Sequential dependencies in pitch judgments

Dorothée Arzounian,[1,a)] Mathilde de Kerangal,[2] and Alain de Cheveigné[1,b)]

[1]*Laboratoire des Systèmes Perceptifs, Département d'Etudes Cognitives, Ecole normale supérieure, PSL Research University, CNRS, 29 rue d'Ulm, Paris, 75005, France*

[2]*The Ear Institute, University College London, 332 Grays Inn Road, Kings Cross, London, WC1X 8EE, United Kingdom*

(Received 16 December 2016; revised 19 October 2017; accepted 19 October 2017; published online 16 November 2017)

Studies that measure pitch discrimination relate a subject's response on each trial to the stimuli presented on that trial, but there is evidence that behavior depends also on earlier stimulation. Here, listeners heard a sequence of tones and reported after each tone whether it was higher or lower in pitch than the previous tone. Frequencies were determined by an adaptive staircase targeting 75% correct, with interleaved tracks to ensure independence between consecutive frequency changes. Responses for this specific task were predicted by a model that took into account the frequency interval on the current trial, as well as the interval and response on the previous trial. This model was superior to simpler models. The dependence on the previous interval was positive (assimilative) for all subjects, consistent with persistence of the sensory trace. The dependence on the previous response was either positive or negative, depending on the subject, consistent with a subject-specific suboptimal response strategy. It is argued that a full stimulus + response model is necessary to account for effects of stimulus history and obtain an accurate estimate of sensory noise.

 https://doi.org/10.1121/1.5009938

[VMR] Pages: 3047–3057

I. INTRODUCTION

Psychophysics attempts to relate a physical dimension of a stimulus (for example, fundamental frequency) to a psychological dimension (for example, pitch) using behavioral methods. Some individuals possessing absolute pitch are capable of accurately identifying the pitch of a musical tone without any preceding reference, but a majority of listeners appreciate tone pitches in a melody by judging their distance relative to previous tones. The physical dimension is then *frequency change* (or ratio) between tones, and the psychological dimension *pitch change* (or interval). A rich literature has probed experimentally the limits of our ability to discriminate small frequency differences (Harris, 1952; Rosenblith and Stevens, 1953; Nordmark, 1968; Moore, 1973; Jesteadt and Sims, 1975; Moore and Glasberg, 1989; Sek and Moore, 1995; Matthews and Stewart, 2008; Dai and Micheyl, 2011; Micheyl *et al.*, 2012). In these studies, listeners typically make judgments on pairs (or triplets, or quadruplets) of tones, and the accuracy of their judgments is assessed as a function of the frequency difference between the tones in the trial. Discrimination thresholds are then interpreted as reflecting the resolution of the sensory representation by which those tones are coded. However, there is evidence that judgments also depend on the history of stimuli that *precede* each trial. In an extreme case, Chambers *et al.* (2017) found that for certain ambiguous stimuli (two successive Shepard tones separated by a tritone interval) responses depended almost entirely on the history of prior stimulation (Chambers and Pressnitzer, 2014; Chambers *et al.*, 2017). To the extent that history effects are uncontrolled, they contribute an unwanted source of variance when measuring the psychophysical relation between the stimulus and the response that it evokes. It is thus of interest to better understand these effects and model their influence.

It is well known that "roving" the overall frequencies of tone pairs can lead to poorer accuracy when judging the frequency difference within each pair. In 2-Interval Forced Choice (2-IFC) designs, where listeners are instructed to identify which of two tones that are presented successively is higher in pitch, frequency discrimination thresholds, sometimes termed Frequency Difference Limens, are lower if the comparison involves a fixed reference tone that appears in all trials, compared to a situation in which the tones can be taken from a set of multiple, spaced frequencies (Harris, 1952; Bull and Cuddy, 1972; Jesteadt and Bilger, 1974) or drawn randomly from a large, continuous frequency range (Demany and Semal, 2005; Mathias *et al.*, 2010; Nahum *et al.*, 2010), two methods that can be referred to as frequency roving. In addition, when a fixed reference is used, thresholds appear to depend on the position (first or second interval) of the reference within trials (Nahum *et al.*, 2010; Raviv *et al.*, 2014). Such effects are often interpreted as reflecting a difference in perceptual *sensitivity* between different experimental configurations, but it has also been argued that they can be accounted for by a history-dependent perceptual *bias* (Raviv *et al.*, 2014).

Detecting a history-dependent bias is important for the study of perceptual mechanisms. First, ignoring it can lead to overestimating the amount of "internal noise" in the sensory representation (Fründ *et al.*, 2014). Second,

a)Electronic mail: dorothee.arzounian@ens.fr

b)Also at: The Ear Institute, University College London, 332 Grays Inn Road, Kings Cross, London, WC1X 8EE, United Kingdom.

the bias itself may inform us about how the brain accumulates perceptual evidence and combines ongoing sensory input with past traces. Last, taking into account inter-trial dependencies might improve the analyses of neural correlates of decision making (Lages and Jaworska, 2012). To better capture these dependencies we made certain methodological choices that differ from those made in previous studies.

In past experiments probing context effects in pitch perception (Ruusuvirta *et al.*, 2008; Raviv *et al.*, 2012, 2014), the subject was presented with a sequence of tone pairs, and answered after each *tone pair* which tone had a higher pitch [two interval two alternative forced choice (2I-2AFC)]. In contrast, we used a one-tone-per-trial procedure in which subjects were presented with a sequence of tones and answered after each *tone* whether it was higher in pitch than the previous tone (sliding 2AFC). The purpose of the new procedure was to ensure a homogeneous sequence of prior tones, in contrast to the classic procedure where the sequence included both reference tones (the first of a pair) and comparison tones (the second). Prior to this study, we established that the new procedure yields similar discrimination thresholds as the old (Arzounian *et al.*, 2017). The task has an analog in melody perception, where each note anchors both the preceding and following interval, or speech intonation where each segment participates in the pitch transitions that precede and follow it.

In a standard adaptive procedure, the interval size for a trial is adjusted based on the correctness of the response to the previous trial. This introduces a strong serial correlation in the sequence of interval sizes, limiting the range of intervals that can precede a trial. To ensure a wider range of history, we interleaved multiple independent tracks, such that the interval on each trial was determined by the response to a trial several steps in the past.

Previous studies of history effects in perception have analyzed these effects as resulting from the recent history of *stimuli* (e.g., Ruusuvirta *et al.*, 2008; Raviv *et al.*, 2012, 2014; Alais *et al.*, 2015; Taubert *et al.*, 2016a, 2016b). However, a subject's judgment is also known to depend on previous *responses* (Fründ *et al.*, 2014). In this study, we analyze our data using a set of models that incorporate stimulus history, response history, or both.

Using this combination of experimental and analysis methodology, we assess the question of history effects on sequential pitch judgments. In brief, we found effects of both stimulation and response, the weight of each factor being subject-dependent. Estimates of internal noise in the model that incorporates these factors were smaller (and arguably more accurate) than those obtained with a history-blind model.

II. METHODS

A. Participants and procedure

Fourteen subjects, seven male and seven female, aged between 19 and 29 years, with no self-reported hearing impairment or history of neurological or psychiatric disorder, participated in the experiment. Among them, seven were inexperienced, two had prior experience in non-auditory psychophysical studies, three had prior experience in psycho-acoustic studies, and two had performed a similar task in a previous study. All gave written, informed consent prior to participation and received a compensation of 20€ per hour of their time. The protocol was approved by the ethics review board of Paris Descartes University (CERES 2013-11).

B. Apparatus, stimuli and task

Participants sat in a double-shielded experimental booth. Visual displays and auditory stimuli were generated by MATLAB (version 2012a). Written instructions and fixation cross were displayed on a computer LCD screen standing in the outside of the booth and visible from the inside through a window. Auditory stimuli were presented through insert earphones (E-A-R-TONE® 3A) at a comfortable level, similar for all participants. As part of a study on the effects of brain state on performance, electroencephalography was recorded using an Active-Two (BioSemi) system with 72 channels (64 channels positioned according to the standard 10/20 layout + 8 additional channels positioned at M1, M2, IO1, IO2, SO1, SO2, EO1, and EO2), sampled at 2048 Hz. The analysis of these data is not reported here.

Auditory stimuli were 100-ms pure tones with 10-ms cosine onset- and offset-ramps. Starting from $f_0 = 1000$ Hz, the frequency varied from one tone to the next with random direction and a step size $|\Delta f|$ determined according to an adaptive procedure (see below). A random walk can produce extreme values; to avoid this situation the probability of an up transition (0.5 at 1000 Hz) decreased linearly in frequency by a factor 0.0003/Hz, so that frequency remained in a region near 1000 Hz (see Sec. III). In this region, up and down transitions were approximately equiprobable.

After each tone, the participant was requested to indicate the direction of pitch change (either "upward" or "downward") by pressing one of two computer keyboard keys. The reaction time on each trial was recorded. The key press triggered the onset of the next tone after an interval of 500 ms.

Before the main block analyzed here, participants were trained on the task with visual feedback (on 60 trials if they were already familiar with the task, on 120 trials if they were not). They then performed two short blocks (120 trials) without feedback, not analyzed here. The main, final block comprised 1080 trials without feedback. Participants were told they could take short breaks when needed by simply holding the response of the current trial until they were ready to continue (such trials were then later excluded from analyses).

C. Multi-track adaptive procedure

The size of the relative frequency step $|\Delta f/f|$ was 10% in the first trial and was then adjusted trial-by-trial according to a weighted up-down procedure (Kaernbach, 1991) with step size limited to at most 30%. During an initial phase, the step size for each trial depended on the success of the previous trial. After a minimum of 20 trials and one reversal, the rule was then changed so that the step size depended on the response four trials in the past, yielding four independent interleaved adaptive tracks (Leek *et al.*, 1991). For an incorrect response, the step size was increased by a factor of 2, otherwise it was decreased by a factor of $\sqrt[3]{2}$. Thus all tracks targeted a 75%-correct performance (Kaernbach, 1991). The adaptive procedure ensures a dense sampling of the psychometric function in the vicinity of the nominal threshold (Dai, 1995), and interleaving reduces the short-term serial correlation of $|\Delta f|$ typically induced by adaptive procedures, thus providing a more balanced distribution of frequencies and step sizes preceding each trial.

D. Analysis of behavioral data

Data were analyzed by fitting to them a series of models of increasing complexity. Previous studies of sequential history effects in perception considered only the *sign* of the relevant stimulus feature in the previous trial (Ruusuvirta *et al.*, 2008; Alais *et al.*, 2015; Taubert *et al.*, 2016b). Here, the preceding signed interval was included as a linear regressor, together with response history, and a fixed bias. Previous studies averaged data over subjects so the analysis could only reveal effects that are consistent across individuals, whereas we fit subject-specific models. We model the probability of reporting an upward change on each trial, rather than the probability of a "correct" response as in many previous studies (Rosenblith and Stevens, 1953; Moore, 1973; Moore and Glasberg, 1989; Sek and Moore, 1995; Dai and Micheyl, 2011). Standard model comparison methods are used to assess the significance of the contribution of each parameter to the model.

1. Models

All models tested here assume the choice probability P of reporting an upward pitch change to be a psychometric function of the form

$$P = \phi_\sigma(X), \tag{1}$$

where ϕ_σ is the cumulative normal distribution function with mean 0 and standard deviation σ, and the decision variable X is determined by the stimulus and, possibly, by an invariant bias and/or by stimulus or response history. This form is in accordance with the framework of Signal Detection Theory, assuming that the internal representation of the frequency change is given by the sum of X and some internal noise that is normally distributed with mean 0 and standard deviation σ. The parameter σ determines discrimination sensitivity.

In the simplest model, which we'll refer to as Baseline model, the value of X at trial n is purely determined by the frequencies f_{n-1} and f_n of the last two tones,

$$X_n = s_n, \tag{2}$$

where $s_n = 12\log_2(f_n/f_{n-1})$ is expressed in semitone units. This first model corresponds to the behavior of an ideal listener in the sense that choice probability only depends on the task-relevant attribute of the stimulation, with the probability being 0.5 when $s_n = 0$, i.e., when consecutive tones have exactly the same frequency. In this ideal case, the discrimination threshold is determined by σ only, and can be calculated as $\phi_\sigma^{-1}(0.75)$, that is $\sim 0.67\sigma$, for a threshold at 75% correct. The parameter σ is assumed to reflect the frequency resolution of the sensory representation.

This Baseline model can be extended by including other potential contributions to the probability of an upward change report, such as the frequency ratio on the previous trial, the response to the previous trial, or a uniform bias

$$X_n = (1-\alpha)\, s_n + \alpha\, s_{n-1} + \beta\, r_{n-1} + b, \tag{3}$$

where r_{n-1} is a binary variable coding for the response in trial $n-1$ (1 in case of an upward report, -1 in case of a downward report), α quantifies the relative influence of the previous trial's frequency change, β weights the contribution of the previous trial's response, and b represents a systematic bias. Figure 1 represents the relations between variables and parameters of this Full model. A non-zero bias ($b \neq 0$) reflects a tendency to report more upward changes (or more downward changes) regardless of the stimuli presented. A non-zero response history parameter ($\beta \neq 0$) might reflect a deliberate or unconscious adaptive response strategy, whereas a non-zero stimulus history parameter ($\alpha \neq 0$) might represent an assimilative ($\alpha > 0$) or else contrastive ($\alpha < 0$) sensory dependency on prior stimulation (Raviv *et al.*, 2012). These contributions are illustrated in Fig. 1.

In addition to the Baseline and Full models, we considered three partial models according to which the parameters of Eq. (3) are non-zero. These models are: *systematic Bias* (B), *systematic Bias + Prior Stimulus* (BPS), and *systematic Bias + Prior Response* (BPR) (Table I). Like the Full model, the last two are history-dependent. It should be noted that the factors prior interval s_{n-1} and prior response r_{n-1} are mutually dependent, as the subject is more likely to have reported an upward change after a positive step. Consequently, with $\beta = 0$, the αs_{n-1} term would indirectly capture a contribution of the preceding response, leading to an incorrect estimate of stimulus history effects. Similarly with $\alpha = 0$, the βr_{n-1} term would capture some contribution of the previous stimulus.

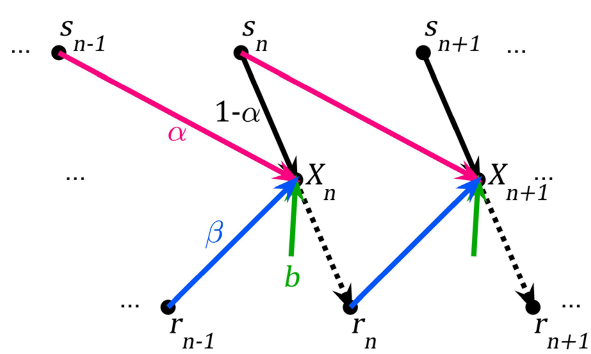

FIG. 1. (Color online) Structure of the Full model. Solid arrows represent the additive contributions of current (s_n) and previous (s_{n-1}) intervals, previous response (r_{n-1}) and choice bias (b) to the decision variable (X_n). Parameters α and β weight the relative contributions of previous interval and previous response, respectively. The dashed arrow symbolizes the stochastic dependence of the response r_n on X_n.

TABLE I. Models. The five models differ based on which of the additive terms in Eq. (3) are assumed to be zero. All models have in common a free parameter σ corresponding to the standard deviation of internal noise. Each model has between 0 and 3 additional free parameters: b representing systematic bias, α weighting the contribution of the previous interval to the decision, and β weighting the contribution of the previous report. Non-free parameters are set to 0. Models in which $\alpha \neq 0$ or $\beta \neq 0$ are history-sensitive.

		Parameters			
	Model	σ	b	α	β
History-Insensitive	Baseline	free	0	0	0
	B	free	free	0	0
History-Sensitive	BPS	free	free	free	0
	BPR	free	free	0	free
	Full	free	free	free	free

Obviously, these constitute only a subset of plausible models. In particular, we consider only linear dependencies [Eq. (3)], we ignore the possibility that parameters might vary over the session, for example, due to a change in the subject's strategy, and we ignore potential contributions of earlier history (trials $n-2$, etc.).

2. Data analysis

Thresholds were computed in a first analysis estimating the classic psychometric function relating probability of a *correct answer* to the logarithm of the absolute value $|\Delta f/f|$ of the relative frequency difference (Dai and Micheyl, 2011). This accuracy psychometric function is implicit in most studies that estimate a discrimination threshold using adaptive methods (Levitt, 1971; Kaernbach, 1991). A logistic function varying between 50% (chance level performance) and 100% correct was fit to individual subjects' data and individual discrimination thresholds were defined as the interpolated value of $|\Delta f/f|$ yielding 75%-correct accuracy.

As the sign of the frequency difference is not taken into account by the accuracy psychometric function, it is inadequate to highlight choice biases that may affect performance. A second analysis estimated instead the choice psychometric function relating probability of an *"up" report* [Eq. (1)] to the decision variable, which was log frequency ratio in the Baseline model, and some combination of this factor, bias and prior stimulus, and response in the four other models. Each model was fit individually to the behavioral data collected from each participant. Using the notation $y_n = (1 + r_n)/2$ for convenience ($y_n = 1$ if the listener reported an upward change, $y_n = 0$ if they reported a downward change), parameters were estimated with the MATLAB (version R2015b) *glmfit* function performing a generalized linear regression of the binary responses y_n on the predictor variables (s_n, s_{n-1}, r_{n-1}) or a subset of these depending on the model, using a probit link function. For all except the Baseline model, predictors included an additional constant term to capture the systematic bias (b); 95%-confidence intervals for the free parameters were computed using a bootstrap procedure with 1000 resampling iterations.

The quality of each fit was assessed by three different metrics. First, the Mean Squared Residual (MSR) was computed as

$$\mathrm{MSR} = \frac{1}{N}\sum_{n=1}^{N}\left[\phi_\sigma(X_n) - y_n\right]^2 \tag{4}$$

where N is the number of trials. Second, the Mean Log Likelihood (MLL) was computed as

$$\mathrm{MLL} = \frac{1}{N}\sum_{n=1}^{N}\log\left\{\left[\phi_\sigma(X_n)\right]^{y_n}\left[1-\phi_\sigma(X_n)\right]^{(1-y_n)}\right\}. \tag{5}$$

Last, a Receiver Operating Characteristics (ROC) and the Area Under the ROC curve (AUROC) were computed using MATLAB's *perfcurve* function. The ROC describes the performance of the model when it is used as a binary classifier predicting the subject's response y_n ("up" versus "down") based on X_n. By changing the probability threshold to be reached by $\phi_\sigma(X_n)$ to label a response as "up" rather than "down," one changes the "true positive" (i.e., model predicts "up" response, subjects responds "up") and "false positive" (i.e., model predicts "up" response, subjects responds "down") rates, similarly to what happens in Signal Detection Theory when a subject changes criterion. The AUROC reflects the "sensitivity" of the model, i.e., its capacity to discriminate between the two classes of trials.

Nested models were compared by F-tests with a 5% false-rejection probability (Motulsky and Christopoulos, 2004), with correction for multiple comparisons as needed.

III. RESULTS

Main blocks lasted between 19 and 29 min (mean duration was 24 min), except for one participant who took more than 38 min. Trials in which the reaction time was shorter than 300 ms or exceeded 1500 ms were excluded. This resulted in the exclusion of between 11 and 124 trials depending on the participant, leaving between 956 and 1069 trials to include for model fitting, except for the slowest participant for which 448 trials were excluded leaving 632 for analysis. Frequencies followed a random walk with opposite biases above and below 1000 Hz preventing large deviations from this center frequency (see Sec. II). Resulting frequency distributions had an average (over subjects) center of 1000.9 $\pm$41 Hz [mean $\pm$ standard deviation (s.d.)] and an average (over subjects) s.d. of 89$\pm$74 Hz.

A. Classic discrimination threshold analysis

The accuracy psychometric function relating probability of a correct response to log absolute value of the relative frequency step was fit individually for each subject, and the abscissa at 75% correct was taken as the discrimination threshold. Thresholds are plotted for each subject as open symbols in Fig. 2. On average over subjects, the threshold was 0.10 semitones (geometric mean) with a deviation factor of 2.0 (geometric s.d.).

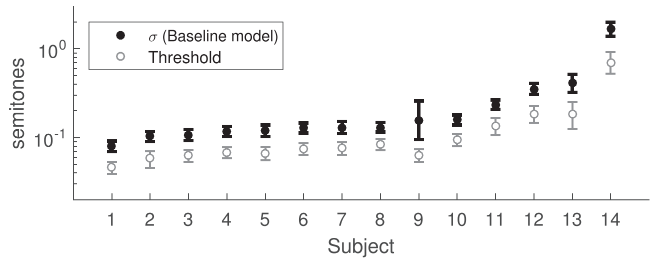

FIG. 2. Estimated noise parameter σ for the Baseline model (closed symbols) derived from a fit of the choice psychometric function and thresholds (open symbols) derived from a fit of the accuracy psychometric function for each subject. Subjects are ordered on the horizontal axis by increasing noise.

B. Baseline model

The choice psychometric function relating probability of an "up" report to log frequency ratio was fit individually for each subject using the Baseline model [Eqs. (1) and (2)], yielding an estimate of the σ parameter for each subject. Values of σ for each subject are plotted as closed symbols in Fig. 2. For convenience, subjects are ordered according to increasing σ. The average value of σ was 0.18 semitones (geometric mean) with a deviation factor of 2.2 (geometric s.d.). Assuming a Gaussian model with zero response bias, discrimination thresholds are expected to be $\sim 0.67\sigma$, and thus these values appear consistent with the previous analysis. The wide scatter of values across subjects (more than an order of magnitude) is typical of previous studies (e.g., Micheyl *et al.*, 2006). In the absence of bias and history effects, discrimination thresholds are assumed to reflect sensory noise, so it is reassuring that the two estimates seem to agree.

C. Full model

The probability of "up" versus "down" reports was also fit using the Full model [Eqs. (1) and (3), $b \neq 0$, $\alpha \neq 0$, $\beta \neq 0$] that includes factors *prior stimulus*, *prior response*, and *systematic bias*, as well as reduced models with subsets of these factors. Adding factors improves the fit, as reflected by the MLL scores plotted in Fig. 3(a). The scores indicate a better fit for the Full model (plotted rightmost, average MLL is -0.46 ± 0.03, mean $\pm$ s.d.) than for the Baseline model (plotted leftmost, average MLL is -0.52 ± 0.03, mean $\pm$ s.d.), with intermediate scores for the reduced models, and similar trends were found for AUROC and MSR measures (not shown). However, an obvious concern is whether the more complete models are justified given their complexity. Pairs of nested models were compared using the F-statistic with a p-value threshold of 0.05, and for each pair the number of subjects for which the more complete model was superior is reported in Fig. 3(b). Detailed statistics can be found in the Appendix. Models BPS and BPR are not compared with each other because they are not nested. According to this analysis, for all factors, a model that includes that factor is better than a model that excludes it for most subjects. The benefit was cumulative, and the Full model was superior to all other models for 13 of the 14 subjects. After applying the Bonferroni correction of

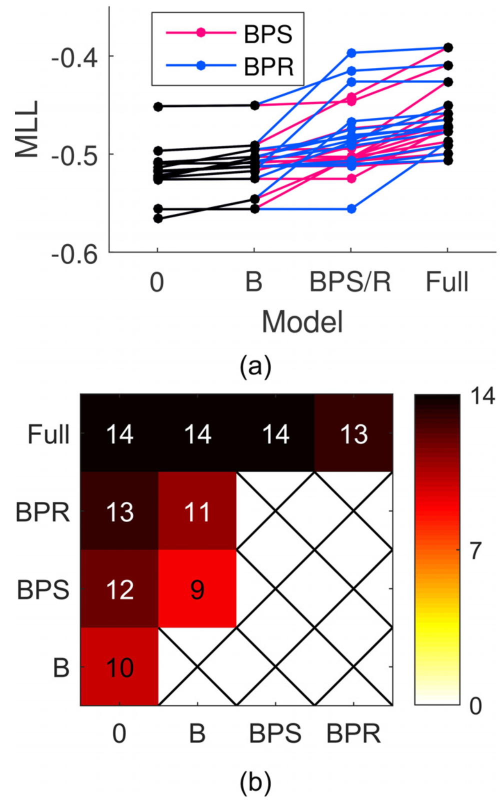

FIG. 3. (Color online) Summary of model quality of fit and model comparisons. Tested models were 0, B, BPS, BPR, and Full. (a) MLL per model. Lines correspond to individual subjects. Models are ordered by increasing complexity on the horizontal axis, MLL of BPS and BPR models have the same number of parameters and are plotted in distinct colors for better readability. (b) Summary of all pair-wise model comparisons. Each box displays the number of participants for which the model on the vertical axis was retained against the model on the horizontal axis after F-test. BPR and BPS models were not compared because they are not nested.

p-value thresholds for multiple comparisons, the Full model was still superior to all others for 12 of the 14 subjects.

Figure 4 shows parameter estimates of the Full model for all subjects (parameters b and β are normalized by the value of σ obtained for this model, for visual convenience). Estimates of σ (geometric mean: 0.12 semitones, geometric s.d.: 2.2) were significantly lower under the Full model than under the Baseline model (based on bootstrap resampling) for nine subjects [Fig. 4(a)]. Smaller values of σ are to be expected as the factors controlled for by the Full model appear as noise in the Baseline model. The bias parameter b [Fig. 4(b)] was significantly different from zero for ten subjects (positive for eight and negative for two). The average normalized bias b/σ was 0.08 ± 0.18 (mean $\pm$ s.d.). The parameter α [Fig. 4(c)] was significantly different from zero for all subjects but one, with a positive value suggesting an *assimilative* effect of prior stimulation. The average value of α was $20\% \pm 8.5\%$ (mean $\pm$ s.d.). The parameter β [Fig. 4(d)] was significant for most subjects, but with a sign that differed between subjects, suggesting different behavior strategies (also suggested by the between-subject differences in systematic bias b). The average value of β/σ was -0.14 ± 0.45 (mean $\pm$ s.d.).

As mentioned earlier, the factors prior stimulus and prior response are mutually dependent but not collinear: if

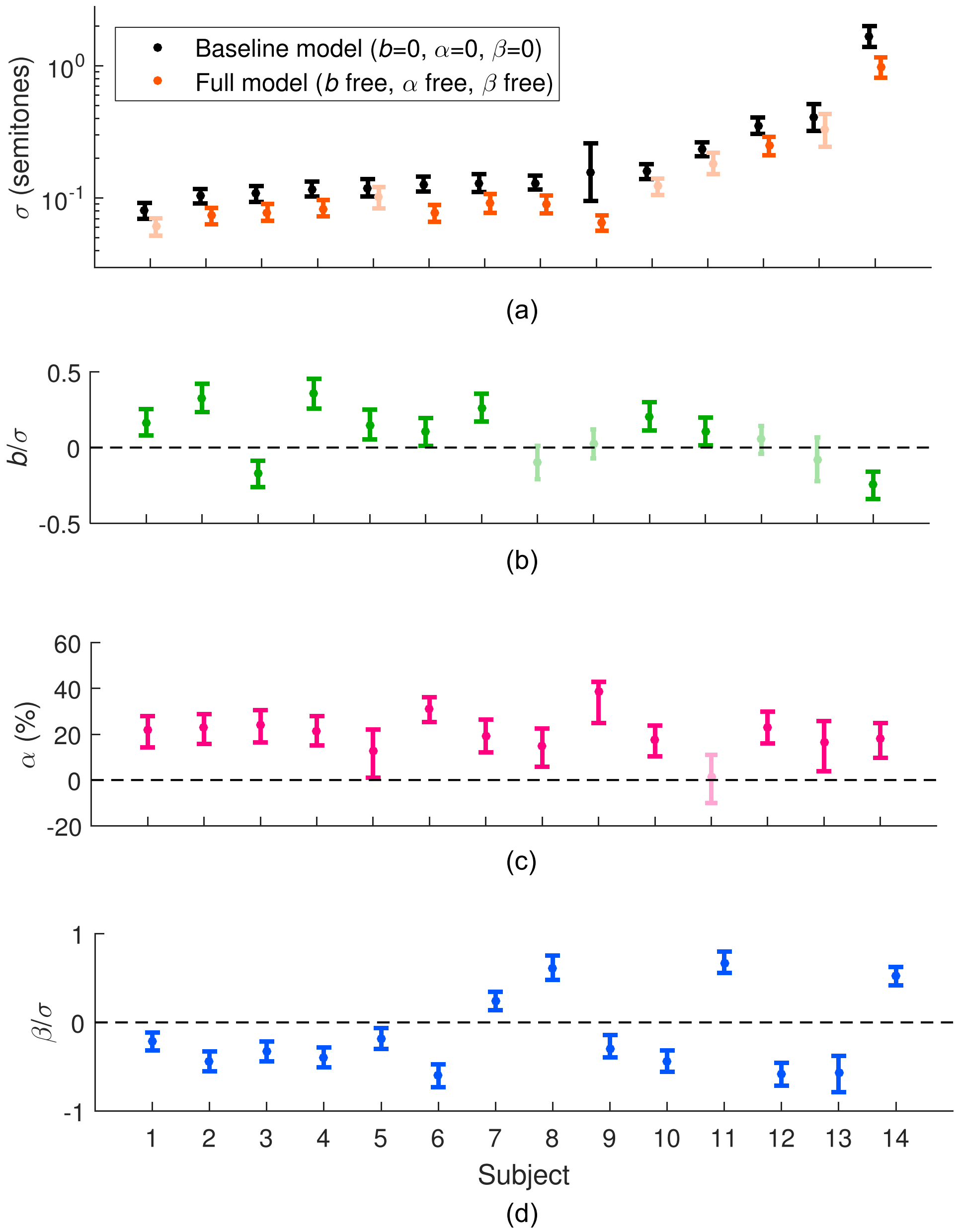

FIG. 4. (Color online) Parameter estimates of the Full model. (a) Noise parameter σ obtained under the Full model (orange), and under the Baseline model (black). (b) Parameter $\boldsymbol{b}$ quantifying systematic choice bias. (c) Parameter α weighting the relative contribution of the previous interval. (d) Parameter β weighting the relative contribution of the previous response. Bars represent bootstrap estimates of 95-% confidence intervals, faded symbols denote no significant difference from the Baseline model. Subjects are ordered by increasing internal noise (σ) as estimated under the Baseline model (same as Fig. 2).

either is removed, the model quality is reduced for most subjects. If one factor is removed, the variability associated to it is captured by the other factor, overestimating its true effect when both factors act in the same direction, underestimating it otherwise. Thus, response history effects should be controlled for, as we do, even in studies where the goal is to measure purely sensory effects (e.g., Raviv *et al.*, 2012). To illustrate this point, Fig. 5 shows the amount by which α is overestimated when the contribution of the prior response is ignored (model BPS) relative to the Full model as a function of normalized β. The parameter is overestimated for β positive, and underestimated otherwise.

IV. DISCUSSION

This study probed the influence on sequential pitch judgments of the history of frequency changes and responses preceding each trial in a sliding 2-AFC task. In most psychophysical studies of pitch perception, the "physical" dimension considered is the interval between tones within a trial, based on the assumption that the subject can ignore previous trials. This is at best an approximation: discrimination thresholds are known to be higher if frequencies are roved, implying that the ability to discriminate two tones within a trial is impaired if the frequencies of previous trials fluctuated over a wide range (Mathias *et al.*, 2010). This might be attributed to the wider distribution of frequencies preceding a trial, for example, because the subject cannot focus on a restricted frequency region. Alternatively, it might be explained by a stimulus context-dependent bias, such as we found in this study, that leads to non-optimal performance and thus elevated thresholds. In the presence of roving, the large frequency interval preceding each trial might more strongly bias the decision on that trial, leading to higher thresholds. Effects of stimulus history on performance in a 2-interval task have recently been

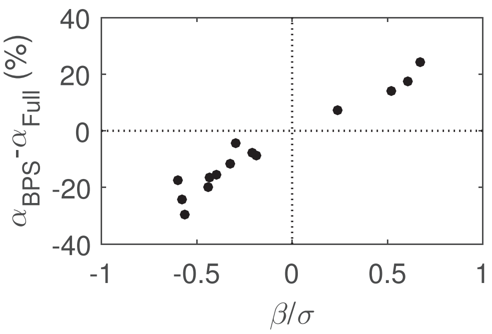

FIG. 5. Error on the α estimate when the contribution of the prior response is ignored. The difference between the estimate of α obtained by a fit of the BPS model and that obtained by a fit of the Full model is plotted against the ratio of β and σ estimates obtained with the Full-model fit. Dots correspond to individual participants.

examined more closely (Nahum *et al.*, 2010) and modeled as an integration of frequencies of prior tones (Raviv *et al.*, 2012, 2014). Our study extends those studies using a different methodology. Of course, one should beware that the effects described here might be specific to this methodology and not apply to more traditional experimental procedures, in particular to 2-interval procedures. However, the agreement of these effects with the roving effect and with the history effects previously found by Raviv *et al.* in 2-interval tasks suggests that similar mechanisms may be at play in these different procedures.

A. Discrimination acuity

The measurement of perceptual limits is a primary goal of psychophysics (Fechner, 1966), and numerous psychophysical studies have been devoted to quantifying frequency discrimination thresholds (Harris, 1952; Rosenblith and Stevens, 1953; Nordmark, 1968; Moore, 1973; Jesteadt and Sims, 1975; Moore and Glasberg, 1989; Sek and Moore, 1995; Matthews and Stewart, 2008; Dai and Micheyl, 2011; Micheyl *et al.*, 2012). According to Signal Detection Theory (Green and Swets, 1966; Macmillan and Creelman, 2005), perceptual limits are determined by the amount of noise within the sensory dimension underlying the task, and with appropriate assumptions (Gaussian noise, no bias) the noise magnitude can be inferred from the measured threshold. The threshold is defined as the abscissa of the point at which the psychometric function, relating frequency step size to percentage correct, reaches a criterion value, for example, 75%. This point can be determined *explicitly* from a psychometric function estimated by fitting the density of correct and incorrect responses to a range of stimuli, or else *implicitly* from the rule associated with an adaptive procedure (Levitt, 1971; Kaernbach, 1991). For the classic correct-response analysis (Sec. II), we chose to estimate thresholds from the psychometric function relating percent correct to $\log(|\Delta f/f|)$ sampled at values chosen by the adaptive procedure. Thresholds that we obtain in this fashion (Fig. 2, open symbols) are consistent with the literature of frequency discrimination (Moore, 1973; Emmerich *et al.*, 1989; Moore and Glasberg, 1989; Sek and Moore, 1995; Micheyl *et al.*, 2006) although markedly smaller than those reported by Nahum *et al.* (2010). The wide spread of values across subjects (more than an order of magnitude) is also consistent with other studies (e.g., Micheyl *et al.*, 2006). Under the assumption of Gaussian sensory noise and no bias, the standard deviation σ of the noise is inferred to be $\sim$1.5 times the threshold at 75%.

For the alternative up-response analysis (Sec. II), the noise magnitude σ is inferred from the slope of the psychometric function relating the density of "up" reports to $\log(f_n/f_{n-1})$. Again assuming Gaussian noise, σ is proportional to the inverse of the slope at 50% of this curve. In the absence of bias the two methods for estimating variance are equivalent, and indeed fitting our data with the model without bias terms leads to estimates of σ (Fig. 2, closed symbols) that are consistent with the thresholds estimated with the classic analysis (ratio close to 1.5).

In the presence of bias, the two approaches are no longer equivalent. With the first approach (psychometric function relating percent correct to the *unsigned* relative frequency difference), sensory noise is overestimated, whereas with the second (psychometric function relating percentage of "up" reports to the *signed* interval), bias contributions appear as opposite effects on rising and falling intervals and can be factored out to obtain an accurate estimate of sensory noise. Estimates of σ obtained in this way are indeed smaller (Fig. 4 top, red symbols) and arguably more reliable than with the first approach. An accurate estimate of sensory noise is important for studies that compare human performance to theoretical limits such as the Gabor tradeoff, (Moore, 1973; Oppenheim and Magnasco, 2013), or to physiological data and models (e.g., Heinz *et al.*, 2001).

B. Effects of stimulus and response history

As pointed out earlier, the existence of stimulus history effects can be surmised from higher thresholds observed with roving in 2-interval tasks. Likewise, the lower thresholds obtained using a fixed standard (e.g., Bull and Cuddy, 1972; Nahum *et al.*, 2010) can be interpreted either as a beneficial "perceptual anchor" effect of a fixed reference (Durlach and Braida, 1969; Matthews and Stewart, 2008; Nahum *et al.*, 2010) or as a deleterious confusing or bias effect of a roving reference (Mathias *et al.*, 2010, 2011). The latter was probed in recent studies (Raviv *et al.*, 2012, 2014).

However, the methodology of prior studies limits their ability to conclude. Studies that evidenced effects of roving by comparing thresholds or percent correct between conditions conflate bias with sensory noise. They cannot distinguish between a history-dependent sensory noise (e.g., a deleterious effect of roving or beneficial effect of a fixed perceptual anchor), or a history-dependent bias (e.g., regression of new sensory traces to the mean of prior traces). Likewise, studies that model effects of stimulus but not response history miss the opportunity to factor out that source of variance and risk producing misleading estimates of sensory history effects due to dependencies between the two factors. In contrast, our methods allow us to control both for stimulus and response history, and indeed we find evidence for both.

We found a significant and positive contribution of stimulus history for all subjects, except one for which the contribution was not distinguishable from zero [Fig. 4(c)]. We also found a contribution of *response* history that was significant for all subjects, and with a sign that depended on the subject [Fig. 4(d)]. A possible explanation is that near threshold subjects adopt a guessing strategy (conscious or unconscious) that takes into account their previous response, and that this strategy is subject-dependent. For example, a subject might try to avoid a series of identical responses, judged unlikely, or else stick to the same decision as was made on the previous trial in the absence of a strong sensory cue to decide otherwise. Systematic bias was also subject-dependent [Fig. 4(b)]. As we pointed out, the factors stimulus and response history are mutually dependent, and the variance of one is likely to be absorbed by the other unless both are included in the model. A model including both factors was superior for all subjects to a model containing only stimulus history (Fig. 3), and omitting response history strongly affected estimated weights α of stimulus history (Fig. 5). Thus, including both factors leads to a more accurate estimate of the contribution of stimulus history.

C. Sensory integration

Once the confounding effect of response history has been controlled for (Full model), the contribution of prior stimulation appears to be positive (assimilative) for all subjects but one for which it did not differ from zero. One interpretation is that an internal representation of the *interval* on that trial is integrated with that of the previous interval. A representation of frequency change by Frequency-Shift Detectors was hypothesized by Demany and Ramos (Demany and Ramos, 2005; Demany and Semal, 2005; Demany *et al.*, 2009, 2011; Carcagno *et al.*, 2011). Another is that internal representations of the two *frequencies* that determine the interval associated with a trial are affected by those of previous tones. These two hypotheses are compatible with the same model of Eq. (3), so we cannot distinguish them on the basis of the data.

Integration across trials might be due to an inability to rapidly discard past sensory traces and follow fast change (sluggishness). Alternatively, it could reflect a mechanism of temporal integration or evidence accumulation designed to counteract sensory noise, or to smooth out irrelevant stimulus fluctuations. Sensory integration is an effective way of reducing noise in stimulus representations when the world tends to remain constant (Burr and Cicchini, 2014; Cicchini *et al.*, 2014), i.e., when stimulus changes are small in comparison to noise fluctuations in internal representations. Here, all frequency changes were near threshold for pitch discrimination, possibly promoting greater integration. The weight of the interval preceding a trial was indeed quite large, 20% on average (Fig. 4). An interesting question is whether this weight might change with the statistics of frequency changes preceding each trial. In 2-interval tasks, there is some evidence that the threshold elevation due to roving is larger for an intermediate than a large frequency range (Matthews and Stewart, 2008). The weight of prior sensory evidence is also expected to increase with decreasing salience of sensory evidence within the current trial. As an extreme example, the direction of pitch change between ambiguous stimuli (Shepard tones spaced by half an octave) was found to be almost entirely dependent on prior stimulation (Chambers and Pressnitzer, 2014; Chambers *et al.*, 2017). It has been suggested that updating of the perceptual weight of previous observations might occur on a rapid time scale, according to a process that can be modeled as a Kalman filter (Burr and Cicchini, 2014). It has also been suggested that information can be accumulated over repetitions of a recurring "reference" stimulus to form a *perceptual anchor* (Durlach and Braida, 1969; Matthews and Stewart, 2008; Nahum *et al.*, 2010). Both hypotheses imply a model more complex than embodied by Eq. (3) that assumes fixed weights.

The model proposed here considers only the contribution of the previous interval (s_{n-1}) to the decision variable [Eq. (3)]. We tested linear models including even earlier intervals (s_{n-2}, s_{n-3}, not reported here) and found that effects can extend over several trials, with effects decreasing with anteriority, as reported in previous studies of auditory and visual perception (Raviv *et al.*, 2012; Alais *et al.*, 2015; Cicchini *et al.*, 2014; Fischer and Whitney, 2014; Fründ *et al.*, 2014; Taubert *et al.*, 2016a). In the extreme, we could hypothesize that subjects performed the task by comparing each new tone to a weighted average of all preceding stimuli (e.g., Morgan *et al.*, 2000). Other models can be proposed to capture sensory integration over multiple stimuli, such as the implicit memory model of Raviv *et al.* (2012), or the Bayesian integration model of Cicchini *et al.* (2014). Such models lead to similar predictions as the generalized linear model (as shown by Raviv *et al.*, 2012).

The sliding 2AFC procedure used in this study differs from the classic two-interval procedure used in most pitch discrimination studies in two important ways. The first is that each tone was involved in two successive comparisons, first as a test tone and then as a reference tone. In particular, the need to store the representation of tone $n-2$ (to compare it with $n-1$ on the previous trial) might have increased its salience and weight within the current trial. The second is that each tone was followed by a response in our procedure, whereas no response follows the reference tone in the classic procedure. The response to tone $n-1$ might require additional cognitive resources, reducing its weight relative to tone $n-2$. However, in both cases we would expect a performance difference between procedures, whereas our previous study found none (Arzounian *et al.*, 2017). The motivation for the new procedure was to ensure effects such as these, if they exist, affect uniformly all tones in the sequence, so that tones preceding a trial differ only by their rank. It would be useful to repeat our analyses with data from the classic procedure to clarify these issues.

Sensory integration has an assimilative effect, as if the representation of an interval were attracted towards that of the previous interval(s). This differs from *adaptation* effects, where past stimuli bias the perception of following stimuli in the opposite direction (Gibson, 1937). Alais *et al.* (2015) found such negative aftereffects when listeners had to detect continuous, directional frequency modulations (sweeps) in a series of modulated and non-modulated tones. It is still unclear which types of perceptual traits are dominated by positive assimilation or by negative aftereffects, and in which conditions (Cicchini and Kristjánsson, 2015). It has been suggested that naturally stable stimulus attributes be prone to assimilation, while naturally changeable attributes be subject to adaptation (Taubert *et al.*, 2016a). The intervals presented in our study were sampled around listeners' thresholds and rarely exceeded ± 2 semitones, whereas Alais *et al.* used frequency sweeps with amplitudes ranging from 0 to ± 3 octaves.

D. Differences between subjects

Thresholds differed widely across subjects (Fig. 2, open symbols) as found in previous studies of frequency discrimination (Amitay *et al.*, 2005; Micheyl *et al.*, 2006; Kidd *et al.*, 2007). Using the Full model, subjects were also found to differ in the *magnitude and sign of the systematic bias* [Fig. 4(b)], the *magnitude and sign of the previous response factor* [Fig. 4(d)], and the *magnitude of the previous stimulus weight* [Fig. 4(c)]. After factoring out these effects, *internal noise* [Fig. 4(a)] also differed between subjects by an order of magnitude, which quashes any speculation that differences in pitch discrimination ability result only from differences in ability to ignore trial history. The weight of interval s_{n-1} ranged from non-significant for subject 11 to 40% for subject 9, suggesting a difference in ability to "isolate" the current sensory trace from previous traces. Mathias *et al.* (2010) previously found that subjects differed in their susceptibility to frequency roving in 2AFC tasks. Although they used a different procedure, those inter-individual differences might be linked to the differences we report here. It has been observed that some subjects can perceive a pitch change but have difficulty assigning a *direction* to it (Semal and Demany, 2006; Mathias *et al.*, 2010). It would be interesting to extend our experimental and modeling framework to address this situation.

E. Limitations of our study

The frequency step preceding a trial was restricted to a few times the listener's threshold. This limits the applicability of the present findings to situations where the frequency steps are larger, e.g., roving over a wide range. We mainly consider history limited to the previous trial, for ease of exposition, and because our trial selection criterion based on response time (Sec. II) reduces the number of longer sequences. In any case, because we used only four interleaved tracks, interval sizes in trial n and trial $n-4$ are highly correlated, which limits our ability to assess the influence of a deeper history. We considered only a linear model [Eq. (3)]. It is possible that choice probability is better predicted by a non-linear transform of the factors, in which case our linear model captured the best linear approximation of this dependence. In particular, we might expect interactions between factors, for example, a stronger weight of response history when stimulus evidence is weak, or an increase in sensory noise with roving due to confusion. Other factors might conceivably affect responses, like absolute frequency. A systematic exploration of all possible factors, transforms, and interactions is tedious and prone to overfitting. For the same reason, we did not explore the likely hypothesis that model parameters vary as a function of time, for example, due to adjustments of response strategy. Although testing for such a hypothesis and adjusting the model accordingly is theoretically feasible, it requires assumptions about the lifespan of a given set of parameters, and the number of data points available for each fit would be reduced, reducing confidence in parameter estimates. For simplicity, the model was therefore assumed to be stationary over the entire duration of the block.

V. SUMMARY

This study investigated how responses in a sliding 2-AFC pitch discrimination task are affected by factors other than the stimuli to be compared, and in particular by the preceding pitch interval and the report made about this preceding interval. We found a significant influence of interval history, of assimilative nature for all subjects, suggesting that the sensory trace of each new stimulus might be integrated with the memory trace of previous stimuli. We also found a significant influence of response history, with a sign that was subject-specific, that might reflect a conscious or unconscious response strategy based in part on the previous response. Because the two factors were correlated (the previous response was affected by the previous interval), a model that contains only one would have incorrectly estimated the weight of the other. In particular, ignoring response history would have led to misleading conclusions with respect to stimulus history effects. Factoring out effects of interval and response history as well as systematic bias (also subject-dependent) led to an estimate of sensory noise that was smaller (and arguably more reliable) than that obtained with a simpler model, or from the measure of a pitch discrimination threshold. The level of sensory noise varied widely across subjects, suggesting that perceptual acuity was highly subject-dependent, as suggested by previous studies. Subjects also differed in the weight assigned to the previous interval (suggesting differences in ability to "isolate" the current interval), and response (suggesting differences in ability to resist the influence of previous decisions). Future work may clarify if these context effects occur similarly in more traditional 2-interval tasks. To obtain these results we used several methodological refinements (choice psychometric function, continuous tone series, interleaved tracks, model including response history and bias) that may be of use in future studies.

ACKNOWLEDGMENTS

This work has received support under the program ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*. The first author was supported by the Région Ile-de-France in the framework of the "DIM Cerveau et Pensée" and by the Bettencourt-Shueller Foundation via the "Frontières du vivant" Ph.D. program.

APPENDIX

TABLE II. Statistics for nested model comparisons. The first column corresponds to subject (*S*) rank as in Fig. 2 and Fig. 4. The second column lists the number (*N*) of data points available for model fitting after trial rejection based on reaction times. Following columns list the *F*-statistics (*F*, first sub-column) and *p*-values (*p*, second sub-column) for the test comparing the two models indicated in the heading line (the reader is referred to Table I for a definition of model acronyms). The symbols *** indicate $p < 0.001$. Figure 3(b) summarizes the number of tests with $p < 0.05$ for each model pair.

		B vs 0		BPS vs 0		BPS vs B		BPR vs 0		BPR vs B		Full vs 0		Full vs B		Full vs BPS		Full vs BPR	
S	N	F	p	F	p	F	p	F	p	F	p	F	p	F	p	F	p	F	p
1	1049	18.3	***	15.9	***	13.2	***	10.3	***	2.27	0.13	42.1	***	30.8	***	98.1	***	27.8	***
2	990	38.1	***	21.7	***	5.1	0.02	34.7	***	30.2	***	22.5	***	46.3	***	18.1	***	42.1	***
3	1042	12.5	***	9.24	***	5.85	0.02	15.6	***	18.4	***	42.1	***	30.8	***	98.1	***	35.6	***
4	1058	44.2	***	22.9	***	1.45	0.2	39.1	***	32.5	***	22.5	***	46.3	***	18.1	***	30.8	***
5	854	6.79	0.01	4.04	0.02	1.28	0.3	5.15	0.006	3.48	0.06	42.1	***	30.8	0.004	98.1	0.001	7.72	0.006
6	1001	2.1	0.1	7.31	***	12	***	28.3	***	53.7	***	22.5	***	46.3	***	18.1	***	98.1	***
7	1043	61.3	***	69.8	***	73.8	***	69.5	***	73.3	***	42.1	***	30.8	***	98.1	***	23.7	***
8	961	16	***	73.9	***	129	***	138	***	255	***	22.5	***	46.3	***	18.1	***	12.1	***
9	1026	0.46	0.5	29.8	***	59	***	0.15	0.9	-0.2	1	42.1	***	30.8	***	98.1	***	95.5	***
10	990	12.4	***	6.28	0.002	0.13	0.7	31.9	***	50.7	***	22.5	***	46.3	***	18.1	***	0.006	***
11	1045	14.5	***	42.2	***	68.8	***	144	***	270	***	42.1	***	30.8	***	98.1	***	22.5	0.76
12	956	2.22	0.1	1.08	0.3	-0.1	1	43.3	***	84.1	***	22.5	***	46.3	***	18.1	***	46.3	***
13	458	1.09	0.3	1.19	0.31	1.28	0.26	21.8		42.3	***	42.1	***	30.8	***	98.1	***	12.4	***
14	1021	28.7	***	64.4	***	97.3	***	119	***	203	***	22.5	***	46.3	***	18.1	***	18.1	***

Alais, D., Orchard-Mills, E., and Van der Burg, E. (**2015**). "Auditory frequency perception adapts rapidly to the immediate past," Atten. Percept. Psychophys. **77**, 896–906.

Amitay, S., Hawkey, D. J. C., and Moore, D. R. (**2005**). "Auditory frequency discrimination learning is affected by stimulus variability," Percept. Psychophys. **67**, 691–698.

Arzounian, D., de Kerangal, M., and de Cheveigné, A. (**2017**). "A sliding two-alternative forced-choice paradigm for pitch discrimination," J. Acoust. Soc. Am. **142**, 167–172.

Bull, A. R., and Cuddy, L. L. (**1972**). "Recognition memory for pitch of fixed and roving stimulus tones," Percept. Psychophys. **11**, 105–109.

Burr, D., and Cicchini, G. M. (**2014**). "Vision: Efficient adaptive coding," Curr. Biol. **24**, R1096–R1098.

Carcagno, S., Semal, C., and Demany, L. (**2011**). "Frequency-shift detectors bind binaural as well as monaural frequency representations," J. Exp. Psychol. Hum. Percept. Perform. **37**, 1976–1987.

Chambers, C., Akram, S., Adam, V., Pelofi, C., Sahani, M., Shamma, S., and Pressnitzer, D. (**2017**). "Prior context in audition informs binding and shapes simple features," Nat. Commun. **8**, 15027.

Chambers, C., and Pressnitzer, D. (**2014**). "Perceptual hysteresis in the judgment of auditory pitch shift," Atten. Percept. Psychophys. **76**, 1271–1279.

Cicchini, G. M., Anobile, G., and Burr, D. C. (**2014**). "Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform," Proc. Natl. Acad. Sci. **111**, 7867–7872.

Cicchini, G. M., and Kristjánsson, Á. (**2015**). "Guest Editorial: On the possibility of a unifying framework for serial dependencies," i-Perception **6**, 1–16.

Dai, H. (**1995**). "On measuring psychometric functions: A comparison of the constant-stimulus and adaptive up–down methods," J. Acoust. Soc. Am. **98**, 3135–3139.

Dai, H., and Micheyl, C. (**2011**). "Psychometric functions for pure-tone frequency discrimination," J. Acoust. Soc. Am. **130**, 263–272.

Demany, L., Pressnitzer, D., and Semal, C. (**2009**). "Tuning properties of the auditory frequency-shift detectors," J. Acoust. Soc. Am. **126**, 1342–1348.

Demany, L., and Ramos, C. (**2005**). "On the binding of successive sounds: Perceiving shifts in nonperceived pitches," J. Acoust. Soc. Am. **117**, 833–841.

Demany, L., and Semal, C. (**2005**). "The slow formation of a pitch percept beyond the ending time of a short tone burst," Percept. Psychophys. **67**, 1376–1383.

Demany, L., Semal, C., and Pressnitzer, D. (**2011**). "Implicit versus explicit frequency comparisons: Two mechanisms of auditory change detection," J. Exp. Psychol. Hum. Percept. Perform. **37**, 597–605.

Durlach, N. I., and Braida, L. D. (**1969**). "Intensity perception I: Preliminary theory of intensity resolution," J. Acoust. Soc. Am. **46**, 372–383.

Emmerich, D. S., Ellermeier, W., and Butensky, B. (**1989**). "A reexamination of the frequency discrimination of random-amplitude tones, and a test of Henning's modified energy-detector model," J. Acoust. Soc. Am. **85**, 1653–1659.

Fechner, G. (**1966**). *Elements of Psychophysics*, Vol. I (Holt, Rinehart and Winston, New York).

Fischer, J., and Whitney, D. (**2014**). "Serial dependence in visual perception," Nat. Neurosci. **17**, 738–743.

Fründ, I., Wichmann, F. A., and Macke, J. H. (**2014**). "Quantifying the effect of intertrial dependence on perceptual decisions," J. Vis. **14**(7), 1–16.

Gibson, J. J. (**1937**). "Adaptation with negative after-effect," Psychol. Rev. **44**, 222–244.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (John Wiley and Sons, New York), 504 pp.

Harris, J. D. (**1952**). "The decline of pitch discrimination with time," J. Exp. Psychol. **43**, 96–99.

Heinz, M. G., Colburn, H. S., and Carney, L. H. (**2001**). "Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve," Neural Comput. **13**, 2273–2316.

Jesteadt, W., and Bilger, R. C. (**1974**). "Intensity and frequency discrimination in one- and two-interval paradigms," J. Acoust. Soc. Am. **55**, 1266–1276.

Jesteadt, W., and Sims, S. L. (**1975**). "Decision processes in frequency discrimination," J. Acoust. Soc. Am. **57**, 1161–1168.

Kaernbach, C. (**1991**). "Simple adaptive testing with the weighted up-down method," Percept. Psychophys. **49**, 227–229.

Kidd, G. R., Watson, C. S., and Gygi, B. (**2007**). "Individual differences in auditory abilities," J. Acoust. Soc. Am. **122**, 418–435.

Lages, M., and Jaworska, K. (**2012**). "How predictable are 'Spontaneous Decisions' and 'Hidden Intentions'? Comparing classification results based on previous responses with multivariate pattern analysis of fMRI BOLD signals," Front. Psychol. **3**, 1–8.

Leek, M. R., Hanna, T. E., and Marshall, L. (**1991**). "An interleaved tracking procedure to monitor unstable psychometric functions," J. Acoust. Soc. Am. **90**, 1385–1397.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.

Macmillan, N. A., and Creelman, C. D. (**2005**). *Detection Theory: A User's Guide* (Lawrence Erlbaum Associates Publishers, Mahwah, NJ), 492 pp.

Mathias, S. R., Bailey, P. J., Semal, C., and Demany, L. (**2011**). "A note about insensitivity to pitch-change direction," J. Acoust. Soc. Am. **130**, EL129–EL134.

Mathias, S. R., Micheyl, C., and Bailey, P. J. (**2010**). "Stimulus uncertainty and insensitivity to pitch-change direction," J. Acoust. Soc. Am. **127**, 3026–3037.
Matthews, W. J., and Stewart, N. (**2008**). "The effect of stimulus range on two-interval frequency discrimination," J. Acoust. Soc. Am. **123**, EL45–EL51.
Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A. J. (**2006**). "Influence of musical and psychoacoustical training on pitch discrimination," Hear. Res. **219**, 36–47.
Micheyl, C., Xiao, L., and Oxenham, A. J. (**2012**). "Characterizing the dependence of pure-tone frequency difference limens on frequency, duration, and level," Hear. Res. **292**, 1–13.
Moore, B. C. J. (**1973**). "Frequency difference limens for short-duration tones," J. Acoust. Soc. Am. **54**, 610–619.
Moore, B. C. J., and Glasberg, B. R. (**1989**). "Mechanisms underlying the frequency discrimination of pulsed tones and the detection of frequency modulation," J. Acoust. Soc. Am. **86**, 1722–1732.
Morgan, M. J., Watamaniuk, S. N. J., and McKee, S. P. (**2000**). "The use of an implicit standard for measuring discrimination thresholds," Vision Res. **40**, 2341–2349.
Motulsky, H., and Christopoulos, A. (**2004**). *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting* (Oxford University Press, Oxford, UK), 353 pp.
Nahum, M., Daikhin, L., Lubin, Y., Cohen, Y., and Ahissar, M. (**2010**). "From comparison to classification: A cortical tool for boosting perception," J. Neurosci. **30**, 1128–1136.
Nordmark, J. O. (**1968**). "Mechanisms of frequency discrimination," J. Acoust. Soc. Am. **44**, 1533–1540.
Oppenheim, J. N., and Magnasco, M. O. (**2013**). "Human time-frequency acuity beats the Fourier uncertainty principle," Phys. Rev. Lett. **110**, 044301.
Raviv, O., Ahissar, M., and Loewenstein, Y. (**2012**). "How recent history affects perception: The normative approach and its heuristic approximation," PLoS Comput. Biol. **8**, e1002731.
Raviv, O., Lieder, I., Loewenstein, Y., and Ahissar, M. (**2014**). "Contradictory behavioral biases result from the influence of past stimuli on perception," PLOS Comput Biol **10**, e1003948.
Rosenblith, W. A., and Stevens, K. N. (**1953**). "On the DL for frequency," J. Acoust. Soc. Am. **25**, 980–985.
Ruusuvirta, T., Wikgren, J., and Astikainen, P. (**2008**). "Proactive interference in a two-tone pitch-comparison task without additional interfering tones," Psychol. Res. **72**, 74–78.
Sek, A., and Moore, B. C. J. (**1995**). "Frequency discrimination as a function of frequency, measured in several ways," J. Acoust. Soc. Am. **97**, 2479–2486.
Semal, C., and Demany, L. (**2006**). "Individual differences in the sensitivity to pitch direction," J. Acoust. Soc. Am. **120**, 3907–3915.
Taubert, J., Alais, D., and Burr, D. (**2016a**). "Different coding strategies for the perception of stable and changeable facial attributes," Sci. Rep. **6**, 32239.
Taubert, J., Van der Burg, E., and Alais, D. (**2016b**). "Love at second sight: Sequential dependence of facial attractiveness in an on-line dating paradigm," Sci. Rep. **6**, 22740.