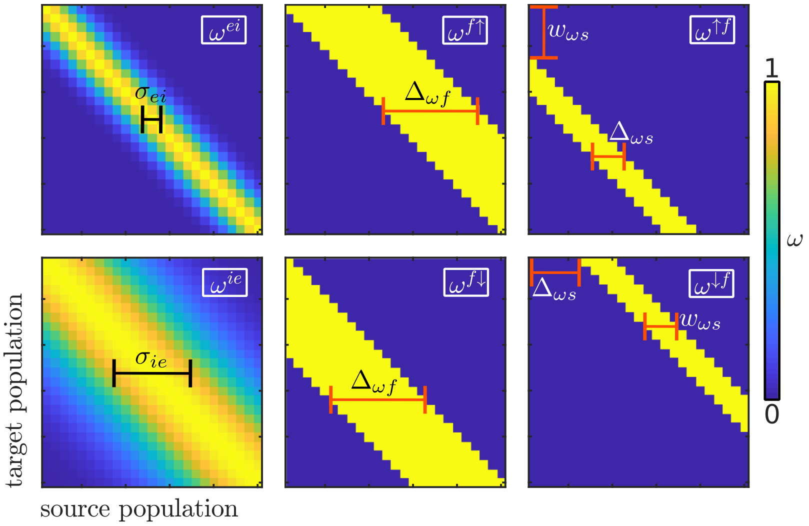

Fig 13. Connectivity matrices. Matrices show the connection between the first 25 ensembles of each source-target group. From left to right, matrices correspond to: excitatory-to-inhibitory ω^{ei}, inhibitory-to-excitatory ω^{ie}; bottom-up AMPA connections spectral-to-up $\omega^{f\uparrow}$, spectral-to-down $\omega^{f\uparrow}$; and feedback NMDA connections up-to-spectral $\omega^{\uparrow f}$, down-to-spectral $\omega^{\downarrow}$ f. Labels are encircled in a white square in the top right of each plot. The free parameters of each connectivity matrix are defined geometrically in the plots.

https://doi.org/10.1371/journal.pcbi.1008787.g013

The gap $w_{\omega s} > 0$ is enforced to avoid resonances between sweep-selective and spectral populations with the same centre frequency during the encoding of pure tones. The free parameters were initialised to standard values (the NMDA conductivity J^{NMDA} to the value recommended by [62], and the connectivity parameters $w_{\omega s}$ and $\Delta_{\omega s}$ to $0.1N$) and manually tuned so that the pitch predictions of the model (as computed in Eq (10)) matched the empirical data.

Supporting information

S1 Fig. Subject specific estimations of the linear fits between the pitch shift Δ*p* and Δ*f*. Each plot shows the slopes m of the linear fit $f_{\text{perceived}} \sim \bar{f} + m\,\Delta f$ for each subjects, for the single sweep (left) and sweep train (right) stimuli; error bars mark the 95% confidence intervals of the estimations.
(EPS)

S2 Fig. Effect of the presentation order on Δ*p*. Kernel density estimations of the difference between the perceived pitch evaluated when the sweep was presented before the probe tone $f^{\leftarrow}_{\text{perceived}}$ and the perceived pitch evaluated when the probe tone was presented before the sweep $f^{\rightarrow}_{\text{perceived}}$; no systematic effect of the presentation order was found for any of the conditions. Each sample of the distributions corresponds to the difference of the average perceived pitch between presentation orders of the same Δf for a given subject and centre frequency ($N = 8 \times 3 = 24$). Error bars show the average and the standard error of the groups. Average difference across Δf was 145Hz ± 227Hz, largely overlapping 0.
(EPS)

S1 Table. Summary statistics on the relationship between the perceived pitch and the frequency span for single sweeps and sweep trains. The slope of the linear fit, Pearson's correlation r_p, and Spearman's correlation r_s for the relationship between $f_{\text{perceived}}$ and Δf are

presented for each centre frequency $\bar{f}$ and direction of the presentation (probe before sweep, →; and sweep before probe, ←). Spearman's correlation is systematically larger than Pearson's, indicating that the elicited pitch is related to Δf in a non-linear monotonic way.
(TEX)

S1 Sounds. Stimuli used in the experiments. Each waveform corresponds to each of the single-sweeps and sweep-trains used in the first and second experiment, respectively.
File names indicate the properties of the stimulus as follows: `[sweep/train]_fbar<`$\bar{f}$`>_delta<`Δf`>.wav`; e.g., `train_fbar1200Hz_delta-333Hz.wav` is the sweep-train with $\bar{f} = 1200$ Hz and $\Delta f = -333$Hz used in the second experiment.
(ZIP)

Acknowledgments

The authors would like to thank Shih-Cheng *Vincent* Chien for his enlightening suggestions during the writing of the manuscript.

Author Contributions

Conceptualization: Alejandro Tabas, Katharina von Kriegstein.

Data curation: Alejandro Tabas.

Formal analysis: Alejandro Tabas.

Funding acquisition: Katharina von Kriegstein.

Investigation: Alejandro Tabas.

Methodology: Alejandro Tabas.

Software: Alejandro Tabas.

Supervision: Katharina von Kriegstein.

Validation: Alejandro Tabas.

Visualization: Alejandro Tabas.

Writing – original draft: Alejandro Tabas, Katharina von Kriegstein.

Writing – review & editing: Alejandro Tabas, Katharina von Kriegstein.

References

1. Kent RD, Kim Y. Acoustic Analysis of Speech. In: Ball MJ, Perkins MR, Mller N, Howard S, editors. The Handbook of Clinical Linguistics. Oxford, UK: Blackwell Publishing Ltd.; 2008. p. 360–380.
2. Hu B. Functional organization of lemniscal and nonlemniscal auditory thalamus. Experimental Brain Research. 2003; 153(4):543–549. https://doi.org/10.1007/s00221-003-1611-5
3. Geis HRAP, Borst JGG. Intracellular responses to frequency modulated tones in the dorsal cortex of the mouse inferior colliculus. Frontiers in Neural Circuits. 2013; 7(4):2002–2016.
4. Li AA, Zhang AY, Chen QC, Wu FJ. Effects of modulation range and presentation rate of FM stimulus on auditory response properties of mouse inferior collicular neurons. Sheng li xue bao: [Acta physiologica Sinica]. 2010; 62(3):210–8.
5. Hage SR, Ehret G. Mapping responses to frequency sweeps and tones in the inferior colliculus of house mice. European Journal of Neuroscience. 2003; 18(8):2301–2312. https://doi.org/10.1046/j.1460-9568.2003.02945.x
6. Kuo RI, Wu GK. The Generation of Direction Selectivity in the Auditory System. Neuron. 2012; 73(5):1016–1027. https://doi.org/10.1016/j.neuron.2011.11.035

7. Lui B, Mendelson JR. Frequency modulated sweep responses in the medial geniculate nucleus. Experimental Brain Research. 2003; 153(4):550–553. https://doi.org/10.1007/s00221-003-1618-y

8. Issa JB, Haeffele BD, Young ED, Yue DT. Multiscale mapping of frequency sweep rate in mouse auditory cortex. Hearing Research. 2016; 344:207–222. https://doi.org/10.1016/j.heares.2016.11.018

9. Trujillo M, Carrasco MM, Razak K. Response properties underlying selectivity for the rate of frequency modulated sweeps in the auditory cortex of the mouse. Hearing Research. 2013; 298:80–92. https://doi.org/10.1016/j.heares.2012.12.013

10. Ye Cq, Poo Mm, Dan Y, Zhang Xh. Synaptic mechanisms of direction selectivity in primary auditory cortex. Journal of Neuroscience. 2010; 30(5):1861–1868. https://doi.org/10.1523/JNEUROSCI.3088-09.2010

11. Zhang LI, Tan AYY, Schreiner CE, Merzenich MM. Topography and synaptic shaping of direction selectivity in primary auditory cortex. Nature. 2003; 424(6945):201–205. https://doi.org/10.1038/nature01796

12. Schofield BR. Central Descending Auditory Pathways. In: Ryugo D, Fay R, editors. Auditory and Vestibular Efferents. Springer Handbook of Auditory Research; 2011. p. 261–290.

13. Winer JA, Schreiner CE. The Inferior Colliculus. 1st ed. New York: Springer; 2005.

14. Skorheim S, Razak K, Bazhenov M. Network models of frequency modulated sweep detection. PLoS ONE. 2014; 9(12):1–25. https://doi.org/10.1371/journal.pone.0115196

15. Moore RC. Integration of speech with natural language understanding. Proceedings of the National Academy of Sciences of the United States of America. 1995; 92(22):9983–9988. https://doi.org/10.1073/pnas.92.22.9983

16. Chambers C, Akram S, Adam V, Pelofi C, Sahani M, Shamma S, et al. Prior context in audition informs binding and shapes simple features. Nature Communications. 2017; 8. https://doi.org/10.1038/ncomms15027 PMID: 28425433

17. Brady PT, House AS, Stevens KN. Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency. Journal of the Acoustical Society of America. 1961; 33(10):1357–1362. https://doi.org/10.1121/1.1908439

18. Nabelek I, Nabelek A, Hirsh IJ. Pitch of Short Tone Bursts of Changing Frequency. The Journal of the Acoustical Society of America. 1970; 45(1):293–293. https://doi.org/10.1121/1.1970857

19. Rossi M. La perception des glissandos descendants dans les contours prosodiques. Phonetica. 1978; 35(1):11–40. https://doi.org/10.1159/000259920

20. D'Alessandro C, Castellengo M. The pitch of short-duration vibrato tones. Journal of the Acoustical Society of America. 1994; 95(3):1617–1630. https://doi.org/10.1121/1.408548

21. D'Alessandro C, Rosset S, Rossi JP. The pitch of short-duration fundamental frequency glissandos. The Journal of the Acoustical Society of America. 1998; 104(4):2339–48. https://doi.org/10.1121/1.423745

22. Zilany MSa, Bruce IC, Carney LH. Updated parameters and expanded simulation options for a model of the auditory periphery. The Journal of the Acoustical Society of America. 2014; 135(1):283–286. https://doi.org/10.1121/1.4837815

23. Meddis R, O'Mard L. A unitary model of pitch perception. The Journal of the Acoustical Society of America. 1997; 102(3):1811–1820. https://doi.org/10.1121/1.420088

24. Meddis R, O'Mard LP. Virtual pitch in a computational physiological model. The Journal of the Acoustical Society of America. 2006; 120(6):3861. https://doi.org/10.1121/1.2372595

25. de Cheveigné A. Pitch Perception Models. In: Plack CJ, Fay RR, Oxenham AJ, Popper AN, editors. Pitch: Neural Coding and Perception. Springer New York; 2005. p. 169–233.

26. Razak KA, Fuzessery ZM. Facilitatory Mechanisms Underlying Selectivity for the Direction and Rate of Frequency Modulated Sweeps in the Auditory Cortex. Journal of Neuroscience. 2008; 28(39):9806–9816. https://doi.org/10.1523/JNEUROSCI.1293-08.2008

27. Mumford D. On the computational architecture of the neocortex II: The role o f cortico-cortical loops. Biological Cybernetics. 1992; 66(3):241–251. https://doi.org/10.1007/BF00198477

28. Friston K. A theory of cortical responses. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2005; 360(1456):815–36. https://doi.org/10.1098/rstb.2005.1622

29. Oxenham AJ. Revisiting place and temporal theories of pitch. Acoustical Science and Technology. 2013; 34(6):388–396. https://doi.org/10.1250/ast.34.388

30. Erfanian Saeedi N, Blamey PJ, Burkitt AN, Grayden DB. An integrated model of pitch perception incorporating place and temporal pitch codes with application to cochlear implant research. Hearing Research. 2017; 344:135–147. https://doi.org/10.1016/j.heares.2016.11.005

31. Balaguer-Ballester E, Denham SL, Meddis R. A cascade autocorrelation model of pitch perception. The Journal of the Acoustical Society of America. 2008; 124(4):2186–95. https://doi.org/10.1121/1.2967829

32. Wiegrebe L. Searching for the time constant of neural pitch extraction. The Journal of the Acoustical Society of America. 2001; 109(3):1082–1091. https://doi.org/10.1121/1.1348005

33. Pollak GD, Gittelman JX, Li N, Xie R. Inhibitory projections from the ventral nucleus of the lateral lemniscus and superior paraolivary nucleus create directional selectivity of frequency modulations in the inferior colliculus: a comparison of bats with other mammals. Hearing research. 2011; 273(1-2):134–44. https://doi.org/10.1016/j.heares.2010.03.083

34. Gittelman JX, Li N. FM velocity selectivity in the inferior colliculus is inherited from velocity-selective inputs and enhanced by spike threshold. Journal of Neurophysiology. 2011; 106(5):2399–2414. https://doi.org/10.1152/jn.00250.2011

35. Fuzessery ZM, Razak KA, Williams AJ. Multiple mechanisms shape selectivity for FM sweep rate and direction in the pallid bat inferior colliculus and auditory cortex. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology. 2011; 197(5):615–623. https://doi.org/10.1007/s00359-010-0554-0

36. Williams AJ, Fuzessery ZM. Multiple mechanisms shape FM sweep rate selectivity: complementary or redundant? Frontiers in Neural Circuits. 2012; 6(August):1–14.

37. Morrison JA, Valdizón-Rodríguez R, Goldreich D, Faure PA. Tuning for rate and duration of frequency-modulated sweeps in the mammalian inferior colliculus. Journal of Neurophysiology. 2018; 120(3):985–997. https://doi.org/10.1152/jn.00065.2018

38. Friston KJ, Price CJ. Dynamic representations and generative models of brain function. Brain Research Bulletin. 2001; 54(3):275–285. https://doi.org/10.1016/S0361-9230(00)00436-6

39. Salin PA, Bullier J. Corticocortical Connections in the Visual System: Structure and Function. Physiological reviews. 1995; 75(1):107–154. https://doi.org/10.1152/physrev.1995.75.1.107

40. Rauschecker JP. Cortical control of the thalamus: top-down processing and plasticity. Nature Neuroscience. 1998; 1(3):179–180. https://doi.org/10.1038/625

41. Brunel N, Wang XJ. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. Journal of Computational Neuroscience. 2001; 11(1):63–85. https://doi.org/10.1023/A:1011204814320

42. Ostojic S, Brunel N. From spiking neuron models to linear-nonlinear models. PLoS Computational Biology. 2011; 7(1):e1001056. https://doi.org/10.1371/journal.pcbi.1001056

43. Shore SE, Nuttall AL. High-synchrony cochlear compound action potentials evoked by rising frequency-swept tone bursts. Journal of the Acoustical Society of America. 1985; 78(4):1286–1295. https://doi.org/10.1121/1.392898

44. McCormick DA, Connors BW, Lighthall JW, Prince DA. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. Journal of Neurophysiology. 1985; 54(4):782–806. https://doi.org/10.1152/jn.1985.54.4.782

45. Uppenkamp S, Fobel S, Patterson R. The effects of temporal asymmetry on the detection and perception of short chirps. Hearing research. 2001; 158:71–83. https://doi.org/10.1016/S0378-5955(01)00299-4

46. von Helmholtz HLF. On the Sensations of Tone. Dover Publications; 1954.

47. Tabas A, Andermann M, Schuberth V, Riedel H, Balaguer-Ballester E, Rupp A. Modeling and MEG evidence of early consonance processing in auditory cortex. PLoS Computational Biology. 2019; 15(2):1–28. https://doi.org/10.1371/journal.pcbi.1006820

48. Krumbholz K, Patterson RD, Seither-Preisler A, Lammertmann C, Lütkenhöner B. Neuromagnetic evidence for a pitch processing center in Heschl's gyrus. Cerebral Cortex. 2003; 13(7):765–772. https://doi.org/10.1093/cercor/13.7.765

49. Suga N. Tuning shifts of the auditory system by corticocortical and corticofugal projections and conditioning. Neuroscience and Biobehavioral Reviews. 2012; 36(2):969–988. https://doi.org/10.1016/j.neubiorev.2011.11.006

50. Tabas A, Mihai G, Kiebel S, Trampel R, Von Kriegstein K. Abstract rules drive adaptation in the subcortical sensory pathway. eLife. 2020; 9:1–19. https://doi.org/10.7554/eLife.64501

51. Malmierca MS, Anderson LA, Antunes FM. The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: a potential neuronal correlate for predictive coding. Frontiers in systems neuroscience. 2015; 9(March):19.

52. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 1999; 2(1):79–87. https://doi.org/10.1038/4580

53. Bastos AM, Usrey WM, Adams Ra, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. Neuron. 2012; 76(4):695–711. https://doi.org/10.1016/j.neuron.2012.10.038

54. Denham SL, Winkler I. Predictive coding in auditory perception: challenges and unresolved questions. European Journal of Neuroscience. 2018.

55. Nabelek AK, Czyzewski Z, Crowley H. Cues for perception of the diphthong /ai/ in either noise or reverberation. Part I. Duration of the transition. The Journal of the Acoustical Society of America. 1994; 95(5):2681–2693. https://doi.org/10.1121/1.409837

56. Divenyi P. Perception of complete and incomplete formant transitions in vowels. The Journal of the Acoustical Society of America. 2009; 126(3):1427–1439. https://doi.org/10.1121/1.3167482

57. Levitin DJ. Experimental Design in Psychological Research. In: Cook PR, editor. Music, Cognition, and Computerized Sound. The MIT Press; 1999. p. 299–328.

58. Oxenham AJ, Micheyl C, Keebler MV, Loper A, Santurette S. Pitch perception beyond the traditional existence region of pitch. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(18):7629–34. https://doi.org/10.1073/pnas.1015291108

59. Hart Jt, Collier R, Cohen A. A Perceptual Study of Intonation. Cambridge University Press; 1990.

60. Zilany MSA, Bruce IC, Nelson PC, Carney LH. 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. 2009; 126(5):2390–2412. https://doi.org/10.1121/1.3238250

61. Wiegrebe L, Meddis R. The Representation of Periodic Sounds in Simulated Sustained Chopper Units of the Ventral Cochlear Nucleus. The Journal of the Acoustical Society of America. 2004; 115(3):1207–1218. https://doi.org/10.1121/1.1643359

62. Wong KF, Wang XJ. A recurrent network mechanism of time integration in perceptual decisions. The Journal of Neuroscience. 2006; 26(4):1314–1328. https://doi.org/10.1523/JNEUROSCI.3733-05.2006

63. Deco G, Ponce-Alvarez A, Mantini D, Romani GL, Hagmann P, Corbetta M. Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. The Journal of neuroscience: the official journal of the Society for Neuroscience. 2013; 33(27):11239–52. https://doi.org/10.1523/JNEUROSCI.1091-13.2013

64. Golding NL. Neuronal Response Properties and Voltage-Gated Ion Channels in the Auditory System. In: Trussell LO, Popper AN, Fay RR, editors. Springer Handbook of Auditory Research. vol. 41 of Springer Handbook of Auditory Research. New York, NY: Springer New York; 2012. p. 7–41.