Neural coding of fast frequency modulated sweeps

Alejandro Tabas[12] and Katharina von Kriegstein[2]

[1]Chair of Cognitive and Clinical Neuroscience, Faculty of Psychology, Technische Universität Dresden, 01062 Dresden, Saxony, Germany

[2]Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstr 1a, 04107 Leipzig, Saxony, Germany

Abstract

Frequency modulation (FM) is a basic constituent of vocalisation. Formant transitions in speech are characterised by short rising and falling FM-sweeps in the kilohertz frequency range. These sounds elicit a pitch percept that deviates from their average frequency. This study uses this perceptual effect, termed here the sweep pitch shift, to inform a model characterising the neural encoding of FM. First, a reexamination of the classical effect, consisting of two perceptual experiments, provides a quantitative characterisation of the dependence of the sweep pitch shift with the properties of the sweeps. Next, simulations carried on the new experimental data show that classical temporal and spectral models of pitch processing cannot explain the pitch shift. Conversely, a modified spectral model considering a predictive interaction between frequency and FM encoding fully reproduces our and previous experimental data. The model introduces a feedback mechanism that modulates the neurons that are expected to respond to future portions of the sweeps, accelarating their onset response. Combined, the experimental and modelling results suggest that predictive feedback modulation plays an important role in the neural encoding of FM even at early stages of the processing hierarchy.

1 Introduction

Frequency modulation (FM) is a basic acoustic feature of music, animal vocalisation, and speech. For example, consonants preceding and following a vowel can be acoustically characterised by formant transitions: a series of simultaneous fast FM tones of around 50 ms duration that start or finish in the frequencies characterising the vowel [1]. Despite the importance of FM sounds for perception and communication, a mechanistic account of FM encoding in humans is to-date unavailable.

In this study, we use a classical effect from psychoacoustics to inform a computational model of FM-sweep encoding. The effect, that we call here *sweep pitch shift*, was first reported by Brady and colleagues [2]. They measured the pitch elicited by fast rising and falling FM sweeps with a rich spectral contour, and discovered a tendency to judge up sweeps as eliciting a higher pitch than down sweeps with the same average fundamental frequency. These findings were later replicated in further experiments [3, 4]. More recently, d'Alessandro and colleagues proposed that the pitch of a sweep corresponds to a weighted average of the sweep fundamental frequency across time, where later frequencies receive a stronger weight due to a fixed-size temporal integration window [5, 6]. D'Alessandro and colleagues used their phenomenological model to assess the pitch elicited by up sweeps, down sweeps, and vibrato tones; however, they found that different integration weights were necessary to explain different partitions of their data. Thus, despite the crucial role that FM encoding plays in speech perception, the mechanisms responsible for the sweep pitch shift are still unknown. Whether classical models of pitch processing (see [7] for a review) can explain the sweep pitch shift has so far not been tested.

The aim of this study is to develop a biophysically plausible model of FM-sweep encoding. We consider the sweep pitch shift phenomenology to inform the mechanisms responsible for the interplay between FM neural selectivity and frequency representation in the processing hierarchy of the model. We approach this aim in three parts. In the first part, we reexamine and quantify the sweep pitch shift and test if the experimental data can be explained by three existing models: (i) a classical spectral model of pitch [8], (ii) a family of classical temporal models of pitch [9, 10, 11], and (iii) D'Alessandro's phenomenological model [5, 6]. A key feature of the classical models of pitch is that they are bottom-up driven; i.e., they only use sensory information from lower representations of the sound to process pitch. However, the human auditory pathway is endowed with massive feedback connections that have complex repercussions on the way sounds are processed by, for instance, modulating the properties of the receptive fields [12, 13]. Thus it is likely that a computational model of FM-sweep encoding can only fully explain perceptual data if it includes feedback modulation.

The second part of this work is committed to build a hierarchical model of frequency and FM sweep direction processing to test this hypothesis. Neural encoding of frequency and FM has been extensively studied in the mammal auditory system. Frequency is spatially represented along the tonotopic axis in all the stations of the auditory pathway [14]. Neural selectivity to FM direction and rate has also been repeatedly reported in rats and mice in the inferior coliculus [15, 16, 17], medial geniculate body [18, 19], and auditory cortex [20, 21, 22, 23]. The bottom-up components of the model are based on the results of these studies in animals. The top-down architecture is grounded in the basis of generative hierarchical models and predictive coding [24, 25] and informed by the human psychophysics results from the first part of the study.

In the third and last part of this work, a new set of stimuli termed *sweep trains* are used to further validate the model. These stimuli, consisting of a concatenation of five sweeps, preserve the same acoustical features of the original sweeps but elicit 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 this new stimuli illustrates the generalisation power of the neural mechanisms proposed in this work.

2 The sweep pitch shift revisited

2.1 Experimental methods: bottom-up models of pitch

2.1.1 Participants

8 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 (< 25 dB SPL) according to pure tone audiometry (Micromate 304, Madsen Electronics). All reported at least five years of musical training, but none of them was a professional musician.

The 8 participants were derived from a larger set of 22 candidates. Candidates were screened by a first behavioural test assessing their capacity to match pure tones against pure tones, and then by a second test measuring their consistency when matching sweeps against pure tones (see details bellow). From the 14 excluded participants, one failed the first test and 13 failed the second test. 6 of the excluded participants reported no previous musical experience; the remaining 8 had at least five years of musical training.

2.1.2 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 and carried out in 40 ms. Stimuli were ramped-in and damped-out with 5 ms Hanning windows overlapping the sections with constant frequency.

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

2.1.3 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 increments of $\pm\epsilon = \pm 25\,\mathrm{Hz}$, bringing the pitch of the sound closer to the participants 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 gaps) × 3 (3 average frequencies) × 2 (probe played first or last) factorial design.

2.1.4 Experimental procedure

Before the experiments, all potential participants performed a brief training to ensure that they had understood the task. The trial structure of the training was exactly the same as in the experiment, but both probe and target consisted of pure tones. During the training, the software provided feedback after each trial informing the participant whether the response was correct or incorrect. The training 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 training in the first batch.

After the training, participants were evaluated on their response consistency to sweeps. To do that they undertook a block of 12 trials consisting in 4 repetitions of 3 sweeps with diverse properties and small frequency gaps: $\{\Delta f = 67\,\mathrm{Hz}, \bar{f} = 900\,\mathrm{Hz}\}$, $\{\Delta f = -200\,\mathrm{Hz}, \bar{f} = 1200\,\mathrm{Hz}\}$, and $\{\Delta f = -67\,\mathrm{Hz}, \bar{f} = 1500\,\mathrm{Hz}\}$. Trial ordering was randomised, and the relative order between probe tone and the target sweep was reversed in half of the trials. After the completion of this block, the participant's pitch matching consistency was calculated as the inverse of the average of the absolute differences between the reported pitch in each sweep type and presentation order. Only participants with an average deviation smaller than twice the frequency increment step $2\epsilon = 50\,\mathrm{Hz}$ were included in the experiment. This prevented us from adding data consisting in random guesses that would bias the sweep pitch shift towards the average frequency of the sweep.

The 8 included participants undertook 4 additional blocks of 27 trials. Each block contained a single instance of each sweep type. The order of the sweeps within each trial was randomised and the relative position of the probe tone with respect to the target stimulus was pseudorandomised so that half of the trials in each block were presented in each direction. 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 showed a high self-consistency in the main experiment (i.e., a smaller variance than $2\epsilon = 50\,\mathrm{Hz}$ within each sweep type). Only sweeps with frequency gaps $\Delta f \in [-200, 200]\,\mathrm{Hz}$, which were expected to yield the most unequivocal pitch sensation according to Hart's law [26], were used to compute the overall self-consistency, although the participants were unaware of this. Participants typically completed the experiment between 1.5 and 3 hours.

$\bar{f}$ =	900 Hz		1200 Hz		1500 Hz	
probe ↔ sweep	←	→	←	→	←	→
slope (±0.03)	0.36	0.30	0.38	0.42	0.41	0.40
r_p ($p < 10^{-22}$)	0.88	0.68	0.86	0.89	0.90	0.88
r_s ($p < 10^{-33}$)	0.91	0.78	0.92	0.96	0.94	0.96

Table 1: **Summary statistics on the relationship between the perceived pitch and the frequency gap for single-sweep stimuli.** The slope of the linear fit, Pearson's correlation r_p, and Spearman's correlation r_s for the relationship between $f_{\mathrm{perceived}}$ 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.

2.1.5 Data analysis

The perceived pitch corresponding to each stimuli was summarised using the *pitch shift*, $\Delta p = f_{\mathrm{perceived}} - \bar{f}$, where $f_{\mathrm{perceived}}$ is the frequency of the pure tone matched to the corresponding stimulus. Distributions of Δp were drawn pooling the data corresponding to each stimulus across trials and participants. Thus, there were $4 \times 8 = 32$ data points for each stimuli, 16 points for each of the sweep-probe relative order.

2.2 Experimental results: bottom-up models of pitch

The pitch shift Δp depended on the size of the gap Δf (Table 1 and Figure 1). The exact dependence was consistent across listeners for sweeps with $\Delta f \leq 333\,\mathrm{Hz}$ lying in the vicinity of the linear fit $f_{\mathrm{perceived}} \simeq \bar{f} + m\,\Delta f$ (with an average deviance from the fit of 46 Hz). Sweeps with larger frequency gaps resulted in wider distributions of $f_{\mathrm{perceived}}$ due to higher inter- and intra- subject variabilities (Figure 2). Presentation order did not systematically affect the perceived pitch[1].

In their classical study, Brady and colleagues [2] 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

[1] See Supplementary Figure S1 for a more formal description.

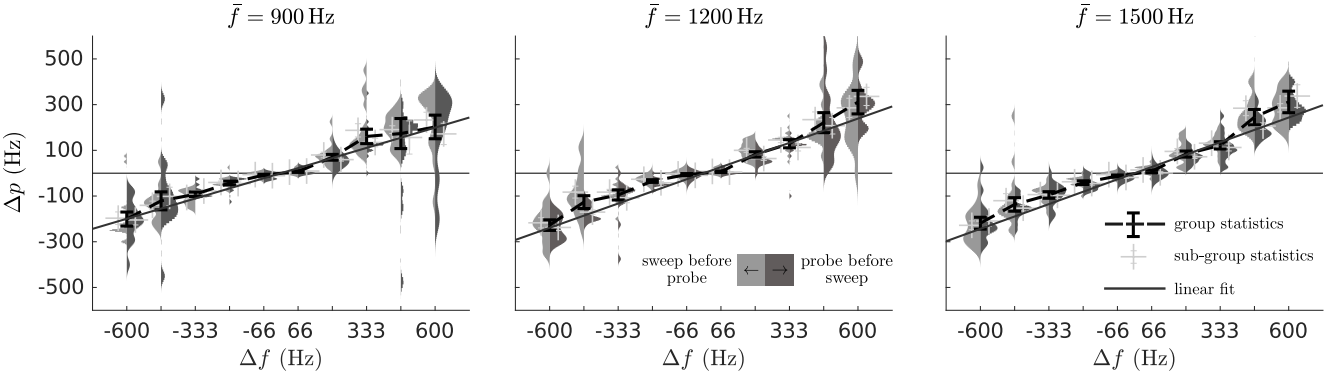

Figure 1: **Sweep pitch shift.** Kernel density estimations on the perceived pitch are plotted separately for each of the 30 sweeps used in the experiment. The y-axis of each plot shows the magnitude of the sweep pitch shift Δp. The x-axis list the gaps of each of the sweeps. Distributions are plotted separately for trials where the probe was presented before ($\rightarrow$) and trials where presented after ($\leftarrow$) the sweep. Light-gray error bars show the separate average and standard error of the $\rightarrow$ and $\leftarrow$ subgroups. Black error bars and dashed lines show the average and the standard error of the data pooled across presentation order. The dark thick grey solid lines show the group linear fit of the data.

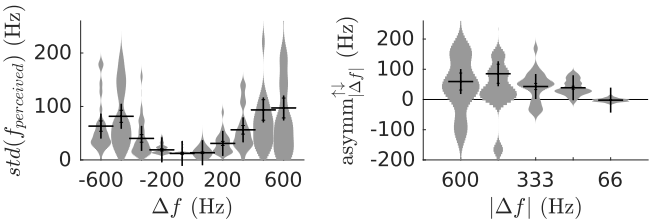

Figure 2: **Variance of the perceived pitch and up/down asymmetry.** Left: Kernel density estimations of the intra-subject standard deviation of the sweep pitch shift Δp, plotted separately for the different frequency gaps Δf. Each point in the distributions corresponds to the standard deviation of the 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 variance is monotonically correlated to the absolute gap $|\Delta f|$ ($r_s = 0.63$, $p < 10^{-27}$). Right: Kernel density estimations of the up/down asymmetry distributions as defined in Equation (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 ($N = 8 \times 3 = 24$). Error bars show the average and the standard error of the groups.

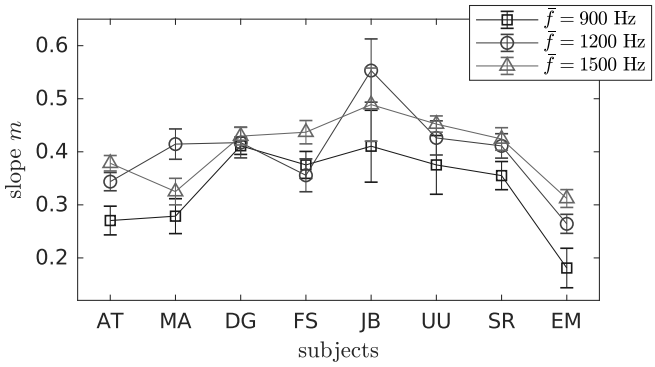

Figure 3: **Subject specific estimations of the linear fits between the pitch shift Δp and Δf.** The plot shows the slope m of the linear fit $f_{\text{perceived}} \sim \bar{f} + m\,\Delta f$ for each of the 8 subjects; error bars mark the 95% confidence intervals of the estimations.

colleagues [3] showed the reversed effect. To test if our data replicates any of these previous findings we draw, for each absolute frequency gap $|\Delta f|$, the distribution of the differences between the pitch shift in up and down sweeps:

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

Results shows a general trend in the direction of the observations from Nabelek and colleagues (Figure 2, right). The effect is significant for $|\Delta f| \leq 200\,\text{Hz}$ ($p < 2 \times 10^{-5}$) but not for $|\Delta f| = 66\,\text{Hz}$ ($p = 0.77$), according to two-tailed rank-sum tests ($N = 96$). Up sweeps have been consistently found to be easier to discriminate from pure tones than down sweep in a wide range of experimental conditions [27, 28, 29, 30], probably because auditory nerve responses to up sweeps compensate, at least partially, for the low-frequency processing delay of the basilar membrane [31] provoking more salient neural responses than their down counterparts.

Last, we tested if the dependence of the sweep pitch with Δf was robustly replicated across subjects. The slope of the linear fit between $f_{\text{perceived}}$ and Δf, similar in magnitude in all participants, are plotted in Figure 3.

2.3 Modelling methods: bottom-up models of pitch

The experimental results were compared with the predictions of three families of existing models: 1) a spectral model of

pitch processing based on the tonotopic arrangement of the cochlear output at the beginning of the auditory nerve [32], 2) a temporal model based on the principles of the summary autocorrelation function that measures pitch according to the phase-locked response in the auditory nerve [9, 11], and 3) a phenomenological model of frequency integration specifically designed to predict the pitch of FM sounds [5, 6].

2.3.1 Spectral models of pitch processing

The predictions of the spectral model of pitch processing are based on the spectral decomposition computed at the periphery of the auditory system by the basilar membrane. First, a realistic model of the peripheral auditory system [8, 33] computes 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 $f_{\min} = 125\,\text{Hz}$ to $f_{\max} = 10\,\text{kHz}$.

The peripheral output of each of the $N = 100$ channels is then integrated by a neural ensemble following mean-field neural ensemble dynamics [34]. Although these ensembles were first formulated to describe dynamics in cortical regions dedicated to visual decision making, they have been successfully used to describe the dynamics of many different cortical areas (e.g., [35]). Each ensemble in the array of integrators receive inputs from a single cochlear channel. Inputs were modelled with α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid-receptor (AMPA) synaptic dynamics [36]. 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., [37]). The firing rate $h_n(t)$ of the nth ensemble follows the dynamics of a leaky integrator:

$$\tau^{\text{memb}}\,\dot{h}_n(t) = -h_n(t) + \phi(J_{\text{in}}^{\text{AMPA}}\,S_{\text{in},n}^{\text{AMPA}}(t)) \qquad (2)$$

Similarly, the synaptic gating variable $S_{\text{in},n}^{\text{AMPA}}(t)$ follows [36]:

$$\tau^{\text{AMPA}}\,\dot{S}_{\text{in},n}^{\text{AMPA}}(t) = -S_{\text{in},n}^{\text{AMPA}}(t) + p_n(t) \qquad (3)$$

Time constants $\tau^{\text{memb}} = 10\,\text{ms}$ and $\tau^{\text{AMPA}} = 2\,\text{ms}$ where taken from the literature [36]. The effective conductivity $J_{\text{in}}^{\text{AMPA}} = 0.38\,\text{nA}$ was manually tuned within the realistic range such that the peripheral system would elicit firing rates on the range $5\,\text{Hz} \geq h_n(t) \geq 100\,\text{Hz}$ in the integrator ensembles. The transfer function $\phi(x) = (cx - I_0)/(1 - e^{-g(cx-I_0)})$ and its parameters, empirically derived for networks of integrate-and-fire neurons, were taken from [34].

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

$$E[k] = \sum_n n\rho_n \quad \text{with} \quad \rho_n = \frac{\int_0^{T_d} dt\, h_n(t)}{\sum_n \int_0^{T_d} dt\, h_n(t)} \qquad (4)$$

The frequency corresponding to the expected channel $E[k]$ was computed according to a logarithmic fit $f_{\text{predicted}}(E[k]) = e^{a_1 E[k] + a_0}$ fitted to match the response of 5 probe pure tones with frequencies linearly spaced between 600 Hz and 1900 Hz (i.e., in the range of the frequency sweeps used in the experimentation).

All dynamic systems described in this work were integrated using Euler's method with a time step of $dt = 0.1\,\text{ms}$. To assess possible variabilities due to the stochasticity of the peripheral model, statistics of the predictions were obtained by running each model 10 times with different random seeds.

2.3.2 Temporal models of pitch processing

The summary autocorrelation function (SACF) was used as the representative of the family of temporal models of pitch. The SACF used in this work follows the original formulation by Meddis and O'Mard [9, 10]. 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:

$$\tau_m^{\text{SACF}}\,\dot{A}_m(t) = -A_m(t) + \sum_n p_n(t)\,p_n(t - \delta t_m) \qquad (5)$$

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 $\delta t_m = 0.5\,\text{ms}$ and $\delta t_m = 30\,\text{ms}$, which allows the model to capture periodicities corresponding to frequencies between 2 kHz and 135 Hz up to four lower harmonics. The integration constant τ_m^{SACF} depends linearly on δt_m (see details in [11, 38]).

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 $\delta t_m = 2\,T_0 = T_1$, $\delta t_m = 3\,T_0 = T_2$, etc. The first four peaks of the harmonic series were used to obtain a robust estimation of the pitch predicted by the model, so that:

$$f_{\text{predicted}} = \left(\frac{1}{4}\sum_{i=0}^{3}\frac{T_i(t)}{i+1}\right)^{-1} \qquad (6)$$

2.3.3 Phenomenological models of frequency integration

Previous studies attempted to explain the sweep pitch shift with heuristic numerical procedures of less [2] or more [5] complexity. In general, these methods assume that pitch is evaluated as a weighted integral of the frequency of the tone:

$$f_{\text{predicted}} = \frac{\int_0^{T_d} d\tau\omega(\tau)f(\tau)}{\int_0^{T_d} d\tau\omega(\tau)} \qquad (7)$$

where T_d is the stimulus duration and $f(t)$ is the frequency of the stimulus at the instant t. Unlike the temporal and spectral models introduced before, this formula considers directly the stimulus properties rather than the cochlear output elicited by the stimulus and lacks of a biological rationale. However, it is to our knowledge the only attempt to

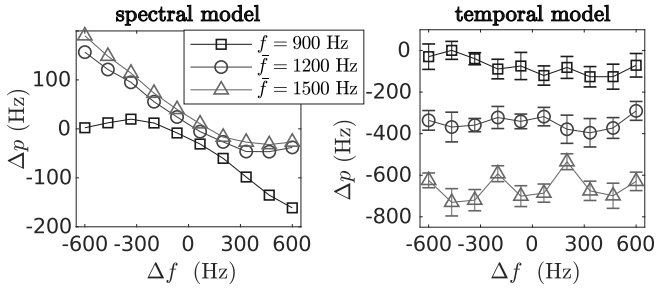

Figure 4: **Predictions of the spectral and temporal models.** Predictions of the spectral model (left) were precisely consistent across runs and thus only the nominal value is plotted. To account for variability of the temporal model (right), we plotted the average results across ten runs; error bars represent standard errors. Cf. Figure 1.

more formally explain the sweep pitch shift and it is therefore informative to determine if such simple relationship can account for our experimental results.

D'Alessandro and Castellengo [5] proposed that $\omega(\tau)$ could give a stronger weight to the final part of the sound via a fixed integration window τ_{int}: $\omega(\tau) = e^{\frac{\tau - T}{\tau_{\text{int}}}} + \beta$, and found a fit with $\alpha = 22$, $\beta = 0.20$ for the vibrato tones used in their experiment. However, when trying to explain the results from a more recent experiment [6], D'Alessandro and Rosset needed to use two different sets of parametrisations for the evaluation of up and down sweeps. Thus, rather than using a specific instance of their multiple sets of parameters to try to predict the pitch of our stimuli, here we attempt to fit the weighting parameters to our perceptual results independently for each stimulus length (i.e., single or sweep-train stimuli) and centre frequency f_0. We would conclude that the model provides for a parsimonious explanation of our data if there exist a set of parameters yielding a satisfactory account of all the data.

2.4 Modelling results: bottom-up models of pitch

2.4.1 Models of pitch processing

Neither the spectral nor the temporal model of pitch replicated our experimental observations (Figure 4). Predictions of the spectral model show a dependence of $f_{\text{perceived}}$ with Δf in the opposite direction than the empirical data. This intriguing effect is a consequence of adaptation in the auditory nerve: responses are the strongest around stimulus' onset, provoking a perceptual bias towards the frequencies present at the beginning of the stimuli. Predictions of the temporal model lay within $f_{\text{perceived}} \simeq 800\,\text{Hz}$ independently of the centre frequency, resulting in negative Δp for increasing $\bar{f}$ values. This is most likely a consequence of the SACF being unable to decode rapidly changing frequencies with such short stimuli. Note that this is the case even for the conditions with the smaller Δf.

2.4.2 Phenomenological model

Let us first rewrite the model (Equation (7)) in the following form:

$$f_{\text{perceived}} = \bar{f} + \Delta f \frac{\int_0^{T_d} d\tau \omega(\tau) \left(\frac{\tau}{T_d} - \frac{1}{2}\right)}{\int_0^{T_d} d\tau \omega(\tau)} \tag{8}$$

For the model to explain our data, we would need to find two numbers τ_{int} and β such that the rightmost fraction in Eq (8) with $\omega(\tau) = e^{\frac{\tau - T}{\tau_{\text{int}}}} + \beta$ is constant across conditions. Our data for single sweeps can be approximated by this function with parameters $\alpha = 1/\tau_{\text{int}} = \simeq 0.10\,\text{ms}^{-1}$ and $\beta \simeq 0$.

However, the model cannot account for the up/down asymmetry observed in the absolute pitch shift $|\Delta p|$: since $f_{\text{perceived}}(-\Delta f) - \bar{f} = -(f_{\text{perceived}}(\Delta f) - \bar{f})$ (cf. Equation (8)), the absolute pitch shift would be the same in up and down sweeps. Moreover, since this model only attempts to describe the phenomenology of the effect, it lacks of biological plausibility and does not explain the neural mechanisms underlying the integration process. In the next section, we will derive a mechanistic ensemble model of FM encoding that explains the experimental results mechanistically.

3 FM encoding and its role in the sweep pitch shift

3.1 Formulation of the model

In this section we introduce a hierarchical model of FM-encoding, termed here *FM-feedback spectral model*, with two levels (Figure 5). 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 main hypothesis introduced in the 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 next in the spectral layer. This parsimonious mechanism qualitatively explains why the latest parts of the sweep are given a higher weight during perceptual integration and quantitatively reproduces the exact dependence of pitch with Δf observed in our data.

3.1.1 Modelling FM direction selectivity

At least three mechanisms for FM direction selectivity have been identified in the animal literature: asymmetric sideband inhibition [15, 39, 40], duration sensitivity [41, 18, 39], and delayed excitation [40, 22, 42]. In order to contain the dimensionality of the model's parameter space, we focus here on delayed excitation, a straightforward mechanism where neurons with different best frequencies output to the direction selective neuron with different delays; e.g., an up-selective neuron will receive delayed inputs from a neuron tuned to low frequencies and instantaneous inputs from a neuron tuned to high frequencies, so that an up sweep results in simultanous

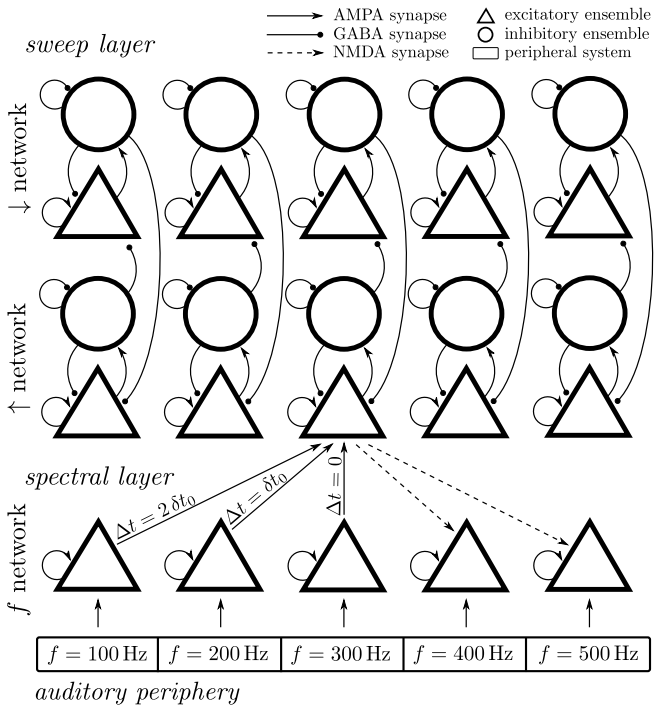

Figure 5: **Diagram of the FM-feedback spectral model.** The model consists of three layers: first, the auditory periphery; second: the *spectral layer*, with a single network integrating the spectral information of the sound (f network); and third, the *sweep layer*, with two networks specialised in detecting up (↑ network) and down (↓ network) sweeps, respectively. Afferent connections allow the spectral layer to integrate information from the periphery, and the sweep layer to decode the direction of the sweep from the spectral information of the f network. Feedback connections from the sweep layer modulate the time constants of the populations in the spectral layer that are expected to follow 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.

excitation from both of them. Any related mechanism showing FM direction selectivity should yield similar overall results [43].

In the FM-feedback spectral model, delayed excitation is implemented by introducing 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 selectively to up sweeps will receive increasingly delayed inputs from the spectral populations centred at $f < f_0$ (Figure 5). 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: $\delta t_{nm} = |n - m|\delta t_0$.

The spectral layer is modelled following the same principles of the spectral model of pitch introduced in Section 2.3.1:

$$\tau^{\text{pop}}\, \dot{h}_n^f(t) = -h_n^f(t) + \phi(I_n^f(t)) \tag{9}$$

with

$$I_n^f(t) = J_{\text{in}}^{\text{AMPA}} \sum_k \omega_{nk}^{\text{in}} S_{\text{in},k}^{\text{AMPA}}(t) \tag{10}$$

Note that we used the index ^{f} to denote variables in the spectral layer. Pitch decisions are taken according to the spectral information encoded in the neural ensembles h_n^f according to Equation (4).

We allowed some dispersion in the propagation from the peripheral model to the spectral layer by using a Gaussian-shaped connectivity matrix

$$\omega_{nm}^{\text{in}} = \frac{1}{\sqrt{\sigma_{\text{in}}}} e^{-\frac{(m-n)^2}{2\sigma^{\text{in}}}} \tag{11}$$

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 ensembles used in this model are endowed with an adaptive time constant:

$$\tau_{e,i}^{\text{pop}}(h, I) = \tau_{e,i}^{\text{memb}}\, \Delta_T \frac{\partial_x \phi(x)|_{x=I}}{h} \tag{12}$$

where $\Delta_T = 1\,\text{mV}$ is the size of the spike initialisation of the neural model and τ_e^{memb} and τ_i^{memb} are the neural membrane time constants for excitatory and inhibitory populations, respectively. Using adaptive integration time constants allows the system to react faster to changes when the target populations are already active and the synaptic input is not too large, a behaviour often reported in tightly connected populations of neurons [44]. This component plays an active role in the feedback activation mechanism that we will describe in the next section. The analytic formulation of $\tau^{\text{pop}}(h, I)$ stems from a theoretical study of networks of exponential-integrate-and-fire neurons [44].

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 consist of N columns, each comprising an excitatory and an inhibitory population (Figure 5). The instantaneous firing rate of each *up* $(h_n^{\uparrow e}(t), h_n^{\uparrow i}(t))$ and *down* $(h_n^{\downarrow e}(t), h_n^{\downarrow i}(t))$ population follows the same dynamics described in Equation (9), with *up*

$(I_n^{\uparrow e}(t), I_n^{\uparrow i}(t))$ and *down* $(I_n^{\downarrow e}(t), I_n^{\downarrow i}(t))$ synaptic inputs, respectively. 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 [34] (Table 2).

Excitatory and inhibitory inputs to populations in the *sweep layer* are modelled according to AMPA-like and GABA-like synaptic gating dynamics [36]:

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

where ξ is an uncorrelated Gaussian noise sampled independently for each synapse and instant t, and $\sigma = 0.0007\,\text{nA}$ is the amplitude of the noise [34]. The total synaptic input for each population is then:

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

where I^E_{bkg} and I^I_{bkg} are constant background inputs putatively sourced in external neural populations [34].

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

$$\omega_{nm}^{\alpha} = e^{-\frac{(n-m)^2}{2\sigma_\alpha}}, \quad \alpha = ei, ie \tag{13}$$

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:

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

The free parameters were initialised to standard values (the effective conductivities J_f^{AMPA}, J^{GABA}, and J_s^{AMPA}, according to [34]; the baseline delay δt_0 to 2 ms/channel; and the dispersion constants σ_{in}, σ_{ei}, σ_{ei}, and $\Delta_{\omega f}$, to $0.1\,N$) and manually tuned to achieve 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 2.

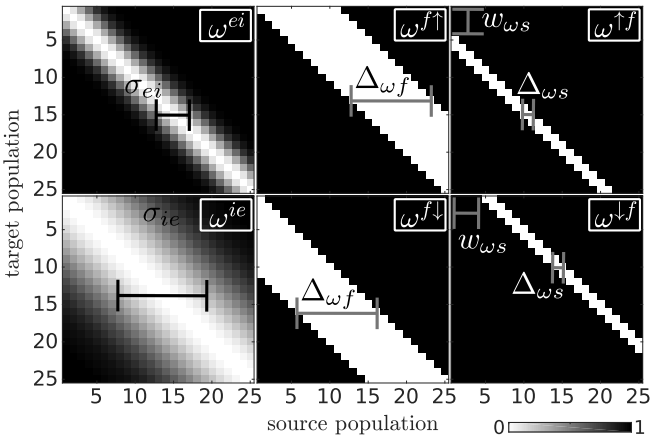

Figure 6: **Connectivity matrices.** Matrices show the connection between the first 25 ensembles of each source-target group. From left to right, matrices correspond to: excitatory-to-inhibitory ω^{ei}, inhibitory-to-excitatory ω^{ie}; bottom-up AMPA connections spectral-to-up $\omega^{f\uparrow}$, spectral-to-down $\omega^{f\uparrow}$; 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.

3.1.2 Feedback from the sweep to the spectral layer

Once neurons in the sweep layer encode the sweep direction, feedback connections targeting the spectral layer facilitate the encoding of expected frequencies. 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 $f \le f_0$. Although in some occasions the up sweep will culminate in f_0, in most of the cases f_0 will be only an intermediate step in the ascending succession of the sweep and thus the 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 top down mechanism of the model, encoded in the feedback projections stemming from the sweep layer and targeting the spectral layer, reduces the temporal constant of these populations using low-current feedback excitatory signals. Similarly, feedback connections stemming from a population j in the down-network that receives 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. Feedback current intensity is kept low in comparison to the bottom-up driver by enforcing $J^{\text{NMDA}} \ll J^{\text{AMPA}}$ (see Table 2).

The low-current feedback signal modulates the population to elicit only a subtle higher firing rate than a not modulated population. The subtle activation results into a significantly lower effective time constant τ^{pop} (Figure 7; cf. Equation (12)), causing the population to react faster to changes in the bottom-up input. This increased readiness reduces the

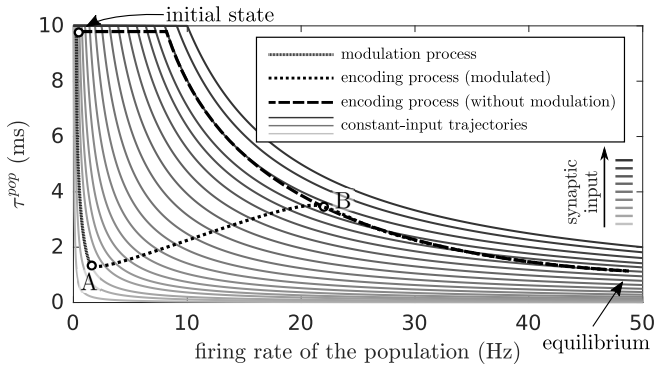

Figure 7: **Effect of the predictive feedback mechanism on the population time constants.** Solid lines show the dependence of $\tau^{\text{pop}}(h, I)$ with h for several synaptic inputs (cf. Equation (12)). The thin-dotted lines (encoding process, modulated) correspond to the trajectory of τ^{pop} associated to a population during the predictive feedback modulation. The thick-dashed and thin-dotted lines correspond to the trajectory after the onset of peripheral input of a non-modulated and a modulated population, respectively. During the modulation, the low synaptic input decreases τ^{pop} rapidly without substantially increasing its firing rate (point A). Once modulated, the population reacts quickly to the onset of the peripheral input and arrives to the point B in a much shorter time than an equivalent non-modulated population (compare the trajectory from the initial state to B with the trajectory from A to B).

metabolic cost of encoding expected frequencies and, since the population will spend more time in the high-firing-rate regime than increasing its firing rate to match the input, it indirectly results in a stronger contribution of the frequencies expressed in the last part of the sweep.

NMDA receptors are typically responsible for conveying feedback excitatory information in the cerebral cortex [45, 46]; specifically, NMDA-deactivation results in a reduced feedback control in the auditory pathway [47]. Thus, feedback connections were modelled according to NMDA-like synaptic gating dynamics with a finite rising time constant [36]:

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

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 (10):

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

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 centre frequency of the source and target ensembles (see Figure 6, right):

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

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 [34], and the connectivity parameters $w_{\omega s}$ and $\Delta_{\omega s}$ to $0.1\,N$) and manually tuned so that the pitch predictions of the model (as computed in Equation (4)) matched the empirical data.

3.2 Model predictions

3.2.1 FM direction selectivity

Example responses of the excitatory populations of the model to up and down sweeps are shown in Figure 8.

To quantify direction selectivity, we used the standard (e.g., [23]) direction selectivity index (DSI), 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:

$$\text{DSI}^{\alpha} = \frac{\sum_n \int dt \left([h^{\alpha e}_n(t)]_{+\Delta f} - [h^{\alpha e}_n(t)]_{-\Delta f} \right)}{\sum_n \int dt \left([h^{\alpha e}_n(t)]_{+\Delta f} + [h^{\alpha e}_n(t)]_{-\Delta f} \right)} \quad \alpha = \uparrow,\downarrow \tag{14}$$

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 gap Δf. 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 (Figure 9); systematically increasing DSI magnitudes were elicited by increasing $\bar{f}$ and $|\Delta f|$. Network selectivity to FM direction was robust across reparametrisations of the model, although deactivation of the feddback connections resulted in a 8.7($\pm$1.5)% average decrease in DSI$^{\uparrow}$ and in a 9.7($\pm$1.4) average increase in DSI$^{\downarrow}$, indicating that the feedback connections sharpen direction selectivity [2].

3.2.2 Reproduction of the sweep pitch shift

Here we assess the ability of the FM-feedback spectral model to explain the sweep pitch shift (Figure 10). To compare the model responses and the experimental data we fitted a logarithmic function that estimates the expected channel corresponding to a pure tone of a given frequency following the same procedure as in Section 2.3.1; the model explains $R^2 = 0.88$ of the variance of the experimenal data.

[2] See Supplementary Figure S3 for a study on the dependence of the DSI across the parameter space.

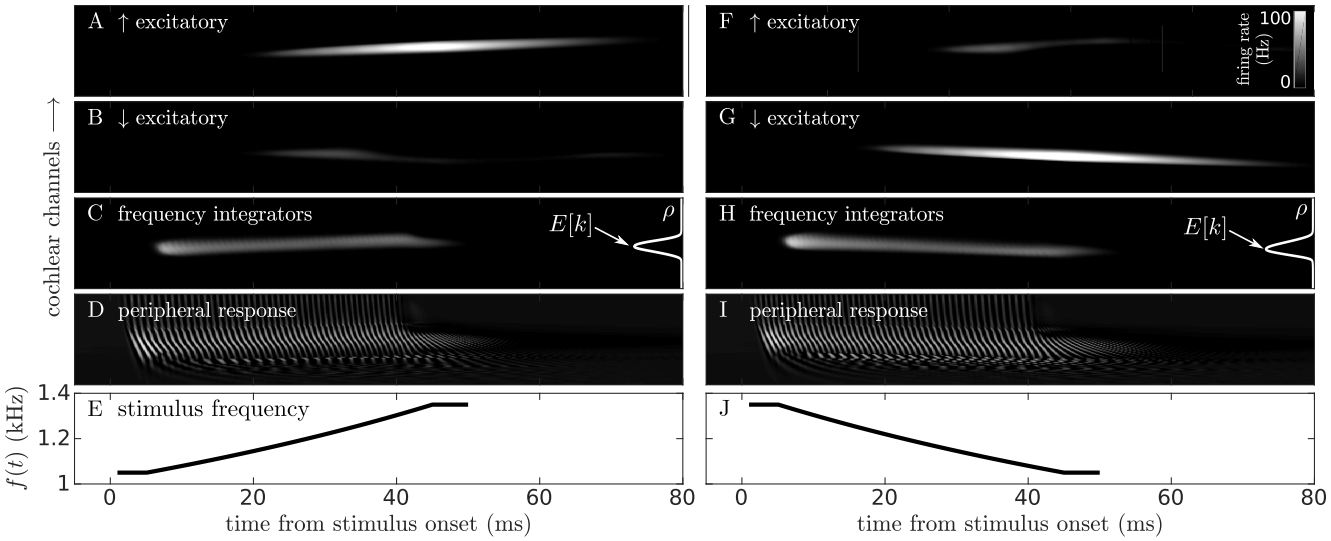

Figure 8: **Model responses to an up and a down sweep.** A-E show the responses to an up sweep, and F-J to a down sweep. From top to bottom: (A/F) the instantaneous firing rate of the up-selective excitatory populations $h_n^{\uparrow e}(t)$; (B/G) the instantaneous firing rate of the down-selective excitatory populations $h_n^{\downarrow e}(t)$; (C/H) the instantaneous firing rate of the frequency integrators $h_n^f(t)$ and, in the right of the panels, the probability distribution ρ derived from the integral of $h_n^f(t)$ over the duration of the stimulus and the associated expected channel $E[k]$ (cf. Equation (4)); (D/I) the firing rate at the auditory nerve according to the peripheral system $p_n(t)$; (E/F) the instantaneous frequency of the sweeps along time $f(t)$. In all panels except for E/J, y-axis represents the cochlear channel n, ordered from bottom to top. The stimuli were an up and a down single sweep as defined in Section 2.1.2 with $\Delta = \pm 300$ Hz and $\bar{f} = 1200$ Hz.

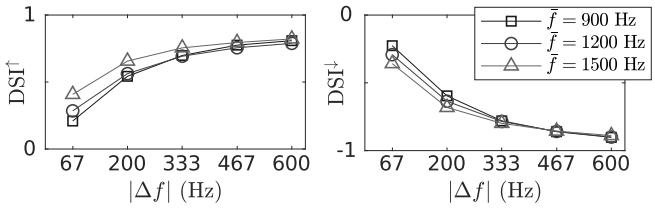

Figure 9: **Direction selectivity indices for the sweeps of the experiment.** DSIs (as in Equation (14)) in the up and down network. Each point corresponds to a given $\bar{f}$ and $|\Delta f|$.

To avoid the error introduced by the approximate transformation from the frequency to the channel space of the experimental results, we next computed Pearson correlations between the model predictions in the channel space and the sweep pitch shift measured in the experiments (Figure 10B). The model predictions were strongly correlated with the data ($r_p = 0.98, p < 10^{-22}$); moreover, there was a significant correlation between the variance of the model responses and the standard error of the experimental data ($r_p = 0.60, p = 0.0005$; Figure 10C), indicating that the larger variability in the pitch shift observed for the larger Δf can be understood as a consequence of a wider spread activation across the spectral populations. Last, the model also reproduced the up/down asymmetry (Figure 10D).

To study the dependence of the fitness with the parameter choice we recomputed the explained variance R^2 across the parameter space of the model (Figure 11). The model explained the experimental data in a wide section of the parameter space, with an average R^2 across a 5-point diameter sphere around the final parameters of $E[R^2] = 0.78 \pm 0.03$. It is also interesting to consider the effect of fixing the effective population time constant to $\tau = \tau^{\text{memb}}$. It is clear that, even considering lower τ^{memb} than the physiologically valid nominal value $\tau^{\text{memb}} = 20$ ms, without an adaptive τ the feedback mechanism of the model remains unactive (i.e., much stronger NMDA currents ($J^{\text{NMDA}} \sim J^{\text{AMPA}}$) are necessary to drive the spectral distribution towards the experimental results.

3.2.3 Reproduction of previous results in the literature

Last, we tested whether the FM-feedback spectral model was able to predict the pitch shift of additional data from the earlier study by Brady and colleagues [2]. We chose their stimuli because this was the only study that investigated the dependence of the pitch shift with properties different than Δf. Specifically, in the *experiment II* from the original paper, Brady and colleagues considered FM-sweeps with a fixed 20 ms transition between 1000 Hz and 1500 Hz that was located at different positions within a 90 ms stimulus (see Figure 12, left). In the *experiment III*, they used FM-sweeps in the same Δf but with transitions of varying durations (Figure 12, right).

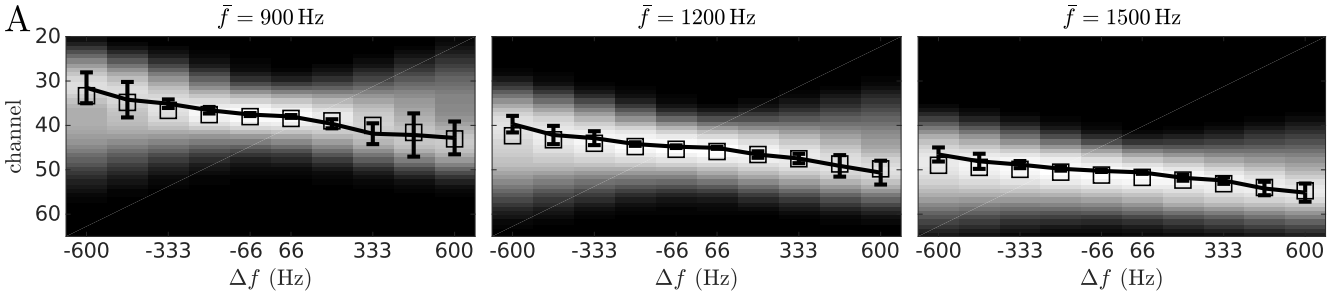

Figure 10: **Predictions of the FM-feedback spectral model for FM-sweeps.** (A): Shading matrices show the distribution of the activation across channels (y-axis) for different sweep Δf (x-axis). Squares printed over the distributions mark the expected channel $E[k]$ as defined in Equation (4). Solid error bars are estimations of the experimental results in the channel space. The expected value agrees with the experimental data. Moreover, stimuli with larger Δf seem to elicit wider activation distributions than stimuli with smaller Δf, mirroring the generally larger variance observed in the data corresponding to the larger Δf. (B/C): Scatter plots show the correlation between the perceived pitch and the expected channel of the model response (B) and the correlation between the experimental standard error and the variance of the model response (C). Grey dashed lines are the least-squares lines of the pooled data. (D): Error bars show the model predictions of the up/down asymmetry coefficient asymm↑↓ (see Equation (1)). Errorbars are estimations of the standard error calculated based on the dispersion of the centroids for different $\bar{f}$ and the variance of the spectral distribution ρ of each condition. Experimental data in the background is the same as in Figure 2, right.

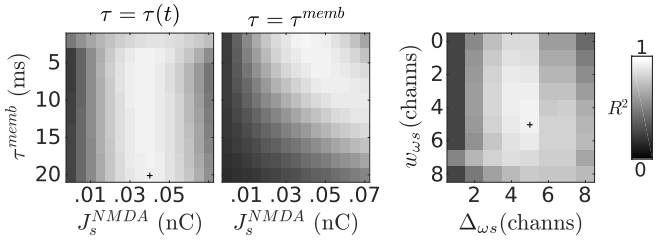

Figure 11: **Experimental fit in relation to the model parametrisation.** Shading matrices show the explained variance of the experimental data R^2 (white means $R^2 = 1$, black means $R^2 = 0$) for different points in the parameter space. Unless stated otherwise, parameters not varied in the matrices correspond to the values in Table 2. The two leftmost plots show the dependence of R^2 with J_s^{NMDA} and the dynamics of the excitatory population time constants. Different values of τ^{memb} were used to illustrate that the dynamic effect (rather than the resulting shorter time constant) is crucial to explain the experimental results; however, τ^{memb} was constrained to $\tau^{\mathrm{memb}} = 20$ ms based on to physiological observations [48]. The rightmost plot shows the dependence of R^2 on the width ($w_{\omega s}$) and the scope ($\Delta_{\omega s}$) of the feedback connections. Black crosses in the parameter space signal the final parametrisation.

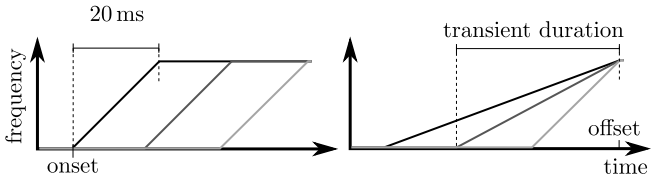

Figure 12: **Schematic view of the stimuli from [2].** All stimuli had the same duration (90 ms) and frequency span (1000-1500 Hz). In the first family, plotted on the left, the transient was fixed to a 20 ms duration and its onset was systematically varied so that the transition falls at different segments of the stimulus. In the second family, plotted on the right, the stimulus offset was fixed at 90 ms and the transient's onset was varied between 10 and 50 ms, resulting in transients of different durations. In these last stimuli, we extended the duration to 95 ms to prevent the ramping at the end of the stimulus from overlapping with the FM transient.

parameter	value (unit)	source
N	100 channels	ad-hoc
dt	0.1 ms	ad-hoc
periph dt	0.01 ms	ad-hoc
periph $f_{\min}$	125 Hz	[33]
periph $f_{\max}$	10000 Hz	[33]
τ_e^{memb}	20 ms	[48]
τ_i^{memb}	10 ms	[48]
Δ	1 mV	[44]
$c^{\text{excitatory}}$	310 (V nC)$^-$1	[34]
$I_0^{\text{excitatory}}$	125 Hz	[34]
$g^{\text{excitatory}}$	0.16 s	[34]
$c^{\text{inhibitory}}$	615 (V nC)$^-$1	[34]
$I_0^{\text{inhibitory}}$	177 Hz	[34]
$g^{\text{inhibitory}}$	0.087 s	[34]
I_{bkg}^E	0.23 nA	[34]
I_{bkg}^I	0.10 nA	[34]
σ	0.0007 nA	[34]
γ	0.641	[36]
τ^{AMPA}	2 ms	[36]
τ^{GABA}	5 ms	[36]
τ^{NMDA}	100 ms	[36]
$J_{\text{in}}^{\text{AMPA}}$	0.38 nC	tuned (1)
J_f^{AMPA}	0.55 nC	tuned (2)
J_s^{AMPA}	0.67 nC	tuned (2)
J^{GABA}	0.30 nC	tuned (2)
J^{NMDA}	0.04 nC	tuned (3)
σ_{in}	0.1 N channels	tuned (2)
σ_{ie}	0.5 N channels	tuned (2)
σ_{ei}	0.03 N channels	tuned (2)
Δt_0	3 ms/channel	tuned (2)
$\Delta_{\omega f}$	0.05 N channels	tuned (2)
$\Delta_{\omega s}$	0.05 N channels	tuned (3)
$w_{\omega s}$	0.03 N channels	tuned (3)

Table 2: **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 (see Section 2.3.1); 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.

We compared the predictions of the FM-feedback spectral model with the experimental results reported in the original paper (Figure 13). Although the dependence of pitch with the sweep properties is much smoother in the model predictions than in the experimental data, the trend and extreme values are well reproduced by the model. Predictions showed a strong Pearson's correlation with the reported pitch shift across both experiments ($r_p = 0.87$, $p < 10^{-6}$; Figure 14) and a weaker correlation between the variance of the activation distribution ρ and the experimental standard error ($r_p = 0.46, p = 0.03$).

The FM-feedback spectral model provides for a mechanistic interpretation of these results. In Brady's experiment II, the transient duration is kept constant but its onset is varied across the simulus duration. When the transient is located near the beginning of the stimulus, the greatest part of the sounds excites neurons encoding frequencies near the ending side of the transient pushing the distribution of the responses ρ towards the ending frequencies of the sweep f_1. This shift is larger than it would be expected for a sound without a transient because of the feedback modulation of the later frequencies exerted by the sweep network. When the transition 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 that is varied. The decreased sweep pitch shift observed for shorter transition durations is a consequence of the stimuli presenting a larger segment with the initial frequency, thus shifting ρ towards f_0. Larger transients covering more extended parts of the stimuli present the same pitch shift observed in our experiments.

4 Sweep trains and further stimuli

The results described so far are in favour of the hypothesis that a feedback system between direction-encoding and frequency-encoding populations are responsible for the sweep pitch shift. To validate this findings, this section introduces a new set of stimuli specifically designed to constest this hypothesis. The new stimuli, called here *sweep trains*, were a continuous concatenation of several sweeps with the same properties as the stimuli used in Section 2 and present the same acoustical properties as the single sweeps. Thus, it would be reasonable to hypothesise that, if the stimuli are still perceived as a single acoustical object, they would elicit the same pitch percept as their single-sweep subcomponents. The FM-feedback spectral model, however, 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. This prediction is tested in a perceptual experiment analogous to that of Section 2.

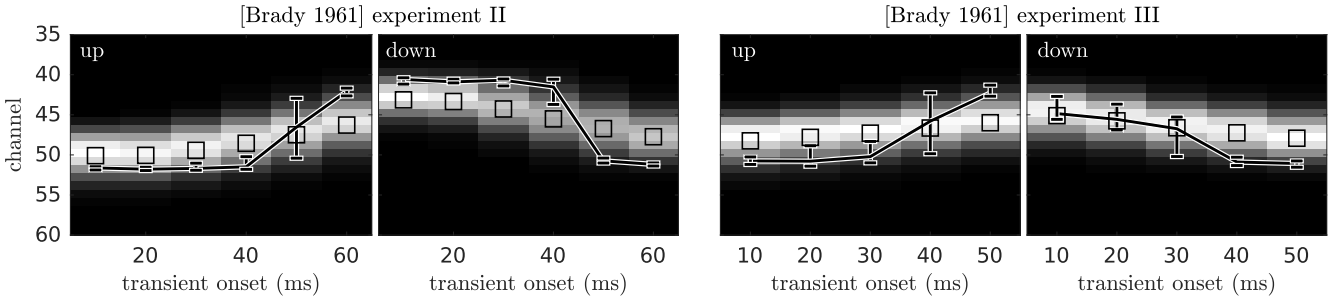

Figure 13: **Predictions of the FM-feedback spectral model for Brady's stimuli.** Shading matrices show the distribution of the activation across channels (y-axis) for different transient onsets (x-axis). Squares printed over the distributions mark the expected value with respect to the distribution. Solid error bars are estimations of the experimental results in the channel space.

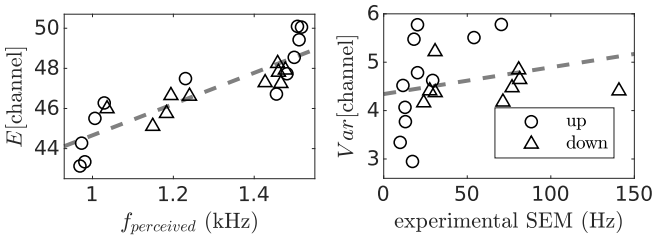

Figure 14: **Correlations between the FM-feedback spectral model predictions and Brady's experimental results.** Scatter plots show the correlation between the perceived pitch and the expected channel of the model response (left) and the correlation between the experimental standard error and the variance of the model response (right). Grey dashed lines are the least-squares lines of the pooled data.

4.1 Experimental methods

4.1.1 Participants

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

4.1.2 Stimuli

Stimuli were concatenations of 5 sweeps adding up to a total of 250 ms (sweep trains; see Figure 15). The sweeps were taken from a subset of 18 elements from the first experiment with 6 different frequency gaps $\Delta f \in [-333, 333]$ Hz. To ensure continuity of the stimulus waveform, the sweeps were concatenated in the frequency domain. 5 ms Hanning windows were applied only at the very beginning and very end of the sweep trains.

4.1.3 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 gaps) × 3 (average frequencies) × 2 (probe played first or last) factorial design.

4.1.4 Experimental procedure

Since the participants were already familiar with the task, the experiment contained no training. Four repetitions of the 18 sweep-trains were distributed across 5 blocks following the same principles as described in Section 2.1. Participants typically completed the second experiment between 1 and 2 hours.

4.2 Experimental results

As expected, the magnitude of the pitch shift depended on the size of the gap (Figure 16, Table 3). However, as qualitatively predicted by the FM-feedback spectral model, the effect sizes of the correlation were lower than in the single-sweep experiment (Table 3). Data also showed much higher inter- and intra-subject variability than in the single-sweep experiment[3]. After completing the experiment, participants reported 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[4].

Sweep-train stimuli show only a subtle up/down asymmetry that did not reach statistical significance.

[3] See also Supplementary Figure S2.

[4] Stimuli from the first and from the second stimuli are available in the Supplementary Materials.

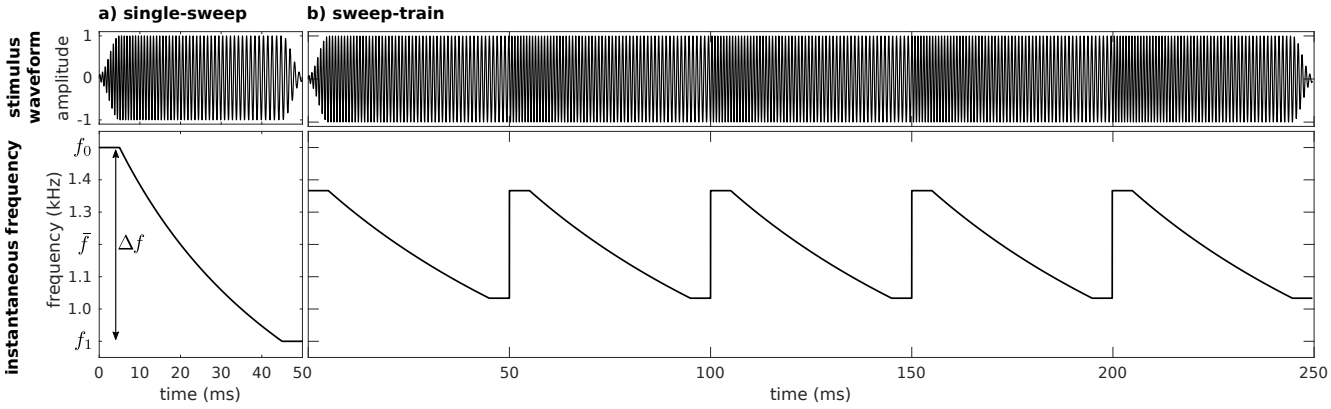

Figure 15: **Schematics of the stimuli.** Waveform $s(t)$ and instantaneous frequency $f(t)$ of the a) $\{\Delta f = -600\,\text{Hz}, \bar{f} = 1200\,\text{Hz}\}$ single-sweep stimulus from the first experiment, in comparison to b) the $\{\Delta f = -333\,\text{Hz}, \bar{f} = 1200\,\text{Hz}\}$ sweep-train stimulus from the second experiment.

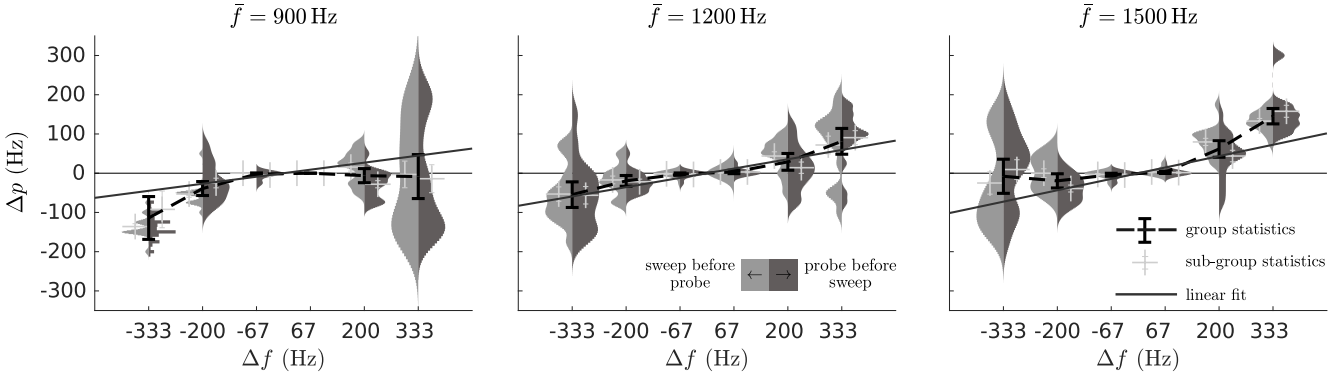

Figure 16: **Sweep pitch shift for sweep trains.** Kernel density estimations on the perceived pitch are plotted separately for each of the 30 sweeps used in the experiment. The y-axis of each plot shows the magnitude of the sweep pitch shift Δp. The x-axis list the gaps of each of the sweeps. Distributions are plotted separately for trials where the probe was presented before ($\rightarrow$) and trials where presented after ($\leftarrow$) the sweep. Light-gray error bars show the separate average and standard error of the $\rightarrow$ and $\leftarrow$ subgroups. Black error bars and dashed lines show the average and the standard error of the data pooled across presentation order. The dark thick grey solid lines show the group linear fit of the data.

$\bar{f}$ =	900 Hz		1200 Hz		1500 Hz	
probe ↔ sweep	←	→	←	→	←	→
slope (±0.04)	0.19	0.08	0.17	0.18	0.22	0.21
r_p ($p < 10^{-7}$)	0.54	n.s.	0.60	0.61	0.66	0.59
r_s ($p < 10^{-3}$)	0.55	0.26	0.58	0.60	0.66	0.56

Table 3: **Summary statistics on the relationship between the perceived pitch and Δf for 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 train, →; and sweep train before probe, ←).

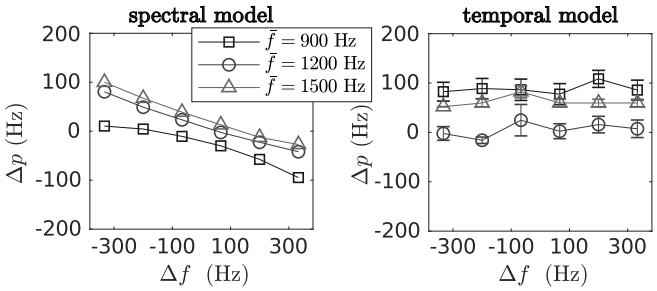

Figure 17: **Predictions of the spectral and temporal models for sweep trains.** Predictions of the spectral model (left) were precisely consistent across runs and thus only the nominal value is plotted. To account for variability of the temporal model (right), we plotted the average results across ten runs; error bars represent standard errors. Cf. Figure 16.

4.3 Predictions from bottom-up models of pitch

Neither the temporal nor the spectral models were able to successfully explain the sweep-train data (Figure 17). The results were, however, generally more stable than in the simulations with the single sweeps: the trends of the negative dependence of the pitch shift with the frequency gap are consistent across different $\bar{f}$, and the temporal model predictions are much closer to the average frequency of the sweeps. This gain in stability is probably a consequence of the longer duration of the trains, which allows for a more reliable integration of the spectral information of the stimuli.

When using the optimal parameters derived for the single sweep stimuli, D'Alessadro's phenomenological model predicted the opposite effect of that observed: a negative dependence between $f_{\text{perceived}}$ and Δf ($m \in [-1.12, -0.31]$). Attempting to fit the parameters to the sweep trains data resulted in $\alpha \simeq 0.06\,\text{ms}^{-1}$, which resulted in equally unsatisfying results in the single-sweep condition. Since we failed to find a set of parameters that simultaneously explained the data of both, single sweeps and sweep trains, we concluded that this model is unable to account for the experimental results in a parsimonious way.

4.4 Predictions of the FM-feedback spectral model

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 $R^2 = 0.83$ of the variance of the data (Figure 18A) and the response distribution's expected value was strongly correlated to the observed pitch shift ($r_p = 0.99, p < 10^{-18}$; Figure 18B).

In the single-sweep data, the FM-feedback spectral model related the variability of the sweep pitch shift to a wider spread activation across the spectral populations during the processing of sweeps encompassing lager frequency gaps. Since this gap is related to the absolute frequency gap $|\Delta f|$ but not to FM direction, the model would qualitatively predict similarly higher variabilities in the sweep pitch shift for trains with the larger $|\Delta f|$. The experimental results show that this is indeed the case[5]. Moreover, analogously to the single-sweep stimuli, the variance of the experimental data was strongly correlated to the width of the model responses ($r_p = 0.60, p = 0.0005$; Figure 18C).

Last, we tested whether the different up/down asymmetry (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 pitch shift magnitude $|\Delta p|$ for up sweeps because, due to the compensation for the delay introduced by the basilar membrane in response to low frequencies, these elicit a more synchronous and stronger peak activation in the auditory nerve [49], resulting in larger feedback currents. Qualitatively, a much weaker asymmetry was expected in the sweep-train data, since the effects of the feedback system are virtually absent during the processing of the ending four fifths of the stimuli.

Modelling results on the up/down asymmetry showed an outstanding resemblance with the empirical data (Figure 18D), fully explaining the observed differences between the two families of stimuli. Note that this is not an obvious result of the model fitting for the single sweep data, as the expected difference between the absolute deviance $f_{\text{perceived}} - \bar{f}$ for up and down sweeps $E[\text{asymm}^{\uparrow\downarrow}] \simeq 24\,\text{Hz}$ is significantly smaller than the average error of the model predictions with respect to the data ($E[\text{error}] \simeq 54\,\text{Hz}$).

5 Conclusion

In this work we used a perceptual effect involving FM and pitch, the sweep pitch shift, to study how these two processes interact with each other. Our results, in contrast with the classical view of FM encoding as a bottom-up process [43], suggested the presence of a predictive feedback modulation stemming from neurons encoding FM direction and targeting neurons encoding the spectral properties of the stimuli. Besides explaining several variants of the sweep pitch shift, the suggested predictive mechanism increases the efficiency

[5] See Supplementary Figure S2, left.

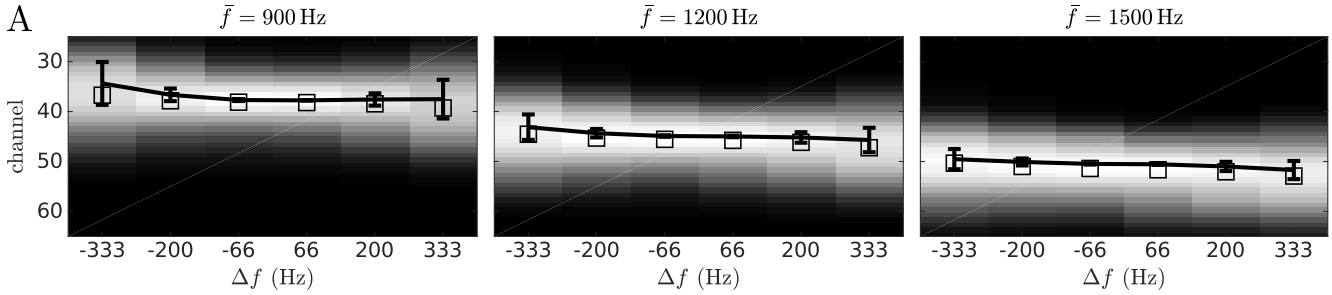

Figure 18: **Predictions of the FM-feedback spectral model for sweep trains.** (A): Shading matrices show the distribution of the activation across channels (y-axis) for different sweep Δf (x-axis). Squares printed over the distributions mark the expected channel $E[k]$ as defined in Equation (4). Solid error bars are estimations of the experimental results in the channel space. The expected value agrees with the experimental data. Moreover, stimuli with larger Δf seem to elicit wider activation distributions than stimuli with smaller Δf, mirroring the generally larger variance observed in the data corresponding to the larger Δf. (B/C): Scatter plots show the correlation between the perceived pitch and the expected channel of the model response (B) and the correlation between the experimental standard error and the variance of the model response (C). Grey dashed lines are the least-squares lines of the pooled data; scatter plots for the single sweep conditions are included for comparison. (D): Error bars show the model predictions of the up/down asymmetry coefficient asymm$\uparrow\downarrow$ (see Equation (1)). Errorbars are estimations of the standard error calculated based on the dispersion of the centroids for different $\bar{f}$ and the variance of the spectral distribution ρ of each condition.

of FM encoding by decreasing its metabolic cost [50], shortening its processing time [51, 52], and enhancing direction selectivity.

5.1 Relation to predictive coding and hierarchical processing strategies

The presence of predictive feedback modulation in the subcortical sensory pathway been shown before in humans [13] and non-human mammals [53]. Previous studies often interpreted it in the context of the predictive coding framework [24, 54, 25], 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 frequency network, and the sweep network) and each layer performs predictions on the sensory input incoming at the immediately lower representation of the hierarchy. In the case of the frequency network, the temporal integration can be interpreted as the prediction that the input's distribution across cochlear channels will change with a much longer time constants as that of the fluctuations introduced by neuronal noise. However, unlike the classical predictive coding microcircuit where predictions and prediction error are kept in separate neural ensembles [55], the frequency and sweep network simultaneously hold a representation that is both, descriptive for their own representation and predictive for the immediately lower representation of the hierarchy.

Combining predictions and representations in the same neural code solves some of the open questions of classical predictive coding architectures recently summarised by Denham and Winkler [56]: 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 hardcoded in the top-down 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 if they were naturally formed through synaptic plasticity after sufficient exposure to the stimuli. Last, 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 and its metabolic cost lower, 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 (as in, for instance, stimulus-specific adaptation [57, 53]).

5.2 Bottom-up pitch models and pitch codes

Two codes of pitch-related information are available in the auditory nerve at early stages of the auditory pathway: 1) the *place-code* or spatial information, produced by the spectral decomposition of the stimuli performed by the basilar membrane; and 2) the *time-code* or temporal information, comprised in the spike timings of the neurons across the auditory nerve that are phase locked to the stimulus waveform (see [58] for a review).

Our simulations showed that the time-code does not suffice to explain the pitch of the 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 [59, 60], coinciding with the duration threshold for accurate pitch discrimination [61]. However, our stimuli presented an average change of $\sim 25\,\text{Hz}$ across four repetitions of their average frequency, making this integration virtually impossible. Thus, the FM-feedback spectral model assumes that the pitch of FM sweeps and pure tones is encoded in a place-code, siding with the idea that spatial information can still play a crucial role in pitch processing.

The bottom-up integration of the place-code, cornerstone of the classical place theories of pitch [32], predicted a sweep pitch shift in the opposite direction of the experimental data; i.e., a shift towards the frequencies expressed at the beginning of the sweep. This is a direct consequence of the global adaptation effects experienced in the auditory nerve after the first few milliseconds of the stimuli [33]. Even without such adaptation, the plain integration proposed by the place models would predict a null sweep pitch shift. Feedback modulation facilitating the encoding of the predictable parts of the sweeps is thus crucial to account for the experimental data.

5.3 Comparison with previous measurements of the sweep pitch shift

Our experimental findings qualitatively replicated the sweep pitch shift effect 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 [2], and that the perceptual bias is monotonically related to the frequency gap Δf [3, 4]. 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 [2] ($m \simeq 0.34$ with transitions of 50 ms) and Nabelek's [3] ($m \simeq 0.32$ with transitions of 40 ms) reports, and significantly higher than Rossi's [4] ($m = 1/6 \simeq 0.17$ 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: Whilst only the very ending segment of the stimulus is facilitated in the short sweeps, since the time to decode FM direction is independent of sweep duration, in a long sweep the facilitation currents would affect a much larger portion of the sound, potentially including frequencies occurring before $\bar{f}$.

The more subtle disagreement with Brady's and Nabelek data has three possible explanations: 1) the differences are a result of the three studies having relatively low sample sizes in comparison with the high inter-subject variability of the effect (see Figure 3); 2) Brady's and Nabelek's studies do not report any participant selection criteria: perhaps the inclusion of listeners that were unable to perform the match resulted in experimental results biased towards a null effect (i.e., towards $m = 0$); and 3) Nabelek and Brady used analogue synthesisers to produce their stimuli, resulting in sweeps with a richer spectral contour than our digital FM-sinusoids, which might have resulted in a weaker effect (Fig 6 in [2] indicates that the spectral properties of the sweep do indeed affect the pitch shift: sweeps of the same duration, spectral scope and Δf produced different sweep pitch shift magnitudes).

5.4 FM encoding and physiological location of the sweep and spectral layers

FM direction selectivity was modelled according to the principles of delayed excitation [42, 22, 18]. Although both delayed excitation and sideband inhibition contribute to direction selectivity in the mammal auditory pathway [39, 40, 15], the two mechanisms are often redundant and yield equivalent results when embedded in a neuronal model [43]. We chose to use delayed excitation alone in order to restrain the number of free parameters of the model.

Although we did not attempt to model FM rate selectivity, the FM-feedback spectral model's DSIs monotonically increased with Δf, a property that could be exploited in further developments of the model to encode modulation rate. FM rate encoding has been reported in mice [15, 21], rats [19] and more extensively in bats (e.g., [62, 40]).

The earliest neural centre within the auditory pathway showing FM direction selectivity in mammals is the inferior colliculus [18, 15, 16, 17], although subsequent nuclei (medial geniculate body [18, 19] and auditory cortex [20, 21, 22, 16, 23]) 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. Thus, the putative physiological location of the mechanisms described here remains an open question.

In this work we have harnessed a well-established perceptual phenomenon to inform a model of FM direction encoding. We have shown that neither phenomenological nor mechanistic bottom-up models of auditory processing are able to explain the experimental data. We concluded that FM direction-selective neurons at a higher stage of the auditory processing hierarchy must alter the way that spectral information is encoded. The main contribution of this work is a specific theory of how this feedback modulation might be ex-

erted. Given the paramount role played by fast FM-sweeps in speech, the predictive mechanisms described here could be part of a larger hierarchical network responsible for the encoding of speech sounds in the human auditory pathway.

Acknowledgments

This research was funded by the ERC Consolidator Grant SENSOCOM 647051. he authors would also like to thank Shih-Cheng Chien for his enlighted suggestions during the writting of the manuscript.

References

[1] R. D. Kent and Y. Kim, "Acoustic Analysis of Speech," in *The Handbook of Clinical Linguistics* (M. J. Ball, M. R. Perkins, N. Mller, and S. Howard, eds.), pp. 360–380, Oxford, UK: Blackwell Publishing Ltd., mar 2008.

[2] P. T. Brady, A. S. House, and K. N. Stevens, "Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency," *Journal of the Acoustical Society of America*, vol. 33, no. 10, pp. 1357–1362, 1961.

[3] I. Nabelek, A. Nabelek, and I. J. Hirsh, "Pitch of Short Tone Bursts of Changing Frequency," *The Journal of the Acoustical Society of America*, vol. 45, pp. 293–293, jan 1970.

[4] M. Rossi, "La perception des glissandos descendants dans les contours prosodiques," *Phonetica*, vol. 35, no. 1, pp. 11–40, 1978.

[5] C. D'Alessandro and M. Castellengo, "The pitch of short-duration vibrato tones," *Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1617–1630, 1994.

[6] C. D'Alessandro, S. Rosset, and J. P. Rossi, "The pitch of short-duration fundamental frequency glissandos.," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2339–48, 1998.

[7] A. de Cheveigné, "Pitch Perception Models," in *Pitch: Neural Coding and Perception* (C. J. Plack, R. R. Fay, A. J. Oxenham, and A. N. Popper, eds.), ch. 6, pp. 169–233, Springer New York, 2005.

[8] M. S. a. Zilany, I. C. Bruce, and L. H. Carney, "Updated parameters and expanded simulation options for a model of the auditory periphery," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 283–286, 2014.

[9] R. Meddis and L. O'Mard, "A unitary model of pitch perception.," *The Journal of the Acoustical Society of America*, vol. 102, pp. 1811–1820, sep 1997.

[10] R. Meddis and L. P. O'Mard, "Virtual pitch in a computational physiological model," *The Journal of the Acoustical Society of America*, vol. 120, no. 6, p. 3861, 2006.

[11] E. Balaguer-Ballester, S. L. Denham, and R. Meddis, "A cascade autocorrelation model of pitch perception.," *The Journal of the Acoustical Society of America*, vol. 124, pp. 2186–95, oct 2008.

[12] S. Shamma and J. Fritz, "Adaptive auditory computations," *Current Opinion in Neurobiology*, vol. 25, pp. 164–168, 2014.

[13] N. Suga, "Tuning shifts of the auditory system by corticocortical and corticofugal projections and conditioning," *Neuroscience and Biobehavioral Reviews*, vol. 36, no. 2, pp. 969–988, 2012.

[14] B. Hu, "Functional organization of lemniscal and nonlemniscal auditory thalamus," *Experimental Brain Research*, vol. 153, no. 4, pp. 543–549, 2003.

[15] H.-R. A. P. Geis and J. G. G. Borst, "Intracellular responses to frequency modulated tones in the dorsal cortex of the mouse inferior colliculus," *Frontiers in Neural Circuits*, vol. 7, pp. 2002–2016, feb 2013.

[16] A.-A. Li, A.-Y. Zhang, Q.-C. Chen, and F.-J. Wu, "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]*, vol. 62, no. 3, pp. 210–8, 2010.

[17] S. R. Hage and G. Ehret, "Mapping responses to frequency sweeps and tones in the inferior colliculus of house mice," *European Journal of Neuroscience*, vol. 18, no. 8, pp. 2301–2312, 2003.

[18] R. I. Kuo and G. K. Wu, "The Generation of Direction Selectivity in the Auditory System," *Neuron*, vol. 73, no. 5, pp. 1016–1027, 2012.

[19] B. Lui and J. R. Mendelson, "Frequency modulated sweep responses in the medial geniculate nucleus," *Experimental Brain Research*, vol. 153, pp. 550–553, dec 2003.

[20] J. B. Issa, B. D. Haeffele, E. D. Young, and D. T. Yue, "Multiscale mapping of frequency sweep rate in mouse auditory cortex," *Hearing Research*, vol. 344, pp. 207–222, 2016.

[21] M. Trujillo, M. M. Carrasco, and K. Razak, "Response properties underlying selectivity for the rate of frequency modulated sweeps in the auditory cortex of the mouse," *Hearing Research*, vol. 298, pp. 80–92, 2013.

[22] C.-q. Ye, M.-m. Poo, Y. Dan, and X.-h. Zhang, "Synaptic mechanisms of direction selectivity in primary auditory cortex," *Journal of Neuroscience*, vol. 30, no. 5, pp. 1861–1868, 2010.

[23] L. I. Zhang, A. Y. Y. Tan, C. E. Schreiner, and M. M. Merzenich, "Topography and synaptic shaping of direction selectivity in primary auditory cortex," *Nature*, vol. 424, pp. 201–205, 2003.

[24] D. Mumford, "On the computational architecture of the neocortex II: The role o f cortico-cortical loops," *Biological Cybernetics*, vol. 66, pp. 241–251, jan 1992.

[25] K. Friston, "A theory of cortical responses.," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 360, pp. 815–36, apr 2005.

[26] J. t. Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation*. Cambridge University Press, 1990.

[27] H. Luo, A. Boemio, M. Gordon, and D. Poeppel, "The perception of FM sweeps by Chinese and English listeners," *Hearing Research*, vol. 224, no. 1-2, pp. 75–83, 2006.

[28] M. Gordon and D. Poeppel, "Inequality in identification of direction of frequency change (up vs. down) for rapid frequency modulated sweeps," *Acoustics Research Letters Online*, vol. 3, pp. 29–34, jan 2002.

[29] J. P. Madden and K. M. Fire, "Detection and discrimination of frequency glides as a function of direction, duration, frequency span, and center frequency," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2920–2924, 1997.

[30] M. J. Collins and J. K. Cullen, "Temporal integration of tone glides," *The Journal of the Acoustical Society of America*, vol. 63, pp. 469–473, feb 1978.

[31] S. Uppenkamp, S. Fobel, and R. Patterson, "The effects of temporal asymmetry on the detection and perception of short chirps," *Hearing research*, vol. 158, pp. 71–83, 2001.

[32] H. L. F. von Helmholtz, *On the Sensations of Tone*. Dover Publications, 1954.

[33] M. S. A. Zilany, I. C. Bruce, P. C. Nelson, and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve : Long-term adaptation with power-law dynamics," *The Journal of the Acoustical Society of America.*, vol. 126, no. 5, pp. 2390–2412, 2009.

[34] K.-F. Wong and X.-J. Wang, "A recurrent network mechanism of time integration in perceptual decisions.," *The Journal of Neuroscience*, vol. 26, no. 4, pp. 1314–1328, 2006.

[35] G. Deco, A. Ponce-Alvarez, D. Mantini, G. L. Romani, P. Hagmann, and M. Corbetta, "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*, vol. 33, pp. 11239–52, jul 2013.

[36] N. Brunel and X. J. Wang, "Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition," *Journal of Computational Neuroscience*, vol. 11, no. 1, pp. 63–85, 2001.

[37] N. L. Golding, *Neuronal Response Properties and Voltage-Gated Ion Channels in the Auditory System*, pp. 7–41. New York, NY: Springer New York, 2012.

[38] L. Wiegrebe and R. Meddis, "The Representation of Periodic Sounds in Simulated Sustained Chopper Units of the Ventral Cochlear Nucleus," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1207–1218, 2004.

[39] A. J. Williams and Z. M. Fuzessery, "Multiple mechanisms shape FM sweep rate selectivity: complementary or redundant?," *Frontiers in Neural Circuits*, vol. 6, pp. 1–14, 2012.

[40] Z. M. Fuzessery, K. A. Razak, and A. J. Williams, "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*, vol. 197, no. 5, pp. 615–623, 2011.

[41] J. A. Morrison, R. Valdizón-Rodríguez, D. Goldreich, and P. A. Faure, "Tuning for rate and duration of frequency-modulated sweeps in the mammalian inferior colliculus," *Journal of Neurophysiology*, vol. 120, pp. 985–997, sep 2018.

[42] K. A. Razak and Z. M. Fuzessery, "Facilitatory Mechanisms Underlying Selectivity for the Direction and Rate of Frequency Modulated Sweeps in the Auditory Cortex," *Journal of Neuroscience*, vol. 28, pp. 9806–9816, sep 2008.

[43] S. Skorheim, K. Razak, and M. Bazhenov, "Network models of frequency modulated sweep detection," *PLoS ONE*, vol. 9, no. 12, pp. 1–25, 2014.

[44] S. Ostojic and N. Brunel, "From spiking neuron models to linear-nonlinear models," *PLoS Computational Biology*, vol. 7, no. 1, p. e1001056, 2011.

[45] K. J. Friston and C. J. Price, "Dynamic representations and generative models of brain function," *Brain Research Bulletin*, vol. 54, no. 3, pp. 275–285, 2001.

[46] P.-A. Salin and J. Bullier, "Corticocortical Connections in the Visual System : Structure and Function," *Physiological reviews*, vol. 75, no. 1, pp. 107–154, 1995.

[47] J. P. Rauschecker, "Cortical control of the thalamus : top-down processing and plasticity," *Nature Neuroscience*, vol. 1, no. 3, pp. 179–180, 1998.

[48] D. A. McCormick, B. W. Connors, J. W. Lighthall, and D. A. Prince, "Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex," *Journal of neurophysiology*, vol. 54, no. 4, pp. 782–806, 1985.

[49] A. Rupp, S. Uppenkamp, A. Gutschalk, R. Beucker, R. D. Patterson, T. Dau, and M. Scherg, "The representation of peripheral neural activity in the middle-latency evoked field of primary auditory cortex in humans," *Hearing Research*, vol. 174, no. 1-2, pp. 19–31, 2002.

[50] Z. Alexandre, S. Oleg, and P. Giovanni, "An information-theoretic perspective on the costs of cognition," *Neuropsychologia*, vol. 123, pp. 5–18, 2018.

[51] S. Jaramillo and A. M. Zador, "The auditory cortex mediates the perceptual effects of acoustic temporal expectation," *Nature Neuroscience*, vol. 14, no. 2, pp. 246–253, 2011.

[52] L. Mazzucato, G. Camera, and A. Fontanini, "Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli," *Nature Neuroscience*, vol. 22, pp. 787–796, 2019.

[53] M. S. Malmierca, L. A. Anderson, and F. M. Antunes, "The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: a potential neuronal correlate for predictive coding.," *Frontiers in systems neuroscience*, vol. 9, p. 19, 2015.

[54] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, pp. 79–87, jan 1999.

[55] A. M. Bastos, W. M. Usrey, R. a. Adams, G. R. Mangun, P. Fries, and K. J. Friston, "Canonical microcircuits for predictive coding.," *Neuron*, vol. 76, pp. 695–711, nov 2012.

[56] S. L. Denham and I. Winkler, "Predictive coding in auditory perception: challenges and unresolved questions," *European Journal of Neuroscience*, jan 2018.

[57] N. Ulanovsky, L. Las, and I. Nelken, "Processing of low-probability sounds by cortical neurons," *Nature Neuroscience*, vol. 6, no. 4, pp. 391–398, 2003.

[58] A. J. Oxenham, "Revisiting place and temporal theories of pitch," *Acoustical Science and Technology*, vol. 34, no. 6, pp. 388–396, 2013.

[59] A. Tabas, M. Andermann, V. Schuberth, H. Riedel, E. Balaguer-Ballester, and A. Rupp, "Modeling and MEG evidence of early consonance processing in auditory cortex," *PLoS Computational Biology*, vol. 15, no. 2, pp. 1–28, 2019.

[60] L. Wiegrebe, "Searching for the time constant of neural pitch extraction," *The Journal of the Acoustical Society of America*, vol. 109, pp. 1082–1091, mar 2001.

[61] K. Krumbholz, R. D. Patterson, A. Seither-Preisler, C. Lammertmann, and B. Lütkenhöner, "Neuromagnetic evidence for a pitch processing center in Heschl's gyrus," *Cerebral Cortex*, vol. 13, no. 7, pp. 765–772, 2003.

[62] J. X. Gittelman and N. Li, "FM velocity selectivity in the inferior colliculus is inherited from velocity-selective inputs and enhanced by spike threshold," *Journal of Neurophysiology*, vol. 106, no. 5, pp. 2399–2414, 2011.

Supplementary figures

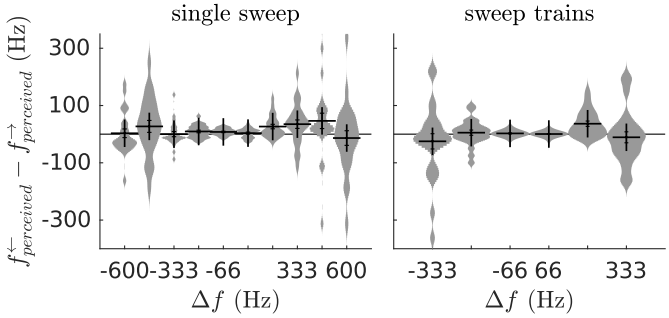

Figure S19: **Effect of the presentation order on Δp.** Kernel density estimations of the difference between the perceived pitch evaluated when the sweep was presented before the probe tone $f^{\leftarrow}_{\text{perceived}}$ and the perceived pitch evaluated when the probe tone was presented before 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.

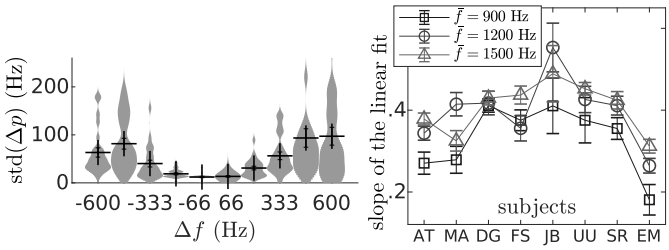

Figure S20: **Variance of the perceived pitch in sweep trains.** Left: Kernel density estimations of the intra-subject standard deviation of the sweep train pitch shift magnitude Δp, plotted separately for the different frequency gaps Δf. Each point in the distributions corresponds to the standard deviation 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 variance is monotonically correlated to the absolute gap $|\Delta f|$ ($r_s = 0.75$, $p < 10^{-27}$). Right: Slope m of the linear fit $f_{\text{perceived}} \sim \bar{f} + m\,\Delta f$ for the sweep train stimuli, independently for each of the eight subjects; error bars mark the 95% confidence intervals of the estimations.

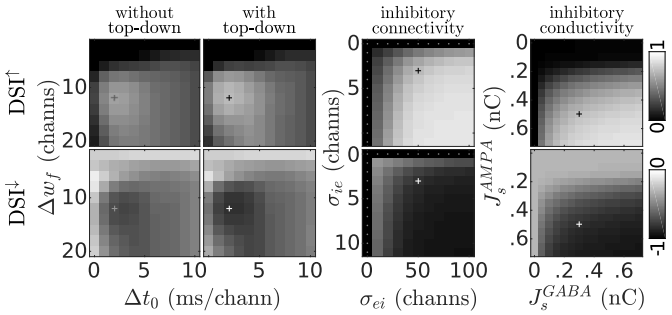

Figure S21: **Direction selectivity dependence with the model parametrisation.** Shading matrices show the average DSI in the up (top; white means $\text{DSI}^{\uparrow} = 1$) and down (bottom; black means $\text{DSI}^{\downarrow} = -1$) network for different locations of the parameter space. Averages were performed across different $\bar{f}$ and $|\Delta f|$. Unless stated otherwise, parameters not varied in the matrices correspond to the values described in Table II in the main text. The first two columns show the dependence of the DSI on the baseline delay δt_0 and the width of the $\Delta_{\omega f}$ with (left) and without (right; here we set $J^{\text{NMDA}} = 0$) top-down connections. The third column shows the dependence of the DSI on the width of the excitatory-to-inhibitory (σ_{ei}) and of the inhibitory-to-excitatory (σ_{ie}) connectivity matrices. The rightmost column shows the dependence of the DSI with the inhibitory-to-excitatory conductivity J_s^{AMPA} and the excitatory-to-inhibitory J_s^{GABA} effective conductivities. Black/white crosses in the parameter space signal the final parametrization. Expectedly, DSI increases monotonically with the amount and extend of mutual inhibition between the up and down networks but, since an overly wide excitatory-to-inhibitory connection would prevent the network from decoding simultaneous up and down sweeps occurring at different frequency ranges, we kept $\sigma_{ei} \lesssim 0.05\, N$.