Buildup and bistability in auditory streaming as an evidence accumulation process with saturation

Quynh-Anh Nguyen[1¤], John Rinzel[2,3], Rodica Curtu[1,4*]

1 Department of Mathematics, The University of Iowa, Iowa City, Iowa, United States of America

2 Center for Neural Science, New York University, New York, New York, United States of America

3 Courant Institute of Mathematical Sciences, New York University, New York, New York, United States of America

4 The Human Brain Research Laboratory, University of Iowa, Iowa City, Iowa, United States of America

¤Current Address: Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis (IUPUI), Indianapolis, Indiana, United States of America

* Corresponding author:

E-mail: rodica-curtu@uiowa.edu (RC)

Abstract

A repeating triplet-sequence $ABA_$ of non-overlapping brief tones, A and B, is a valued paradigm for studying auditory stream formation and the cocktail party problem. The stimulus is "heard" either as a galloping pattern (integration) or as two interleaved streams (segregation); the initial percept is typically integration then followed by spontaneous alternations between segregation and integration, each being dominant for a few seconds. The probability of segregation grows over seconds, from near-zero to a steady value, defining the buildup function, BUF. Its stationary level increases with the difference in tone frequencies, DF, and the BUF rises faster. Percept durations have DF-dependent means and are gamma-like distributed. Behavioral and computational studies usually characterize triplet streaming either during alternations or during buildup. Here, our experimental design and modeling encompass both. We propose a pseudo-neuromechanistic model that incorporates spiking activity in primary auditory cortex, A1, as input and resolves perception along two network-layers downstream of A1. Our model is straightforward and intuitive. It describes the noisy accumulation of evidence against the current percept which generates switches when reaching a threshold. Accumulation can saturate either above or below threshold; if below, the switching dynamics resemble noise-induced transitions from an attractor state. Our model accounts quantitatively for three key features of data: the BUFs, mean durations, and normalized dominance duration distributions, at various DF values. It describes perceptual alternations without competition per se, and underscores that treating triplets in the sequence independently and averaging across trials, as implemented in earlier widely cited studies, is inadequate.

Author summary

Segregation of auditory objects (auditory streaming) is widely studied using ambiguous stimuli. A sequence of repeating triplets $ABA_$ of non-overlapping brief pure tones, A and B, frequency-separated, is a valued stimulus. Studies typically focus on one of two behavioral phases: the early (say, ten seconds) buildup of segregation from the default integration or later spontaneous alternations (bistability) between seconds-long integration and segregation percepts. Our experiments and modeling encompass both. Our novel, data-driven, evidence-accumulation model accounts for key features of the observations, taking as input recorded spiking activity from primary auditory cortex (as opposed to most existing, more abstract, models). Our results underscore that assessing individual triplets independently and averaging across trials, as in some earlier studies, is inadequate (lacking neuronal-accountability for percept duration statistics, the underlying basis of buildup). Further, we identify fresh parallels between evidence accumulation and competition as potential dynamic processes for choice in the brain.

Introduction

Stimulus sequences of interleaved A and B pure tones have been widely used in studying segregation of distinct objects in an auditory scene (auditory streaming), in human psychophysics [1–6], invasive neurophysiology [1, 3, 7, 8], or in experiments implementing both [9]. A valued stimulus is triplet-streaming $ABA_$ with the tone frequency difference, DF, as a tunable parameter [10] (Fig 1). For small DF human listeners most likely perceive integration (one galloping rhythm); for DF large, segregation dominates (two simultaneously heard parallel streams). The initial percept is typically integration but within seconds the probability of segregation increases ("the buildup phase") and perceptual switching eventually occurs ("perceptual bistability"). Alternating percepts have variable durations, described by either gamma or lognormal distributions [2]. Time courses of spiking activity (macaque, primary auditory cortex, A1, [1]) show dynamical features (adaptation over 1-2 seconds) that were interpreted as neural correlates of buildup, although the behavioral and physiological experiments were not conducted together [1, 3].

Dynamics of buildup and/or perceptual alternation for ambiguous auditory stimuli were described by computational models based on signal processing [1, 3, 11–13], competition dynamics [6, 14], coupled-oscillator patterning [15, 16], evidence accumulation [5], and statistical descriptions [17]; also reviews by [18] and [19]. However, with few exceptions (e.g. [1, 6]) these models did not incorporate neurophysiological data. Furthermore, experimental and modeling studies primarily focused on either buildup, describing the probability of segregation during short, tens of seconds, trials [1, 3, 4, 20, 21], or on the stationary phase of alternations, characterizing the statistics of percept durations over long, several minutes, trials [2, 5, 6].

Here we designed the experiment (30 s trials with many trials per condition/subject) so that we could characterize these features simultaneously. Then we proposed a model

that takes spike-recordings from A1 as input, and accounts for both the behavioral time course of buildup and the observed duration statistics during alternations, over a range of DF values: 3, 5, 7 semitones. Our model is neuromechanistic-like, transforming the neuronal input for processing in two evidence-accumulators downstream of A1. From the input-sensory level, sampling of spike counts across A1-units provides a measure for the contribution of each triplet to the evidence-accumulation stage; if evidence against the current percept exceeds a threshold then a perceptual switch occurs and accumulation resets. This approach parallels in spirit Barniv and Nelken's model [5] although that was implemented from a Bayesian-viewpoint. Our model is data-driven: input is neural-based; initial parameters are estimated from our behavior data (mean probability of segregation) then fine-tuned to match the gamma-distributed percept durations.

We propose that although the model is not competition-based it shares some features of such approaches: Adaptation is key in competition dynamics; evidence accumulation might be viewed as recovery from adaptation. Matching duration statistics with competition requires some balancing of noise and adaptation [22]; its analogue is the interplay between accumulation and noise. Adaptation strength, when set near the boundary between noise-free oscillatory and noise-driven attractor dynamics, constrains dominance durations [23]; comparably, our accumulators have a novel feature of saturation which if set below but near the switching threshold, produces observed statistics only if adequate noise is present.

Importantly, our modeling highlights that accounting for the duration statistics of behavioral data is key when studying auditory bistable perception. With quantitative matches to these data the buildup phase is then naturally reproducible by an alternating renewal process [17]. We show that a widely cited signal-detection approach [1, 3, 12, 21], based on treating each triplet independently without accumulation, that overlooked this crucial feature does not account for the single-trial percept duration statistics. We

argue for caution when applying it to test neural-inspired behavioral hypotheses.

Results

Model-based analyses of ramping trial-averaged neuronal responses have recently come under scrutiny by consideration of single-trial data [24]. Arguments were made, admittedly still under debate [25, 26], that trial-averaged smooth time courses of evidence accumulation during decision-making might arise from temporally "discrete steps" rather than from continuous ramping dynamics. This work motivated us to revisit an earlier model for auditory bistable perception [1] that applied a signal detection scheme to individual stimulus elements and then averaged data across multiple trials to generate slowly increasing buildup of segregation functions. We combined neural data from [1] with behavioral data from our experiments (see section A) to investigate if the signal detection model could yield percept durations in a self-consistent fashion. We found it did not; and moreover that it was unable to fit buildup functions that, for different stimulus conditions, were not widely spread (section C). We then developed a neural-based evidence accumulation-like explanation of the observed data, as alternative to explicit competition, and with the advantage of being intuitive (section B).

A. Auditory triplet-streaming

A.1. Experimental protocol

Fifteen human subjects with normal hearing listened to sequences of repeating $ABA_$ triplets and were instructed to continuously report their ongoing percept by selectively pressing one of two different buttons on a keypad. Subjects began reporting their percept typically 2 s after stimulus onset as integration (I; a single, coherent stream $ABA_ABA_$) or segregation (S; two distinct streams $A_A_A_A_$ and $_B___B_$). Stimuli were sequences of triplets $ABA_$ that consisted of alternating high (A) and low

(B) pure tones followed by a 125 ms silent pause "_" (Fig 1A-B). In total, triplets were 500-ms in duration and were repeated 60 times per trial. Tones were separated in frequency by DF semitones chosen from three conditions (DF= 3, 5, 7) with each condition being presented five times per experimental block (nine blocks total). This resulted in group data with 675 30-s trials for each of three DF values.

A.2. Behavioral task performance

For each DF condition the buildup function was constructed by computing the probability of segregation from trial-averaging (Fig 1 B-C). The buildup functions started at zero and increased over time before stabilizing to certain DF-dependent asymptotic values, similar to reports by [1, 3, 5, 12]. They started at zero due to the latency period (when no percept was identified) and not because the initial percept was I; see *Methods*, also [4]. While I first percepts were indeed more likely, S first percepts were reported too. The proportion of segregation as initial percept increased with DF from 103 out of 675 trials at DF=3 to 137 at DF=5 and 220 at DF=7. The probability of segregation increased faster and reached higher levels at larger DF, with transient times of approximately 16, 10, 5 s after stimulus onset and with asymptotic values 0.45, 0.6, and 0.65 at DF= 3, 5, 7 respectively.

We computed distributions of normalized phase durations for subsequent durations, separately for each DF, and found them to be gamma-like, consistent with previous results on subsequent percepts [2, 5, 6]. Herein we report that duration distributions of the first percept are also gamma-like (Fig 1C; see also S1 Fig). We used statistical bootstrapping to compute the shape parameter α of each gamma distribution (see *Methods*), and determined that $\alpha \approx 2$ for normalized first durations and $\alpha \approx 2.6$ for subsequent durations. The distributions satisfied the scaling property $\gamma_1 \approx 2CV$ with skewness γ_1 and coefficient of variation $CV \approx 0.7$ and $CV \approx 0.6$ respectively, similar to reports by [27]. First percept durations were found to be longer in the mean than

subsequent percept durations in all cases but segregation at DF=3. Mean durations of first I-percept were 10.9, 5.3 and 3.1 s, decreasing with DF= 3, 5, 7. Mean durations of first S-percept were 3.5, 6.6, 8.1 s, increasing with DF. Similar trends were observed for subsequent percepts with means of 5.4, 3.4, 3.1 s for I and 4.9, 5.2, 5.6 s for S at DF= 3, 5, 7.

B. Auditory streaming as an evidence accumulation process

Herein we propose an **EV**idence **A**ccumulation (EVA) model (Fig 2; see details in *Methods*) that accounts for the observed dynamical features of buildup and alternations: gamma-like distributions for first and subsequent durations, DF-dependent mean durations, and psychometric buildup functions (Fig 1C-D and Figs 3-4). In short, data-based [1] estimates of spike counts of neurons in the primary auditory cortex (area A1) are sampled by a population of units and their summed responses lead to a population vote and to an increment of evidence "for" and "against" the current percept. When enough evidence has built up against the current percept, there is a switch to the opposite percept. Current increments can be positive or negative but only when the accumulated evidence is adequate, does a switch occur.

B.1. Properties of the EVA model

The model is structured as a three-layer network. It takes Poisson spike counts from tone-A-selective A1-neurons (the Input Layer, IL) [1, 7, 8] and passes them through binary units in the Sampler Layer, SL (Fig 2A). Only spike counts recorded during tone B are included (Fig 2B). Each SL-unit compares the averaged spike count across a small number N_{in} of input units to a fixed threshold C_{th} and places the outcome into either state 0 (for I) or 1 (for S). High activation in IL (above C_{th}) is assumed to support percept I while low activation facilitates percept S (Fig 2A-B). The proportions $p_I(t)$, $p_S(t)$ ($p_I = 1 - p_S$) of SL-units in states 0 and 1, together with stochastic noise terms

$\xi_I(t)$, $\xi_S(t)$, modulate the activity of the Accumulation Layer, ACC (Fig 2C). Two accumulators representing evidence for the percepts drift towards two targets. Their activities x_I, x_S are updated at discrete time steps determined by the position t of each triplet in the $ABA_$ sequence. One unit accumulates evidence for the current percept (e.g. x_I during integration) in the presence of additive "neural" noise defined by a Gaussian process of strength σ_I=σ_f, and approaches target T_I=T_f. The other accumulator works against the current percept (x_S during integration). It experiences stronger noise level σ_a, and approaches another target, T_a. In the deterministic (noise free) case, alternations between percepts are not possible given that both T_a and T_f are subthreshold targets ($T_f < T_a < 1$), a novel feature of our accumulation model. Instead, the ACC system is bistable with accumulators x_I, x_S reaching either steady state (T_a, T_f) or (T_f, T_a) depending on the initial conditions (Fig 2C, dotted lines in blue and red). In the presence of noise, however, the accumulator against the current percept reaches the decision threshold; a switch to the other percept occurs, the accumulators are reset, the targets are swapped (T_S=T_f, σ_S=σ_f and T_I=T_a, σ_I=σ_a), then another accumulation cycle begins (Fig 2C, traces for x_I, solid blue, and x_S, solid red; the percept's type is identified by the background color, blue for I, red for S. See also S2 Fig). It is essential that the accumulators are subjected to noise in order for the distribution of threshold crossing events idealizing the percept durations to exist.

B.2. Computational advantages of the EVA model

The model is pseudo-neuromechanistic; it takes A1 responses as input, it allows for attractor-states, and it includes accumulators that are saturating akin to synaptic currents. The spike counts are in accordance with neurophysiological data from A1 [1] and provide input to the computation of perception dominance downstream, as in the conceptual population-separation model of [7] and in competition-based model of [6]. The model incorporates fast habituation (after one triplet or so, Fig 2B) as in [1, Fig.3]

and it accounts for the decrease in response amplitudes and in spatial activity patterns evoked by tone B at tone A tonotopic locations observed as DF increases [1, 7]. Indeed, if most of tone-A-selective A1-neurons are active, the model predicts a large proportion of samplers in SL to be active (p_I large) and thus favors I percept. If the opposite happens and A1 is mostly inactive, a large proportion of samplers are inactive (p_S is large) and the model favors S percept. Activation in IL decreases with larger DF (fewer IL-units have mean spike counts above threshold C_{th}) and so does $p_I(t)$; Fig 2A-B, compare DF= 3, 5, 7; see also [1, Fig.3] and [7, Fig.11]. This affects the dynamics of the accumulators since $x_I(t)$, $x_S(t)$ gather evidence about percepts with increment rates proportional to $p_I(t)$, $p_S(t)$ respectively, while also being modulated by a certain level of noise (Fig 2C, Eqs).

The accumulators resemble discrete time versions of the leaky integrate-and-fire neuron model with conductance-based synaptic input [28], $dV = (V_R - V)\mathcal{D}dt + noise$. The voltage-like variable (here, normalized, by the threshold value for switching) has a maximum amplitude of one. The reversal potential V_R is set to either target T_a or T_f depending on the type of the dominant percept and the type of the accumulator. The synaptic drive $\mathcal{D}$ consists of feedforward input from SL and is analogous to the reciprocal of the time constant. Finally, Gaussian white noise represents input from other brain sources or internal to ACC. Its strength σ needs to reach an appropriate level for the statistics of percept durations generated by the EVA model to match the behavioral data (see *Methods*).

Our EVA model is data-driven. Initial conditions are set using the latency periods and the proportion of first S-percepts from the experimental data. Poisson spike counts of IL neuronal units at each triplet t and semitone difference DF are generated using mean values $m_{t;DF}$ derived from macaque A1 multi-unit spiking neural data [1] (Fig 5A; see *Methods*). Parameters N_{in} and C_{th} are obtained by least-squares fit between the

probability of a sampler to support an S percept and the behavioral buildup at all three DF values (Fig 5B; see also *Methods*). Neuronal granularity as a suitable substrate for perceptual representations [27] is implemented through SL. The number of samplers is chosen flexibly from a wide range of values ($N_{sl} \geq 1$). There are few other free parameters, T_a, σ_a, T_f, σ_f, b (baseline), but only the former two are major players in fitting the model to data (see below).

B.3. EVA model captures DF-dependence of mean durations

Numerical simulations of the EVA model followed the experimental setup with N_{tr}=675 repetitions (trials) per DF. The model-generated mean durations were computed separately for each percept type (I, S; first, subsequent) and DF. They approximated well their counterparts from behavioral data (Figs 3B and 4B). They also captured two important DF-related trends reported by other studies. First, at each DF mean durations of the first I percept were found to be longer than those of subsequent I percepts [2, 6]. The same was true for S percepts except at DF=3. Secondly, at either first or subsequent percepts, mean durations for I decreased with DF while mean durations for S increased with DF, similar to results from [6]. The model was robust to noise as demonstrated by 100 Monte Carlo runs of each DF simulation that yielded consistent results in terms of average values and 95% CI (Figs 3B and 4B; error bars).

In EVA model, the switch to a new accumulation cycle occurred when the ACC-unit that accumulated evidence against the current percept reached the decision threshold. Target-against T_a, noise level σ_a, and increment rate $p_S(t)$ determined the trajectory of the suppressed unit x_S and the length of the corresponding dominant percept I. Similarly, T_a, σ_a and $p_I(t)$ determined the duration of percept S. We studied the effect of T_a and σ_a on the model-generated mean durations at each DF by varying their values while keeping all other parameters fixed (see *Parameter values used in model simulations*, in Methods). At a given T_a, EVA model exhibited no alternations if σ_a was

small (Fig 6; region in gray). For moderate σ_a values, perceptual switches occurred but yielded percepts of mean durations much longer than those found experimentally (in warm colors); then, for larger σ_a, simulated durations became comparable to (in green) and then much shorter than (in cool colors) the behavioral mean durations. Similar results were obtained for σ_a fixed when varying T_a. As a general rule, the smaller the target-against, the stronger the noise level had to be in order for the accumulator's trajectory to be pushed above the threshold and to generate acceptable statistical approximations of the data (Fig 6, in green; also black dots; within one standard error to the experimental mean, SEM).

The decrease in mean durations of I percepts with increasing DF (Figs 3B and 4B, blue) stemmed from the increase in probability of a sampler to support segregation (Fig 5B) which led to an increase of increment rate p_S of accumulator x_S (see *Statistical properties of SL-activation*, in Methods). Mean durations of S percepts increased with DF (Figs 3B and 4B, red) due to the decrease of increment rate p_I of accumulator x_I. These DF-dependent properties of p_S, p_I, inherited from A1 spike counts, enabled the EVA model to capture the correct qualitative trend of the experimental means across all percept-types and DF conditions. Suitable quantitative agreements were then obtained by fine-tuning the value of target-against T_a (Fig 6, red diamond; error between numerical and behavioral results was restricted to 0.1 SEM. See also S6 Fig).

B.4. EVA-modeled first and subsequent percept durations match observations

The model reproduces an important statistical feature of the behavioral data, the distributions of normalized durations for all first and subsequent I, S percepts at $DF=$ 3, 5, 7. Histograms were drawn and fitted by gamma probability density functions of shape parameters α (see Eq (1) in Methods) whose values agreed with those from the behavioral experiment. The shape of distributions was tested and confirmed statistically

by 100 Monte Carlo runs of the model for each DF condition separately (Figs 3B and 4B; error bars indicate 95% CI around α-mean values). Exemplar distributions for first and subsequent durations are shown in Figs 1D and 4B at DF=5. For other DF values, see S1 Fig.

Since alternations were caused in the model by the accumulator that gathered evidence against the current percept, the distribution of threshold crossing event times depended on T_a and σ_a. In particular, for a given T_a, EVA model generated percept durations that were normally distributed for σ_a small (Fig 6, region N; α was much bigger than 3) and exponentially distributed for σ_a large (Fig 6, region E; α was close to 1). For intermediate σ_a, the distributions were gamma-like matching those fit to the observed data (Fig 6, region G; model-generated α values were similar to those determined from experiments, α_{exp}, at relative error up to 20-30%). The closer T_a was to the decision threshold 1, the easier it was to find σ_a that yielded gamma-like distributions. With decreasing T_a, the transition to a narrower region G was either sharp-edged (e.g. $DF = 7$, first I) or rather smooth ($DF = 7$, first S). Percept durations that approximated well both the distribution shape and the mean duration of the experimental data were obtained by using parameters from region G that overlapped with the black dotted sheet.

B.5. EVA model captures DF-dependence of stream segregation buildup

The model-generated buildup functions captured both the rising and the asymptotic phases of the behavioral buildup for each DF= 3, 5, 7 (Fig 1D). These trends were a consequence of already having simulated percept durations and percept means well-fit to behavioral data, in accordance with previous works describing the buildup of stream segregation as a byproduct of an alternating renewal process [17].

C. Signal detection algorithm yields fast buildup and unrealistic percept durations

C.1. Modeling the buildup with Micheyl's model for auditory streaming

The signal detection algorithm of Micheyl et al. [1] has been extensively cited in the auditory streaming research [3–6, 12, 21] in regard to computing time-varying probabilities of stream segregation from neuronal responses in A1. The model was based on choosing a threshold number, C_{th}, for mean spike count (first, trial-averaged; then averaged across sampled neurons) to classify each triplet as I or S; doing this for each $ABA_$ in the sound sequence generated a time course, a "neurometric" function. Briefly, for a given triplet position, a probability distribution was constructed for the B-tone responses measured at and convolved over A1 neurons whose best frequency was that of the A-tones. The area under the probability distribution to the right of C_{th} determined the probability that tone B was detected and, consequently, that the triplet belonged to I percept; the complementary probability was associated with S percept. The algorithm classified each triplet independently and assumed no memory among nearby triplets.

A simplified view of this procedure is to consider the distribution of trial-averaged counts for the neurons as straddling the mean for each triplet in the time course of an $ABA_$ sequence [1, Fig.3]. Conceptually, one chooses a level C_{th} that will cut across the distributions, say for DF=3, and correspond to low probability of S for early time and correspond to the asymptotic level (from behavior) for late time (e.g. C_{th}, thin horizontal line, in S5 Fig). This classification could likely provide a decent fit for DF=3 but for DF=6 the spike counts will fall below the threshold, leading to an overestimate of probability of S. The remaining cases will yield extreme classifications: for DF=1, spike counts for each triplet in the sequence will be above C_{th} and probability of S will be estimated as near zero; for DF=9, all spike counts will be below C_{th} (except maybe for the initial triplet) and therefore probability of S will be near one for the time course.

The spread of the behavioral time courses in [1, Fig.4], one lying intermediate for DF=3 and the others at very low or quite high levels, provided an opportunity for reasonable fitting with a single C_{th} level [1]. For details on numerical fitting with such signal detection algorithm, see S5 Fig.

In the case of our data, the conditions were DF= 3, 5, 7 and the behavioral buildup functions lay in more intermediate levels and clustered around the asymptotic value of 0.5 (Fig 1C; also Fig 5B, dotted lines). It thus became challenging to fit the buildup functions (especially the early, slower rising portions for multiple DF values, 3 and 5 semitones) using Micheyl's model with a unique C_{th}.

We attempted to meet this challenge by applying the signal detection algorithm to our behavioral data while using interpolated and Poisson distributed spike counts based on the neural data from [1]. The approach was equivalent to the computation of the probability of a sampler to support segregation from the EVA model, observing only the input layer, IL, and passing its output through one single sampler, N_{sl}=1. While the mean spike counts over a pool of N_{in} A1-neurons did not change significantly with N_{in}, the standard error to the mean did (Fig 5A). The decrease in the spike count error to the mean made the horizontal line C_{th} intersect fewer local distributions and biased the data-fitting towards the behavioral curve for a certain DF, at the expense of others. Choosing more A1 neuronal units (larger values of N_{in}; Fig 5A) led to larger spread in the simulated neurometric functions and poorer fitting (Fig 5B; at DF= 5, 7 the neurometric functions (solid lines) plateaued at probability approximately 1 after triplet-sequence onset, as N_{in} increased; e.g. case N_{in}=100). The best approximation of the asymptotic levels of the behavioral buildup for all DF conditions was found at a relatively low N_{in} however the rising transients of the neurometric functions were still much faster than in the experiment.

C.2. Modeling percept durations with Micheyl's model

We extended the work from [1] by using the signal detection algorithm to generate "percepts" and characterize their distributions. For each DF, adjacent triplets of the same type (I or S) were grouped together to create percept phase durations and construct frequency graphs (Fig 7). Theoretical calculations showed that subsequent percept durations generated by Micheyl's model were exponentially distributed, as opposed to gamma-like. During subsequent durations we could assume that the buildup functions of stream segregation had reached an asymptotic level p (Fig 7A-B, upper panel) and that the activity in the A1 pool was independent at each triplet. Then the probability that percept S consisted of n-triplets could be calculated as $Prob(D_S) = p^n(1-p) = (1-p)e^{n \ln p}$, depending exponentially on n. Likewise the probability that I consisted of n-triplets was $Prob(D_I) = p(1-p)^n = pe^{n \ln(1-p)}$. Similar results were obtained from numerical simulations of Micheyl's model. The probability density functions were found to be discrete versions of exponential curves and the mean durations were small at about 1 s (Fig 7A-B, middle and lower panels), suggesting that the signal detection algorithm is not appropriate to describe perceptual alternations and percept durations, key aspects of bistable stream segregation. (Compare to Fig 1C; also S1 Fig and [2, 5, 6].)

Discussion

We developed a new evidence accumulation model for auditory streaming of triplet sequences $ABA_ABA_\ldots$ that takes as input neuronal responses of primary auditory cortex, A1 (macaque, [1]). Our neural-like model accounts for the (human) behavior we observed under three conditions (tone frequency difference, DF). During trials, subjects reported spontaneous alternations (bistability) between integration, I, and segregation, S. The first percept was usually I; the probability of S built up over time rising from

near zero and plateaued within a few seconds to a level that increased with DF. In the model, switching between I and S occurred when noisy accumulation of evidence against the current percept exceeded threshold. Our simulations matched both buildup time-courses and percept-duration distributions.

Our model draws inspiration from the population separation hypothesis of [7] and focuses primarily on the B-tone responses of A-tone selective neurons. Micheyl et al [1] used similar principles to compute "neurometric" functions for segregation buildup. Their signal-detection model was applied to A1 and to sub-cortical neuronal spike count data to conclude that perceptual organization of auditory streams was present in early stages of the auditory pathway [3, 21]. It treated each triplet as independent of the previous ones, without an accumulation process from triplet to triplet. The only time dependent mechanism was adaptation of A1 neurons that was nearly complete after 2-3 triplets – too fast to account for buildup. Herein we show that Micheyl's model behaves as if classification is like coin-tossing with possible bias. Simulated durations are therefore like run-lengths in coin-tossing, exponential-like and very brief, contrary to the observed data (gamma or lognormal-like).

Our approach underscores the essential significance of duration distributions as characterizations of streaming and switching, a constraint overlooked by previous analysis [1]. It emphasizes that neuronal-based modeling of behavioral data that goes beyond trial-averaged behavior may need to involve an evidence accumulation process in order to account for the statistics of single trials.

Novel features

Our model is intuitively straightforward. It describes the accumulation of evidence, incremental from each triplet, for or against the current percept. The estimated A1-spike counts are passed through a sampler layer, SL, each of whose units sample a

few A1-neurons. SL-units vote 0 or 1 if the summed spike counts for the current triplet are above or below threshold. The fraction p_I (p_S) of sampler-votes 0 (1) represents the net output which favors integration (segregation), transmitted to the accumulators. After multiplicative weighting, p_I, p_S are used together with additive noise to update the accumulators. Of significant novelty, the weighting factor is state dependent, proportional to the difference, $T-x$, between the current accumulator value x and a target T. Accumulation slows when x is closer to T and, importantly, we can choose $T<1$ in which case our model mimics noise-driven attractor competition dynamics [23]. Further, if T is close to one (i.e. accumulation saturates below, near threshold), gamma-like distributed threshold-crossing times are more robustly obtained with modest noise levels [22, 29].

Our model is a hybrid: it incorporates some neuro-based phenomenology (A1 neuronal responses as input, saturating driving force, escape dynamics) but it is non-committal to specific neuronal mechanisms of inhibition and adaptation. Moreover, key parameters are not directly linked to neuromechanistic processes but rather determined by fitting model dynamics (simulated threshold-passage times) to observed duration distributions.

Fitting of model to data

We assigned different values of noise and targets to "against" and "for" accumulators to ensure switching was caused by the against-unit crossing the decision threshold. Intuitively, as the accumulator-against saturates around target-against T_a (subthreshold), enough noise σ_a guarantees threshold-crossing. The closer T_a is to one, the less noise is required to produce alternations. Within the switching domain different combinations of T_a and σ_a yield different distributions and means of percept durations. Our model reproduces the experimental data when T_a, σ_a are taken from a restricted

parameter region. With T_a constant across conditions we captured the observed trend of mean durations although some values were off. With fine-tuning of T_a across conditions (but σ_a constant) we match the observed duration distribution shapes and means. This approach is analogous to obtaining the proper balance between noise and adaptation necessary for alternations in other models for bistable perception [22, 30]. Noteworthy, our model shows switching behavior when tuned in other parameter regimes, including with T_a>1. However, in such a drift-dominated regime although noise is not needed for alternations, we found that matching the statistical features and behavioral trends required a substantially higher (unacceptable) noise level (not shown here).

Comparison with other models

Barniv and Nelken (2015) and Cao et al (2016) also modeled auditory bistable perception as evidence accumulation based. The former's model used Bayesian assignments of B-tones to either the same class as A-tones (integration) or to a different class (segregation). Its noise-free version shows periodic alternations, as does our system for T_a>1, but the dynamics do not reset. Instead, our accumulators undergo discontinuous resetting after each switch. Most importantly, in contrast to [5], the parameters in our model are interpretable, functionally if not physiologically. Cao et al. formulated a stochastic accumulator that reproduced (like ours) several scaling properties of bistable behavior but without a description of switching and of, possibly asymmetric, alternations. Our work differs from both models by incorporating directly A1-spiking data as input. In this sense it is more akin to [6], a literal competition model. Notably, our approach predicts that neuronal computation for percept representation and evidence accumulation takes place beyond A1; its input (activity from A1 devoid of switch-dynamics) implicitly includes inhibition, adaptation and noise that occur within A1 and preceding stages.

Dynamic competition models commonly include two or more units representing response patterns associated with different percepts, and share mechanistic features of mutual inhibition, adaptation, and noise [6, 14]. In our two-process model only one percept is currently dominant thereby realizing mutual exclusivity. However, inhibition is not explicitly invoked; rather, our model performs as if a firm choice is made at the switch time, further accumulation of evidence in favor of the fresh percept is prevented; the in-favor accumulator is reset and targeted to a low value, T_f, despite continued incoming evidence from SL.

In oscillator-based alternations, switching may be determined by strongly rising adaptation in the dominant unit, leading to "release" from inhibition, or by stronger recovery from adaptation in the suppressed unit, leading to "escape" from inhibition [31–34]. Our model has no explicit adaptation variable as negative feedback. However, the accumulation of evidence against the current percept may be viewed as recovery of salience of the non-dominant percept. The rise and eventual take-over of dominance is therefore analogous to the escape from suppression in competition models.

In such models if adaptation is weak, changes in dominance may be represented by noise-driven switches between stable states in attractor state dynamics [22, 23]. These insights motivated our choice of an evidence-against accumulator that saturates just-subthreshold. Dominance durations are longer with reduced noise, and no switches occur in the noise-free idealization. Further, gamma-like duration distributions are more robustly obtainable with this mechanism: rise to saturation and wait for switch-favoring fluctuations to induce a switch. Satisfactory results are also obtainable with $T_a > 1$, but if T_a exceeds threshold by too much, acceptable duration statistics seemed to require strong noise, and accumulator time courses were noise-dominated.

Limitations, extensions, predictions

Reports on triplet-streaming are conflicted about correlations between successive I, S durations, showing either statistical independence [2] or small positive correlations [5]. In our model both accumulators are reset after a switch approximately to target T_f so correlation between successive percepts is weak. However, we could likely match the reported correlations [5] by changing the resetting to generate continuous dynamics of accumulators.

Our model could be extended to mimic the transient behavior of buildup by relaxing the constraints on initial conditions and treating the baseline as DF-dependent. Two hypotheses may be tested: that integration emerges with first percept probability as in the behavioral data and that early adaptation of A1-responses accounts for longer first, than subsequent, I-durations [35].

With minimal modifications to our model we could test for behavior at other DF values or for dependence on presentation rate. Assuming lower target-against levels T_a for faster presentations, we predict at constant DF similar mean I-durations but longer S-durations, and higher probability of segregation [6]. With increased presentation rate mean spike counts for B-tones will decrease [8] and lead to lower vote counts p_I and lower effective accumulation rate, $T_a p_I$. Although p_S (=1-p_I) would increase, the increase would be compensated by the decreased T_a leaving $T_a p_S$ relatively unchanged.

To conclude, we propose an evidence accumulation model for auditory bistable perception with neurally-plausible mechanisms that accounts for statistics of behavioral data. In principle, it could be extended to study dynamics induced by transient perturbations (deviants/distractors; [35]) or associated with multiple percepts [14]; implementations of such generalizations remain as open topics for future research.

Methods

Experimental design and statistical analyses

Participants

Fifteen human subjects with normal hearing (8 female and 7 male; ages 18-45 yrs.; median 22 yrs.) were included in the behavioral study. They listened to sequences of repeating $ABA_$ triplets and were instructed to continuously report their ongoing percept by selectively pressing one of two different buttons on a keypad. Subjects began reporting their percept typically 2 s after stimulus onset as integration (I; a single, coherent stream, the galloping pattern $ABA_ABA_$) or segregation (S; two simultaneous distinct streams $A_A_A_A_$ and $_B___B_$) (Fig 1 A-B).

Stimuli

Stimuli were 30-s long sequences of triplets $ABA_$ that consisted of alternating high (A) and low (B) pure tones gated with 10 ms raised cosine ramps and followed by a 125 ms silent pause "$_$" (Fig 1A). In total, triplets were 500-ms in duration and were repeated 60 times per trial. Tones were separated in frequency by DF semitones chosen from three conditions (DF= 3, 5, 7) with each condition being presented five times per experimental block. To prevent habituation to a certain frequency, for each DF the tones were generated by roving through variants of frequencies taken 0, ±1 or ±2 semitones apart from their geometric mean (middle pair in the list below); see also [4]. Frequencies (f_A, f_B), in Hz, were chosen as: $(494, 415)$, $(523, 440)$, $(554, 466)$, $(587, 494)$ or $(622, 523)$ Hz at DF=3; $(523, 392)$, $(554, 415)$, $(587, 440)$, $(622, 466)$, $(659, 494)$ at DF=5; and $(554, 370)$, $(587, 392)$, $(622, 415)$, $(659, 440)$, $(698, 466)$ at DF=7. Stimuli were digitally generated in Matlab at 48 kHz sampling rate and were delivered through earphones in a soundproof isolated room. Subjects had the sound volume adjusted to

their comfortable hearing level.

Experimental protocol

Each subject performed 9 experimental blocks. Each DF condition was randomly presented 5 times per experimental block, using different combinations of frequencies for tones A and B, without repetition (see *Stimuli*). A Latin square design was used to determine the order of presentation of each condition in each block. This resulted in group data with 675 30-s trials for each of three DF values. The frequency separation values (DF= 3, 5, 7) were chosen to fall within the range of ambiguity of the van Noorden diagram in which listeners can perceive both integration and segregation [10, 36]. All subjects underwent a training session in which they were given verbal explanations and auditory illustrations of the two possible percepts, and they practiced distinguishing between them. Then, during the recording session, listeners were instructed to press and hold one key on a keypad when they perceived stimuli as I, and to release it while pressing another key when they perceived stimuli as S, and so on. They were encouraged to respond as soon as they heard the change in percept. The key-response data were converted to binary vectors with value 0 assigned to I (and to the latency period defined as the time before identification of either percept) and 1 to S, for further analysis. Experiments were performed in a dedicated soundproof booth in the Human Brain Research Laboratory, Neurosurgery Department at The University of Iowa. Written informed consent was obtained from all subjects. Research protocols were approved by the University of Iowa Institutional Review Board.

Build-up functions and the latency period

The time course of S percept after stimulus onset was computed from key-pressed data, 0 for I and 1 for S. Those were sampled at 1 ms to create vectors of binary values corresponding to appropriate percept type at a particular time instance. At each DF

condition, binary vectors were averaged across 675 trials to obtain the build-up function of S; bootstrapping was also used to compute the 95% confidence interval (CI) around the mean (Fig 1 B-C). The time course of I was computed with the same procedure but over key-pressed data labeled as 1 for I and 0 for S. At a particular time t, the proportion of trials classified as I, S or neither (during the latency period/the first few seconds after stimulus onset) were $p_{int}(t)$, $p_{seg}(t)$ and $p_{Lat}(t)$, and summed up to $p_{int}(t) + p_{seg}(t) + p_{Lat}(t) = 1$. The first percepts were typically I. However, for larger values of tone frequency separation, subjects tended to report S as first percept more often which led to an increase in $p_{seg}(t)$.

First durations and subsequent durations

All complete percept durations across trials and conditions were included in the behavioral analysis. Unfinished percepts and button presses recorded after the end of stimulus presentation were discarded. For each DF= 3, 5, 7, the statistics was evaluated over four subsets of data, separately: first I, first S, subsequent I and subsequent S. For each DF and each of these four percept types, the mean dominance duration was computed in two steps: first, it was computed per subject, say μ_i for subject $i = 1, ..., 15$; then the mean duration μ of the group data was defined (and reported) as the unweighted average across all subjects, $\mu = (\mu_1 + \mu_2 + \cdots + \mu_{15})/15$ (e.g. Fig 1C; mean μ is shown for first duration distributions at DF=5). By this approach, any potential bias of the calculation towards fast switchers who might contribute more durations to the pool and concurrently spend less time in a particular percept, was mitigated. Error bars at 95% CI of the mean were also determined; error bars corresponded to 1.96 SE; standard error SE= $std_{exp}/\sqrt{15}$ was computed from the standard deviation std_{exp} over the group of means μ_i. For analysis of grouped data we used subject-specific normalized (by individual subject mean) percept durations as follows: at each DF condition and each percept type (first/subsequent, I/S), any raw

percept duration D of subject i was normalized by the corresponding mean μ_i to $\tilde{D} = D/\mu_i$. Histograms of normalized phase durations $\tilde{D}$ for each DF condition and percept type were computed and fit by gamma distributions with density functions

$$f(\tilde{D}|\alpha, \tilde{\mu}) = \frac{\alpha/\tilde{\mu}}{\Gamma(\alpha)} \left(\alpha \tilde{D}/\tilde{\mu}\right)^{\alpha-1} e^{-\alpha \tilde{D}/\tilde{\mu}}, \quad \tilde{\mu} \approx 1. \tag{1}$$

Mean $\tilde{\mu}$ was well-fit to 1 due to normalization. Then the coefficient of variation $CV = 1/\sqrt{\alpha}$ and the skewness $\gamma_1 = 2/\sqrt{\alpha}$ depended exclusively on the shape parameter α. If α was large then (1) was equivalent to a normal distribution. If $\alpha \approx 1$ then (1) was equivalent to an exponential distribution. On the other hand, for $\alpha \approx 2$ (as observed for first durations in behavioral data) and $\alpha \approx 2.6$ (as observed for subsequent durations), the distributions satisfied the scaling property $\gamma_1 = 2CV$ with $CV \approx 0.7$ and $CV \approx 0.6$ respectively. The latter case was similar to the findings of [27] that described the statistics of percept durations for other examples of perceptual bistability.

Distribution testing of phase durations

The fitting of the experimental (and numerical) data was obtained by calculating the values of α and $\tilde{\mu}$ with the Maximum Likelihood Estimation (MLE) algorithm. The goal was to determine α and $\tilde{\mu}$ that yielded the maximum product $\prod_k y_k$ of all y_k gamma-likelihood values of the normalized percepts $\tilde{D}_k$ counted by index k for each run of the experiment. This was equivalent to maximizing the log-likelihood $LL = \ln \prod_k y_k = \sum_k \ln y_k = \sum_k \left((\alpha - 1) \ln \tilde{D}_k - \frac{\alpha}{\tilde{\mu}} \tilde{D}_k + \alpha \ln \frac{\alpha}{\tilde{\mu}} - \ln \Gamma(\alpha)\right)$ based on formula (1). The optimization of LL was implemented numerically with MATLAB function fminsearch. Distribution testing on normalized durations was done by statistical bootstrapping. We generated 10000 bootstrapping sets of gamma distributions with fitted parameters α and $\tilde{\mu}$ and constructed the distribution of maximum log likelihood values for those sets. The test statistics LL was compared to this distribution to obtain the probability of log likelihood to be less than LL (the

p-value). The normalized durations were well fit by a gamma distribution with the optimal values α and $\tilde{\mu}$ (as null hypothesis) at significance level 0.05 if p-value ≥ 0.05.

Statistical analysis of model-generated data

The histograms of first and subsequent durations I and S in trials generated by the model (see below), and their fitting by gamma distributions, were computed in a similar manner as for the experimental data. Likewise, build-up functions for the model were constructed as those for the behavioral data.

The evidence accumulation model

Our proposed EVA model is a feedforward network of 3 layers: the input layer of spiking units, the sampler layer of binary response units, and the accumulation layer of two accumulators. The time-unit of the model is discrete and defined as the position of the triplet in the auditory sequence. Every DF-dependent numerical simulation of the EVA model consisted of N_{tr}=675 repetitions (trials) to mimic the setup from the behavioral experiment. The trials were then used to generate the statistics of the percept durations in terms of mean values, shape of distributions, and buildup functions. Finally, in order to test for the model's robustness in the presence of noise (not for sensitivity to model parameter values), this numerical procedure was run 100 times for each DF= 3, 5, 7 condition separately, and the results were characterized by averaged values and their 95% CI.

Input layer (IL)

The IL-units were assumed to be tone-A selective neurons from primary auditory cortex (A1) as described in [1]. The averaged (over trials) spike counts $m_{t,DF}$ of the IL-units were derived from data (see section *Data-driven parameters for EVA model*) and depended on the position t of the triplet in the $ABA_$ sequence (t=1, . . . , 60 for a

60-triplet long stimulus; 30 s in duration) and on the semitone difference DF. The model was simplified by focusing only on the spike counts during the B-tone presentation at A-tone selective neurons in A1. As reported by multi-unit recordings in monkeys, such A1-neurons adapted strongly and rapidly during presentation of triplet-repeating auditory sequences [1, total of 91 neurons]. Temporal correlations between the means of an A1-neuron from triplet to triplet were captured in the model by the trend of $m_{t,DF}$ that decreased exponentially with t (Fig 2B). Trial-to-trial variability of the dynamics of IL units as well as unit variability in IL during a single trial were implemented using Poisson point processes. (We used this approach because we could extract mean spike counts from published data of [1] but did not have access to the original spike times.) For an A1-neuron with mean spike count $\lambda = m_{t,DF}$ we supposed that its instantaneous spike count k ($k = 0, 1, 2, \ldots$) at triplet t and condition DF, was randomly generated from a Poisson distribution with probability $P(X = k) = \lambda^k \, e^{-\lambda}/k!$. The Poisson spike counts are generated independently for each neuronal unit in IL, each triplet and each semitone difference condition. Note that the model could be generalized by assuming neuronal heterogeneity, with averaged spike counts $m^j_{t,DF}$ at neuron j chosen from a normal distribution $\mathcal{N}(m_{t,DF}\,,\, s_{t,DF})$ of mean $m_{t,DF}$ and standard deviation $s_{t,DF}$ derived from [1]. However, the impact of heterogeneity on the model's outcome would be negligible given that EVA model was formulated to use mean spike counts over IL neuronal pools as input rather than mean spike counts of individual neurons (see below).

Sampler layer (SL)

The SL-units were tasked with summating and classifying spike counts from subsets of N_{in} IL units (A1-neurons). Consider that a trial of length N_t (N_t=60 triplets) during a certain DF condition was simulated by EVA model: each sampler summed the input of a pool of N_{in} neuronal units from IL; weighted by N_{in}, this gave the mean spike count

for B-tones of the corresponding pool of A-tone selective A1-neurons, $\bar{X}_t = (\sum_{j=1}^{N_{in}} X_t^{(j)})/N_{in}$ for triplet t; then $\bar{X}_t$ was compared to a DF-independent, fixed, neuronal threshold C_{th}. If the averaged spike count was large ($\bar{X}_t \geq C_{th}$) then the subset activity was high and the pool was assumed to support, for this triplet, the integration percept I. The sampler's response at triplet t was classified as "0". If the averaged spike count was small ($\bar{X}_t < C_{th}$), the subset activity was low and the IL-pool was said to support the segregation percept S. The sampler's response was classified as "1" (Fig 2A). Therefore, at each triplet t, each sampler behaved like a biased coin being flipped with probabilities $p_{0;t,DF}$ and $p_{1;t,DF}$ over the binary probability space of 0 and 1. Since neuronal units in each IL-pool followed independent Poisson distributions of parameters $m_{t,DF}$, the pool itself was also a Poisson process defined by the product $N_{in}\, m_{t,DF}$. Then, the samplers were binary signal detectors with probabilities

$$p_{1;t,DF} \quad = \quad Prob\Big(\bar{X}_t < C_{th}\Big) = \sum_{0 \leq k < C_{th} N_{in}} \frac{\big(N_{in}\, m_{t,DF}\big)^k\, e^{-N_{in}\, m_{t,DF}}}{k!} \qquad (2)$$

and $p_{0;t,DF} = 1 - p_{1;t,DF}$ calculated over 675 repetitions of the model in order to maintain similarities to the behavioral experiment, and with expected value and variance $E[\bar{X}_t] = p_{1;t,DF}$ and $Var[\bar{X}_t] = p_{1;t,DF}\,(\,1 - p_{1;t,DF}\,)$.

Three DF-independent parameters were associated with SL: N_{in}, the number of A1 inputs to a sampler unit; C_{th}, the neuronal counting threshold that categorizes ensemble activity in A1 as high (class 0) or low (class 1); and N_{sl}, the number of neuronal units in SL. The values of N_{in} and C_{th} were obtained by least-squares fit between probabilities $p_{1;t,DF}$ and the "asymptotic" levels of the psychometric buildup functions (last 15 seconds of trial duration) for all DF= 3, 5, 7 (Fig 5B; also section *Data-driven parameters for EVA model*). The psychometric buildup represented the fraction (over the trials) of the segregation percept S reported by all subjects at each time point and DF during the 30-s long trial. Through this optimization procedure (Fig 5; optimal values obtained for $N_{in} = 5$, $C_{th} = 4.21$) probabilities $p_{1;t,DF}$ of a

sampler to be in a state that supported percept S were estimated – at least for triplets several seconds from the stimulus onset – as

$$p_{1;t,DF} \approx p_{seg;DF} = 0.4,\ 0.6,\ 0.75 \quad \text{at} \quad DF = 3, 5, 7. \tag{3}$$

The inclusion of the sampler layer in the model ($N_{sl} > 1$) ensured neuronal granularity that was found by other studies to be a suitable substrate for perceptual representations [27]. In particular, at each triplet t, some of the N_{sl} samplers were in class 1 showing low A1 spiking and thereby associated with segregation [1, 7] while others were in class 0 supporting integration. The percentages $p_S(t)$ and $p_I(t) = 1 - p_S(t)$ of such samplers were taken herein as stochastic (over trials) output of SL (Fig 2A).

Accumulation layer (ACC)

The accumulation layer consisted of two units whose dynamic states x_I and x_S changed from triplet to triplet according to Eqs

$$\begin{aligned} x_I(t+1) &= x_I(t) + (T_I - x_I(t))\, p_I(t) + \sigma_I \xi_I(t), \\ x_S(t+1) &= x_S(t) + (T_S - x_S(t))\, p_S(t) + \sigma_S \xi_S(t). \end{aligned} \tag{4}$$

The accumulator for x_I gathered evidence that favored integration (through input $p_I(t)$ from SL) while the accumulator for x_S gathered evidence that favored segregation (through input $p_S(t)$ from SL). Importantly, their states were influenced by the perceptual context as well (Fig 2C, solid lines in blue and red illustrated traces for x_I and x_S; the background color showed the percept's type, blue for I and red for S). In particular, if *segregation* was the current dominant percept then x_I accumulated evidence *against* segregation and aimed to reach target T_I=T_a; simultaneously, x_S accumulated evidence *for* the current percept and it drifted instead towards target T_S=T_f. Discrete-time Gaussian white noise processes $\sigma_I \xi_I(t)$, $\sigma_S \xi_S(t)$ with zero mean and standard deviations σ_I=σ_a and σ_S=σ_f, interacted with the stochastic inputs from

SL to produce certain levels of fluctuations. The additive stochastic terms in ACC were target-dependent with $\sigma_a > \sigma_f$ for $T_a > T_f$ (Fig 2C). On the contrary, if the current percept was integration then x_I accumulated evidence for I and approached T_I=T_f while x_S accumulated evidence against I and approached T_S=T_a. The level of local noise was adjusted accordingly to values σ_I=σ_f and σ_S=σ_a.

Switches and resetting conditions

In the experiment, subjects identified the dominant percept by pressing a certain button on the keypad. Equivalently, in EVA model, the switch from one dominant percept to the next occurred when either $x_I(t)$ or $x_S(t)$ crossed the ACC threshold set to 1. Herein, the alternations were initiated by the accumulator that observed how many samplers in SL opposed the current percept at each triplet t. Its trace was attracted to target T_a that lay near the threshold, then was pushed across the threshold by the noise of strength σ_a. In the meantime, the accumulator in favor of the current percept hovered around T_f with fluctuations set by σ_f. For example during percept I, accumulator x_S was the first to reach threshold 1 producing a switch to percept S; then, during S, x_I reached threshold 1 leading to another switch to subsequent percept I, and so on (Fig 2C). At every change in percept, the accumulators were reset to new levels x_I^+, x_S^+. These were defined as x_*^- where x_*^- was the value that the evidence-for accumulator x_* ($* = I, S$ during current percept I, S respectively) reached just before the switch.

The simulations took into account the proportion of first percepts reported as segregation at each DF during the behavioral experiment as well as the latency period during each trial. The initial conditions of the accumulators were set to a DF-independent baseline value b and kept constant during the entire latency period (calculated in length of T_{Lat} triplets) of any given trial. We defined $x_I(t) = x_S(t) = b$ for all triplets t between 1 and T_{Lat}; then at $t = T_{Lat}$ the type of the current first percept, I or S, was imported from the behavioral data; the dynamics of the accumulators for

$t \geq T_{Lat}$ were then determined according to Eqs (4) and the associated reset conditions.

The choice of parameters (σ_f small) and of reset conditions ($x_a^+ = x_f^+ = x_f^-$ where x_a, x_f are accumulators "against" and "for" the dominant percept) ensured that the switch was triggered by the dynamics of x_a. Rare events when x_f might have crossed the threshold ahead of x_a were disregarded. Another possible implementation of resetting would allow for correlations between consecutive percepts; it could depend on each accumulator state just before a switch, a simple interchange of roles such that ACC variables remained continuous, $x_f^+ = x_a^- = 1$, $x_a^+ = x_f^-$ (not shown in this paper).

Model analysis

Parameter values used in model simulations

All figures were generated with the following parameter values: $N_{in} = 5$, $C_{th} = 4.21$, $N_{sl} = 20$, $b = 0.7$, $T_f = 0.6$, $\sigma_a = 0.085$, $\sigma_f = 0.03$ and decision threshold $\theta = 1$, unless otherwise stated in their caption. (For parameter values $m_{t,DF}$ associated with IL, see *Data-driven parameters for EVA model.*) Target T_a was initially chosen equal to 0.9 and then was fine-tuned to best fit the mean dominance durations of the first and subsequent integration and segregation percepts from the behavioral data – within $\pm 10\%$ standard error (SE) of the experimental mean values (Figs 3 and 4; also S3 Fig). Its values changed with DF, and with classification (first or subsequent) and type (I or S) of the percept. We used the notation T_{aI1} for "target against integration, first percept", T_{aS1} for "target against segregation, first percept", T_{aI2} for "target against integration, subsequent percepts", and T_{aS2} for "target against segregation, subsequent percepts", respectively. Therefore, in the model, whenever acting as target-against, $T_S = T_{aI1}$ or T_{aI2} while $T_I = T_{aS1}$ or T_{aS2}. In simulations we used the following values (see red diamonds in Fig 6):

At DF=3: $T_{aI1} = 0.8273$, $T_{aS1} = 0.9273$, $T_{aI2} = 0.8924$, $T_{aS2} = 0.8924$;

At DF=5: $T_{aI1} = 0.9000$, $T_{aS1} = 0.8909$, $T_{aI2} = 0.9288$, $T_{aS2} = 0.9106$;

At DF=7: $T_{aI1} = 0.9348$, $T_{aS1} = 0.8773$, $T_{aI2} = 0.9242$, $T_{aS2} = 0.9318$.

Selection of target-against T_a values

The mean duration μ obtained by numerical simulations of EVA model was compared to its behavioral counterpart μ_{exp} for first and subsequent I, S percepts, and each DF= 3, 5, 7. Behavioral results from 15 subjects were characterized by group mean data μ_{exp} and 95% CI with CI corresponding to 1.96 SE (Fig 3A and Fig 4A, lower panel). Then the model was considered to provide a good approximation of the experimental data if μ belonged to a narrow band within 1 SE from μ_{exp} (Fig 6, green region and black dots). This was equivalent to the relative error $|\mu/\mu_{exp} - 1| \leq CV/\sqrt{15}$ where $CV = std_{exp}/\mu_{exp}$ was the coefficient of variation computed over the group of subjects. Parameters T_a, σ_a used for the numerical simulations of EVA model (Fig 6, red diamond) were chosen as follows: $\sigma_a = 0.085$ was kept fixed while values of T_a were determined by restricting the error magnitude to only 10% SE (i.e. $|\mu - \mu_{exp}| \leq 0.1$ SE); then, among the latter set we selected the value T_a that yielded the least error in shape of the gamma-fit distributions (see Eq (1)).

Statistical properties of SL-activation

At any fixed triplet position t in the $ABA_$ sequence presentation, each of the N_{sl} samplers was equivalent to an independent Bernoulli process (during repeated trials) with probability of success $p_{1;t,DF}$ and probability of failure $p_{0;t,DF} = 1 - p_{1;t,DF}$. Likewise, the state of SL described a binomial process equivalent to the random variable, over trials, $N_{sl}\, p_S(t)$ where $p_S(t)$ represented the percentage of samplers in class 1 that favored segregation at triplet t. The stochastic process had mean $N_{sl}\, p_{1;t,DF}$ and variance $N_{sl}\,(1 - p_{1;t,DF})\, p_{1;t,DF}$. As a result, the first two moments of the output $p_S(t)$ and $p_I(t)$ of SL were well-approximated, for sufficiently large

triplet-indexes ($t \geq 30$; see Eq (3) and Fig 5B, 2nd panel), by

$$E\big[p_S(t)\big] = p_{seg;DF}\,, \qquad E\big[p_I(t)\big] = 1 - p_{seg;DF}\,,$$
$$Var\big[p_S(t)\big] = Var\big[p_I(t)\big] = \frac{1}{N_{sl}}\,(1 - p_{seg;DF})\,p_{seg;DF}. \tag{5}$$

In particular, for large N_{sl} the variance of $p_S(t)$ and $p_I(t)$ became negligible while their means remained unchanged.

Selection of noise level σ_a

Our EVA model features accumulation that could saturate, given that target-against T_a was assumed to be subthreshold ($T_a < 1$). Hence the noise level σ_a has to be sufficiently large in order for the trajectory of the accumulator drifting towards T_a to reach the ACC-threshold. A theoretical lower-bound estimate for σ_a was obtained by assuming N_{sl} large and focusing only on the properties of the subsequent percept durations. Under such assumptions, $p_S(t)$ and $p_I(t)$ were approximately constant as demonstrated by Eqs (3) and (5). Then, after each switch, both accumulators satisfied an equation of the form $X_{n+1} = X_n + (T - X_n)p + \varepsilon_{n+1}$ with $X_0 \approx T_f$; T, σ taken as either T_a, σ_a or T_f, σ_f; and independent random variables $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$. Such an equation describes a stationary first order autoregressive model with parameter λ=1-p [37]. Therefore, at the n-th triplet during a percept immediately following a switch, the states of the accumulators followed a normal distribution with mean $E[X_n]$=T-(T-T_f)λ^n and variance $Var[X_n]$=$\sigma^2(1-\lambda^{2n})/(1-\lambda^2)$. In particular, at the n-th triplet during integration, the mean and variance of x_S (x_I) that accumulated evidence against (for) the percept were $E[x_S] = E_a$, $Var[x_S] = V_a$, $E[x_I] = E_f$, $Var[x_I] = V_f$ where

$$E_a = T_a - (T_a - T_f)p^n, \quad V_a = \frac{1-p^{2n}}{1-p^2}\sigma_a^2, \quad E_f = T_f, \quad V_f = \frac{1-(1-p)^{2n}}{1-(1-p)^2}\sigma_f^2$$

with $p = 1 - p_{seg;DF}$. Then at the n-th triplet during segregation they were $E[x_I] = E_a$, $Var[x_I] = V_a$, $E[x_S] = E_f$, $Var[x_S] = V_f$ with E_a, V_a, E_f, V_f defined as above but computed with $p = p_{seg;DF}$. Given that 3 times the standard deviation from the mean

accounts for 99.73% of values in a normal distribution, if σ_a was too small the accumulators could cross the threshold 1 only with very small probabillity. From the calculation above, a lower bound for σ_a in the model was estimated at $\sigma_a > \sigma_{a,min}$ with $\sigma_{a,min} = (1 - T_a)\sqrt{1 - p_M^2}/3$, where p_M was the maximum of $p_{seg;DF}$ and $1 - p_{seg;DF}$ for all DF= 3, 5, 7. For example, if $T_a = 0.9$ then a necessary condition for switching was $\sigma_a > 0.022$.

Statistical properties of ACC-activation

As explained in the previous section, for small σ_a the accumulators in the EVA model could not cross threshold 1 (Fig 6; na, gray region). Numerical simulations showed that when σ_a increased to the right of curve $\sigma_a = \sigma_{a,min}$ in the (σ_a, T_a)-plane, alternations between percepts occurred and the dominant durations were distributed according to: normal distributions at σ_a small (the parameter α in Eq (1) was very large) or to exponential distributions at σ_a large (α in (1) was near 1); see Fig 6, regions labeled "N" and "E", respectively. At intermediate values σ_a, the distributions were gamma-like with shape close to that found experimentally (Fig 6, region labeled "G" between the two white curves). In the latter case, parameter α in (1) was either near 2 (for first percept durations) or near 2.6 (for subsequent durations), and it differed from α_{exp} by relative error up to 20%, $|\alpha/\alpha_{exp} - 1| \leq 0.2$ (except for first and subsequent I at DF=7 at which the range for α was extremely narrow and we allowed for a larger error, up to 30% instead). The range for σ_a that led to gamma-like distributions varied slightly with N_{sl} with the biggest difference being identified at $N_{sl} = 1$; see S6 Fig.

Data-driven parameters for EVA model

Parameter mean values $m_{t,DF}$ were used to generate Poisson spike counts of IL-units at each triplet t ($t = 1, 2, \ldots, 60$) in the $ABA_$ sequence and for each semitone difference DF=3, 5, 7; see section *Input Layer* and Eq (2). They were derived from multi-unit

spiking neural data recorded from macaque monkey primary auditory cortex A1 by [1], using a procedure that combined exponential fitting with numerical interpolation. First, mean spike counts $m_{t,DF}$ at A-tone selective neurons in A1 during presentation of tones A, B and A in $ABA_$ were extracted from [1] for each triplet $t \leq 20$ in the sequence and for each DF=1, 3, 6, 9; See scatter points in [1, Fig.3]; also S5 Fig. The mean spike counts at each tone decreased from value $m_{1,DF}$ measured at the first triplet to some level m^*_{DF} at which they stabilized after a few seconds since stimulus onset. They were fitted by functions

$$m_{t,DF} \approx m^*_{DF} + (\, m_{1,DF} - m^*_{DF}\,)e^{-1.1(t-1)} \tag{6}$$

with parameters m^*_{DF} chosen to minimize the least-squares error between the extracted mean spike counts $m_{t,DF}$ and the corresponding exponential curve (S5 Fig, solid curves). In particular, significant differences in mean spike counts at different DF values were observed only during tone-B presentation with fitting (6) achieved for parameter values $m_{1,1} = 7.25$, $m_{1,3} = 6.25$, $m_{1,6} = 6$, $m_{1,9} = 5.25$ (according to data from [1]) and $m^*_1 = 6.09$, $m^*_3 = 4.57$, $m^*_6 = 3.95$, $m^*_9 = 3.44$ respectively. Secondly, simulations of EVA model were performed for DF= 3, 5, 7 instead of 1, 3, 6, 9, and for a total of 60 rather than 20 triplets. We implemented these constraints in two steps: for DF=3 we chose the mean spike counts $m_{t,DF}$ as in [1] for $t \leq 20$ and as m^*_{DF} for $t > 20$. Then for DF=5 and DF=7 and each triplet t we defined the mean spike counts by interpolation using the power function $m_{t,DF} = a_t\, DF^{b_t}$ whose coefficients a_t, b_t satisfied the least-squares fit between this curve and the points (DF , $m_{t,DF}$) defined by (6) for all DF=1, 3, 6, 9 at any fixed t.

The mean spike counts for B-tones of any pool of A-tone selective IL neuronal units, as well as the standard error to the mean, were computed from simulation of Poisson processes with parameter $m_{t,DF}$ while varying N_{in} (Fig 5A). Then a threshold value C_{th} was chosen to minimize the squared differences error between the model-based

probabilities (2) of a sampler to support the segregation percept for all DF=3, 5, 7 and the behavioral buildup functions, applied to the last 30 triplets (15 seconds) of the stimulus (Fig 5B). The pair of parameter values $N_{in} = 5$, $C_{th} = 4.21$ that generated the minimum error was then used in numerical simulations of the EVA model.

EVA model versus classical drift-diffusion models

To gain some intuition about the accumulation process in our EVA model and about the timing of switch events, we considered an approximation of the stochastic Eqs (4) in continuous time. For that, we assumed the drift in (4) to be constant and neglected its dependence on activity x. Then percept durations corresponded to the first-passage time of the ACC-unit that accumulated evidence against the current percept. Its equation could be interpreted as the constant-drift continuous-time diffusion model (DDM) $dx = \gamma_a dt + \sigma_a \, dW_t$ with positive drift rate γ_a, noise amplitude σ_a, Gaussian white noise dW_t, and decision threshold $\theta = 1$. In this DDM, the likelihood of first-passage at time t follows an inverse Gaussian distribution [29] with density function $f(t) = \frac{1}{\sigma_a \sqrt{2\pi t^3}} \exp\left(-\frac{(t-1/\gamma_a)^2}{2t(\sigma_a^2/\gamma_a^2)}\right)$ and mean $1/\gamma_a$, variance $\gamma_a + \sigma_a^2$, and coefficient of variation $CV = \sqrt{1/\gamma_a + \sigma_a^2/\gamma_a^2}$. Moreover, the inverse Gaussian resembles a gamma distribution for large CV but converges to a normal distribution as σ_a decreased in relation to drift rate γ_a [29]. To some extent, the dynamics of the discrete-time ACC (4) share similarities with DDM above. Numerical simulations of our EVA model showed that gamma-like distributions of percept durations were possible only for σ_a chosen in a restricted parameter range, given fixed targets T_a and T_f (see section *Statistical properties of ACC-activation*). Outside this range, percept durations followed either normal distributions (for lower values of σ_a) or exponential distributions (for larger values of σ_a). However, the accumulation process in the EVA model is different than in the DDM for several reasons: Eqs (4) are discrete-time drift diffusion models; they

include leakage; their deterministic version admits bistable non-oscillatory solutions (no threshold crossing); and the input drive from SL is itself stochastic with fluctuations described by (5).

Acknowledgments

The authors thank Xiayi Wang, Haiming Chen and Kirill V. Nourski for help with data acquisition, and Dan Tranchina for suggestions on data analysis.

References

1. Micheyl C, Tian B, Carlyon R, Rauschecker J. Perceptual Organization of Tone Sequences in the Auditory Cortex of Awake Macaques. Neuron. 2005;48:139–148.

2. Pressnitzer D, Hupé JM. Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. Current Biology. 2006;16:1351–1357.

3. Pressnitzer D, Sayles M, Micheyl C, Winter IM. Perceptual organization of sound begins in the auditory periphery. Current Biology. 2008;18:1124–1128.

4. Deike S, Heil P, Böckmann-Barthel M, Brechmann A. The buid-up of auditory stream segreagtion: a different perspective. Frontiers in Psychology. 2012;3:461:1–7.

5. Barniv D, Nelken I. Auditory streaming as an online classification process with evidence accumulation. PLoS ONE. 2015;10(12):e0144788:1–20.

6. Rankin J, Sussman E, Rinzel J. Neuromechanistic model of auditory bistability. PLoS Computational Biology. 2015;DOI:10.1371/journal.pcbi.1004555:1–34.

7. Fishman YI, Reser DH, Arezzo JC, Steinschneider M. Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hearing Research. 2001;151:167–187.

8. Fishman YI, Arezzo JC, Steinschneider M. Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration. J Acoust Soc Am. 2004;116(3):1656–1669.

9. Curtu R, Wang X, Brunton BW, Nourski KV. Neural Signatures of Auditory Perceptual Bistability Revealed by Large-Scale Human Intracranial Recordings. Journal of Neuroscience. 2019;39(33):6482–6497.

10. van Noorden LPAS. Temporal coherence in the perception of tone sequences. Doctoral Dissertation, Eindhoven University of Technology; 1975.

11. Beauvois M, Meddis R. Computer simulation of auditory stream segregation in alternating-tone sequences. J Acoust Soc Am. 1996;99(4):2270–2280.

12. Micheyl C, Carlyon RP, Gutschalk A, Melcher JR, Oxenham AJ, Rauschecker JP, et al. The role of auditory cortex in the formation of auditory streams. Hearing Research. 2007;229:116–131.

13. Krishnan L, Elhilali M, Shamma S. Segregating Complex Sound Sources through Temporal Coherence. PLoS Computational Biology. 2014;10(12):e1003985:1–10.

14. Mill RW, Böhm TM, Bendixen A, Winkler I, Denham SL. Modelling the Emergence and Dynamics of Perceptual Organisation in Auditory Streaming. PLoS Computational Biology. 2013;9(3):e1002925.

15. Almonte F, Jirsa V, Large E, Tuller B. Integration and segregation in auditory streaming. Physica D. 2005;212:137–159.

16. Wang D, Chang P. An oscillatory correlation model of auditory streaming. Cognitive Neurodynamics. 2008;2(1):7–19.

17. Steele S, Tranchina D, Rinzel J. An alternating renewal process describes the buildup of perceptual segregation. Frontiers in Computational Neuroscience. 2015;8:166:1–13.

18. Szabó BT, Denham SL, Winkler I. Computational Models of Auditory Scene Analysis: A Review. Frontiers in Neuroscience. 2016;10:524: 1–16.

19. Rankin J, Rinzel J. Computational models of auditory perception from feature extraction to stream segregation and behavior. Current Opinion in Neurobiology. 2019;58:46–53.

20. Hill KT, Bishop CW, Miller LM. Auditory grouping mechanisms reflect a sound's relative position in a sequence. Frontiers in Human Neuroscience. 2012;6(158):1–7.

21. Scholes C, Palmer A, Sumner CJ. Stream segregation in the anesthetized auditory cortex. Hearing Research. 2015;328:48–58.

22. Shpiro A, Moreno-Bote R, Rubin N, Rinzel J. Balance between noise and adaptation in competition models of perceptual bistability. Journal of Computational Neurosience. 2009;27(1):37–54.

23. Moreno-Bote R, J R, N R. Noise-induced alternations in an attractor network model of perceptual bistability. Journal of Neurophysiology. 2007;98:1125–1139.

24. Latimer K, Yates J, Meister M, Huk A, Pillow J. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. Science. 2015;349 (6244):184–187.

25. Shadlen MN, Kiani R, Newsome WT, Gold JI, Wolpert DM, Zylberberg A, et al. Comment on "Single-trial spike trains in parietal cortex reveal discrete steps during decision-making". Science. 2016;351 (6280):1406.

26. Zoltowski DM, Latimer KW, Yates JL, Huk AC, Pillow JW. Discrete Stepping and Nonlinear Ramping Dynamics Underlie Spiking Responses of LIP Neurons during Decision-Making. Neuron. 2019;102:1249–1258.

27. Cao R, Pastukhov A, Mattia M, Braun J. Collective activity of many bistable assemblies reproduces characteristic dynamics of multistable perception. Journal of Neuroscience. 2016;36 (26):6957–6972.

28. Ermentrout GB, Terman DH. Mathematical Foundations of Neuroscience. Springer. 2010;35.

29. Simen P, Rivest F, Ludvig EA, Balci F, Killeen P. Timescale Invariance in the Pacemaker-Accumulator Family of Timing Models. Timing & Time Perception. 2013;1:159–188.

30. Shpiro A, Curtu R, Rinzel J, Rubin N. Dynamical characteristics common to neuronal competition models. Journal of Neurophysiology. 2007;97:462–473.

31. Wang XJ, Rinzel J. Alternating and synchronous rhythms in reciprocally inhibitory model neurons. Neural Computation. 1992;4:84–97.

32. Skinner F, Kopell N, Marder E. Mechanisms for oscillation and frequency control in reciprocally inhibitory model neural networks. Journal of Computational Neurosience. 1994;1:69–87.

33. Curtu R, Shpiro A, Rubin N, Rinzel J. Mechanisms for frequency control in neuronal competition models. SIAM Journal on Applied Dynamical Systems. 2008;7(2):609–649.

34. Curtu R, Rubin J. Interaction of canard and singular Hopf mechanisms in a neural model. SIAM Journal on Applied Dynamical Systems. 2011;10(4):1443–1479.

35. Rankin J, Osborn Popp PJ, Rinzel J. Simulus pauses and perturbations differentially delay or promote the segregation of auditory objects: psychoacoustics and modeling. Frontiers in Neuroscience. 2017;11 (198):1–12.

36. Bregman AS. Auditory scene analysis: the perceptual organization of sound. The MIT Press; 1990.

37. Novikov A, Kordzakhia N. Martingales and first passage times of AR(1) sequences. Stochastics: An International Journal of Probability and Stochastic Processes. 2008;80 (2-3):197–210.

Supporting information

S1 Fig. The evidence accumulation (EVA) model captures experimental mean duration (μ) and shape of gamma-like distributions (α) for first and subsequent percept durations at other DF values. Distributions of normalized phase durations are shown for A: DF=3 and B: DF=7. They are obtained from numerical simulations of the EVA model (columns 2,4) and compared to those derived from behavioral data (columns 1,3).

S2 Fig. Exemplar time courses of accumulators in the EVA model, shown for DF=3 (top) and DF=7 (bottom). In only a few trials, 103 out of 675 for DF=3 and 220 out of 675 for DF=7, the first percept is segregation (see panels 2,4). During a cycle, the suppressed unit accumulates evidence against the current percept until it reaches the switching threshold. Then, a perceptual switch occurs and accumulators are reset to the same value. In the noise free case, the accumulators

stabilize to their corresponding target values and there are no alternations. Such trajectories are depicted by dashed lines.

S3 Fig. EVA model with fixed target-against value T_a across all conditions and percept types captures some, but not all, characteristics of perceptual alternations. For comparison, mean durations and shape parameter α of gamma distributions are shown for A: Experimental data; B: EVA model with optimized values for target-against (see Methods, Parameter values used in model simulations). EVA-generated results are identical to those in Figs 3 and 4; and C: Non-optimized EVA simulated with $T_a = 0.9$ across all DF= 3, 5, 7 and first and subsequent I, S. All other parameters are as in panel B. The mean durations from simulations follow the trend of experimental data which is decreasing/increasing with DF for I/S respectively. However, they fail to approximate well the entire set of behavioral data (e.g. approximations of mean first durations at DF=3 and DF=7 are inaccurate). On the other hand, gamma-fit shape values α are comparable to those from panels A and B. This is not surprising given that α depends mostly on the noise-level σ_a, as shown in Fig 6.

S5 Fig. The signal detection algorithm for constructing a neurometric function (the probability of segregation as a function of time) generates acceptable buildup fits at DF= 1, 3, 6, 9. For comparison, see Micheyl et al (2005) [1]. Upper panel: mean spike counts $m_{t;DF}$ (scatter points) at A-tone selective neurons in A1 during tone B were extracted from [1, (Fig.3A)]. They correspond to conditions DF=1 (blue), 3 (green), 6 (red), 9 (cyan), based on 10 s (20 triplets) long trials. The mean spike counts decrease exponentially and stabilize within a few seconds (solid curves for the exponential fits). The algorithm generates spike counts during B-tone by using Poisson processes of means $m_{t;DF}$, and then average them over N_{in}

neuronal units. The average values of the mean spike counts, including asymptotic values (written in parenthesis) at each DF, and the standard error to the mean (SEM) are computed over 675 trials. Lower panel: The signal detection algorithm constructs neurometric functions using numerical data from all N_{in} neuronal units. Parameters N_{in} and C_{th} are chosen to yield SEM similar to those observed in the spike count data [1, (Fig.3A)] and to yield the least-squares error of the experimental buildups (dashed, extracted from [1, (Fig.4)] and the computer-simulated neurometric functions (solid) for DF= 1, 3, 6, 9. The best approximation is obtained for $N_{in} = 30$, $C_{th} = 4.64$. Note: Statistics of percept durations were not reported in [1]; this prevented us from comparing these aspects of behavioral data from [1] to our numerically-generated duration distributions at DF=1, 3, 6, 9.

S6 Fig. Mean percept durations (μ) and shape parameter values (α) in the EVA model are largely unaffected by changes in N_{sl}, the number of units in the sampler layer. Some important differences occur, however, at $N_{sl} = 1$ (e.g. for subsequent durations). A bifurcation diagram of the dependence of μ and α on target-against T_a and noise strength σ_a is shown for DF=5 and varying N_{sl} for A: first percepts and B: subsequent percepts. All parameters are chosen as described in Methods, except for N_{sl} (here $N_{sl} = 1$, 5 or 10). For comparison, see Fig 6, middle column; $N_{sl} = 20$ at DF=5. Red diamonds correspond to same parameter choices as in Fig 6 for $N_{sl} = 20$, as well. The heat map represents the ratio μ/μ_{exp} between model-generated μ and mean duration μ_{exp} from the behavioral data. Regions of no alternations (na) are colored in gray. Mean durations are much longer than their experimental counterparts μ_{exp} (region in warm colors), much shorter than μ_{exp} (region in cool colors), or close to μ_{exp} (within one standard error to μ_{exp}; in green; black dots depict a discrete selection of values in the green region). The distributions of normalized percept durations are characterized by three distinct regions: for small σ_a the

distributions are normal (region N, to the left of dashed-white line; $\alpha \gg 3$); for large σ_a the distributions are exponential (region E, to the right of solid-white line; α near 1); for intermediate values σ_a the distributions are gamma-like with shape close to that found experimentally (region G, between white contours; $\alpha \approx 2$ for first percepts and $\alpha \approx 2.6$ for subsequent percepts; α differs from α_{exp} by relative error up to 20% except for integration at DF=7 where it is up to 30%). As in Fig 6, middle column, the intersection of white contours with the sheet of black dots identifies parameter values that yield well-fit data. Note that at $N_{sl} = 1$ this intersection is empty for both subsequent percepts I and S (panel B, first column).

Fig 1. Stimulus, buildup, and distribution of first percept durations in auditory streaming of triplets. A: Stimulus paradigm (left) used for behavioral experiments and corresponding percept types (right). Stimulus consists of sequences of high (A) and low (B) pure tones presented as repeated triplets $ABA_$ where '$_$' denotes silent gap. Depending on DF between tones A and B, there are two fundamental percepts: integration (I; blue), one connected stream with galloping rhythm, and segregation (S; red), two parallel streams of high tone $A_A_A_A_$ and low tone $_B___B__$ occurring simultaneously. B: Computation of the buildup function (time course of probability of S) obtained by averaging binary response vectors over all trials at each time point τ up to 30 s (45 trials for each DF and subject; 15 subjects). Values are assigned as 1 for S (red) and 0 for non-S. The latter includes both latency (gray) and I (blue) states. Due to latency, the buildup function always starts at 0 even though the first percept is not necessarily I. For example, in Trial 2, the first percept is S. C: Experimental-based psychometric buildup function (upper panel) and distribution of first percept durations (middle and lower panels). Buildup functions are computed for DF=3 (green), 5 (red), and 7 (cyan). The error bars indicate 95% CI around the mean using statistical bootstrapping. Durations are normalized by dividing by mean duration. Likelihood ratio test confirms that normalized first percept durations are gamma distributed – shown here at DF=5 for I (N=533, p=0.49) and S (N=114, p=0.47). The shape parameters α, obtained by Maximum Likelihood Estimation (MLE), and the mean durations μ are indicated in the graphic. D: Model-based simulated buildup function (upper panel) and distribution of normalized first percept durations (middle and lower panels). Buildup functions from the evidence accumulation model (EVA; solid) closely resemble those from the behavioral experiment (dashed, also in C). Normalized first percept durations are gamma distributed (shown at DF=5). Similar results are obtained for other DF values; see S1 Fig.

Fig 2. Accumulation model as feed-forward auditory network of 3 layers. A: Input layer and sampler layer of evidence accumulation model. Input layer comprises A1 units with mean spike counts presented in panel B. Sampler layer has N_{sl}=20 binary neuronal units, either in state 0 (blue; favoring I percept) or state 1 (red; favoring S percept). Each unit samples a small number of input units (N_{in}=5) and the averaged spike count across the units is compared to C_{th} (see panel B) to determine the unit's appropriate perceptual state. B: Mean spike counts (scatter plot) for tone B of tone-A-selective neurons, and exponential fit (solid) of mean spike counts. These values are interpolated for our specific DF=3,5,7 using data from cortical area A1 of awake macaque extracted from [1]. A Poisson spike count is generated using the mean value at each triplet. Asymptotic values of mean spike count (printed in parenthesis next to corresponding DF values) are used to generate spike counts after the 20-th triplet. Poisson spike counts are averaged across sets of N_{in}=5 neuronal units, and the resulting values are subject to a binary neural threshold C_{th} (black horizontal line). The error bars indicate the standard errors of the mean spike counts. C: Accumulation layer has 2 accumulators drifting over successive triplets towards their own target values T_a and T_f where $T_a > T_f$. Their activities are governed by input factors from the sampler layer and stochastic factors. The noise level depends on the target ($\sigma_a > \sigma_f$). During a cycle, the suppressed unit accumulates evidence against the current percept. A switch to the other percept occurs when the accumulator of the suppressed unit reaches the switching threshold of 1. A new cycle starts, with accumulators reset to appropriate values, and targets values switched to corresponding perceptual states. Shown for DF=5. For other DF values, see S2 Fig.

Fig 3. EVA model yields realistic first percept durations. A: Mean percept durations (top) and fitted α value (bottom) from gamma distribution of first I (blue) and first S (red) percepts from behavioral experiment for DF=3,5,7. The error bars indicate 95% CI around the mean and are obtained from statistical bootstrapping (see *Methods*). The mean durations of I decrease with DF while those of S increase with DF. The shape parameters α from gamma-fit using MLE for DF=3,5,7 are also presented here. There is no observed trend for α values. B: Mean percept durations (top) and fitted shape parameter α (bottom) from gamma distribution of first I (blue) and first S (red) states from EVA model. The error bars are 95% CI obtained from 100 Monte Carlo runs to show the robustness of the model. The mean values of duration follow the similar trend as those from experiment. Also, the shape parameters show a close resemblance to those from the experiment. Related results are included in S3 Fig.

Fig 4. EVA model yields realistic subsequent percept durations and distributions. A: Distribution of normalized subsequent percept durations (top) and other properties (bottom) from behavioral experiment. (Top) Likelihood ratio test confirms that both subsequent I (blue; N=1642, p=0.49) and S (red; N=1785, p=0.49) percepts follow a gamma distribution, shown here for DF=5. The shape parameters α computed using MLE, and the mean durations μ are shown in the graphic. (Bottom) Mean subsequent durations for I (blue) and S (red) for DF=3,5,7. The error bars indicate 95% CI around the mean, computed using statistical bootstrapping (see *Methods*). Similar to the first percept, mean durations of subsequent I decrease with DF and mean of subsequent S increase with DF. The shape parameters α from MLE for DF=3,5,7 are also presented, and no trend for α values is found. B: Distribution of normalized subsequent percept durations (top) and properties (bottom) from EVA model. (Top) Normalized subsequent percepts are gamma distributed for DF=5 with mean durations μ and shape parameters α shown in the figure; similar results are obtained for other DF values, see S1 Fig. (Bottom) Mean percept durations and fitted shape parameters α for DF=3,5,7 from EVA model. Mean subsequent durations follow the same trend and the shape parameters have similar values as compared to those from the experiment. The error bars are 95% CI obtained from 100 Monte Carlo runs of the model. The result shows the robustness and consistency of the model. Related results are shown in S3 Fig.

Fig 5. Parameter fitting for input and sampler layers in the EVA model. The signal detection algorithm for constructing a neurometric function (the probability of a sampler to support the segregation percept) utilizes spike count time courses as shown in panel A (data extracted from [1] and interpolated for the cases DF=3, 5, 7); see below for more detail. The behavioral buildup functions (dashed, in panel B) occupy intermediate ranges of probability of S, and show slow initial rise for DF=3, 5. The simulated functions (solid, in panel B) do not capture the slow-rising phase of behavior buildup, and the spread between the neurometric curves increases unacceptably at larger N_{in}. For an optimal choice of parameters N_{in}, C_{th}, the algorithm yields well-fit asymptotic values of behavioral data. A: Mean spike counts $m_{t;DF}$ are interpolated at DF=3, 5, 7 st from data in [1], and then extended for triplets $t \leq 60$; see Methods. (Note: In [1, Fig.3] mean spike count data were shown for A-tone selective neurons in A1 during triplet tones at DF=1, 3, 6, 9. They decreased exponentially and stabilized within a few seconds. Mean spike counts changed with DF only during B-tone.) Herein, spike counts during B-tone are generated using Poisson processes of means $m_{t;DF}$ (DF=3, 5, 7) and then averaged over N_{in} neuronal units of the input layer IL (e.g. N_{in}= 1, 5, 30, 100). The average values of the mean spike counts and the standard error to the mean (SEM) are computed over 675 trials. Averages, including asymptotic values (written in parenthesis, at each DF), do not change with N_{in} but SEM decreases with a factor of $1/\sqrt{N_{in}}$. B: The signal detection algorithm [1] generates neurometric functions using numerical data from IL-pools of N_{in} neuronal units; parameter C_{th} is chosen to yield the least-squares error of the experimental buildups and the computer-simulated neurometric functions for DF=3,5,7. If N_{in} is small the neurometric curves tend to bunch together due to overlapping and large SEM regions across conditions. As N_{in} gets bigger, the neurometric curves are pushed apart. The best approximation to the set of psychophysical buildups is obtained for $N_{in} = 5$, $C_{th} = 4.21$.

Fig 6. Dependence of mean percept durations and shape of distributions on target T_a and stochastic term σ_a, in the EVA model. Diagrams show the difference between the mean duration μ derived from numerical simulations of the model and mean μ_{exp} from the behavioral data, represented as ratio μ/μ_{exp}; see the color scheme. Results are shown for A: first and B: subsequent integration and segregation percepts at conditions DF= 3,5,7. Given a fixed value for T_a, the dynamics changes from no alternations between percepts at small σ_a (na; in gray); to alternations of mean durations much longer than the experimental mean (region in warm colors); to mean durations that approximate well the corresponding experimental values (within one standard error to μ_{exp}; in green; black dots depict a discrete selection of values in the green region); then, to mean durations much shorter than μ_{exp} (region in cool colors). In each diagram, σ_a, T_a that were used to generate model-based results are identified by a red diamond (see Methods, $\sigma_a = 0.085$, T_a varies). Besides mean durations, the shapes of the gamma-like distributions that fit normalized percept durations depend on T_a and σ_a as well (α is the shape-parameter in the gamma-fit; see Eq (1) in Methods). There are three main regions that characterize α and they are delineated by the dashed-white and solid-white curves. Low-level of noise σ_a yields normal distributions (region N, to the left of dashed-white line; $\alpha \gg 3$) while high-level of noise yields exponential distributions (region E, to the right of solid-white line; α near 1). For intermediate level of noise, the distributions are gamma-like with shape close to that found experimentally (region G, between white contours; $\alpha \approx 2$ for first percepts and $\alpha \approx 2.6$ for subsequent percepts; α differs from α_{exp} by relative error up to 20% except for integration at DF=7 where it is up to 30%). The parameter range where both model-generated mean duration and shape of distribution are good approximations of their corresponding experimental observations is found at the intersection of region G with the sheet of black dots. Related results are shown in S6 Fig.

Fig 7. Signal detection algorithm adapted from [1] yields exponential distributions and unrealistic mean durations of percepts. (Top) Binary threshold C_{th} is chosen to yield the least-squares error between neurometric buildups (solid) and behavioral buildups (dashed) at DF=3,5,7. Poisson spike counts are averaged across a sample of N_{in}=5 neuronal units and compared to C_{th} to classify a triplet either as I (0) or S (1). Trial-averaging these binary responses produces the neurometric functions. The threshold value is determined by least-squares fit for A: the first 15 seconds of the stimulus to match the transients, or for B: the last 15 seconds of the stimulus to match the asymptotic level of the behavioral buildup; See also Fig 5. (Bottom) Trial-by-trial applications of the signal detection algorithm from [1] with A: C_{th}=4.01 and B: C_{th}=4.21 yield exponentially distributed subsequent percept durations for I (blue) and S (red). Their mean values μ are significantly smaller than those reported in the experiment. Note: Same parameter values N_{in}=5, C_{th}=4.21 were used in the EVA model for activation of the sampler layer SL (Fig 2A-B) and obtain gamma-like distributions of percepts (Figs 1D and 4B).

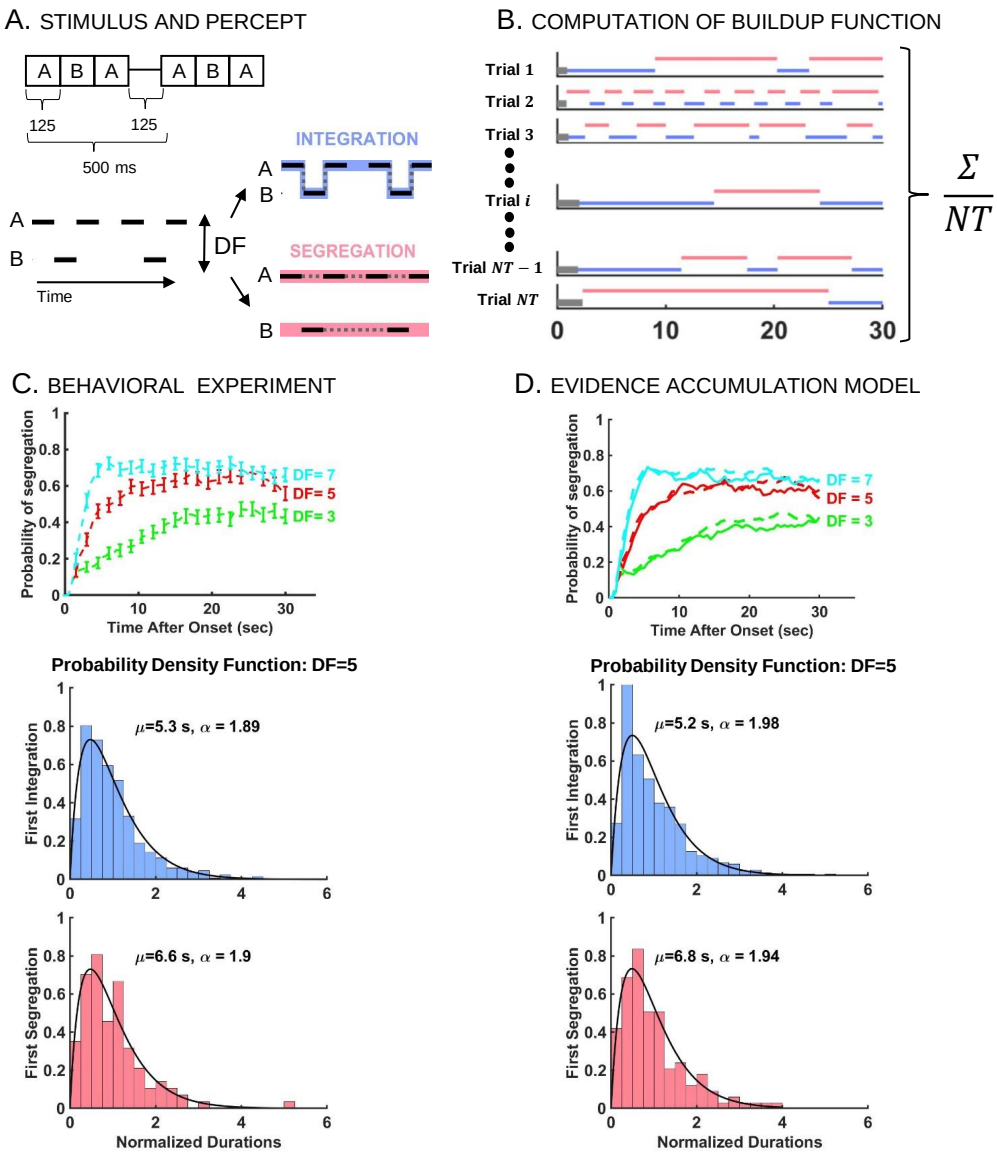
A. STIMULUS AND PERCEPT
A B A A B A
125
125
500 ms
A
B
Time
DF
INTEGRATION
SEGREGATION
B. COMPUTATION OF BUILDUP FUNCTION
Trial 1
Trial 2
Trial 3
Trial i
Trial $NT-1$
Trial NT
0 10 20 30
$\frac{\Sigma}{NT}$
C. BEHAVIORAL EXPERIMENT
Probability of segregation
Time After Onset (sec)
DF= 7
DF= 5
DF= 3
Probability Density Function: DF=5
First Integration
μ=5.3 s, α = 1.89
First Segregation
μ=6.6 s, α = 1.9
Normalized Durations
D. EVIDENCE ACCUMULATION MODEL
Probability of segregation
Time After Onset (sec)
DF = 7
DF = 5
DF = 3
Probability Density Function: DF=5
First Integration
μ=5.2 s, α = 1.98
First Segregation
μ=6.8 s, α = 1.94
Normalized Durations

Fig 1

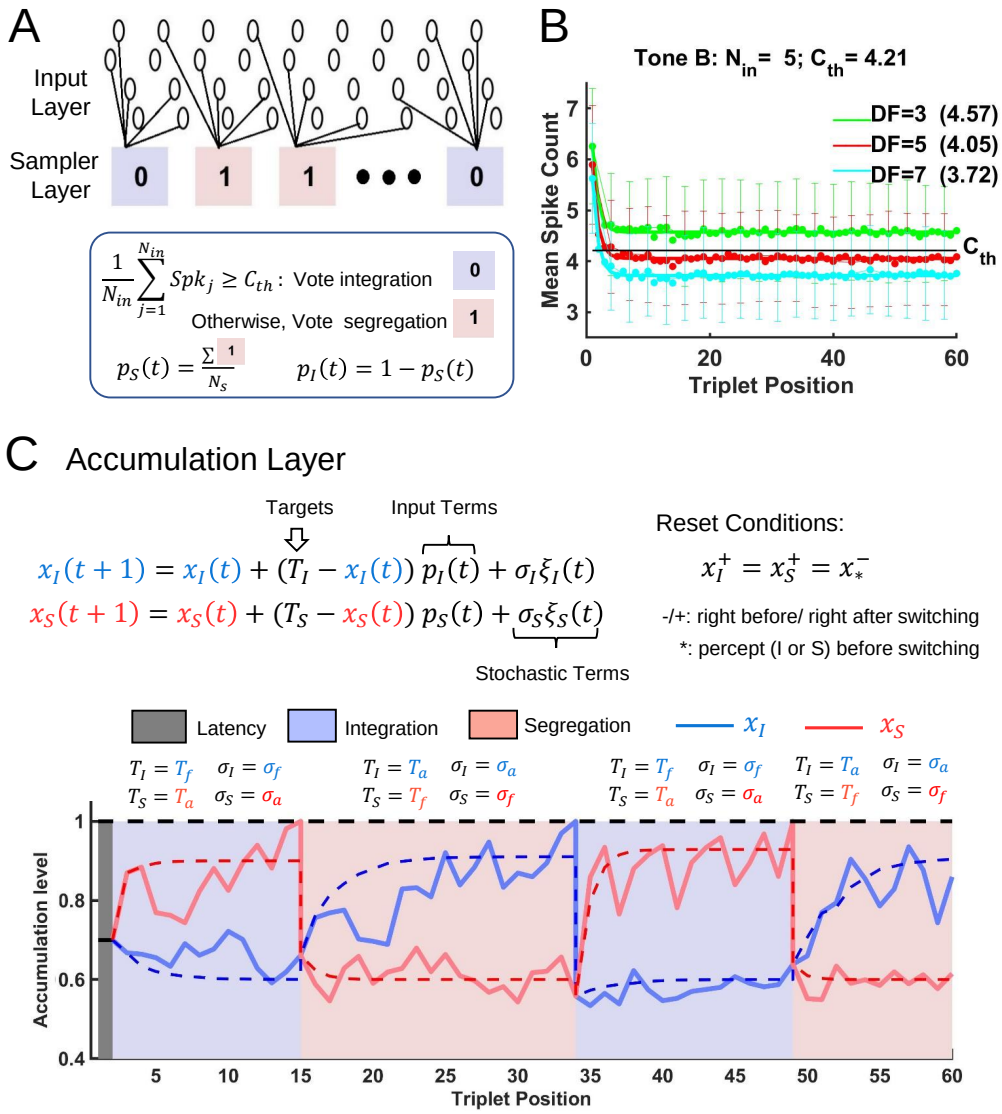
A
Input Layer
Sampler Layer
0
1
1
0
$\frac{1}{N_{in}}\sum_{j=1}^{N_{in}} Spk_j \geq C_{th}$: Vote integration 0
Otherwise, Vote segregation 1
$p_S(t) = \frac{\sum 1}{N_s}$
$p_I(t) = 1 - p_S(t)$
B
Tone B: N_{in}= 5; C_{th}= 4.21
DF=3 (4.57)
DF=5 (4.05)
DF=7 (3.72)
C_{th}
Mean Spike Count
Triplet Position
C Accumulation Layer
Targets
Input Terms
$x_I(t+1) = x_I(t) + (T_I - x_I(t))\, p_I(t) + \sigma_I \xi_I(t)$
$x_S(t+1) = x_S(t) + (T_S - x_S(t))\, p_S(t) + \sigma_S \xi_S(t)$
Stochastic Terms
Reset Conditions:
$x_I^+ = x_S^+ = x_*^-$
-/+: right before/ right after switching
*: percept (I or S) before switching
Latency
Integration
Segregation
x_I
x_S
$T_I = T_f$ $\sigma_I = \sigma_f$ $T_S = T_a$ $\sigma_S = \sigma_a$
$T_I = T_a$ $\sigma_I = \sigma_a$ $T_S = T_f$ $\sigma_S = \sigma_f$
$T_I = T_f$ $\sigma_I = \sigma_f$ $T_S = T_a$ $\sigma_S = \sigma_a$
$T_I = T_a$ $\sigma_I = \sigma_a$ $T_S = T_f$ $\sigma_S = \sigma_f$
Accumulation level
Triplet Position

Fig 2

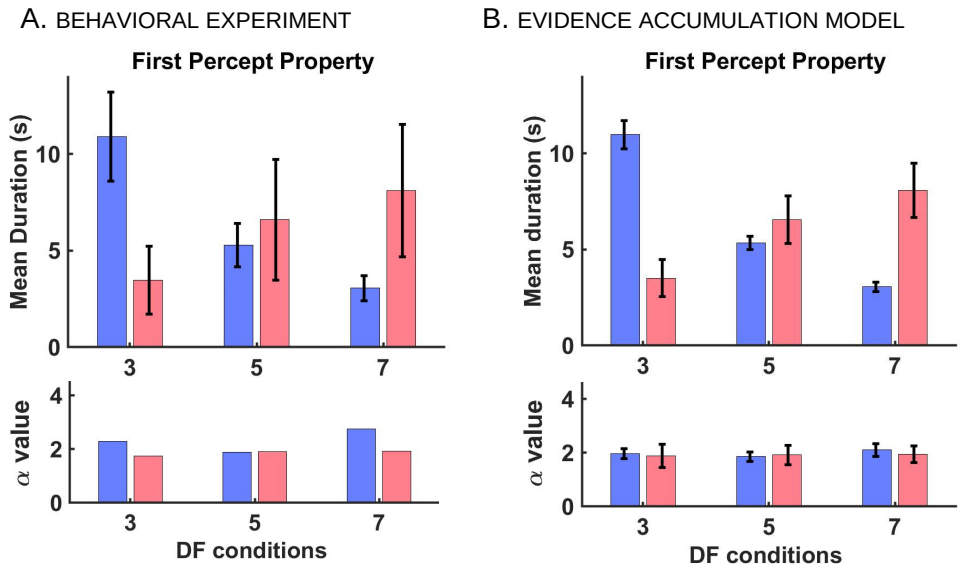
A. BEHAVIORAL EXPERIMENT
B. EVIDENCE ACCUMULATION MODEL
First Percept Property
First Percept Property
Mean Duration (s)
Mean duration (s)
α value
α value
DF conditions
DF conditions

Fig 3

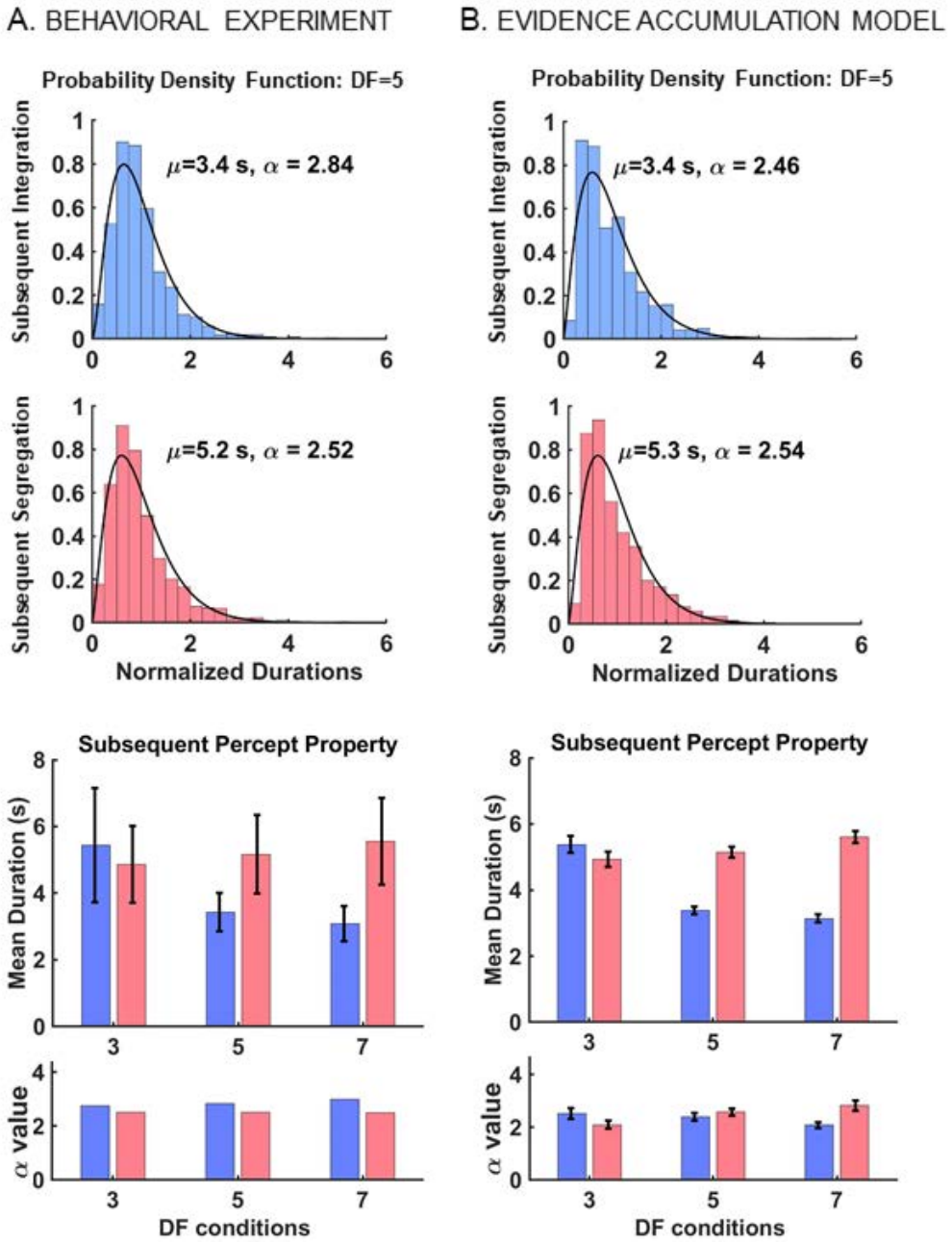
A. BEHAVIORAL EXPERIMENT
B. EVIDENCE ACCUMULATION MODEL
Probability Density Function: DF=5
Subsequent Integration
μ=3.4 s, α = 2.84
μ=3.4 s, α = 2.46
Subsequent Segregation
μ=5.2 s, α = 2.52
μ=5.3 s, α = 2.54
Normalized Durations
Subsequent Percept Property
Mean Duration (s)
α value
DF conditions

Fig 4

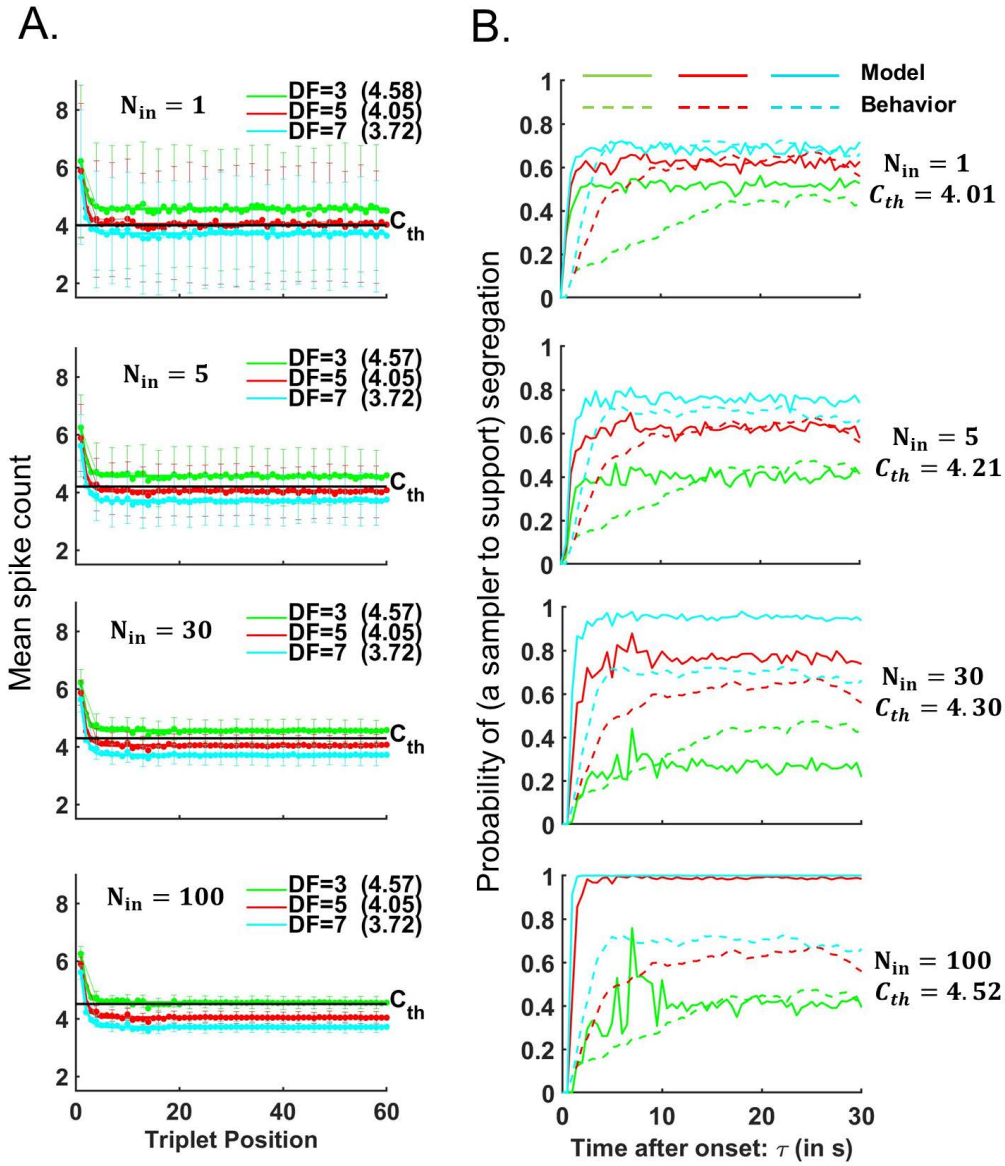
A.
B.
Mean spike count
Triplet Position
$N_{in} = 1$
$N_{in} = 5$
$N_{in} = 30$
$N_{in} = 100$
DF=3 (4.58)
DF=5 (4.05)
DF=7 (3.72)
DF=3 (4.57)
DF=5 (4.05)
DF=7 (3.72)
C_{th}
Model
Behavior
Probability of (a sampler to support) segregation
Time after onset: τ (in s)
$C_{th} = 4.01$
$C_{th} = 4.21$
$C_{th} = 4.30$
$C_{th} = 4.52$

Fig 5

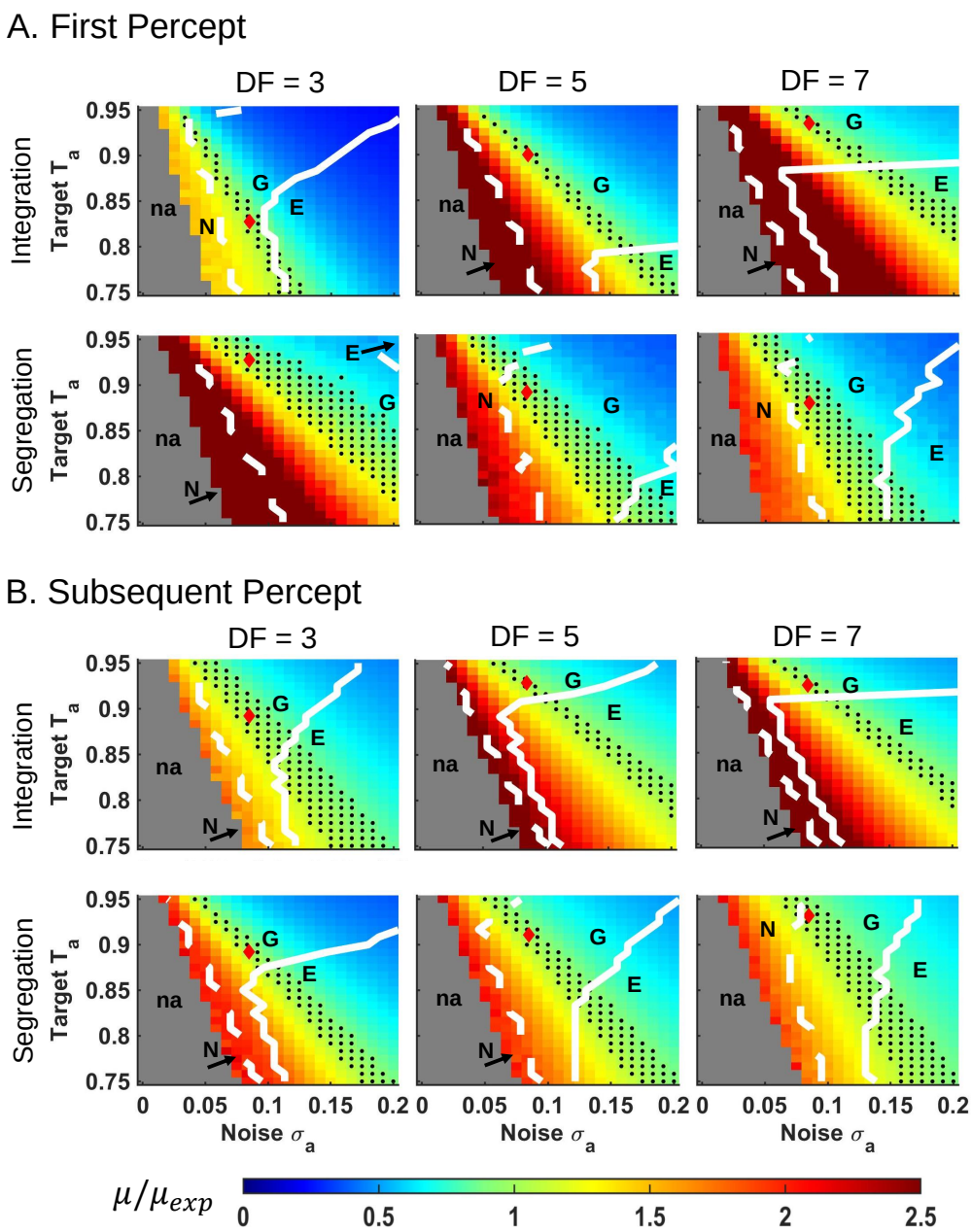
A. First Percept
DF = 3
DF = 5
DF = 7
Integration
Segregation
Target T_a
0.95
0.9
0.85
0.8
0.75
na
N
G
E
B. Subsequent Percept
0
0.05
0.1
0.15
0.2
Noise σ_a
μ/μ_{exp}
0
0.5
1
1.5
2
2.5

Fig 6

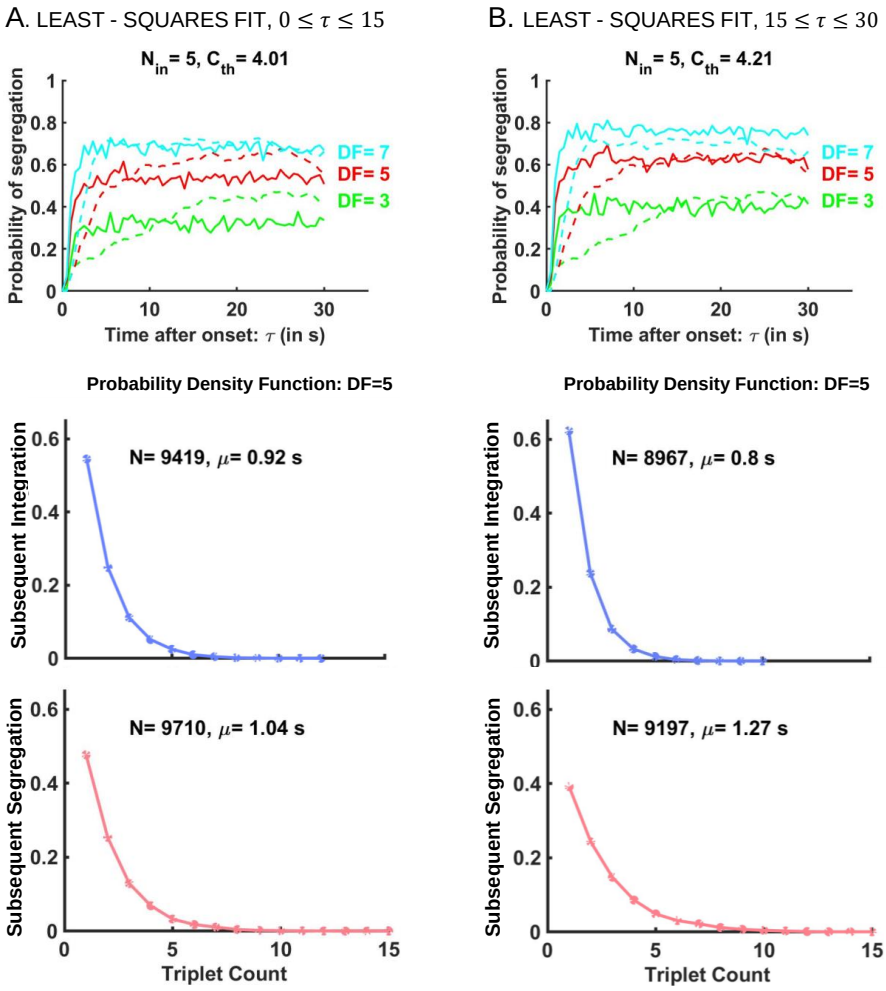
A. LEAST - SQUARES FIT, $0 \leq \tau \leq 15$
B. LEAST - SQUARES FIT, $15 \leq \tau \leq 30$
N_{in}= 5, C_{th}= 4.01
N_{in}= 5, C_{th}= 4.21
Probability of segregation
Time after onset: τ (in s)
DF= 7
DF= 5
DF= 3
Probability Density Function: DF=5
Subsequent Integration
Subsequent Segregation
N= 9419, μ= 0.92 s
N= 8967, μ= 0.8 s
N= 9710, μ= 1.04 s
N= 9197, μ= 1.27 s
Triplet Count

Fig 7

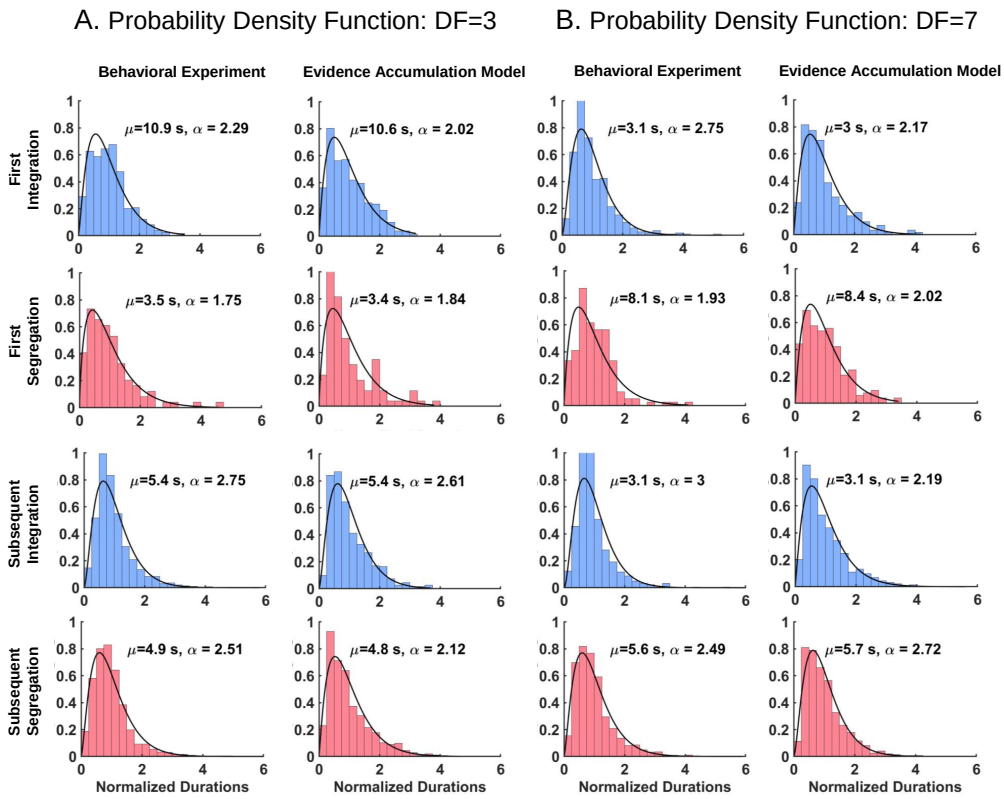
A. Probability Density Function: DF=3
B. Probability Density Function: DF=7
Behavioral Experiment
Evidence Accumulation Model
Behavioral Experiment
Evidence Accumulation Model
First Integration
First Segregation
Subsequent Integration
Subsequent Segregation
μ=10.9 s, α = 2.29
μ=10.6 s, α = 2.02
μ=3.1 s, α = 2.75
μ=3 s, α = 2.17
μ=3.5 s, α = 1.75
μ=3.4 s, α = 1.84
μ=8.1 s, α = 1.93
μ=8.4 s, α = 2.02
μ=5.4 s, α = 2.75
μ=5.4 s, α = 2.61
μ=3.1 s, α = 3
μ=3.1 s, α = 2.19
μ=4.9 s, α = 2.51
μ=4.8 s, α = 2.12
μ=5.6 s, α = 2.49
μ=5.7 s, α = 2.72
Normalized Durations
Normalized Durations
Normalized Durations
Normalized Durations

Fig S1

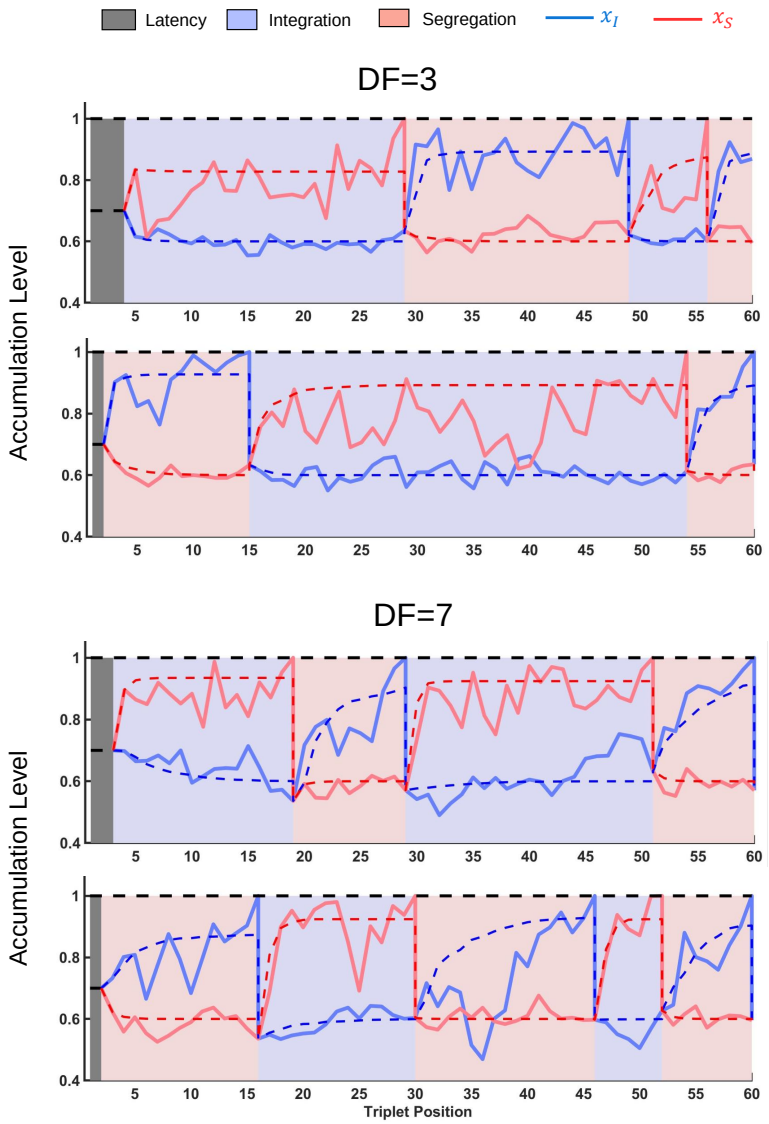
Latency
Integration
Segregation
x_I
x_S
DF=3
DF=7
Accumulation Level
Triplet Position

Fig S2

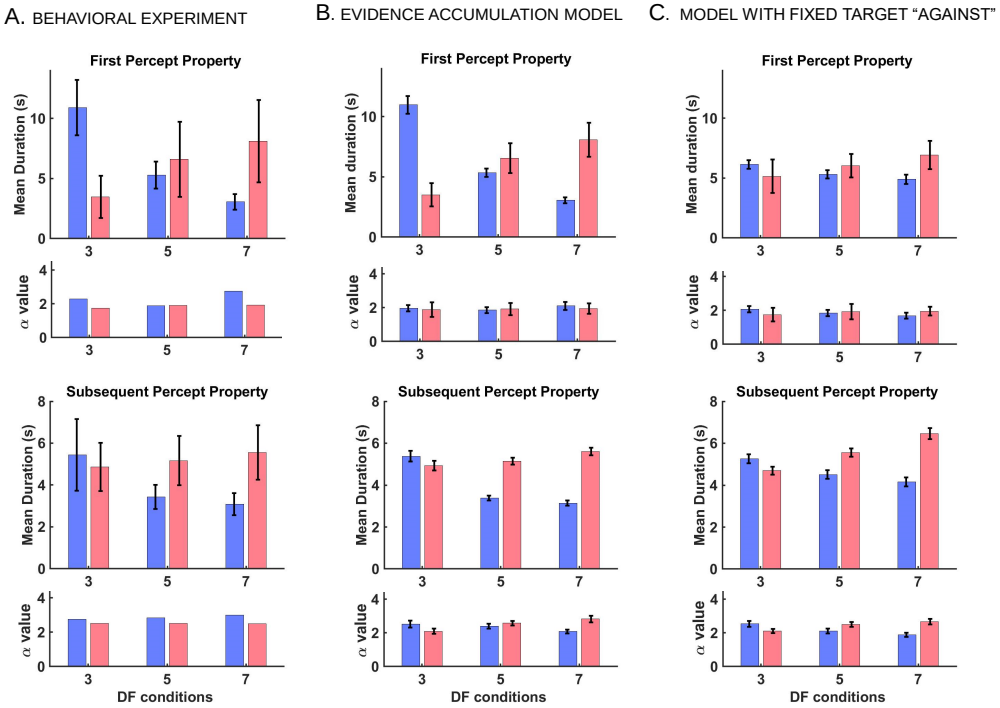
A. BEHAVIORAL EXPERIMENT
B. EVIDENCE ACCUMULATION MODEL
C. MODEL WITH FIXED TARGET "AGAINST"
First Percept Property
Subsequent Percept Property
Mean Duration (s)
Mean duration (s)
α value
DF conditions

Fig S3

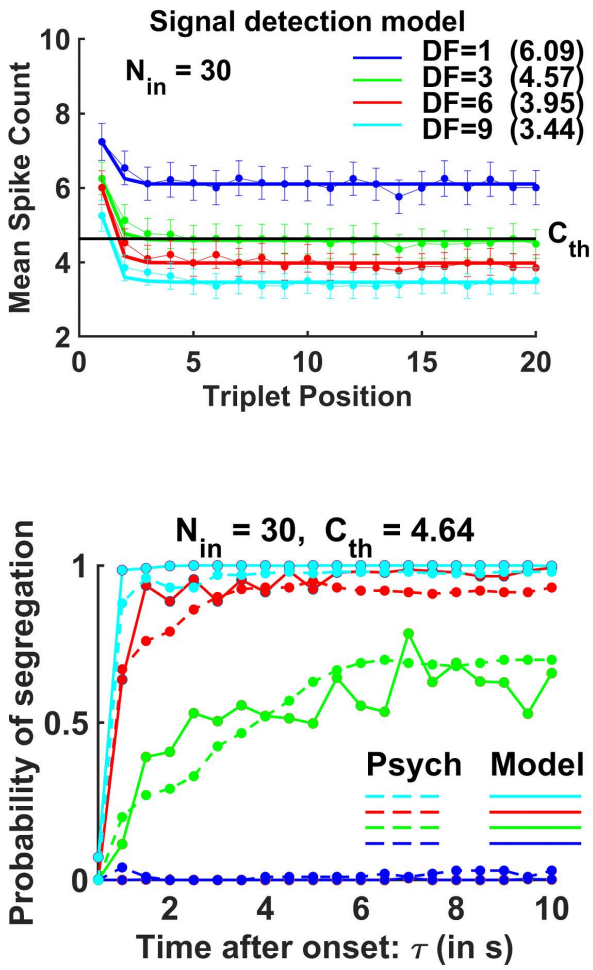
Signal detection model
N_in = 30
DF=1 (6.09)
DF=3 (4.57)
DF=6 (3.95)
DF=9 (3.44)
C_th
Mean Spike Count
Triplet Position
N_in = 30, C_th = 4.64
Probability of segregation
Psych
Model
Time after onset: τ (in s)

Fig S5

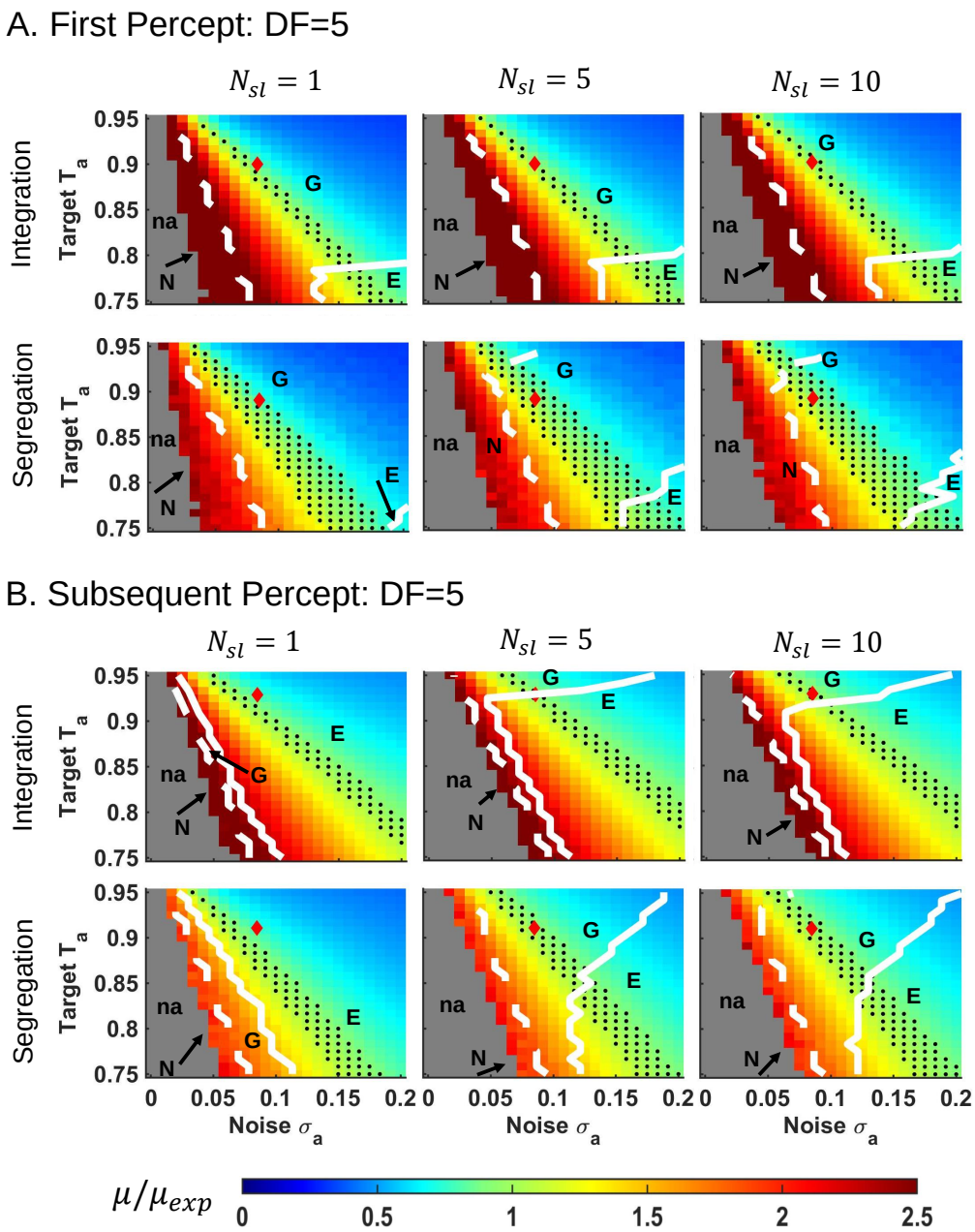
A. First Percept: DF=5
$N_{sl} = 1$
$N_{sl} = 5$
$N_{sl} = 10$
Integration
Segregation
Target T_a
0.95
0.9
0.85
0.8
0.75
na
N
G
E
B. Subsequent Percept: DF=5
0
0.05
0.1
0.15
0.2
Noise σ_a
μ/μ_{exp}
0
0.5
1
1.5
2
2.5

Fig S6