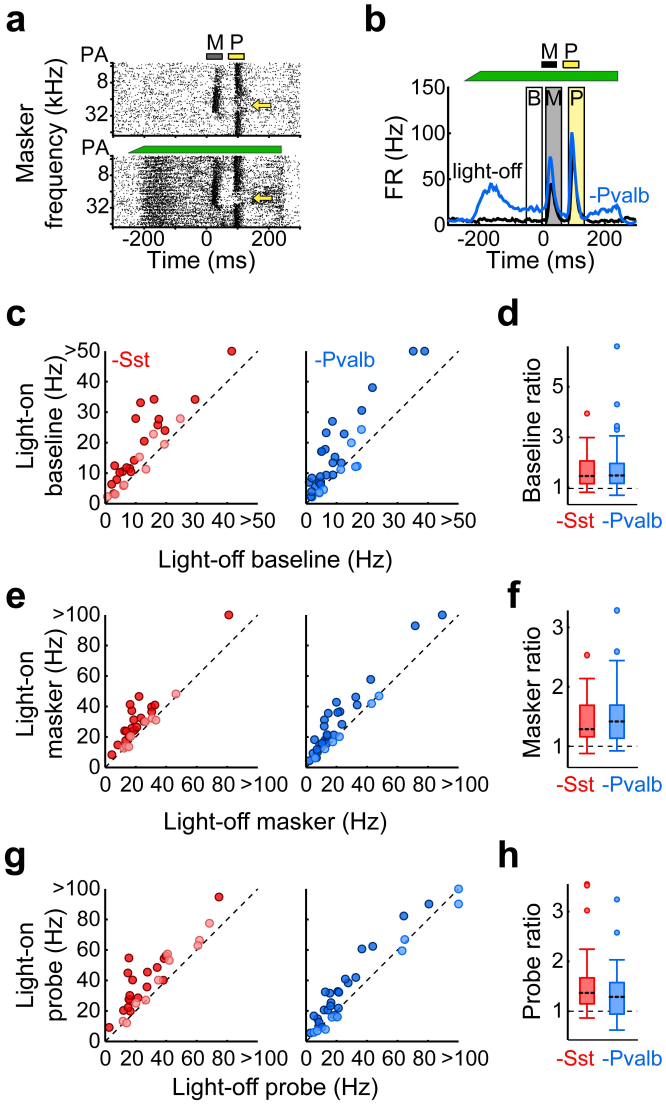
Fig.S1 – Related to Fig.2: Inactivation of Sst+ or Pvalb+ cells increases spontaneous and evoked firing rates.



a) Example SU response rasters without (top) and with (bottom) inactivation of Pvalb+ interneurons.

b) PSTHs of responses, across all stimuli, from light-off (black) or light-on trials (blue). Firing rates in the 50 ms before stimulus onset were used to calculate baseline firing rates. Firing rates in the 50 ms masker and probe response windows (see Methods) were used to calculate masker- and probe-evoked firing rates, respectively.

c) Average baseline firing rates for light-on versus light-off trials with inactivation of Sst+ cells (left) or inactivation of Pvalb+ cells (right). Darker circles: units with significant increases in baseline rates.

d) Ratios of light-on versus light-off spontaneous firing rates with inactivation of Sst+ (red) and Pvalb+ (blue) interneurons are not different (rank-sum, p = 0.88).

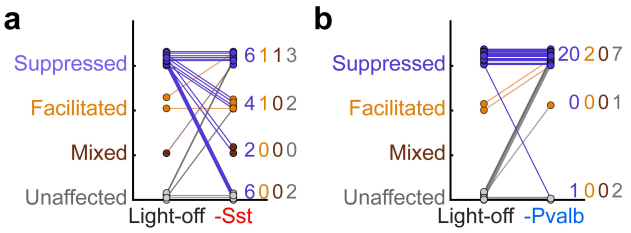
e) As (c) for average responses to the masker tone. Darker circles indicate units for which increases in rates were significant.

f) Ratios of light-on versus light-off masker-evoked firing rates with inactivation of Sst+ (red) and Pvalb+ (blue) interneurons are not different (rank-sum, p = 0.63).

g) As (c) for average responses to the probe tone. Units are colored by whether changes in rates were significant.

h) Ratios of light-on versus light-off probe-evoked firing rates with inactivation of Sst+ (red) and Pvalb+ (blue) interneurons are not different (rank-sum, p = 0.23).

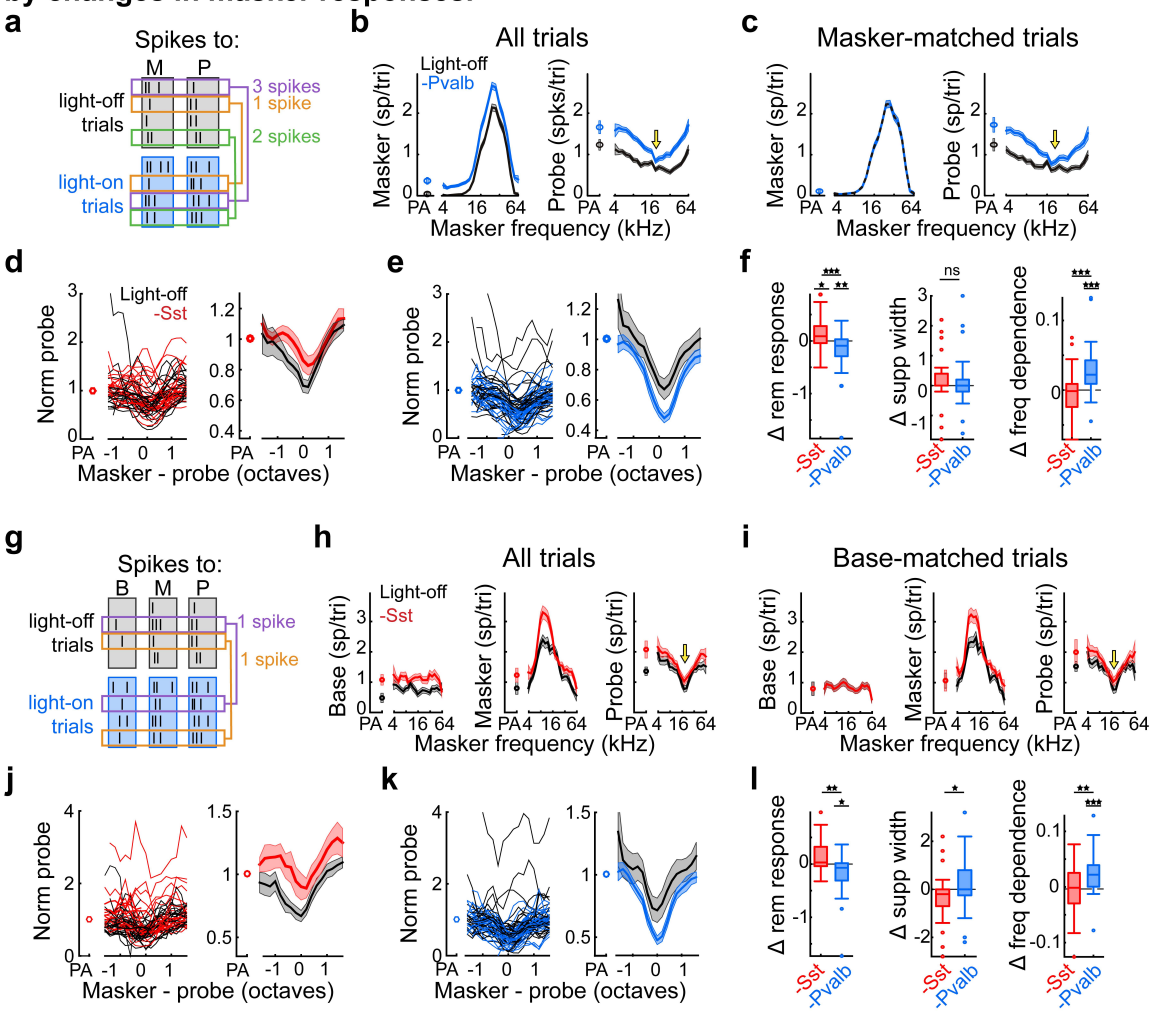
Fig.S2 – Related to Fig.2: Change in quality of forward interactions on a unit-by-unit basis.



a) Change in quality of forward interactions on a unit-by-unit basis with inactivation of Sst+ interneurons. Purple numbers: the number of suppressed units that, with inactivation of Sst+ interneurons, remained suppressed (first row), became facilitated (second row), became mixed (third row), or became unaffected (last row); gold numbers: the number of facilitated units that became suppressed (first row), remained facilitated (second row), became mixed (third row), or became unaffected (last row); brown and gray numbers: the number of mixed and unaffected units, respectively, that became each type of forward interaction as stated above.

b) As (a) for units with inactivation of Pvalb+ interneurons.

Fig.S3 – related to Fig.3: Effects of Sst+ and Pvalb+ inactivation on suppression are not explained by changes in masker responses.



a) Subsets of trials that contain identical numbers of masker-evoked spikes are selected from the light-off and light-on conditions.

b) Example unit responses (including all trials) to the masker (left) and probe (right), with (blue) and without (black) inactivation of Pvalb+ interneurons.

c) As (b), for the subset of trials that are matched based on masker responses.

d) All units' (left) and unit-averaged (right) normalized probe responses from masker-matched trials, as a function of masker – probe distance, without (black) and with (red) inactivation of Sst+ cells.

e) As (d) with inactivation of Pvalb+ interneurons.

f) Changes in remaining response, suppression width, and frequency dependence with inactivation of Sst+ (red) or Pvalb+ (blue) interneurons (*p<0.05, **p<0.01, ***p<0.001).

g) Subsets of trials that contain identical numbers of spikes during the baseline period (50 ms before sound onset) are selected from the light-off and light-on conditions.

h) Example unit responses (including all trials) to the masker (left) and probe (right), with (red) and without (black) inactivation of Sst+ cells.

i) As (h) for the subset of trials matched by baseline firing rate.

j-l) As (d-f) for the subset of trials matched by baseline firing rate.

Supplemental Procedures

Mouse strains. To target either Sst+ or Pvalb+ cells for optogenetic manipulation, we crossed Sst-IRES-Cre or Pvalb-IRES-Cre knock-in lines (JAX stock no. 013044 and 008069, respectively) to the Ai35 line (JAX stock no. 012735), which encodes the light-gated proton pump Archaerhodopsin-3 (Arch) fused to GFP under the CAG promoter after a loxP-flanked STOP cassette.

Details of data acquisition and stimuli. Mice were head-fixed above an air-floated spherical treadmill (Niell and Stryker, 2010) and the silicone plug was removed. The sound pressure from the treadmill was maintained at or below 45 dB and had spectral power mainly at frequencies below 4 kHz. A 16 site linear probe (50 µm spacing, Neuronexus) was inserted perpendicular to the cortical surface. Neural activity was amplified, digitized, and recorded continuously at 24414 Hz with Tucker-Davis hardware.

Sound stimuli were generated in MATLAB and presented through a free-field speaker (ES1, Tucker-Davis) directed toward the mouse's left ear. All sound envelopes were applied with 2 ms linear ramps. Best frequencies (BFs) were determined using 50 ms tones (4 kHz to 64 kHz, 0.2 octave spacing; 0-60 dB, 5 dB increments). The BF of the recording site at 10-15 dB above threshold was used as the frequency of the probe tone for subsequent FWS experiments. The FWS stimulus consisted of two sequential tones, a 50 ms masker followed by a 50 ms probe, separated by a 20 ms gap (stimulus-onset asynchrony of 70ms). The probe remained constant at the BF, 10-15 dB above threshold, while the masker tone randomly varied in frequency (4 kHz to 64 kHz, 0.2 octave spacing) at 15-20 dB above threshold. On a random subset of trials, called probe alone (PA) trials, the probe was presented without a preceding masker. Only units for which the probe frequency was within 0.5 octaves of the BF were analyzed.

On randomly interleaved trials, green light was shined directly above the surface of auditory cortex through a 400 micron fiber. Light turned on 250 ms before sound onset, and the power linearly ramped upwards for 50 ms before reaching maximum (10-15 mW). After sound offset, the light remained on for 120 ms.

Multilayered model. To explore whether synaptic dynamics could explain Sst+ and Pvalb+ cells' distinct effects on FWS, we built a three-layered linear threshold model, where each neuron's response ranges from 0 (no response) to 1 (maximum possible firing), as in (Phillips et al., 2017). Tone frequency (f) is presented in terms of octaves from the probe frequency.

The first layer contains n "thalamic" neurons, who respond to tones of varying frequency in a Gaussian fashion. All thalamic neuron's responses have standard deviations (σ_T) of ¼ octave. Thalamic neurons have evenly spaced center frequencies (five per octave, spanning the range 2 octaves below the probe frequency to 2 octaves above the probe frequency), such that the thalamic responses can be represented as:

$$Response_T(n,f) = e^{-\left(\frac{f(n)-f^2}{2*{\sigma_T}^2}\right)}$$

Thalamic neurons synapse onto three different types of cells in the cortical (second) layer: “Pyramidal” cells, “Sst+” cells, and “Pvalb+” cells. The synaptic output of a thalamic cell onto a cortical cell is proportional to the thalamic response. For example, the synaptic output of a thalamic cell onto a pyramidal cell is:

$$SynOutput_{T\rightarrow Pyr}(n,f) = Gain_{T\rightarrow Pyr} \ * \ Response_T(n,f)$$

Multiple thalamic neurons synapse onto each of the *k* cortical cells of each type in the second layer. These inputs are center-weighted and scaled by a Gaussian connectivity function, such that the connection weights between thalamic neurons and pyramidal cells in the second layer, for example, can be represented as:

$$Weight_{T\rightarrow Pyr}(n,k) = Scale_{T\rightarrow Pyr} * e^{-\left(\frac{(f(n)-f(k))^2}{2*{\sigma_{T\rightarrow Pyr}}^2}\right)}$$

All synapses from thalamic cells onto cortical cells in the second layer are modeled as dynamic. Specifically, synapses onto pyramidal cells, as well as Pvalb+ cells, are modeled as depressing, while the synapses onto Sst+ cells are modeled as facilitating. After a response, these synapses instantaneously depress by an amount proportional to the synaptic output, and they exponentially recover over time t according to a time constant τ. Depression, of inputs from thalamic onto pyramidal cells for example, can then be represented as:

$$Depression_{T\rightarrow Pyr}(n,f,t) = SynOutput_{T\rightarrow Pyr}(n,f) \ * \ e^{-\left(\frac{t}{\tau}\right)}$$

while facilitation, of inputs from thalamic cells onto Sst+ cells, can be represented as:

$$Facilitation_{T\rightarrow Sst}(n,f,t) = SynOutput_{T\rightarrow Sst}(n,f) \ * - e^{-\left(\frac{t}{\tau}\right)}$$

The availability of depressing synapses to respond to a stimulus is either 1 (if there was no prior stimulus) or it is the complement of depression (if there was a prior stimulus). The availability of depressing thalamic connections onto pyramidal cells, for example, can be represented as:

$$Availability_{T\rightarrow Pyr}(n,f,t) = \begin{cases} 1 \ if \ no \ prior \ stimulus \\ 1 - Depression_{T\rightarrow Pyr}(n,f,t) \ if \ prior \ stimulus \end{cases}$$

while the availability of facilitating thalamic connections onto Sst+ cells can be represented as:

$$Availability_{T\rightarrow Sst}(n,f,t) = \begin{cases} 1\ if\ no\ prior\ stimulus \\ 1 - Facilitation_{T\rightarrow Sst}(n,f,t)\ if\ prior\ stimulus \end{cases}$$

Thus, the response of second-layer cortical neurons k, as a function of tone frequency, is the product of the synaptic outputs from the thalamus, their current availability, and their connection weights onto the cortical neuron. For example, the response of second-layer pyramidal cells is represented as:

$$Response_{Pyr}(k,f) = \sum_{n} Weight_{T\rightarrow Pyr}(n,k) \ * \ SynOutput_{T\rightarrow Pyr}(n,f) * \ Availability_{T\rightarrow Pyr}(n,f,t)$$

In the model, responses of interneurons are more broadly tuned than pyramidal cells because they receive a broader distribution of connections from thalamic neurons (i.e., larger σ; parameters below).

Thalamic neurons in the first layer and cortical cells in the second layer all synapse onto a "cortical output" cell CO in the third layer. These synapses are weighted and scaled as described above and exhibit depressing dynamics in the same fashion as thalamic connections onto pyramidal and Pvalb+ cells in the second layer.

The synaptic gains of all connections are:

$$Gain_{T\rightarrow Pyr} = 0.2;\ Gain_{T\rightarrow Pvalb} = 0.2;\ Gain_{T\rightarrow Sst} = 0.2;$$

$$Gain_{T\rightarrow CO} = 0.08;\ Gain_{Pyr\rightarrow CO} = 0.08;\ Gain_{Pvalb\rightarrow CO} = 0.08;\ Gain_{Sst\rightarrow CO} = 0.08$$

The bandwidths of the Gaussian connectivity weight functions are:

$$\sigma_{T\rightarrow Pyr} = 1;\ \sigma_{T\rightarrow Pvalb} = 1.5;\ \sigma_{T\rightarrow Sst} = 1.5;$$

$$\sigma_{T\rightarrow CO} = 1;\ \sigma_{Pyr\rightarrow CO} = 1;\ \sigma_{Pvalb\rightarrow CO} = 1;\ \sigma_{Sst\rightarrow CO} = 1$$

The scaling factors for the Gaussian connectivity weights are:

$$Scale_{T\rightarrow Pyr} = 1;\ Scale_{T\rightarrow Pvalb} = 1;\ Scale_{T\rightarrow Sst} = 1;$$

$$Scale_{T\rightarrow CO} = 1;\ Scale_{Pyr\rightarrow CO} = 1;\ Scale_{Pvalb\rightarrow CO} = -0.4; Scale_{Sst\rightarrow CO} = -0.4$$

Time constants for all synaptic connections are:

$$\tau_{T\rightarrow Pyr} = 100\ ms;\ \tau_{T\rightarrow Pvalb} = 100\ ms;\ \tau_{T\rightarrow Sst} = 100\ ms;$$

$$\tau_{T\to CO} = 100\ ms;\ \tau_{Pyr\to CO} = 100\ ms;\ \tau_{Pvalb\to CO} = 100\ ms;\ \tau_{Sst\to CO} = 100\ ms$$

Time between masker and probe tones (i.e., the gap duration):

$$t = 20\ ms$$

To simulate removing either interneuron from the network, we scaled the connection weights of the removed interneuron onto the cortical output neuron to 0, and (to keep the total amount of inhibition constant) doubled the weight of the other interneuron's connections. For instance, when removing Sst+ interneurons, the new scaling factors were:

$$Scale_{Pvalb\to CO} = -0.8; Scale_{Sst\to CO} = 0$$

And when removing Pvalb+ interneurons, the new scaling factors were:

$$Scale_{Pvalb\to CO} = 0; Scale_{Sst\to CO} = -0.8$$