Current Biology, Vol. 15, 1943–1947, November 8, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.cub.2005.09.040

Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map

Christoph Kayser,* Christopher I. Petkov, Michael Lippert, and Nikos K. Logothetis
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen
Germany

Summary

Our nervous system is confronted with a barrage of sensory stimuli, but neural resources are limited and not all stimuli can be processed to the same extent. Mechanisms exist to bias attention toward the particularly salient events, thereby providing a weighted representation of our environment [1]. Our understanding of these mechanisms is still limited, but theoretical models can replicate such a weighting of sensory inputs and provide a basis for understanding the underlying principles [2, 3]. Here, we describe such a model for the auditory system—an auditory saliency map. We experimentally validate the model on natural acoustical scenarios, demonstrating that it reproduces human judgments of auditory saliency and predicts the detectability of salient sounds embedded in noisy backgrounds. In addition, it also predicts the natural orienting behavior of naive macaque monkeys to the same salient stimuli. The structure of the suggested model is identical to that of successfully used visual saliency maps. Hence, we conclude that saliency is determined either by implementing similar mechanisms in different unisensory pathways or by the same mechanism in multisensory areas. In any case, our results demonstrate that different primate sensory systems rely on common principles for extracting relevant sensory events.

Results

We are frequently exposed to an overabundance of sensory events. Our brain cannot fully process all stimuli at once and neural mechanisms exist for selecting those potentially relevant for behavior [1, 3, 4]. This selection of particular stimuli for thorough analysis is a component of sensory attention and consists of an involuntary and stimulus-driven mechanism and a slower cognitive component, incorporating voluntary control [5–7]. The initial stimulus-driven mechanism provides weighted representations of sensory scenes, biasing perception toward salient events. This mechanism postulates that some features in a scene are conspicuous based on their context and, hence, are salient, and thus attract attention; for example, red cherries in a green tree or a police car's siren amid the rush-hour's noise.

To understand the mechanisms underlying this selection of salient stimuli, the concept of a saliency map has been proposed [2, 3]. These saliency maps can be conceptualized as theoretical models, which employ the hierarchical and parallel extraction of different features and build on existing understanding of sensory processing. For the visual system, such models were shown to replicate several properties of human overt attention [2, 8, 9]. For other sensory systems, however, these attentional mechanisms are largely unexplored. In the following, we develop a saliency map for the auditory system and demonstrate that such models can serve a conceptual basis for comparing the principles underlying this form of attention across sensory systems.

The concept of saliency applies to properties of sensory stimuli and a particular instantiation of these in a sensory scene. As such, saliency describes the potential influence of a stimulus on our perception and behavior. The salient stimuli are those which are more likely to attract our attention or which will be easier to detect. Hence, saliency complements the frequently studied processes of attention and detection and introduces a qualitative description of those stimulus properties relevant for these processes. In addition to being a theoretical concept describing properties of sensory stimuli, the saliency map serves as a basis for understanding the cortical representations and mechanisms which implement this weighting of sensory stimuli [3].

The Auditory Saliency Map

The auditory system segregates sounds in a complex scene based on individual features such as spectral or temporal modulation [10–13], and by relying on such modulations, we are able to detect sounds of interest amidst fairly high levels of "noise" [14]. The auditory saliency map extracts these features in parallel (Figure 1A; see the Supplemental Experimental Procedures available with this article online), representing various levels of sound feature analysis by auditory neurons [15–18]. Different sets of filters are used to quantify sound intensity, frequency contrast, and temporal contrast, and evidence for each feature is compared across scales with a center-surround mechanism [15]. To obtain a feature-independent scale, these maps are normalized using an asymmetric sliding window extending into the past and future in a manner consistent with psychoacoustical masking effects [19, 20]. Finally, the saliency maps from individual features are combined, in analogy to the idea of feature integration [6, 10].

We confirmed that this model replicates basic properties of auditory scene perception as described in the human psychophysical literature [21] (Figure S1): both short and long tones on a noisy background are salient, as are gaps (the absence of frequencies in a broad band noise); long tones accumulate more saliency than short tones; temporally modulated tones are more salient than stationary tones; and in a sequence of two closely spaced tones, the second is less salient in agreement with the phenomenon of forward masking.

*Correspondence: christoph.kayser@tuebingen.mpg.de

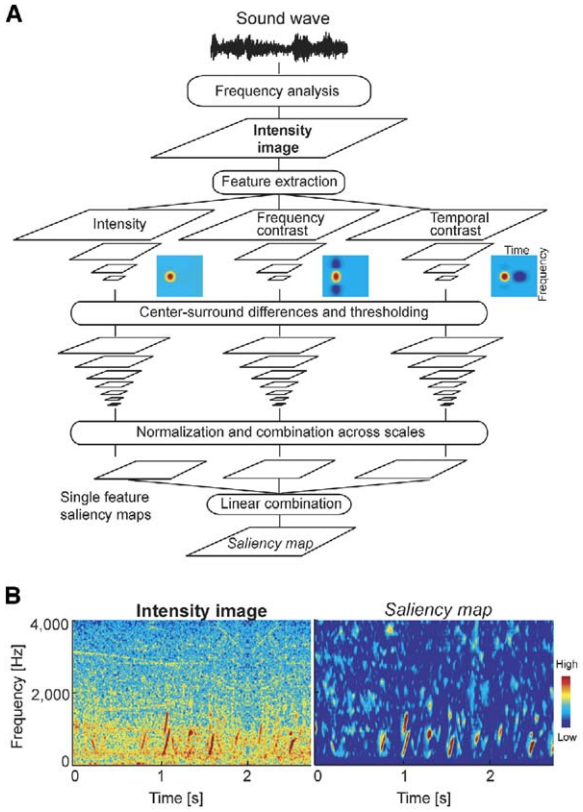

Figure 1. The Auditory Saliency Map

(A) Schematic of the model. Initially the sound wave is converted to a time-frequency representation ("intensity image"). Then important auditory features are extracted on different scales and in parallel streams (intensity, frequency contrast, and temporal contrast) with different sets of filters. These filters are schematized in the colored insets. For each feature, the maps obtained at different scales are compared using a center-surround mechanism and normalized to promote those maps containing highly conspicuous peaks. The center-surround maps are collapsed across scales yielding saliency maps for individual features, which are finally added to yield the saliency map.

(B) Intensity image and saliency map for one example scene (water bubbles on a noisy background).

Human Ratings of Saliency

We experimentally verified that the model captures essential aspects of human judgments of auditory saliency (Figure 2 and Figure S2). As the model describes the stimulus-determined conspicuity of different features and not cognitive aspects of auditory attention, we used a paradigm that minimized the cognitive demand on the subjects and allowed us to ask the same question to the subjects and model. Both were presented with pairs of complex auditory scenes and had to compare the saliency in these by indicating the scene containing the higher saliency (see the Supplemental Experimental Procedures).

A comparison of the subjects' decisions with those of the model yielded a significant correlation for all seven subjects with an average of 0.47 ± 0.1 (Figure 2, left, mean ± SD across subjects). Further, grouping trials according to subjects' responses resulted in a significant effect on the saliency reported by the model (Figure 2, middle): when subjects indicated "equal" saliency, the saliency difference reported by the model was close to zero ($p = 0.22$, t test), but when subjects chose one of the samples the difference was large ($p < 10^{-10}$, t test in both cases). Hence, the model well predicts human perceptual ratings of saliency both when subjects experience a strong difference in saliency and when subjects experience only small differences between scenes.

The saliency map extracts a measure of saliency which cannot be obtained from sound intensity alone. This point is important to establish, as otherwise the saliency map would act as a simple sound level detector, and adding the different feature maps and multiscale analysis would not yield any improvement. To establish this, we used the intensity image instead of the saliency map to compute which scene should be more salient. This prediction correlated with the subjects' decision (0.34 ± 0.85, mean ± SD); however, this correlation was significantly weaker than that obtained from the saliency map ($p < 0.05$, $n = 7$, paired t test).

To verify that the different feature components of the saliency map capture basic perceptual distinctions made by the human observers, we asked the subjects to indicate on which of the three features (intensity, frequency structure, or temporal structure) they had based their saliency decision. We then compared the contribution of each feature channel to the total saliency on trials where this feature was indicated by the subject compared to trials where another feature was indicated (Figure 2, right). For intensity, there was no significant difference between "selected" and "not selected" trials ($p = 0.39$, t test). But the contribution of frequency contrast and temporal contrast to the saliency was significantly larger on trials where subjects reported a reliance on that feature ($p < 0.0001$ and $p < 0.05$, respectively). Thus, the model replicates basic perceptual feature distinctions of human auditory perception. That we did not observe a significant effect for intensity can be understood as any feature is dependent on intensity (zero intensity implies no other features exist). Thus, a feature like frequency or temporal contrast will always be somewhat confounded with intensity.

Human Detection Experiment

The saliency map predicts which features in a complex auditory scene will naturally capture our attention and, hence, are more easily detected. In a second experiment, we directly confirmed that the model replicates detection of salient events in noisy scenes by human subjects. Subjects had to detect monaurally presented sound snippets whose level was varied in relation to a binaural naturalistic background noise (see Supplemental Experimental Procedures).

Overall, subject's performance at detecting these sound snippets was far above chance level (Figure 3A). Using the model to separate sounds into a more salient and a less salient group revealed that the more salient stimuli were more often detected (81% versus 71%). An analysis of the contingency table revealed a significant effect of saliency on detection performance (Figure 3A, Fisher's exact test, $p < 0.01$). Based on the subjects'

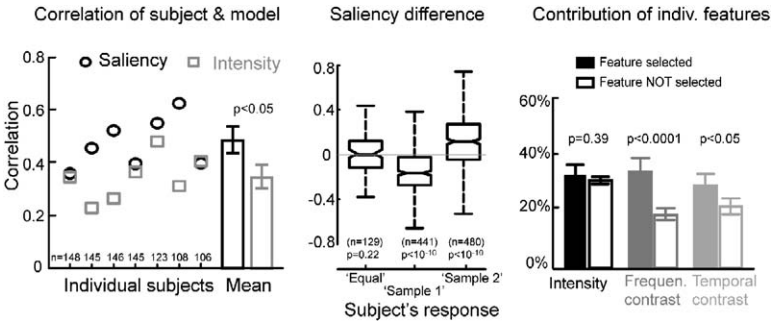

Figure 2. Human Ratings of Saliency

Left: correlation between subjects' and the model's decisions computed upon exclusion of "equal" saliency trials (the number of trials used is indicated for each subject). Similarly, the correlation between predictions based on the intensity of the image and the subject's decisions is shown. Bars and error bars indicate mean and SD across subjects, and the p value refers to a paired t test. Middle: saliency difference reported by the model grouped according to the subjects report. The number of trials in each group is indicated as well as the p value of a t test testing a difference from zero. Right: contribution of individual feature maps to the total saliency. Solid bars indicate the contribution of each feature to the total saliency for trials on which the subject indicated a rely on that feature, and open bars indicate the contribution on all other trials. Bars show the mean and s.d. across subjects. P values refer to t tests.

performance, we determined a detection threshold for each sound, defined as the least intense level of presentation at which the sound was reliably detected across subjects. Figure 3B displays these detection thresholds and demonstrates a significant correlation of these with the saliency reported by the model (Spearman rank correlation, r = 0.56, p < 0.01). Together, these results strongly demonstrate that the saliency predicted by the model well corresponds to a perceptual level of saliency and that the more salient stimuli better attract our attention and are more frequently detected.

These results, as in the above, cannot be explained solely by sound intensity. First, grouping sounds by peak intensity revealed that the detection performance was not significantly different between the more intense and the less intense group (Fisher's exact test, p = 0.16). Further, the correlation of detection threshold and sound intensity was negligible (r = 0.05, p = 0.52).

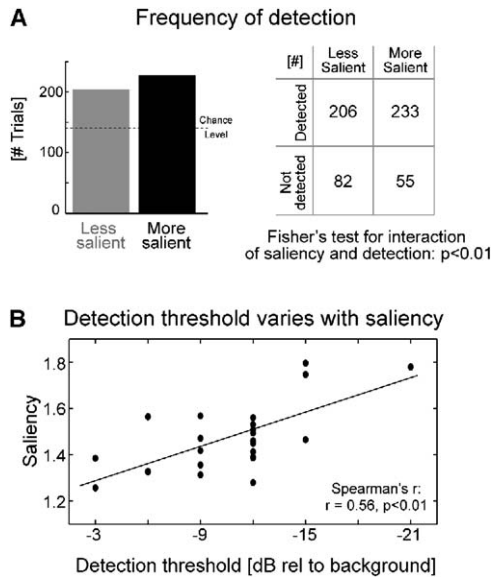

Figure 3. Human Detection Experiment

(A) Frequency of the detection of a sound snippet within ongoing background noise. Sound snippets were grouped according to their saliency as determined from the model, with each group containing half of the sounds. Bars on the left show the total detection frequency, the diagram on the right displays the detection and saliency contingency table. A significant interaction of saliency and detection frequency was determined with Fisher's exact test.

(B) For each sound snippet, a detection threshold was estimated and is indicated as the intensity scaling at which the sound could still be reliably detected; large numbers correspond to a low intensity with respect to the background. The scatter plot demonstrates a significant correlation between detection threshold and sound saliency (Spearman rank correlation).

Monkey Detection Experiment

With human subjects, regardless of instruction, it is difficult to control cognitive aspects of the task imposed. To probe the saliency model in more naive subjects, we performed a similar detection experiment as above with macaque monkeys, exploiting their natural orienting behavior to conspicuous sounds. The animals were exposed to the same background noise as in the human experiments, which were presented from two speakers placed at opposing sides of the animal's head. At irregular intervals, additional sound snippets were presented from one side only, eliciting an orienting behavior toward the source of the sound (invisible speaker). The hypothesis was that more salient stimuli, supposedly those which better attract attention, should lead to a more consistent and, hence, more frequent orienting behavior. We probed a set of six stimuli, three more salient and three less salient, and quantified the frequency of behavioral reaction in a similar way to the human experiment above by using an across-subject design (see Supplemental Materials and Methods).

Figure 4 displays the result of this experiment. In contrast to the human experiment, monkeys exhibited performance at chance level when the less salient stimuli were presented (chi-square test, $\chi^2 = 2.25$, p = 0.13). For the more salient stimuli, however, they oriented toward the source of the sound in the majority of trials ($\chi^2 = 7.1$, p < 0.01), and analysis of the contingency table demonstrated a significant effect of saliency on the detection performance (Fisher's exact test, p < 0.01). Performing the same analysis by using sound intensity instead of saliency to group stimuli yielded only a weak effect of sound intensity on detection performance (Fisher's exact test, p = 0.042). Hence, only the

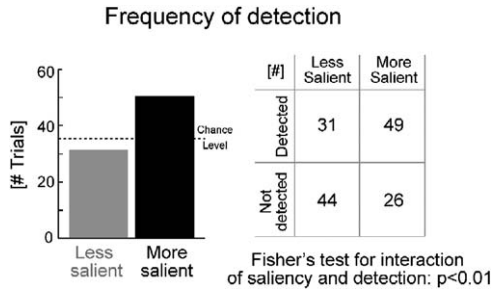

Figure 4. Monkey Detection Experiment

Frequency of the detection of a sound snippet within ongoing background noise by macaque subjects, as indicated by natural orienting movements toward the sound source. Sound snippets were grouped according to their saliency, with each group containing half of the sounds similar to Figure 3.

more salient sounds attracted the animals' attention and led to an overt orienting movement toward the sound source.

Discussion

To master the flood of sensory events, sensory systems use involuntary mechanisms to provide biased representations of the external world [1]. This process is critical for emphasizing behaviorally relevant events and guiding attention to these for more detailed processing. We developed a model for extracting conspicuous events in natural acoustical scenarios based on features important for the analysis of auditory scenes. The auditory saliency map proposed is structurally identical to existing saliency models for the visual system (see the Supplemental Discussion). Hence, our results suggest that the allocation of stimulus-driven attention in different sensory systems involves similar mechanisms.

One purpose of the saliency map is to predict which sensory events attract our attention. The visual model analyzes spatial images and localizes salient features in space so that overt visual attention can be directed toward these; e.g., by virtue of eye movements [2]. The auditory model proposed here analyzes sounds in the time-frequency domain and thus "localizes" salient events in these dimensions. Several properties of audition suggest that these dimensions are important to consider [21–23]: early auditory processing decomposes sounds into their frequencies [24] and attention can be specifically directed to sound frequency and temporal position [21, 22, 25, 26]. Further, spatial location can be encoded as timing or frequency differences between ears, and we can segregate sources even when these appear to come from the same spatial location [27, 28]. Thus auditory feature analysis should prominently rely on feature properties such as sound intensity differences and spectral and temporal contrast that were used in the model. Nevertheless, future versions of an auditory saliency map should explicitly include spatial dimensions. In addition, other possibly more abstract or ecologically relevant features could be incorporated. For example, in vision, letters and nonletters have distinct impacts on our attention as has the emotional impact of stimuli.

In addition to describing properties of sensory stimuli relevant for attentional deployment as well as detection, the saliency map serves as a model of how cortical areas extract these properties from a sensory scene. The cortical substrate of a saliency map is strongly debated. Which areas contribute to such a representation is still an outstanding question and regarding the visual system suggestions range from subcortical structures to association areas in the frontal lobe [29–33]. The present results demonstrate that the mechanisms extracting conspicuous events from a sensory representation are similar in auditory and visual pathways. Therefore, either saliency is extracted by similar mechanisms implemented in both pathways, or saliency for both systems is extracted by the same multimodal cortical areas. Having similar mechanisms extract such events in both pathways could facilitate the integration of saliency maps across sensory systems. Such integration needs to coordinate the reference frames of different sensory systems and could be part of the observed multisensory integration in early sensory areas [34–36]. Alternatively, if visual and auditory saliency maps were extracted by the same multimodal area, one should be able to find evidence for cortical representations of saliency at multimodal sites. Experiments testing this could either involve human fMRI studies by using paradigm similar to those used here or could involve electrophysiological recordings in nonhuman primates. Our finding, that both humans and macaque monkeys seem to follow similar principles for determining stimulus saliency, suggests that these complementary types of experiment should converge to a common cortical substrate for stimulus-driven attention.

Supplemental Data

Supplemental Data include a Supplemental Discussion section, Supplemental Experimental Procedures, and two figures and are available with this article online at http://www.current-biology.com/cgi/content/full/15/21/1943/DC1/.

Acknowledgements

We would like to thank Asif Ghazanfar for inspiring discussions and suggestions on a previous version of this manuscript. This work was supported by the Max Planck Society, the DFG (C.K.; KA-2661/1), and the Alexander von Humboldt foundation (C.P.).

Received: August 10, 2005
Revised: September 9, 2005
Accepted: September 12, 2005
Published: November 7, 2005

References

1. Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. *18*, 193–222.
2. Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. *4*, 219–227.
3. Itti, L., and Koch, C. (2001). Computational modelling of visual attention. Nat. Rev. Neurosci. *2*, 194–203.
4. Simons, D.J., and Levin, D.T. (1997). Failure to detect changes to attended objects. Invest. Opthalmol. Vis. Sci. *38*, 3273–3279.

5. Bergen, J.R., and Julesz, B. (1983). Parallel versus serial processing in rapid pattern discrimination. Nature *303*, 696–698.
6. Treisman, A.M., and Gelade, G. (1980). A feature-integration theory of attention. Cognit. Psychol. *12*, 97–136.
7. Hikosaka, O., Miyauchi, S., and Shimojo, S. (1996). Orienting a spatial attention–its reflexive, compensatory, and voluntary mechanisms. Brain Res. Cogn. Brain Res. *5*, 1–9.
8. Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. *40*, 1489–1506.
9. 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.
10. Bregman, A.S. (1990). Auditory scene analysis (Cambridge: MIT Press.).
11. Yost, W.A. (1992). Auditory Perception and Sound Source Determination. Curr. Dir Psy Sci. *1*, 179–183.
12. Houtgast, T. (1989). Frequency selectivity in amplitude-modulation detection. J. Acoust. Soc. Am. *85*, 1676–1680.
13. 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.
14. Hall, J.W., Haggard, M.P., and Fernandes, M.A. (1984). Detection in noise by spectro-temporal pattern analysis. J. Acoust. Soc. Am. *76*, 50–56.
15. 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.
16. 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.
17. 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.
18. Rauschecker, J.P., Tian, B., and Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. Science *268*, 111–114.
19. Warren, R. (1999). Auditory perception: A new analysis and synthesis (Cambridge: Cambridge University Press).
20. 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.
21. Cusack, R., and Carlyon, R.P. (2003). Perceptual asymmetries in audition. J. Exp. Psychol. Hum. Percept. Perform. *29*, 713–725.
22. Naatanen, R., and Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. Psychol. Bull. *125*, 826–859.
23. Kubovy, M. (1981). Concurrent-pitch segregation and the theory of indispensable attributes. In Perceptual Organization, M. Kubovy and J.R. Pomerantz, eds. (Hillsdale: Erlbaum), pp. 55–98.
24. Moore, B. (1989). An introduction to the psychology of hearing, third edition (London: Academic Press).
25. Mondor, T.A., and Bregman, A.S. (1994). Allocating attention to frequency regions. Percept. Psychophys. *56*, 268–276.
26. Alain, C., and Arnott, S.R. (2000). Selectively attending to auditory objects. Front. Biosci. *5*, D202–D212.
27. Cherry, E. (1953). Some experiments on the recognition of speech with one and with two ears. J. Acoust. Soc. Am. *25*, 975–979.
28. Darwin, C.J., and Hukin, R.W. (1999). Auditory objects of attention: the role of interaural time differences. J. Exp. Psychol. Hum. Percept. Perform. *25*, 617–629.
29. Li, Z. (2002). A saliency map in primary visual cortex. Trends Cogn. Sci. *6*, 9–16.
30. Mazer, J.A., and Gallant, J.L. (2003). Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. Neuron *40*, 1241–1250.
31. Colby, C.L., and Goldberg, M.E. (1999). Space and attention in parietal cortex. Annu. Rev. Neurosci. *22*, 319–349.
32. Robinson, D.L., and Petersen, S.E. (1992). The pulvinar and visual salience. Trends Neurosci. *15*, 127–132.
33. Posner, M.I., and Petersen, S.E. (1990). The attention system of the human brain. Annu. Rev. Neurosci. *13*, 25–42.
34. Macaluso, E., and Driver, J. (2005). Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci. *28*, 264–271.
35. Macaluso, E., Frith, C., and Driver, J. (2000). Selective spatial attention in vision and touch: unimodal and multimodal mechanisms revealed by PET. J. Neurophysiol. *83*, 3062–3075.
36. Schroeder, C.E., and Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. Curr Opin Neurobiol. *15*, 454–458.