Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map

Christoph Kayser, Christopher I. Petkov, Michael Lippert, and Nikos K. Logothetis

Supplemental Discussion

The auditory saliency map described here is structurally identical to saliency maps proposed for the visual system [S21–S25]. The only differences are the shape of the feature detectors, and the normalization procedure. Visual models employ features extracted by neurons in the visual cortex, such as luminance contrast, orientation or color. However, although these features differ from those extracted by the auditory model in their interpretation, mathematically they are very similar: The filters used to extract frequency or temporal contrast can be interpreted as detectors for horizontal and vertical orientations and the filters used to extract sound intensity are identical to those extracting luminance. In this respect, the auditory and visual saliency map differ more in their interpretation than conceptually. A feature of the auditory saliency map is that it incorporates the temporal domain while classical visual models operate on spatial images only. This requires a different normalization procedure for the auditory model. In this case, the temporal domain imposes causality restrictions that are incorporated in the sliding window normalization.

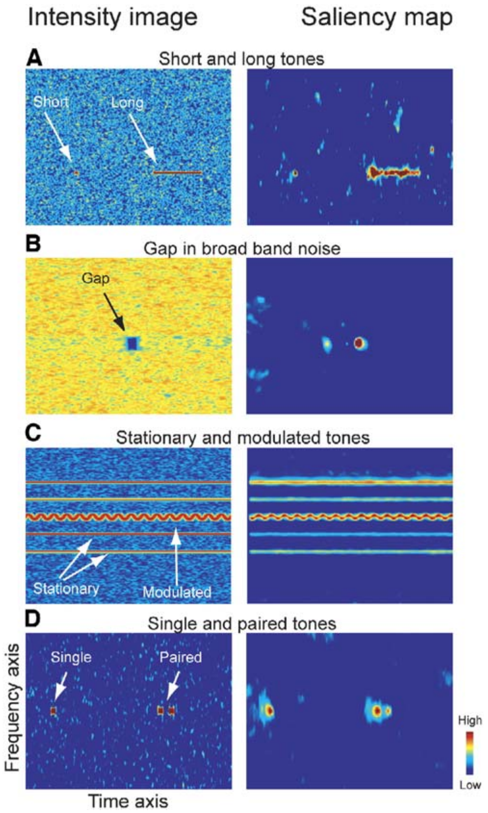

Figure S1. Intensity Images and Saliency Maps for Different Toy Scenarios

These examples demonstrate that the saliency map reproduces basic properties of auditory scene perception [S26]. (A) Tones are salient irrespective of length but longer events accumulate higher saliency in accord with these tones being easier to select from acoustical scenes [S27, S28]. (B) "Missing" parts (gaps) in a broad spectrum are salient. (C) Modulated events achieve higher saliency compared to stationary events in agreement with those being easier to detect [S26, S29]. (D) The saliency map replicates the phenomenon of forward masking. In a sequence of two closely spaced tones the second is less salient in agreement with the phenomenon of forward masking [S15–S17].

Supplemental Experimental Procedures

The Saliency Map

The auditory system segregates sounds in a complex scene based on individual features such as spectral or temporal modulation [S1–S4]. The auditory saliency map incorporates these different features in a hierarchical architecture employing the parallel extraction of features at different scales (Figure 1). (1) At the first stage, an intensity image is created in time and frequency dimensions in analogy to the initial processing by the cochlea and basilar membrane [S5, S6]. Sound samples (sampled at 16 kHz) are preprocessed by using a sliding window Fourier analysis (37 ms windows, 36 ms overlap, 1 msec nominal temporal resolution, 1024 point FFT), resulting in a two-dimensional image with time and frequency as axes. (2) At the second stage, this image is analyzed by feature detectors on different scales, representing various levels of sound feature analysis by auditory neurons [S7–S11]. The features extracted are intensity, frequency structure, and temporal structure. Each feature is

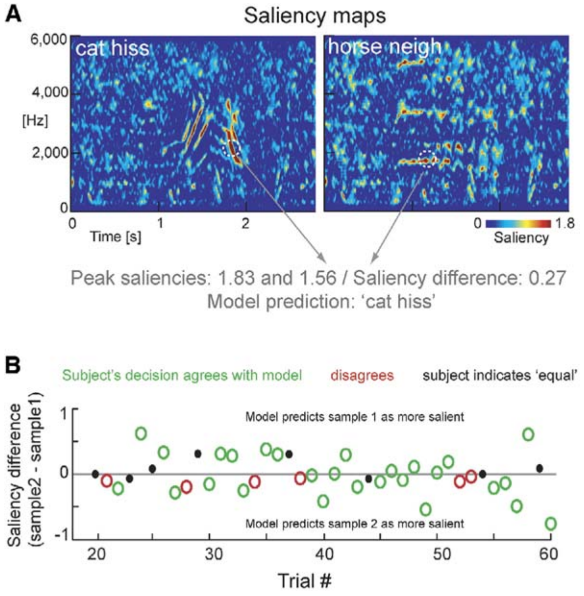

Figure S2. Comparison of Human and Model Ratings of Saliency

(A) Computation of the model's saliency rating. Shown are two example saliency maps together with the locations of peak saliency in each scene (circles). Based on the difference, the model chooses one scene as containing the most salient event (in this case the "cat hiss"). (B) Example data from one subject (40 consecutive trials). Values on the *y* axis indicate the continuous saliency difference reported by the model and the color indicates whether the models' prediction matches the subject's response.

extracted with filters modeled following results from auditory physiology: the intensity filter corresponds to receptive fields with only an excitatory phase, the frequency contrast filters to such with an excitatory phase and simultaneous side band inhibition, and the temporal contrast filters to receptive fields with an excitatory phase and a subsequent inhibitory phase [S8, S12, S13]. On the smallest scale, these filters have a bandwidth of 200 Hz and a duration of 20 ms for both excitatory and inhibitory phases. The peak latency of the postinhibition is shifted to the excitation by 30 ms and the peak strength of the inhibition is half that of the excitation [S12]. These filters are applied on four scales, each being a resampled version (factor 2) of the previous. (3) At the third stage, the maps obtained at different time-frequency scales for each feature are compared by subtracting the coarser from the finer scale and setting negative values to zero. This is a center-surround differentiation, which mimics properties of local cortical inhibition [S7, S14]. (4) Fourth, the resulting maps are then normalized by scaling them with the difference between global and local maxima. Each center-surround map is first scaled to a fixed range [0,1] and then multiplied by $(1 - \langle\text{peak height}\rangle_{\text{all local peaks}})$. In this equation, "$\langle\text{peak height}\rangle$" refers to the average, which is taken over all local peaks. This normalization provides a feature-independent scale and serves to promote maps in which few but highly conspicuous peaks occur compared to maps with many equally sized peaks. If one peak dominates, this factor will be close to one; however, if there are several peaks of similar height, this factor will be small. This normalization is applied to a sliding window of 150 ms duration and the values used for normalization (local and global maxima) are computed using an asymmetric window extending 225 ms into the past and 75 ms into the future. These numbers were chosen based on known properties of forward and backward masking effects [S15–S17]. (5) At the last stage, different center-surround maps are summed for individual features that are finally combined to yield the saliency map, in analogy to the idea of feature integration [S2, S3, S18].

Sounds and Experiments

Auditory scenes consisted of 52 samples of natural sounds such as animal vocalizations (dog, cats, horse, birds, owl, frogs, sheep, turkey, cow, rooster, lions, elephant, monkeys, and puma), nature sounds (bird wings, horse gallop, jungle sounds, ocean waves, water bubbles, river, thunder, wind, storm, and rain), machine noises (train, motorcycle, police sirens, traffic, typewriter, guns, sword, and rocket blast), and the sound of human crowds cheering and applauding, but no speech.

Human Ratings of Saliency

The saliency map describes the stimulus-based conspicuity of different features. To test whether this corresponds to human perceptual judgments of auditory saliency we used a paradigm that minimizes the cognitive demand on the subjects and allows asking the same question to subject and model.

Seven human subjects participated in this experiment following informed consent. The stimuli consisted of pairs of randomly chosen complex auditory scenes. Each scene consisted of a snippet of one of the 52 natural sounds (maximally 1.2 s long), which was presented on a background sound (4 s long). The background consisted of the average of 20 randomly chosen scenes and white noise; hence, the background was not completely white or devoid of structure. The subjects had to compare the saliency within these scenes by making the simplest possible rating: subjects could either choose one of the scenes as containing the most salient event, or indicate that both scenes had similar levels of saliency. Hence, the possible responses were "scene 1," "scene 2,“ or "equal" and subjects indicated their choice by pressing a mouse button. Each subject was presented with 150 random pairs and catch trials were inserted in which the same scene was presented twice. All subjects correctly classified all catch trials as equal. The simplistic nature of this paradigm allowed a direct comparison with the model, which was posed with the same question as the subject. For each pair, the peak levels of saliency for both scenes were compared; both the difference between these values ("saliency difference") as well as the identity which scene had the higher value ("model prediction") were used for a comparison with the human ratings (Figure S2A).

Two types of analysis were carried out to compare the subjects' ratings to the model (Figures 2 and S2B). First, those trials of which the subjects choose either of two scenes were selected (hence, excluding the equal trials, 13% of all trials). For these, the correlation between the subjects' choice and the model prediction were computed (Figure 2, left). Second, the trials were sorted according to the subjects' responses, and the saliency differences reported by the model were compared (Figure 2, middle).

To investigate the contribution of different sound features to these saliency judgments, a second decision was obtained from the subjects. After rating the saliency in these scenes, subjects subsequently had to indicate on which feature (they thought) they had based their decision. Possible choices were "intensity," "frequency structure," and "temporal structure." For analysis, we computed the contribution of the individual features' saliency maps to the total saliency, quantified in percent (Figure 2, right). This contribution was then selectively averaged for those trials were the subjects reported to rely on a particular feature and for all other trials.

Sounds were presented using Matlab (Mathworks, Inc.) and Sennheiser (Sennheiser Inc.) headphones; each subject could adjust the loudness to his/her own comfort level, usually around 55–60 dB SPL.

Human Detection Experiment

This experiment consisted of a measurement of detection thresholds for sound snippets embedded in natural background noise. Twelve human subjects participated in this experiment following informed consent. Subjects were presented with a 4 min long background noise, presented binaurally, on which randomly distributed snippets (0.5 s long) of natural sounds were presented. For this experiment a subset of 25 sounds was used and these snippets were presented every 0.8-6 s (uniform distribution) to only one of the two ears (randomly chosen). The subject's task was to detect these sounds and to indicate on which ear they appeared by pressing a mouse button. For those trials where a sound was detected, subjects' performance at indicating the correct side of presentation was 98.4%. The goal of the experiment was to verify that more salient sounds are detected even if their average (RMS) intensity is low compared to that of the background noise. Hence, sound snippets were presented with different scaling relative to the background (ranging form 0 dB to −20dB SPL from the background noise, in eight steps). Each sound snippet was randomly presented twice to each subject at two different scalings. The intensity distribution was balanced across subjects (across subjects design), yielding three data points per scaling and sound snippet. Two types of analyses were conducted on this data set. First, the snippets were sorted by saliency, as determined from the model, into two groups: a more salient group (the 12 snippets with higher saliency than the median across all 25, which was not grouped) and a less salient group (the 12 snippets with lower than median saliency). Then, the detection performance across subjects was compared for these two groups (Figure 3A). Second, for each sound, a detection threshold was obtained. This was defined as the least intense scaling at which the sound was detected in at least 66% of the presentations. This threshold was then correlated with the peak saliency of the sound snippet (Figure 3B).

Monkey Detection Experiment

This experiment assessed the impact of sound saliency on the natural orienting behavior of macaque monkeys (Macaca mulatta). A total of 17 male animals (age 4–10 years), which are part of a large colony housed at the Max Planck Institute for Biological Cybernetics was tested (eight of these twice, for a total of 25 experiments). All animals are socially housed and provided with enrichments (toys, hammocks, ropes, etc.). All experimental procedures were in accordance with the local authorities (Regierungspräsidium) and the European Community (EUVD 86/609/EEC) for the care and use of laboratory animals. Recent studies support the utility of such behavioral paradigms exploiting the natural behavior of monkeys [S19, S20].

The same background sound as in the human detection experiment was used together with six of those sound snippets. Sounds were presented through speakers placed slightly above and behind on each side of the animals head, at a distance of 60 cm. Sounds were presented at an intensity of 65 dB SPL (as measured by a sound level meter, Bruel & Kjaer 2238). For testing, a subject was seated in a primate chair and placed in the middle between the two speakers. All equipment in the room was concealed with black curtains and the walls were covered with sound attenuating black foam. During testing the room was completely dark and the animal's behavior was monitored using an infrared sensitive CCD camera and an infrared light source placed centrally over the animal. The background sound was played continuously from both speakers. The experimenter monitored the subject's activity from outside the room and initiated a trial whenever the subject's attention was directed centrally by flashing a dim LED light. When the animal was looking centrally, a sound snippet was played randomly on one side and this procedure was repeated for all six sound snippets. The side of presentation was chosen randomly for individual stimuli and balanced across subjects. All behavior was recorded on digital video onto a PC (400 × 300 pixels, 25 Hz) and the stimulus timing, identity, and side of presentation was stored for later analysis. As in all studies examining the spontaneous behavior of animals or human infants, subjects quickly habituate to the testing environment and loose interest. As no reward, feedback, or training was provided, this precluded us from testing a larger number of stimuli.

Data evaluation was done independently by two observers blind to stimulus identity or side of presentation. For each time of stimulus presentation they scored whether the animal showed any clear orienting behavior to either side, and if so, to which side. Scoring whether the animals showed a clear orienting behavior was unambiguous, as the speakers were placed in opposing directions and the animals had to make large head or body turns to direct toward these. This is corroborated by a high interobserver agreement of 94.8% across all trials.

Supplemental References

S1. 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.

S2. Bregman, A.S. (1990). Auditory scene analysis. (Cambridge: MIT Press).

S3. Yost, W.A. (1992). Auditory Perception and Sound Source Determination. Curr. Dir. Psy. Sci. *1*, 179–183.

S4. Alain, C., Arnott, S.R., and Picton, T.W. (2001). Bottom-up and top-down influences on auditory scene analysis: evidence from event-related brain potentials. J. Exp. Psychol. Hum. Percept. Perform. *27*, 1072–1089.

S5. Helmholtz, H. (1863). On the sensations of Tone as a Physiological Basis for the theory of music. (Whitefish, MT: Kessinger Publishing)

S6. Shamma, S. (2001). On the role of space and time in auditory processing. Trends Cogn. Sci. *5*, 340–348.

S7. Schreiner, C.E., Read, H.L., and Sutter, M.L. (2000). Modular organization of frequency integration in primary auditory cortex. Annu. Rev. Neurosci. 23, 501–529.

S8. deCharms, R.C., Blake, D.T., and Merzenich, M.M. (1998). Optimizing sound features for cortical neurons. Science 280, 1439–1443.

S9. Kaur, S., Lazar, R., and Metherate, R. (2004). Intracortical pathways determine breadth of subthreshold frequency receptive fields in primary auditory cortex. J. Neurophysiol. 91, 2551–2567.

S10. Miller, L.M., Escabi, M.A., Read, H.L., and Schreiner, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J. Neurophysiol. 87, 516–527.

S11. Rauschecker, J.P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. Science 268, 111–114.

S12. Valentine, P.A., and Eggermont, J.J. (2004). Stimulus dependence of spectro-temporal receptive fields in cat primary auditory cortex. Hear. Res. 196, 119–133.

S13. Fishbach, A., Yeshurun, Y., and Nelken, I. (2003). Neural model for physiological responses to frequency and amplitude transitions uncovers topographical order in the auditory cortex. J. Neurophysiol. *90*, 3663–3678.

S14. Sillito, A.M., Grieve, K.L., Jones, H.E., Cudeiro, J., and Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. Nature 378, 492–496.

S15. Warren, R. (1999). Auditory perception: a new analysis and synthesis. (Cambridge: Cambridge University Press).

S16. Moore, B., and Glasberg, B. (1983). Growth of forward masking for sinusoidal and noise maskers as a function of signal delay: Implications for suppression in noise. J. Acoust. Soc. Am. 74, 750–753.

S17. Brosch, M., and Schreiner, C.E. (2000). Sequence sensitivity of neurons in cat primary auditory cortex. Cereb. Cortex 10, 1155–1167.

S18. Treisman, A.M., and Gelade, G. (1980). A feature-integration theory of attention. Cognit. Psychol. 12, 97–136.

S19. Maier, J.X., Neuhoff, J.G., Logothetis, N.K., and Ghazanfar, A.A. (2004). Multisensory integration of looming signals by rhesus monkeys. Neuron 43, 177–181.

S20. Ghazanfar, A.A., and Logothetis, N.K. (2003). Neuroperception: facial expressions linked to monkey calls. Nature 423, 937–938.

S21. Itti, L., and Koch, C. (1998). A model of Saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Learning 22, 1254–1259.

S22. Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. 40, 1489–1506.

S23. Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. 4, 219–227.

S24. Parkhurst, D.J., and Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. Eur. J. Neurosci. 19, 783–789.

S25. Parkhurst, D., Law, K., and Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. Vision Res. 42, 107–123.

S26. Cusack, R., and Carlyon, R.P. (2003). Perceptual asymmetries in audition. J. Exp. Psychol. Hum. Percept. Perform. 29, 713–725.

S27. Ehret, G. (1997). The auditory cortex. J. Comp. Physiol. [A] 181, 547–557.

S28. He, J., Hashikawa, T., Ojima, H., and Kinouchi, Y. (1997). Temporal integration and duration tuning in the dorsal zone of cat auditory cortex. J. Neurosci. 17, 2615–2625.

S29. Naatanen, R., and Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. Psychol. Bull. 125, 826–859.