Title

Segregating two simultaneous sounds in elevation using temporal envelope: Human psychophysics and a physiological model.

Journal
The Journal of the Acoustical Society of America, 138(1)

Authors
OConnor, Kevin
Sutter, Mitchell
Johnson, Jeffrey

Publication Date
2015-07-01

DOI
10.1121/1.4922224

Peer reviewed

Segregating two simultaneous sounds in elevation using temporal envelope: Human psychophysics and a physiological model

Jeffrey S. Johnson, Kevin N. O'Connor, and Mitchell L. Sutter[a)]
Center for Neuroscience, University of California at Davis, 1544 Newton Court, Davis, California 95618, USA

(Received 18 March 2014; revised 29 April 2015; accepted 21 May 2015; published online 2 July 2015)

The ability to segregate simultaneous sound sources based on their spatial locations is an important aspect of auditory scene analysis. While the role of sound azimuth in segregation is well studied, the contribution of sound elevation remains unknown. Although previous studies in humans suggest that elevation cues alone are not sufficient to segregate simultaneous broadband sources, the current study demonstrates they can suffice. Listeners segregating a temporally modulated noise target from a simultaneous unmodulated noise distracter differing in elevation fall into two statistically distinct groups: one that identifies target direction accurately across a wide range of modulation frequencies (MF) and one that cannot identify target direction accurately and, on average, reports the opposite direction of the target for low MF. A non-spiking model of inferior colliculus neurons that process single-source elevation cues suggests that the performance of both listener groups at the population level can be accounted for by the balance of excitatory and inhibitory inputs in the model. These results establish the potential for broadband elevation cues to contribute to the computations underlying sound source segregation and suggest a potential mechanism underlying this contribution. *© 2015 Acoustical Society of America.*

[http://dx.doi.org/10.1121/1.4922224]

[VMR] Pages: 33–43

I. INTRODUCTION

The segregation of multiple simultaneous sound sources, exemplified for speech in the cocktail party problem (Cherry, 1953), is an everyday task facing the auditory system. In general, segregation refers to the act of forming separate percepts of simultaneous stimuli, and stands in contrast to integration where simultaneous stimuli are combined into a single percept. Several types of cues can be used to segregate sources in an auditory scene, including spatial location, spectral content, and temporal envelope properties (Bregman, 1990; O'Connor and Sutter, 2000; Woods *et al.*, 2001; Shinn-Cunningham *et al.*, 2007; Hill and Miller, 2010). Spatial location information arising from interaural time and intensity cues is an important contributor to sound segregation for sources that differ in azimuth (i.e., the horizontal plane), but in mammals these cues are not available when sources are separated only in elevation (i.e., the vertical plane). Spatial location in elevation is instead encoded by the presence of frequency-specific decreases in energy ("notches"), which are produced by the spatially dependent spectral filtering of the sounds by the pinnae (Middlebrooks and Green, 1991; Hofman *et al.*, 1998; Tollin and Yin, 2003).

One complication when using spatial location to segregate multiple sources in elevation is that of spectral notch interference: when competing sounds sum at the cochlea, the absence of spectral power at the notch frequencies corresponding to a sound source at one elevation may be masked by spectral power arising from a sound source at another elevation, making detection of the notches, and therefore localization of the sounds, difficult. Two recent studies highlight the difficulty of segregating simultaneous broadband sounds in elevation. Best *et al.* (2004) reported that humans are unable to accurately report whether one or two identical broadband noises are present if both sources are on the vertical midline, suggesting that when simultaneous broadband noise stimuli are distinguished only by elevation cues, they are integrated to a single percept. Integration of broadband sounds was also found in a study where listeners freely reported the vertical location of a brief (50 ms) amplitude-modulated (AM) target in the presence of an isospectral noise masker with both sounds located on the vertical midline (Bremen *et al.*, 2010). When the target and noise were of similar intensity and within 75 degrees of each other, the listeners' reports, intended to localize the target, were better predicted by a weighted average of the two speaker locations than by either of the speaker locations themselves, suggesting that for narrow separations, the two sounds are integrated. At higher separations (60 degrees and above, pooled together) a bimodal response corresponding to the two locations of the speakers emerges, but in this case, the louder speaker is favored regardless of the location of the target. This suggests that for wide separations, the listeners were able to partially segregate the sounds—sufficient to recover the approximate elevations of the speakers–but were unable to fully segregate the sounds in order to assign the

[a)]Electronic mail: mlsutter@ucdavis.edu appropriate respective elevations to the target and the masker, an ability which would have resulted in a unimodal distribution at the veridical location of the target.

Contrary to the Bremen *et al.* (2010) results, preliminary data in our laboratory suggested that human listeners were capable of segregating two simultaneous broadband sounds in elevation on the basis of temporally modulated cues. We designed the present experiment to re-assess this ability across a range of modulation frequencies (MF). Amplitude modulation is a natural manipulation to use here (as it was in Bremen *et al.*, 2010) because it preserves the spectral composition of the original stimuli, allowing us to create target and masker stimuli that have identical spectral carriers. In addition to employing a range of MF, our design differed from the Bremen *et al.* (2010) paradigm in some respects—our stimuli were of longer duration (400 ms), were presented from known, discrete locations, and we removed the analog component of localization and replaced it with a forced choice. When listeners were asked to identify the up/down direction of an amplitude-modulated broadband sound in the presence of a simultaneous unmodulated masker from the opposite direction, they fell into two statistically distinct groups. One group ("veridical") responded to the actual location of the target across a wide range of MF. The other group ("non-veridical") performed near chance for most MF, but tended to identify the modulated sound as coming from the opposite direction at low MF.

To investigate the basis of this bimodal behavioral population, we turned to the known physiology of elevation determination. In mammals, the pathway including the dorsal cochlear nucleus (DCN) (Young *et al.*, 1992) and the inferior colliculus (IC) (Davis *et al.*, 2003) has been posited to play a role in sound elevation processing. Neurons with responses that are tuned for elevation have been reported in the IC in both macaque (Zwiers *et al.*, 2004) and cat (Aitkin and Martin, 1990). Most IC elevation research has been done in the cat, where these cells are termed "type-O" neurons and are selective to the elevation-based spectral notches introduced by the cat's head-related transfer function (HRTF) (May *et al.*, 2008). A detailed circuit diagram for cat type-O neurons, which accounts for notch selectivity, has been proposed (Davis *et al.*, 2003). We developed a model based on this proposed type-O circuit and found that the responses of modeled type-O neurons are capable of accounting for both veridical and non-veridical response groups, at both a population and an individual level, using only adjustments to the gain ratio of excitatory and inhibitory inputs.

II. MATERIALS AND METHODS

A. Experimental protocol

Nine human listeners (eight male, ages 21–55, mean 31.8) participated in the experiment. All listeners gave informed consent. The study was approved by the University of California at Davis Human Subjects institutional review board. Initials have been changed. Two listeners (L.A. and S.T.) are authors on this study; the remaining listeners were naive. Listeners' hearing thresholds were not pre-tested, but all listeners self-reported having no hearing deficiencies.

Two exemplars ("Noise$_A$" and "Noise$_B$") of broadband Gaussian noise (400 ms, 100 kHz sampling rate) were created. These exemplars were presented either unmodulated, or sine-phase 100%-depth AM at MF of 5, 10, 15, 30, 60, 120, 250, 500, 1000, or 2000 Hz. All stimuli were onset- and offset-ramped with a 5-ms sin^2 function.

Stimuli were presented using a CED Power1401 microprocessor controlled by Spike2 (Cambridge Electronic Design, Cambridge, England). Stimuli passed through a passive attenuator (Leader LAT-45, Leader Electronics Corp., Yokohama, Japan) and an active attenuator (Tucker-Davis Technologies PA5, Alachua, FL) before reaching amplifier-equipped studio monitor speakers (Yamaha MSP5, Yamaha Corporation of America, Buena Park, CA). Stimuli were calibrated using a sound level meter (Bruel & Kjaer 2133 with microphone 4155, Brüel & Kjær, Nærum, Denmark) and adjusted to 58 ± 0.3 dB sound pressure level (SPL) using the active attenuator. The same equipment was used to collect speaker transfer functions (STFs).

Listeners sat in a sound-attenuated booth one meter from two visible speakers located on the vertical midline and displaced ±20 degrees from the interaural horizontal plane. Listeners used a joystick for both trial initiation and response. Listeners were familiarized with the AM stimuli at each MF before the experiment. There were two classes of trials: AM-alone (AM stimulus from either speaker, no sound from opposite speaker) and AM + masker (AM stimulus from either speaker, equal-intensity unmodulated noise stimulus from opposite speaker). The two trial classes were randomly interleaved, and listeners were not informed which class of trial would occur. For both classes, listeners were required to indicate, via two-alternative forced-choice, the direction (up/down) from which the AM stimulus originated. For AM + masker trials, different noise carrier exemplars (Noise$_A$/Noise$_B$) were used for the modulated and unmodulated stimuli, counterbalanced over the course of the experiment. Each stimulus was presented 10 times from each speaker in each condition for a total of 800 presentations (2 noise exemplars × 2 speakers × 2 stimulus classes × 10 MFs × 10 trials). Both stimulus classes, both target directions, and all target MF were randomly interleaved. Trial accuracy feedback *was not* provided during either the experiment or during the brief (10–20 stimuli) pre-experimental familiarization period.

HRTFs for upper and lower speaker locations were collected [Etymotic ER-7C (Etymotic Research Inc., Elk Grove Village, IL) probe microphone, in-ear tubes, left and right ear collected simultaneously] for all listeners using 80 dB uniform noise bursts.

B. Modeling

Modeling was done with MATLAB (MathWorks, Natick, MA). All stimulus waveforms were filtered by a STF collected using 80 dB uniform noise bursts, digitally adjusted to the same peak SPL, then filtered by individual HRTFs to produce an estimate of each waveform at each tympanum for three conditions: unmodulated-alone (not used in the psychophysical experiment), AM-alone, and AM + masker (linear sum of AM-alone and unmodulated-alone). Each tympanum estimate was A-weighted and converted to a stimulus power representation using MATLAB's "Spectrogram" function (time window 3 ms, overlap 1.5 ms). A similar non-directional "comparison" was created for each speaker using STF-transformed but non-HRTF-transformed noise.

A filterbank of 20 model IC type-O cells was created with logarithmically spaced best frequencies (BFs) between 7 kHz and 14 kHz; this frequency range was chosen because it spans the frequencies that have been shown to be most useful in determining sound elevation (Roffler and Butler, 1968; Hebrank and Wright, 1974; Asano *et al.*, 1990). Model cells were based on existing IC (Davis *et al.*, 2003) and DCN models (Nelken and Young, 1994; Reiss and Young, 2005), and produced an output ("firing rate," arbitrary units) for each time point in the spectrogram representation of the stimulus. A simplified graphical representation of the model is found in Fig. 2. The model consisted of six cell (or input) types altogether. For the description below, all centering is logarithmic and $\mathrm{BF_O}$ means the BF of the corresponding type-O cell. The bandwidths of model inputs were initially chosen based on a qualitative interpretation of the above IC and DCN models and were not adjusted. Wide band inhibition (WBI) was based on energy in a span of 1.0 octaves centered at the $\mathrm{BF_O}$. Type-II cells were excited by energy in a narrow band of 0.1 octaves centered at 80% of the $\mathrm{BF_O}$, and inhibited by WBI. Type-IV cells were excited by energy in a narrow band of 0.1 octaves centered at $\mathrm{BF_O}$, and inhibited by WBI and type-II cells. Wide band excitation (WBE) was based on energy between 0 kHz and 17 kHz. Narrow band inhibition (NBI) was based on energy in a narrow band of 0.25 octaves centered at 80% of the $\mathrm{BF_O}$. Type-O cells were excited by type-IV cells and WBE, and inhibited by NBI. Connection strengths for the model were initially chosen based on a qualitative interpretation (e.g., "strong connection," "weak connection") of the above IC and DCN models, and were adjusted so that model responses to notched-noise sweeps resulted in reasonable approximations to known physiology. The gain of the NBI connection to the type-O cells was the only parameter varied while modeling the output of type-O cells.

The overall output of the model was as follows:

$$R_{\mathrm{O}} = [(R_{\mathrm{IV}}/2) + (R_{\mathrm{WBE}}/2) - g_{\mathrm{NBI}}R_{\mathrm{NBI}}]^{+},$$

where R_{X} is instantaneous "firing rate" (R_{O}: IC type-O cells; R_{IV}: DCN type-IV cells; R_{WBE}: wide band excitatory input; R_{NBI}: narrow band inhibition), and g_{NBI} is the narrow band inhibitory gain factor, varied between 0.1 and 7.0 in increments of 0.1. The $[\,]^{+}$ operator indicates half-wave rectification. The specific calculations of the "firing rates" for the various inputs were as follows:

$$R_{\mathrm{IV}} = [R_{\mathrm{IV\_EXC}} - (R_{\mathrm{WBI}}/3) - R_{\mathrm{II}}]^{+},$$

$$R_{\mathrm{IV\_EXC}} = 150 \times 2\hat{}\,((^{\mathrm{dB}}S_{\mathrm{BFo}\{0.1\,\mathrm{oct.}\}} - {}^{\mathrm{dB}}C_{\mathrm{BFo}\{0.33\,\mathrm{oct.}\}})/6),$$

$$R_{\mathrm{WBI}} = 25 \times\ 2\hat{}\,((^{\mathrm{dB}}S_{\mathrm{BFo}\{1.0\,\mathrm{oct.}\}} - {}^{\mathrm{dB}}C_{\mathrm{BFo}\{1.0\,\mathrm{oct.}\}})/6),$$

$$R_{\mathrm{II}} = R_{\mathrm{II\_EXC}} - R_{\mathrm{WBI}},$$

$$R_{\mathrm{II\_EXC}} = 25 \times\ 2\hat{}\,((^{\mathrm{dB}}S_{0.8\mathrm{BFo}\{0.1\,\mathrm{oct.}\}} - {}^{\mathrm{dB}}C_{0.8\mathrm{BFo}\{0.33\,\mathrm{oct.}\}})/6),$$

$$R_{\mathrm{WBE}} = 100 \times 2\hat{}\,((^{\mathrm{dB}}S_{\{0\mathrm{kHz}-17\mathrm{kHz}\}} - {}^{\mathrm{dB}}C_{\{0\mathrm{kHz}-17\mathrm{kHz}\}})/6),$$

$$R_{\mathrm{NBI}} = 100 \times 2\hat{}\,((^{\mathrm{dB}}S_{0.8\mathrm{BFo}\{0.25\mathrm{oct.}\}} - {}^{\mathrm{dB}}C_{0.8\mathrm{BFo}\{0.25\mathrm{oct.}\}})/6).$$

(R_{II}: DCN type-II cells; $R_{\mathrm{IV\_EXC}}$: excitatory input to IV; $R_{\mathrm{II\_EXC}}$: excitatory input to II; R_{WBI}: wide band inhibition. ${}^{\mathrm{dB}}S$: dB SPL of stimulus in defined window; ${}^{\mathrm{dB}}C$: dB SPL of "comparison" noise in defined window; subscripts indicate center of octave range with values in curly braces indicating width of octave range, or frequency window if no center is specified. Thus, "${}^{\mathrm{dB}}S_{0.8\mathrm{BFo}\{0.25\ \mathrm{oct.}\}}$" indicates the dB intensity of the test stimulus in the 0.25-octave-wide window centered at a frequency that is 0.8 multiplied by the cell's $\mathrm{BF_O}$.)

Modeling was done separately for each ear. For model readout, the time-domain firing rate output of the model type-O cells was subjected to a frequency-based analysis. In this readout analysis, a 20-element reference distribution, (R_{up}, R_{down}), of the filterbank responses to an unmodulated noise from each speaker was generated (averaged response across both noise exemplars). For each possible stimulus (both AM-alone and AM + masker conditions), the 400-ms firing rate vs time response of each model type-O cell was Fourier transformed into the frequency domain. We selected the maximum value in this fast Fourier transform (FFT) amplitude spectrum as each model cell's response strength, creating a 20-element test distribution (T) for each stimulus. (The frequency corresponding to the maximum FFT amplitude was always the stimulus MF for MF < 500 Hz, the only frequencies at which human performance deviated from chance.) Thus, R_{up} and R_{down} were vectors of typical response strengths for each cell for up/down broadband stimuli, respectively, and T was a vector of the strength of modulation for each cell under test conditions. The mean squared error (MSE) of T with R_{up} and R_{down} was calculated. Model performance was considered correct if the MSE averaged across both noise exemplars and both ears (Hofman and van Opstal, 2003; van Wanrooij and van Opstal, 2007) for the correct speaker (origin of target stimulus) was lower than the corresponding value for the incorrect speaker (that is, if the test distribution was nearer to the reference response vector of the target speaker than the reference response vector of the distracter speaker). Because the model would otherwise produce binary results, uniformly distributed noise (between -0.02 and 0.02) was added to the MSE values (100 repetitions), and an average value was taken so that model results could be compared to behavioral results. The magnitude of this noise was selected "by hand" to reduce quantization in the output, but was not tuned on the basis of model results.

As a control, we implemented a "pinna-only" model. For the pinna-only model, we calculated the spectrogram of our AM-alone and AM + masker stimuli, as detailed in the first paragraph of this section to estimate the power of each stimulus at the tympanum over time. We computed the MSE of each AM + masker spectrogram to the corresponding AM-alone spectrogram for stimuli in the up and down

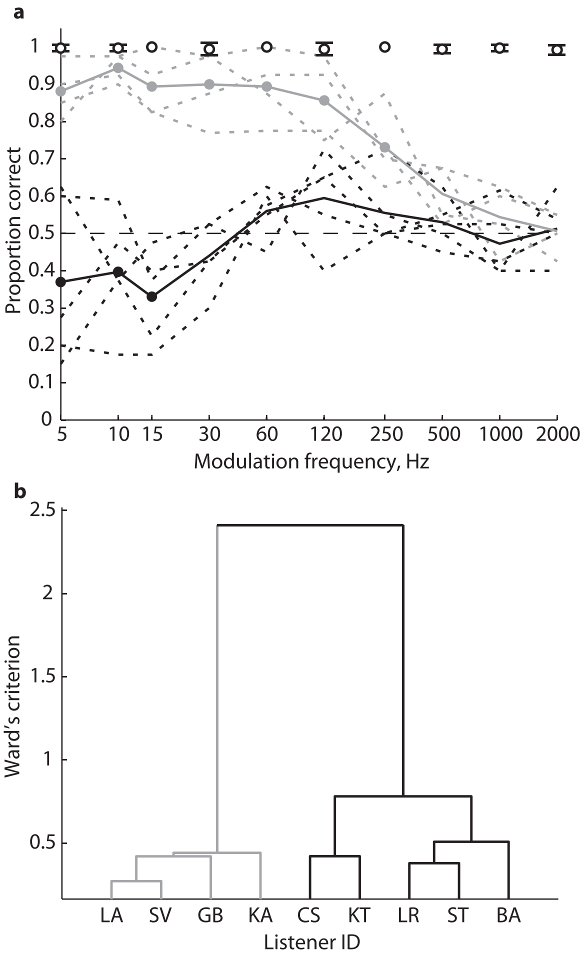

FIG. 1. (a) Accuracy on sound segregation task. Black open circles (line omitted) are the mean across all nine listeners in the AM-alone task. Error bars are standard deviation. Gray lines indicate accuracy on the AM + masker task for veridical responders. Black lines indicate accuracy on the AM + masker task for non-veridical responders. Dashed lines indicate individual performance, solid lines indicate mean performance, and filled circles indicate performance significantly different from chance (0.5, dashed black line) using a binomial test, $p < 0.05$, Bonferroni corrected for ten comparisons. Actual p-values (5–2000 Hz) for the veridical group are: 0, 0, 0, 0, 0, 0, 4.2×10^{-9}, 8.9×10^{-3}, 0.30. 0.94; actual p-values for the non-veridical group are: 2.9×10^{-4}, 4.5×10^{-3}, 1.7×10^{-6}, 0.10, 0.10, 8.7×10^{-3}, 0.14, 0.44, 0.48, 0.78. (b) Dendrogram of cluster analysis of accuracy functions from AM + masker task. Gray lines indicate veridical responders. Black lines indicate non-veridical responders.

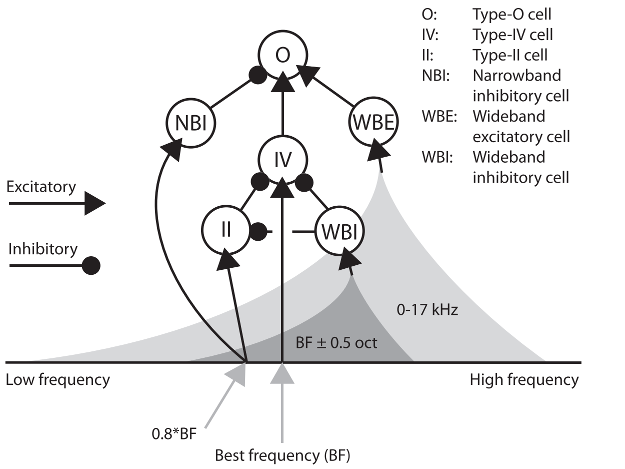

FIG. 2. Model circuit diagram. Connections between the six cell types modeled are shown as arrows (excitatory) and circles (inhibitory). The horizontal black bar corresponds to auditory nerve fiber inputs. Inputs to WBE and WBI cells come across large frequency ranges illustrated with shaded areas. Inputs to type-IV, type-II, and NBI cells are narrower and are illustrated with arrows. Inputs to type-II and NBI cells are centered at a frequency that is 0.8 multiplied by the best frequency (BF) of the corresponding type-IV cell. The weight of the NBI-to-O inhibitory connection ("inhibitory gain") was the only parameter systematically varied. Full connection details can be found in the methods in Sec. II.

positions. Model performance on each repeat was considered correct if the MSE averaged across both noise exemplars and both ears for the correct speaker was lower than the corresponding value for the incorrect speaker. As for the primary model, uniformly distributed noise (in this case between −1000 and 1000 due to different values being compared) was added to the MSE (100 repetitions) and an average value was calculated.

For testing the drift hypothesis, the g_{NBI} levels used in the AM + masker condition to compare the model predicted proportion correct to the psychophysical results were selected on the basis of limits established in the AM-alone condition. To select upper and lower bounds of g_{NBI}, a constant-width gain window (width 0.4 gain units) was used to calculate model predicted proportion correct in the AM-alone condition. For each individual listener, this gain window was centered on $g_{\mathrm{NBI}} = 3.0$ and moved either up (non-veridical listeners) or down (veridical listeners) in g_{NBI} to determine the g_{NBI} bounds where the mean predicted proportion correct across all MFs in the gain window fell below a criterion of 98%. The window chosen was the most extreme window which had a mean predicted proportion correct across all MFs no lower than the 98% criterion.

C. Statistical analysis

Analysis was done with MATLAB (MathWorks, Natick, MA). Separation of listeners into two groups was done on the basis of their behavioral performance curves using K-means clustering. Hierarchical cluster trees were generated on the Euclidean distance between behavioral performance curves using a single-linkage algorithm and Ward's criterion (increase in within-cluster sum-of-squares distance from the cluster centroid when merging clusters). In the clustering Monte Carlo analysis, 100 000 randomized behavioral performance curves were generated by permuting behavioral performance values (within MF) across listeners; the linkage distance between the topmost two groups in each resulting cluster tree was compared to the corresponding value in the observed data to determine the probability that two groups so widely separated might arise by chance.

III. RESULTS

A. Human behavioral performance in the elevation task

In the AM-alone task, all listeners were able to identify the target speaker at ceiling performance regardless of target MF [Fig. 1(a), open circles]. In the AM + masker task, listeners fell into two distinct performance groups. The high-performing group (veridical responders, $n = 4$) identified the target speaker in the presence of the masker at a rate better

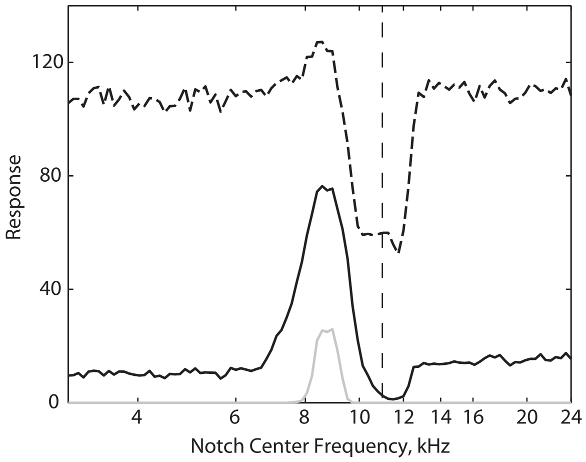

FIG. 3. Model neuron responses to notched-noise sweeps. The black dashed line indicates model DCN type-IV cell response as a function of notch center frequency. The solid black line indicates model IC type-O cell response with an inhibitory gain of 1. The solid gray line indicates model IC type-O cell response with an inhibitory gain of 5. The thin vertical dashed line indicates the BF of all three model cells (11 kHz).

than chance (binomial test, $P < 0.05$ corrected for multiple comparisons to 5.1×10^{-3}) for all MFs between 5 and 250 Hz and dropped to chance performance for MFs of 500 Hz and above [Fig. 1(a), solid gray line, mean; dashed gray lines, individual]. The low-performing group (non-veridical responders, $n = 5$) performed significantly worse than chance for 5–15 Hz AM [Fig, 1(a), solid black line, mean; dashed black lines, individual]. These non-veridical responders rose to chance performance for MFs of 30 Hz and above. *K*-means cluster analysis on the accuracy curves [Fig. 1(b)] revealed two distinct groups, and a Monte Carlo analysis put the likelihood that two groups at least this distinct would arise by chance at $P < 10^{-5}$, indicating that our placement of listeners into veridical and non-veridical groups is not a spurious assignment, but reflects a categorical difference between listeners. By an analogous Monte Carlo, neither the veridical responders ($P = 0.77$) nor the non-veridical responders ($P = 0.36$) could be further broken down into subgroups.

B. Modeling a bimodal listener population using IC

The existence of two distinct listener populations, one with veridical and one with non-veridical responses, is quite unexpected. In the hope of uncovering a mechanism which might account for the segregation of listeners into these two groups, we developed a simple, non-spiking quantitative model of DCN type-IV and IC type-O neurons based on previous qualitative models (Nelken and Young, 1994; Davis *et al.*, 2003; Reiss and Young, 2005). A circuit diagram for our model is shown in Fig. 2. (See the methods in Sec. II for complete model details.)

Type-IV and type-O neurons have characteristic responses to notched-noise stimuli. To verify that our model had normal notched-noise responses, both the output type-O and the intermediate type-IV model neurons were presented with a series of wide-band Gaussian noise stimuli, each with a 30 dB notch (3.2 kHz bandwidth, notch center varied between 3 kHz and 24 kHz) created using Fourier methods. The responses of one model neuron with a BF of 11 kHz are shown in Fig. 3. Response is in arbitrary units (see model details in the methods in Sec. II), scaled to approximate a firing rate in spikes/s. Model cell outputs to notched noise are comparable to observed IC type-O (solid lines) and DCN type-IV (dashed line) cell responses (Davis *et al.*, 2003). Notch-inhibited type-IV neurons respond strongly to noise unless there is a notch at their (BF, in which case their response is reduced. Type-O cells respond weakly to noise unless there is a notch below their BF, in which case their response increases. The dip in response below noise baseline at BF for the type-O cell with an inhibitory gain (g_{NBI}; see the methods in Sec. II) of 1 parallels a similar finding in the physiology [notably Figs. 1(D) and 4(B) from Davis *et al.*, 2003]. For high inhibitory gains, model type-O cells are fully silenced except for the case where there is a notch below their BF. Responses to broadband noise without a notch (not shown) are not distinguishable from the response to notched noise where the notch is distant from the BF (e.g., left- or right-hand tails in Fig. 3).

Because the appropriate ratio of excitation to inhibition was not clear, when comparing our model's performance with human performance, the weight of the model narrow band inhibitory cells' connection to the type-O cells (g_{NBI}, or "inhibitory gain") was varied systematically. For the AM-alone condition, the model result averaged across all listeners accurately determines the origin of the AM stimulus across a wide range of inhibitory gains [Fig. 4(a), red-colored area], consistent with behavioral performance (Fig. 1, open circles). For each listener, we calculated MSE between the behavioral performance in the AM-alone condition and the individual's model predicted performance in a 0.4-unit wide gain window varied across all levels of g_{NBI}. Figure 4(b) depicts these results. The averaged results of veridical responders (green) and non-veridical responders (magenta) indicate that there is a very wide range of g_{NBI} over which the model is capable of reproducing behavioral performance on the AM-alone task at an individual level.

The model result averaged across all listeners for AM + masker stimuli is shown in Fig. 4(c). Unlike the AM-alone condition, for AM + masker stimuli, there is a much narrower inhibitory gain region in which the model accurately determines the origin of the AM stimulus, primarily where the weight of inhibition is less, or only slightly greater, than the weight of the excitatory inputs. Figure 4(d) shows the effect of varying g_{NBI} on the MSE between each individual listener's behavioral performance on the AM + masker task and their individual model performance in a 0.4-unit wide gain window [as in Fig. 4(b)]. Again, our listeners fell into the same two distinct groups found in Fig. 1. Dashed lines indicate individual model fits and solid lines indicate the average of these individual fits for the veridical (green) and non-veridical (magenta) populations. For veridical responders, g_{NBI} levels below or slightly greater than one always resulted in a good match to the behavior, while g_{NBI} levels above ~1.5 resulted in a poor match to behavior for all but one veridical responder. Non-veridical responders showed a different pattern, with g_{NBI} levels below one always resulting in a poor match to behavior and better

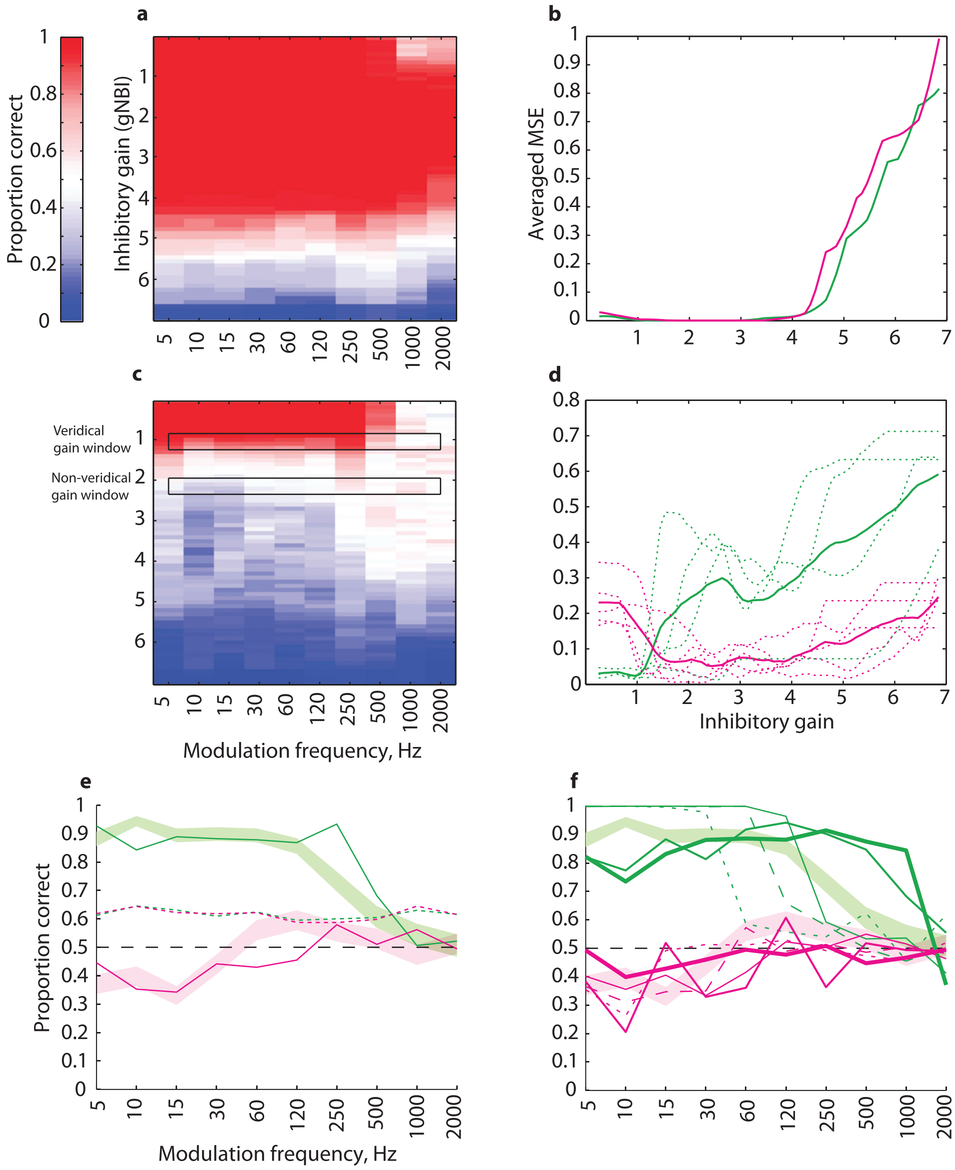

FIG. 4. Model predictions in sound segregation task. (a) Heat map showing predicted proportion correct (mean across all listeners) as a function of inhibitory gain and modulation frequency in model IC type-O cells for the AM-alone task. A scale bar is present to the left of (a). (b) Averaged mean-squared error between model predicted proportion correct and the mean performance of veridical (green) and non-veridical (magenta) responders as a function of inhibitory gain. (c) Same as (a), for the AM + masker task. The upper black box indicates the gain window that best fits the mean performance of veridical responders. The lower black box indicates the gain window that best fits the mean performance of non-veridical responders. (d) Same as (b) for the AM + masker task. Also included are the data for individual listeners (dashed lines). (e) Model predicted proportion correct compared to human performance. The green and magenta solid lines indicate the mean model predicted proportion correct from the upper and lower black boxes in (c), respectively. The green and magenta dotted lines indicate the mean predicted proportion correct from the pinna-only model, for listeners with veridical and non-veridical task performance, respectively. The green and magenta stripes indicate listener performance ± standard error of a proportion for listeners with veridical and non-veridical task performance, respectively, reproduced from Fig. 1(a). (f) Model predicted proportion correct for different model temporal windows. The green and magenta solid and dashed lines indicate the mean model predicted proportion correct as the solid lines in (e), but for data corresponding to different temporal windows [heat maps and gain windows not shown, but created as in (c)]. The line styles correspond to maximum representable AM frequencies as follows: thickest line, 1000 Hz; thick line, 667 Hz; thin line, 100 Hz; wide dashed line, 67 Hz; narrow dashed line, 33 Hz. The green and magenta stripes are as in (e). For all lines in (b), (d), (e), and (f), green corresponds to veridical responders and magenta corresponds to non-veridical responders.

matches at g_{NBI} levels between ∼1 and 4, with a mean optimum ∼2.5.

The best fitting g_{NBI} windows for the mean model results and mean behavioral results are overlaid on Fig. 4(c), and correspond closely with the results of the individual fits in Fig. 4(d). The mean response in each of these windows is shown in Fig. 4(e) (solid lines), along with a plot of the mean accuracy (± standard error, shaded areas) for the veridical (green) and non-veridical (magenta) behavioral groups.

Because these model results could potentially be derived from the pinna-transformed stimulus without requiring the IC step of the model, we also created a pinna-only model. This model attempted to determine, using only the spectrograms of the pinna-transformed sounds, whether the AM in an AM + masker stimulus originated in the up or down position based on similarity to the AM-alone stimuli (see the methods in Sec. II). The results of this pinna-only model are presented in Fig. 4(e) in the dashed green (mean of veridical responders) and magenta (mean of non-veridical responders) lines. Although the pinna-only model performs slightly better than chance, it does not replicate the differences between our veridical and non-veridical responders, suggesting that individual listener differences in the pinna-transformations of the sounds are not sufficient to account for our behavioral results.

The fact that the full model (for veridical responders) continues to perform well to slightly higher AM frequencies than our listeners may be due to our choice of a 3-ms window size in the time domain transformation of the model. With 50% overlap in samples, this resulted in a spectrogram with 667 samples per second, which by Nyquist limits would result in a maximum representable frequency of 333 Hz AM. To investigate the role of changing the window size in the time domain transformation of our model, we ran our model again with windows corresponding to maximum representable AM frequencies of 33, 67, 100, 667, and 1000 Hz AM [plotted in Fig. 4(f), 333 Hz omitted from plot for clarity, but see Fig. 4(e)]. For our veridical responders, we calculated the MSE between the resulting model predicted proportion correct for each temporal window size and the behavioral performance and fit these MSEs with a log-transformed Gaussian. The peak of this Gaussian was at 195 Hz, suggesting that we would have gotten a better match of the data with a slightly larger (~5 ms) temporal window.

C. Toward explaining a bimodal gain distribution

The results from Fig. 4(d) indicate that our bimodal behavioral results in the AM + masker condition could not be accounted for by a single inhibitory gain setting in our model. Although individual estimates of best gain ratio are a bit noisy, these results suggest instead that a bimodal gain distribution would be sufficient for our model to account for both veridical and non-veridical populations. However, the model does not provide a rationale for why gain ratios would be distributed bimodally.

As a first-pass attempt to explain a bimodal gain distribution, we proposed and tested an inhibitory drift hypothesis that was based on individualized model performance in the AM-alone condition. We found that a wide range of inhibitory gains results in veridical model performance for our AM-alone task [Fig. 4(a)], which as a single-sound task exemplifies the majority of real-world elevation tasks. The wide range of inhibitory gains that result in near-zero MSE for the AM-alone task [Fig. 4(b)] make it unlikely that the long-term *in vivo* gain ratio is solely determined by choosing the single optimized value from single-sound elevations. This is because a change in gain would not result in a significant change in the accuracy in the single-source condition. An additional mechanism could also play a role in determining a gain within the large range of possible values. In our drift hypothesis, we proposed that a bimodal gain distribution could arise from an initial random gain drift coupled with a small positive feedback (or feedback that discourages gain ratios toward the center of the range). Under this hypothesis, long-term gain movement toward either high or low gain ratios would eventually be stopped by instructive feedback from elevation localization experience similar to the AM-alone task. When real-world errors in single-sound up/down localization (as modeled here by the AM-alone condition) result, instructive feedback should stop further movement toward higher or lower gain ratios to preserve accurate single-sound localization. This elevation-instructive feedback would therefore establish a ceiling and floor for the gain ratio, which would maintain veridical performance in single-sound tasks, and combined with bidirectional gain drift would result in two populations, one at the ceiling and one at the floor gain ratio.

We implemented this hypothesis in our model by selecting a low-inhibition gain window (for veridical responders) and a high-inhibition gain window (for non-veridical responders) from the AM-alone task (see the methods in Sec. II for details on window selection). We then applied the same gain windows determined by the AM-alone task to AM + masker model responses (similar to the black boxes in Fig. 4(c), but for individual listeners) and calculated the MSE of the model predicted performance in the gain window selected by the drift hypothesis against actual behavioral performance.

Figure 5 shows the results of this analysis, with the MSE calculated from the drift hypothesis plotted against the minimum MSE produced by the model across all possible gain windows. For each individual listener, the model is capable of making an approximation of behavior that is good enough to result in an MSE <0.05. For veridical responders (open circles), the drift hypothesis is also capable of approximating listener behavior to the same standard—in two cases, it comes close to selecting the best possible gain setting from the model's standpoint. However, for non-veridical responders (star symbols) the situation is different. Although the drift hypothesis does a good job for two listeners, for the other three it selects gain windows that are much higher than the optimal gain windows identified by the model, resulting in poor approximations to behavior. Thus, although the drift

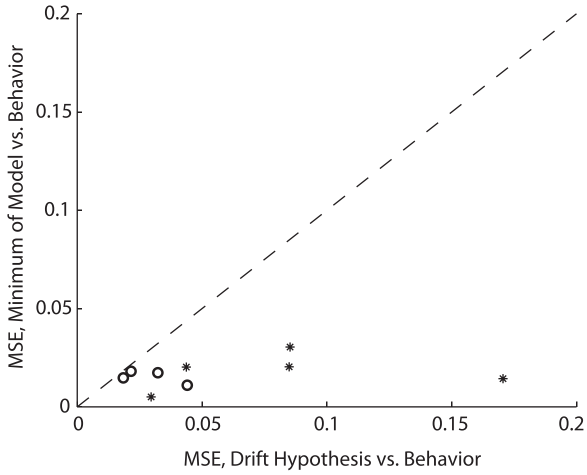

FIG. 5. Model evaluation of drift hypothesis, individual listeners. Comparison of model performance under the drift hypothesis to optimal model performance. MSEs are calculated as in Fig. 4(c), but with individual data rather than averaged data. The open circles indicate veridical responders. The star symbols indicate non-veridical responders. The dashed line is a unity line.

hypothesis, by design, creates a bimodal gain distribution, only one of the two distributions can be reasonably said to have a good correspondence with behavior.

IV. DISCUSSION

For this study, our listeners were asked to perform a simple (up/down only) source localization of an amplitude-modulated target. There is little doubt that in the AM-alone condition, the listeners are doing an elevation localization as requested, in this case without the demand of segregating the target from a masking stimulus. The AM + masker condition is more complicated, not only because the stimulus would seem to require segregation, but also because our listeners unexpectedly form two distinct groups with different response curves. There are two key issues to be considered here. The first is why our listeners should fall into two groups that treat the same stimulus so differently. The second issue is whether our listeners are, in fact, segregating the stimuli in the AM + masker condition. Along with the question of segregation, we will consider differences between the current study and previous studies that might account for different conclusions in what seem to be similar tasks.

A. Accounting for two listener populations

A major finding of this study is that our listeners fall into two distinct populations in the behavioral task. Accounting for the existence of these two populations requires consideration of a range of possibilities, many of these related to different potential listener strategies adopted by the two populations.

Several competing hypotheses about listener strategies can be proposed to account for why our listeners fall into two groups as they do. In order to infer which of these listener strategies was likely used, we should look to two constraints arising from the data. (1) The hypothesis must be consistent with above-chance performance in veridical responders and below-chance performance in non-veridical responders. (2) The hypothesis must be consistent with the frequency dependency and magnitude differences of the effects.

One possible interpretation of listener strategy in the AM + masker condition is that neither listener group used perceived elevation cues to perform the task. The overall data are not consistent with random guessing, but it is possible that a non-elevation cue (i.e., a pinna-imposed spectral feature in the stimulus that does not generate the perception of elevation) could be found to distinguish between the AM-up and -down sounds in the presence of a masker. It is important to note that our listeners were not provided with accuracy feedback and, thus, did not have any way to systematically correct errors, or even to know when an error was made. Therefore, if listeners made decisions based on a non-elevation cue, the mapping between that cue and the AM-up and -down responses should be arbitrary rather than consistent across listeners. About half of the listeners would assign a veridical mapping to the non-elevation cue, and about half would assign a non-veridical mapping. Although we do see an approximately equal split in the number of listeners in each population, we would additionally predict that the veridical and non-veridical populations would have mirror-symmetric curves with the same dependence on modulation frequency and the same magnitude of difference from chance. However, the mirror-symmetric prediction does not hold true, which argues against our listeners using an arbitrary assignment of a single non-elevation cue. Another hypothesis, that the veridical and non-veridical groups might use two different non-elevation cues, could potentially account for differences in modulation frequency dependence and peak accuracy, but cannot account for the absence of within-group mirror-symmetric performance—without feedback, we would again expect an arbitrary cue-response mapping resulting in two mirror-symmetric response curves for each non-elevation cue used. Hence, our data do not support the hypothesis that both groups were using non-elevation cues.

Alternatively, it is possible that veridical responders were able to perceive elevation cues in the AM + masker condition, but non-veridical responders were not. Our data argue against random guessing from our non-veridical responders due to the significant bias to report modulation from the incorrect speaker at low MF. While the performance of two of our subjects might be consistent with random guessing (Fig. 1), three clearly show the bias at all lower MF. As noted above, the absence of mirror-symmetric performance curves among our non-veridical responders argues against their using a non-elevation cue. Thus, the weight of the evidence suggests that both listener groups are using the elevation cues inherent in the stimulus to perform the task.

How our listeners are able to use elevation cues to give two widely different response modes is the next important question. One possible explanation for this could arise from the fact that, during the troughs of the AM noise, there is a brief window when the *masking* sound might be localized without interference. Such "dip-listening" models have been widely used in work on the cocktail party problem (Bacon *et al.*, 1998; Vélez and Bee, 2011) and speech perception (Festen and Plomp, 1990; Gustafsson and Arlinger, 1994). This ability would likely depend on AM frequency because the duration of this window will become shorter as the AM frequency increases. Our bimodal population could result from some listeners misidentifying the masker (localizable during these windows) as the target. If some listeners were dip-listening and others used a different strategy, differences in accuracy and temporal cutoff frequency between the veridical and non-veridical groups such as we observe could arise. This remains an open possibility, though there is little way of assessing whether listeners, in fact, used different strategies, or of making predictions about accuracy and temporal cutoff frequency for these two potential populations.

Our model of the notch-detection circuit, however, does predict these differences in accuracy and temporal cutoff frequency, and it does so without requiring the listeners to adopt different strategies. Instead, it can allow for a bimodal listener population if an appropriate bimodal inhibitory gain distribution is established. Under this model, one simple assumption—that the gain is not primarily set by elevation localization—can suffice to establish a bimodal gain distribution. Our drift hypothesis is one way to imagine how such a bimodal gain distribution might be established. In fact, the drift hypothesis does a good job of selecting low-gain windows (those appropriate for the veridical responders), consistent with the idea that there may be a floor to the inhibitory gain value that is instructively imposed by performance on single-sound vertical localization tasks. However, the drift hypothesis does a poor job of selecting high-gain windows (those appropriate for non-veridical responders) using an analogous method. Thus, if the drift hypothesis for the establishment of a bimodal distribution of gain ratios is conceptually correct, there must be a different source of feedback (perhaps a different auditory task) or some physiological constraint which determines the upper limit of inhibitory gain.

Ultimately, the model suggests that at low and middle MF, listeners with low gains on the inhibitory inputs to IC type-O cells would be provided with strong evidence for the veridical direction of AM, and that at low MF, listeners with high gains would be provided with weaker evidence for the non-veridical direction of AM. Notably, in this case, both populations use the same stimulus features and the same decision criteria (i.e., the same "mapping"). However, whether these result in evidence for or against the veridical location of the AM in the presence of the masker depends on specific aspects of our model—aspects which (1) can easily be changed by long-term activity-dependent (e.g., Hebbian) processes and (2) should not result in differences in everyday, non-masked localization.

B. Segregation of two sounds in elevation

When the auditory system is faced with two simultaneous broadband sounds that are both located on the vertical midline, are they integrated into one sound or can they be segregated into their two constituent sounds? This question has been recently addressed by two studies, with the weight of the evidence favoring integration. Best *et al.* (2004) presented either one or two 150-ms isospectral broadband noises in virtual space and found that when two sounds were both located on the vertical midline, listeners only reported the presence of one sound. This suggests that the two sounds were integrated. Given that there were no onset/offset asynchronies, no envelope differences, and no spectral differences (outside of those imposed by the pinnae) to mark the presence of two sounds, the fact that listeners perceived one sound is sensible.

In a study with more similarities to the one presented here, Bremen *et al.* (2010) played two 50-ms isospectral (but *not* iso-intensity) broadband sounds from speakers on the vertical midline, distinguishing one of the two sounds by applying an AM envelope. Listeners were asked to make a head saccade to identify the location of the AM stimulus. For sounds that were not widely separated (generally, 60 degrees and below), Bremen *et al.* (2010) found that listeners' localizations of the AM target better corresponded with the weighted average of the stimulus intensities than with the actual location of either of the stimuli, again, suggesting that the two sounds were integrated. In this case, the presence of an AM envelope was not sufficient to bring about segregation of the two sounds. For sounds that were widely separated (generally 75 degrees and above), Bremen *et al.* (2010) found an emerging class of bimodal responses, where the listeners' responses corresponded better with the actual location of one of the two stimuli than with the weighted average, but there was no apparent bias for localizing the target. In these cases, the results suggest that the two sounds were partially segregated—enough to recover the two source locations, but not enough to properly assign the AM stimulus to its source, which would have resulted in a unimodal distribution of responses only at the target location.

We do not believe that the results seen at high spatial disparities in Bremen *et al.* (2010) correspond to the results found in the current study for two reasons. First, our stimuli were separated by only 40 degrees, within the range of disparities that resulted in integration in Bremen *et al.* (2010), where the weighted-average prediction was very good up to 45 degrees, and continued to outperform a bimodal prediction up to at least 75 degrees. Second, unlike Bremen *et al.* (2010), where there is no indication that listeners preferred to respond to the location of the target in the bimodal case, for our data (among our veridical responders) responses were effectively unimodal, with the listener selecting the direction of the target sound with, generally, between 80% and 90% accuracy across a wide range of MF.

Does this ability to accurately identify the direction of the target indicate that our listeners are, in fact, segregating the stimuli? We believe that it does. When presented with two simultaneous sounds, ultimately, the listener must either integrate them into one sound (as seen in Best *et al.*, 2004), or segregate them into two. Bremen *et al.* (2010) indicates that the integration of sounds on the vertical midline results in an elevation estimate that is at the weighted average location of the two sounds, which, for our study, would be directly between the two speakers because our AM-target and our noise masker were at the same intensity. We have argued above that our veridical responders are using the elevation information contained in the AM + masker stimulus to perform the task. However, an elevation cue directly between the two speakers, as would occur for an integrated stimulus, would not allow our veridical responders to perform at high accuracy. The raw data from Bremen *et al.* (2010) indicate that saccade distribution endpoints relative to the weighted average span on the order of 20 degrees for sound pairs with a large intensity difference and are even larger when the two sounds have similar sound levels—with this sort of uncertainty in the localization, the elevation cue from an integrated stimulus, even if it were slightly biased from the direct center of the two speakers, would result in random guessing rather than high accuracy. As such, it seems that our listeners must be using an elevation cue that does not arise from a weighted average of the speaker locations, but instead is a result of segregated stimuli. Note that it is not necessarily true that our listeners perform a valid segregation of the stimuli—in fact, it seems our non-veridical responders incorrectly segregate at low MF such that the AM envelope is falsely assigned to the wrong sound location.

It is important to consider why we have found evidence for segregation of sounds on the vertical midline when previous studies have not. Best *et al.* (2004) used identical stimuli, so sound integration is more likely than in our study. Integration at small spatial disparities in Bremen *et al.* (2010) is notable considering the contrary findings here. Bremen *et al.* (2010) used stimuli with similar spectral characteristics (their noise carrier was bandpassed between 0.5 and 20 kHz, while ours was not bandpassed) and their 50 Hz AM target was close to the 30 and 60 Hz AM targets in our experiment, both of which resulted in high-accuracy performance among our veridical responders. There are, however, a few salient differences between the two studies that could potentially account for the differing results. One is the task, which differed in the required precision of the localization that the listeners were asked to perform, as well as the underlying uncertainty of the locations of the stimuli. In the Bremen *et al.* (2010) study, listeners were asked to pinpoint the source of the AM, with many possible spatial configurations of the sounds. In our study, listeners were merely asked to select the source of the AM, which was known to have come from one of two speakers. Although the change from a free localization to a forced-choice decision cannot have impaired our listeners' performance (instead, perhaps, turning a coarse localization estimate, as often seen in Bremen *et al.* (2010), into a discrete up/down decision), we find it unlikely that this change in task can account for the differences between the two studies. In Bremen *et al.* (2010), the contribution of each stimulus to the integrated stimulus was weighted by long-term level of each stimulus independent of the temporal properties of the stimulus. If our listeners were integrating our equal-level stimuli as seen in Bremen *et al.* (2010), the apparent elevation would have been more or less directly between the speakers and, when thresholded to an up/down decision, this would have resulted in a bimodal distribution of responses, not the unimodal distribution seen in our veridical responders.

A second difference between the studies is in the number of speaker configurations. In Bremen *et al.* (2010), there were 72 possible speaker configurations and the listeners had no prior information about where the next pair of sounds was going to come from, whereas in our study, the listeners knew the two locations (one configuration) and could, in principle, attend to particular elevations. This *a priori* information about the source elevations could well assist the segregation of the sounds. For example, in the azimuthal plane, prior information about the location of one source has been shown to aid in the localization of another simultaneous source (Yost and Brown, 2013). Because our task involved only two known speaker locations, a template-matching solution based on specific spectral notch locations may have been easier for our listeners to implement than in a task with multiple speaker configurations.

A third difference between the studies was the duration of the stimuli. Our study used longer stimuli (400 ms vs 50 ms), a factor which has been shown to confer modest benefits in single-sound vertical localization in mammals (ferrets: Bizley *et al.*, 2007; cats: Gai *et al.*, 2013; but, see Hofman and van Opstal, 1998 for inconclusive improvement in human elevation localization between 80 and 500 ms duration) and in AM detection (O'Connor *et al.*, 2011). Another factor when considering the effect of duration on vertical localization is the fact that the integration/segregation process itself takes time (see Bregman, 1990 for examples of duration effects on stream segregation, auditory continuity illusions, and other aspects of auditory segregation), and a brief 50-ms stimulus may not be sufficient to perform a difficult segregation in elevation on the basis of temporal envelope differences alone.

Fourth, our stimuli, which were presented at 58 dB, were >10 dB higher than those in Bremen *et al.* (2010), which were ∼42–45 dB for the case of nearly iso-intensity presentations. Interestingly, in cat IC, Davis *et al.* (2003) show example type-O notch responses that are strong between 50 and 70 dB, but which largely disappear at 40 dB. It is possible that in this application, segregation is an intensity-dependent effect and that the Bremen *et al.* (2010) stimuli may have been too quiet to elicit it. Together, these factors may account for why we were able to demonstrate segregation for simultaneously presented sounds on the vertical midline while others have not.

V. CONCLUSIONS

Here, we show that the ability to properly analyze an "auditory scene" is not limited to sounds along the horizontal plane, but can also extend to the segregation of similar sounds from different elevations. We also report the unexpected finding that this ability appears in only about half of our listeners, with the remainder showing non-veridical source assignment of an AM target that is not consistent with an inability to perform the task. At the population level, both veridical and non-veridical performance on our task can be accounted for by a simple model based on the physiology of IC neurons. Cells in the IC are known to be sensitive to elevation (Aitkin and Martin, 1990, Zwiers *et al.*, 2004) due to their selectivity to spectral notches (Davis *et al.*, 2003). The data presented here suggest two further insights. First, segregation of two similar sounds in elevation is possible through the evaluation of temporal envelopes in conjunction with notch selectivity. Second, veridical or non-veridical identification of elevation direction in a two-sound segregation task may be determined by the synaptic strength of the inhibitory inputs that shape notch selectivity in the IC.

ACKNOWLEDGMENTS

This work was funded by the National Institute on Deafness and Other Communication Disorders Grant Nos. DC-02514 and T32 DC-008072.

Aitkin, L., and Martin, R. (**1990**). "Neurons in the inferior colliculus of cats sensitive to sound-source elevation," Hear. Res. **50**, 97–106.

Asano, F., Suzuki, Y., and Sone, T. (**1990**). "Role of spectral cues in median plane localization," J. Acoust. Soc. Am. **88**, 159–168.

Bacon, S. P., Opie, J. M., and Montoya, D. Y. (**1998**). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech. Lang. Hear. Res. **41**, 549–563.

Best, V., van Schaik, A., and Carlile, S. (**2004**). "Separation of concurrent broadband sound sources by human listeners," J. Acoust. Soc. Am. **115**, 324–336.

Bizley, J. K., Nodal, F. R., Parsons, C. H., and King, A. J. (**2007**). "Role of auditory cortex in sound localization in the midsagittal plane," J. Neurophysiol. **98**, 1763–1774.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), Chap. 3, pp. 213–394.

Bremen, P., van Wanrooij, M. M., and van Opstal, A. J. (**2010**). "Pinna cues determine orienting response modes to synchronous sounds in elevation," J. Neurosci. **30**, 194–204.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and two ears," J. Acoust. Soc. Am. **25**, 975–979.

Davis, K. A., Ramachandran, R., and May, B. J. (**2003**). "Auditory processing of spectral cues for sound localization in the inferior colliculus," J. Assoc. Res. Otolaryngol. **4**, 148–163.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.

Gai, Y., Ruhland, J. L., Yin, T. C. T., and Tollin, D. J. (**2013**). "Behavioral and modeling studies of sound localization in cats: Effects of stimulus level and duration," J. Neurophysiol. **110**, 607–620.

Gustafsson, H. A., and Arlinger, S. D. (**1994**). "Masking of speech by amplitude-modulated noise," J. Acoust. Soc. Am. **95**, 518–529.

Hebrank, J., and Wright, D. (**1974**). "Spectral cues used in the localization of sound sources on the median plane," J. Acoust. Soc. Am. **56**, 1829–1834.

Hill, K. T., and Miller, L. M. (**2010**). "Auditory attentional control and selection during cocktail party listening," Cereb. Cortex **20**, 583–590.

Hofman, P. M., and van Opstal, A. J. (**1998**). "Spectro-temporal factors in two-dimensional human sound localization," J. Acoust. Soc. Am. **103**, 2634–2648.

Hofman, P. M., and van Opstal, A. J. (**2003**). "Binaural weighting of pinna cues in human sound localization," Exp. Brain Res. **148**, 458–470.

Hofman, P. M., van Riswick, J. G. A., and van Opstal, A. J. (**1998**). "Relearning sound localization with new ears," Nat. Neurosci. **1**, 417–421.

May, B. J., Anderson, M., and Roos, M. (**2008**). "The role of broadband inhibition in the rate representation of spectral cues for sound localization in the inferior colliculus," Hear. Res. **238**, 77–83.

Middlebrooks, J. C., and Green, D. M. (**1991**). "Sound localization by human listeners," Annu. Rev. Psychol. **42**, 135–159.

Nelken, I., and Young, E. D. (**1994**). "Two separate inhibitory mechanisms shape the responses of dorsal cochlear nucleus type IV units to narrowband and wideband stimuli," J. Neurophysiol. **71**, 2446–2462.

O'Connor, K. N., Johnson, J. S., Niwa, M., Noriega, N. C., Marshall, E. A., and Sutter, M. L. (**2011**). "Amplitude modulation detection as a function of modulation frequency and stimulus duration: Comparisons between macaques and humans," Hear. Res. **277**, 37–43.

O'Connor, K. N., and Sutter, M. L. (**2000**). "Global spectral and location effects in auditory perceptual grouping," J. Cogn. Neurosci. **12**, 342–354.

Reiss, L. A. J., and Young, E. D. (**2005**). "Spectral edge sensitivity in neural circuits of the dorsal cochlear nucleus," J. Neurosci. **25**, 3680–3691.

Roffler, S. K., and Butler, R. A. (**1968**). "Factors that influence the localization of sound in the vertical plane," J. Acoust. Soc. Am. **43**, 1255–1259.

Shinn-Cunningham, B. G., Lee, A. K. C., and Oxenham, A. J. (**2007**). "A sound element gets lost in perceptual competition," Proc. Natl. Acad. Sci. U.S.A. **104**, 12223–12227.

Tollin, D. J., and Yin, T. C. T. (**2003**). "Spectral cues explain illusory elevation effects with stereo sounds in cats," J. Neurophysiol. **90**, 525–530.

van Wanrooij, M. M., and van Opstal, A. J. V. (**2007**). "Sound localization under perturbed binaural hearing," J. Neurophysiol. **97**, 715–726.

Vélez, A., and Bee, M. A. (**2011**). "Dip listening and the cocktail party problem in grey treefrogs: Signal recognition in temporally fluctuating noise," Anim. Behav. **82**, 1319–1327.

Woods, D. L., Alain, C., Diaz, R., Rhodes, D., and Ogawa, K. H. (**2001**). "Location and frequency cues in auditory selective attention," J. Exp. Psychol. Hum. Percept. Perform. **27**, 65–74.

Yost, W. A., and Brown, C. A. (**2013**). "Localizing the sources of two independent noises: Role of time varying amplitude differences," J. Acoust. Soc. Am. **133**, 2301–2313.

Young, E. D., Spirou, G. A., Rice, J. J., Voigt, H. F., and Rees, A. (**1992**). "Neural organization and responses to complex stimuli in the dorsal cochlear nucleus," Philos. Trans. R. Soc. London B **336**, 407–413.

Zwiers, M. P., Versnel, H., and van Opstal, A. J. (**2004**). "Involvement of monkey inferior colliculus in spatial hearing," J. Neurosci. **24**, 4145–4156.