Music and natural image processing share a common feature-integration rule

Michelle P. S. To (mpst2@cam.ac.uk)
Department of Physiology, Development and Neuroscience, University of Cambridge,
Downing Street, Cambridge CB3 3EG, United Kingdom

Tom Troscianko (tom.troscianko@bris.ac.uk)
Department of Experimental Psychology, University of Bristol,
12a Priory Road, Bristol BS8 1TU, United Kingdom

David J. Tolhurst (djt12@cam.ac.uk)
Department of Physiology, Development and Neuroscience, University of Cambridge,
Downing Street, Cambridge CB3 3EG, United Kingdom

Abstract

The world is rich in sensory information, and the challenge for any neural sensory system is to piece together the diverse messages from large arrays of feature detectors. In vision and auditory research, there has been speculation about the rules governing combination of signals from different neural channels: e.g. linear (city-block) addition, Euclidian (energy) summation, or a maximum rule. These are all special cases of a more general Minkowski summation rule ($\mathrm{Cue}_1^{m}+\mathrm{Cue}_2^{m})^{1/m}$, where m=1, 2 and infinity respectively. Recently, we reported that Minkowski summation with exponent *m*=2.84 accurately models combination of visual cues in photographs [To et al. (2008). Proc Roy Soc B, 275, 2299]. Here, we ask whether this rule is equally applicable to cue combinations across different auditory dimensions: such as intensity, pitch, timbre and content. We found that in suprathreshold discrimination tasks using musical sequences, a Minkowski summation with exponent close to 3 (*m*=2.95) outperformed city-block, Euclidian or maximum combination rules in describing cue integration across feature dimensions. That the same exponent is found in this music experiment and our previous vision experiments, suggests the possibility of a universal "Minkowski summation Law" in sensory feature integration. We postulate that this particular Minkowski exponent relates to the degree of correlation in activity between different sensory neurons when stimulated by natural stimuli, and could reflect an overall economical and efficient encoding mechanism underlying perceptual integration of features in the natural world.

Keywords: Music; Auditory perception; Feature integration; Minkowski Summation; Visual perception.

Introduction

A compound visual or auditory stimulus is easier to detect than either one of its components (e.g. Robson & Graham, 1981; Green, 1958). Several models have been proposed to describe the rules governing combination of signals from different neural channels (Green, 1958; Livingstone & Hubel, 1987; von der Malsburg, 1995; Treisman, 1998; Ghose & Maunsell, 1999): e.g. linear (city-block) addition, Euclidian (energy) summation, or a maximum rule. Now, Minkowski summation (Eqn.1) is widely used to model how the *detection thresholds* of simple and complex visual stimuli depend on the thresholds for the stimulus components (e.g. Stromeyer & Klein, 1975; Mostafavi & Sakrison, 1976; Quick et al, 1978; Robson & Graham, 1981; Rohaly et al, 1997; Watson & Solomon, 1997; Watson & Ahumada, 2005; Párraga, Troscianko & Tolhurst, 2005; Lovell et al, 2006). This might be a special case of a "universal law" of sensory encoding (Shepard, 1987):-

$$S_c = \left(\sum_{i=1}^{n} S_i^{\,m} \right)^{1/m} \qquad \text{Eqn.1}$$

where S_c is the sensitivity (reciprocal of threshold contrast) for the compound stimulus, S_i is the sensitivity to each component stimulus, n is the number of components and *m* is the summating Minkowski exponent. It should be noted that an exponent of unity is simple linear summation (or 'city-block summation'), an exponent *m*=2 is the Energy summation or Euclidian distance (much favored by auditory scientists), whilst the maximum is given by a high exponent (e.g. Li, 2002; Zhaoping & May, 2007).

Recently, we extended the applicability of the Minkowski Summation rule to the perceptual integration of *suprathreshold* features in colored photographs of *natural* visual scenes (To, Lovell, Troscianko and Tolhurst, 2008). In particular, we studied the perception of the difference between paired images that contain visible and recognizable differences, and asked how the perception of two composite differences (e.g. shape and blur) relates to the perception of single differences (shape or blur separately), see Figure 1. Subjective rating for the double change in a natural image stimulus was most accurately modeled by Minkowski summation of the ratings to the single changes:-

$$\textit{predicted } R3 = \left(R1^{m} + R2^{m} \right)^{1/m} \qquad \text{Eqn.2}$$

where *R1* and *R2* are the ratings for each component image pair, *R3* is the rating for the composite stimulus, and *m*=2.84, a value similar to those reported in grating summation experiments (e.g. Graham, 1977; Robson & Graham, 1981; Watson & Solomon, 1997; Watson & Ahumada, 2005).

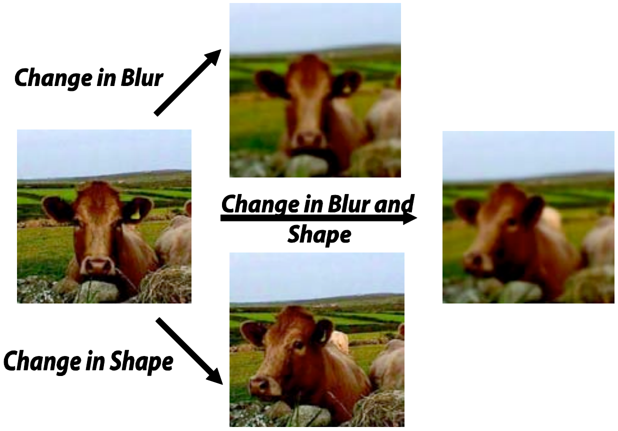

Figure 1: Example of colored image pairs used in our previous vision experiments (To et al., 2008). In this combination set example, image pairs could differ in blur, shape, or both blur and shape.

The purpose of the present study is to determine whether a Minkowski summation rule with exponent $m \approx 3$ is equally applicable to the summation of cues in natural sounds, in particular musical sequences. An Energy-summation model, analogous to a Minkowski summation with exponent 2, has been used to model detection of complex tones (Green, 1958). However, we need to explore whether Minkowski summation with an exponent $m \approx 3$ (as in natural vision, To et al, 2008) might be a closer description of auditory summation. We asked human subjects to discriminate pairs of musical sequences, and to give magnitude estimation ratings for the perceived suprathreshold differences between pairs of stimuli. In addition, we wished to investigate whether a single rule can accurately describe integration across different dimensions: intensity, pitch, timbre and content.

Our experiment differed from previous studies examining integration of auditory features (e.g. Green, 1958; Berg & Green, 1990; Hicks & Buus, 2000) in two ways. First, in contrast to many discrimination studies, the differences presented in this experiment were not only substantially above threshold, but also spanned across a wide range of categories – intensity, pitch, timbre and content (see Methods for examples). This allowed us to investigate how a larger array of cues integrate in more naturalistic stimuli. Second, unlike typical detection and saliency experiments, no thresholds or reaction times were recorded: our subjects were asked to enter magnitude ratings that indicated how they perceive differences between the sound pairs. The present experiment shows that, consistent with our previous findings in vision, a Minkowski summation rule with exponent m=2.95 is most successful in modeling the perceptual feature integration in the processing of musical sequences.

Method

Presentation and Construction of Stimuli

Musical sequences were presented to subjects using a pair of Sennheiser HD 280 pro (64 Ω) headphones. All sounds were played on a DELL laptop XPS M1330 – Window Vista – at level 20 intensity with a sampling rate and bit depth of 44.1 kHz and 16 bit, respectively. The sequences were generated using a free evaluation copy of Notion Demo (Notion Music Software, version 1.5.4.0), a piece of music composition and performance software. Subjects were presented with 160 musical sequence pairs. The sets of stimuli were generated from 16 parent sequences, each matched with 10 variants that differed in one or two of the following dimensions: intensity (by changing the dynamics to pp or ff), timbre (by changing the instrument), pitch (by transposing the sequence upward or downward by various chromatic or diatonic intervals) and/or content (by changing, adding or removing one or more notes). The time signature was the common (4/4) for all the 2 second sequences and the tempo was Vivace – 175 crotchet beats (quarter notes) per minute. Each sequence comprised a single bar (8 eights). The reference sequence was always in the C major key. Examples of sequences and differences are shown in Figure 2A.

The experiment were based around *combination sets*. Starting from one of 16 single reference stimuli, the subjects rated the perceived difference between that stimulus and three others, see Figure 2B. For example, a first pair (component pair) might differ in one dimension such as Intensity, the second pair (a second component pair) might differ in a second dimension such as Pitch (transposition), and the final pair (the composite) would differ in both Intensity and Pitch dimensions. All sound pairs contributed to more than one combination set so that the 160 stimulus pairs made up 96 combination sets.

Participants

The experiment was performed on 15 subjects – 7 female and 8 male. Although some had previously participated in other (visual) rating experiments, they all remained naïve to the purpose of this experiment. Subjects were asked if they were aware of any hearing difficulties they might suffer. Prior to the experiment, they were also presented with many examples of sound stimuli and asked to report any problems with hearing them.

Procedure

Difference ratings were collected for 160 musical sequence pairs from each subject, who was initially instructed during a demonstration session, where they were shown the different types of differences that could be presented to them. A training session then followed the demonstration program. In this phase, subjects were asked to rate 51 musical sound pairs presented in a random order. All sequences used in the demonstration and training phases were different from those to be used in the testing phase proper. During the demonstration and testing phases, subjects were repeatedly presented with the same *standard pair* whose magnitude difference was defined as ‘20’, see Figure 2C. They were instructed that their ratings of the subjective difference between any other test pair should be based on this standard pair: if they perceived the difference between the test pairs to be lesser, equal or greater than the standard pair, their ratings should be less, equal or greater than 20, respectively. They were instructed to use a ratio scale so that, if a given sound pair seemed to have a difference twice as large as that of the reference pair, they would assign a value twice as large to that sound pair (in this case, 40). No upper limit was set so that subjects could rate the differences as highly as they saw fit. Subjects were also told that sometimes sound pairs may be identical and, in such cases, they should set the rating to zero.

The presentation order of musical pairs was randomized differently for each subject. Each block started with the presentation of the standard pair, and this standard was regularly presented after every 10 trials to remind the subjects of the standard difference of ‘20’. The musical sequences lasted 2 seconds each. Because auditory information is processed serially (time-dependent), to avoid the task from being one relying too heavily on memory, subjects were allowed to replay test and standard sequences as often as they liked, before they entered a numerical magnitude rating for that stimulus pair on the computer.

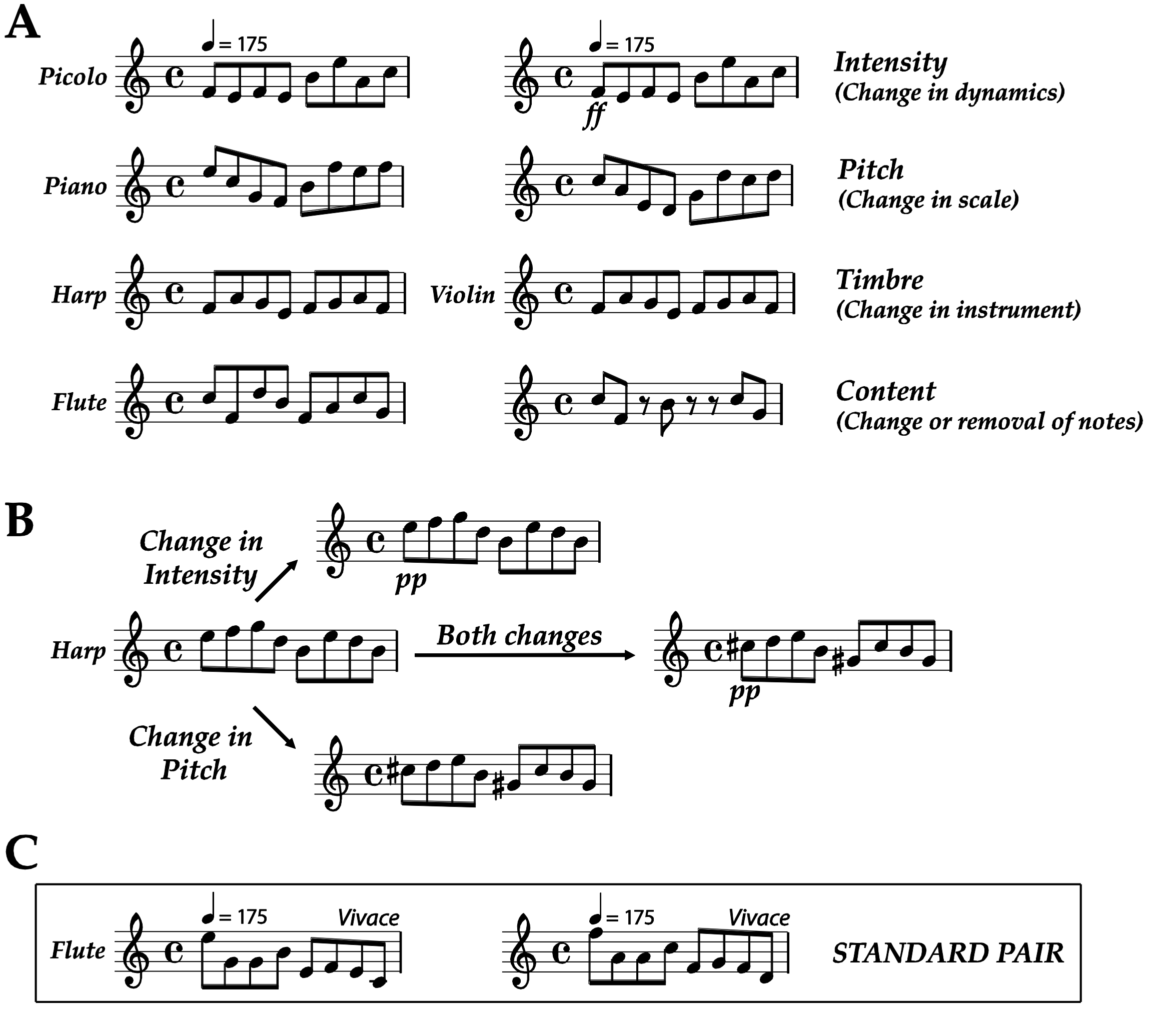

Figure 2: Examples of musical sequences used in the experiment. Panel A shows 4 different sequences (left) changing along four different dimensions: Intensity, Pitch, Timbre and Content. Sequences in the experiment could change along one or two of these dimensions. Panel B presents an example of a combination set: the first pair changes in Intensity, the second changes in Pitch and the third pair changes in both Intensity and Pitch. Panel C shows the specific standard pair used; the difference between these two sequences was defined as having a magnitude of ‘20’.

Results

Fifteen subjects were presented with 160 pairs of musical sequences and were asked to give numerical magnitude estimates for the perceived difference between the stimuli in each pair using a standard pair whose magnitude difference was defined as '20' (Stevens, 1975; Gescheider, 1997, see Figure 2). The robustness of these measures of performance has been assessed in an earlier study [see supplementary materials in To et al. (2008)]. The purpose of this experiment was to determine the performance of the Minkowski Summation with exponent $m\approx3$ model on auditory feature integration and to examine whether a single rule can accurately describe integration across different dimensions.

Sequences could change along one of four dimensions (Intensity, Pitch, Timbre and Content) so that there were 6 different types of dimensional combinations (16 combination sets of each type). The ratings for sequences changing along a single dimension spanned different ranges: Intensity (7.34-24.10, median=17.34), Pitch (19.82-25.18, median=21.29), Timbre (23.72-36.70, median=28.45) and Content (17.02-35.57, median=28.09). We compared the performance of different combination rules – linear summation, Euclidian summation, the Maximum rule and Minkowski summation – in predicting the measured rating ($R3$) to the composite stimulus in each combination set from the separate ratings ($R1$ and $R2$) to its two component sound pairs.

As in our previous studies (To et al., 2008), an iterative search was used to determine the value of the exponent that minimized the sum of squared deviations between the predicted value of $R3$ (Eqn. 2) and the measured value. We found that a Minkowski summation rule with exponent m=2.95 generated the most accurate estimations (see panels A, B, C and D in Figure 3). The correlation coefficients between predicted and measured ratings ranged between 0.85 and 0.87 in all cases. ANOVA revealed that the Minkowski summation model was uniformly efficient in predicting the ratings for all 6 ways of combining any two of the four different dimensions investigated [Intensity, Timbre, Pitch and Content; $F(5,95) = 1.69$, $P = 0.14$] and post hoc Bonferroni analyses found no differences between squared difference between predicted and measured ratings among the different types of combinations.

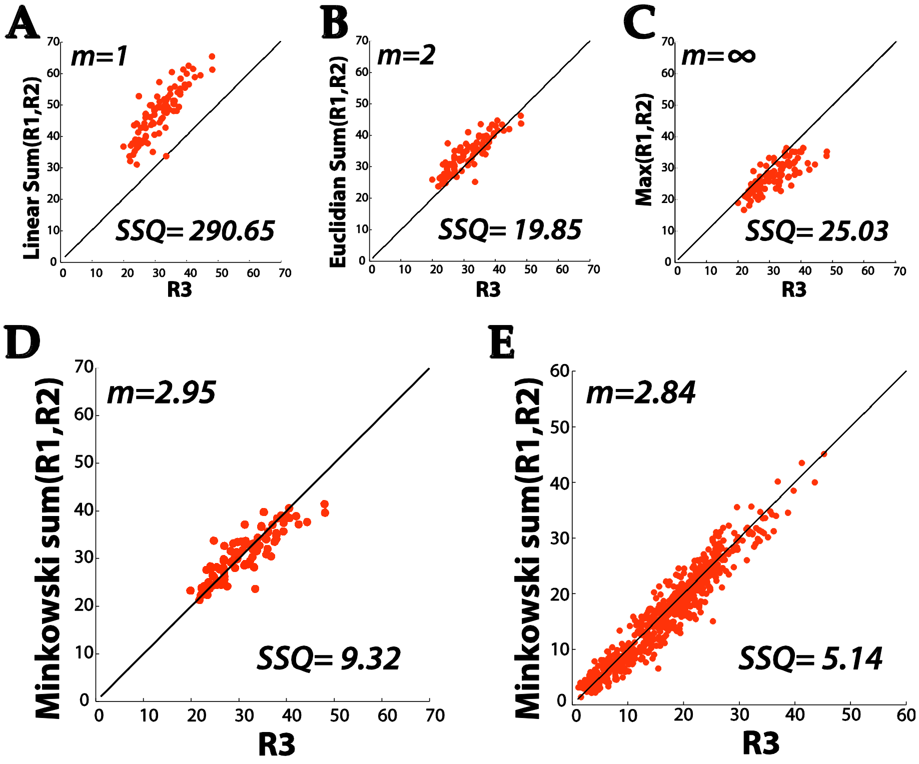

Figure 3: Predictions of the rating ($R3$) given to the composite sound pair in each combination set from the individual ratings ($R1$ and $R2$) to the two separate component sound pairs. In panel A, the linear sum of $R1$ and $R2$ is plotted against the measured $R3$; in panel B, the Euclidian sum (Energy Sum) of $R1$ and $R2$ is plotted against $R3$; in panel C, the maximum of $R1$ and $R2$ is plotted against $R3$; in panel D, the Minkowski sum with exponent m=2.95 of $R1$ and $R2$ is plotted against $R3$. For comparison, the results from our previous vision experiments [encompassing 704 combination sets; To et al. (2008)] are presented in panel E. Lines of equality are shown.

Discussion

The main finding of this study is that the Minkowski Summation rule for cue combination with exponent $m\approx3$ can be extended from vision to audition, its accuracy is consistent for feature integration across different naturally-occurring stimulus dimensions. We have found that Minkowski summation with exponent $m= 2.95$ outperforms city-block, Euclidian and maximum combination rules in describing auditory cue integration across feature dimensions. A similar exponent (m=2.84) was previously reported for visual cue integration in natural scenes (To et al., 2008). For comparison, we have plotted the results from our previous vision experiments (encompassing 704 visual combination sets) in Figure 3E. That the same exponent was found across different dimensions and modalities, suggests the possibility of a universal "Minkowski Summation Law" underlying perceptual integration of features in the natural world.

A long line of research has demonstrated the applicability of the Minkowski Summation rule in the integration of visual information. Hearing experiments have previously studied feature integration in complex auditory sequences (e.g. Melara & Marks, 1990), but Euclidian (energy) summation has been the model of choice to describe auditory cue combination. Indeed, our results (Figure 2B) almost support such a summation model (m=2). However, the results from our music experiment suggest that the Minkowski Summation with exponent $m\approx3$ may just be a superior model. That the integration of auditory features in musical sequences follows the same rule as the integration of visual features in natural scenes supports Shepard's theory of a Universal Law for cue combination (Shepard, 1987).

Origin of 3

Shepard (1987) postulated that summation of cues in simple stimuli might either follow city-block (m=1) or energy (m=2) summation models Why should a slightly higher exponent actually be found. We postulate that the exponent $m\approx3$ relates to the degree of correlation in activity between different sensory neurons when stimulated by natural stimuli: city-block (linear) or Euclidian (energy) summation can be argued to be appropriate if activity is independent, since each neuron conveys a uniquely important signal. On the other hand, if responses were highly correlated, the information given by only one neuron would be sufficient (the maximum rule). If the responses of sensory neurons or channels showed small correlations in their responses to natural stimuli, the most appropriate summating exponent would be slightly greater than expected if cues were coded entirely independently. Yen, Baker and Gray (2007) have recently shown that, when a cat was presented with natural stimuli (movie clips), the signal correlation of neighboring V1 neurons was relatively low but it was greater than zero (r=0.21±0.23 and 0.18±0.20, for neurons recorded using the same or different electrode respectively). This small degree of correlation between actual neuronal responses implies that the "universal" value of the Minkowski summation exponent should be a little greater than suggested by Shepard, but still a lot lower than infinity (maximum rule). Since this degree of correlation is likely to be shaped by the natural statistics of the world, we suspect that this reflects an overall economical and efficient encoding mechanism underlying perceptual integration of features in the natural world (Field, 1994; Laughlin, de-Ruyter-van-Steveninck & Anderson, 1998; Nirenberg, Carcieri, Jacobs & Latham, 2001; Barlow, 2001; Lewicki, 2002).

Future directions

The present findings have raised two questions. First, apart from vision and audition, might the Minkowski Summation rule with exponent m≈3 also apply to feature integration in other modalities? We are currently examining analogous feature integration in the sense of touch. Meredith and Stein (1993) have demonstrated the role of the superior colliculus in the integration of visual, auditory, somatosensory and nociceptive information. In addition, Blakemore (2008) has recently suggested the possibility that normal sensory integration might rely on feedback to early sensory areas from polysensory regions of the cortex, in particular the parietal cortex. These areas are known to be involved in cross-modal integration, but what about within modality integration? If these areas also process integration within individual modalities, then this could explain why feature integration in the visual and auditory systems follow the same Minkowski Summation rule. Assuming that perceptual integration reflects an efficient encoding mechanism shaped by the statistics of natural stimuli and that polysensory regions are involved in the integration of information of all modalities, then feature integration in different modalities, such as touch, might follow the same rule as in vision and audition.

Second, having demonstrated the applicability of the same Minkowski summation rule to visual and auditory stimuli separately, the next step is to study perceptual integration of cues across these two modalities. At present, Bayesian systems are commonly used to model cross-modal integration (e.g. Ernst & Banks, 2002 and Battaglia, Jacobs & Aslin, 2003), however the simplicity of the Minkowski Summation rule could provide an attractive alternative. We are presently investigating this possibility by performing suprathreshold discrimination experiments that present observers with changing auditory and visual stimuli concurrently.

Acknowledgments

This research was supported by EPSRC/Dstl research grants (S56399, S56405) to DJT and TT, under the Joint Grant Scheme. MPST was employed by those grants. We are very grateful for helpful discussions with Dr. R.J. Baddeley, Dr M. Gilmore and Dr. I. Moorhead. We would also like to thank the reviewers for their helpful comments.

References

Barlow, H.B. (2001). Redundancy reduction revisited. *Network*, 12, 241-253.

Battaglia, P.W., Jacobs, R.A. & Aslin, R.N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America*, 20, 1391–1397.

Berg, B.G. & Green, D.M. (1990). Spectral weights in profile listening. *The Journal of the Acoustical Society of America*, 88, 758-766.

Blakemore, C. (2008). Interaction between cortical areas: lessons from synaesthesia. Perception, 37, 169.

Ernst, M.O. & Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415, 429–433.

Field, D.J. (1994). What is the goal of sensory coding? *Neural Computation*, 6, 559–601.

Gescheider, G.A. (1997). *Psychophysics – The Fundamentals*. USA: Lawrence Erlbaum Associates.

Ghose, G.M. & Maunsell, J. (1999). Specialized representations in visual cortex: A role for binding? *Neuron*, 24, 79-85l.

Graham, N.V. (1977). Visual detection of aperiodic spatial stimuli by probability summation among narrowband channels. *Vision Research*, 17, 637-652.

Green, D.M. (1958). Detection of multiple component signals in noise. *The Journal of the Acoustical Society of America*, 30, 904-911.

Hicks, M.L. & Buus, S. (2000). Efficient across-frequency integration: Evidence from psychometric functions. *The Journal of the Acoustical Society of America*, 107, 3333-3342.

Laughlin, S.B., de-Ruyter-van-Steveninck, R. & Anderson, J.C. (1998). The metabolic cost of information. *Nature Neuroscience*, 1, 36–41.

Lewicki, M. S. (2002). Efficient coding of time-varying patterns using a spiking population code. In R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki (Eds), *Probabilistic Models of the Brain: Perception and Neural Function*.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6, 9-16.

Livingstone M.S. & Hubel D.H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7, 3416-3468.

Lovell, P.G., Párraga, C.A., Ripamonti, C., Troscianko, T. & Tolhurst, D.J. (2006). Evaluation of a multi-scale color model for visual difference prediction. *ACM Transactions on Applied Perception*, 3, 155-178.

Melara, R.D. & Marks, L.E. (1990). Interaction among auditory dimensions: timbre, pitch, and loudness. *Perception and Psychophysics*, 48, 169-78.

Mostafavi, H. & Sakrison, D.J. (1976). Structure and properties of a single channel in human visual system. *Vision Research*, 16, 957-968.

Nirenberg, S. , Carcieri, S.M., Jacobs, A.L. & Latham, P.E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, 411, 698-701.

Párraga, C.A., Troscianko, T. & Tolhurst, D.J. (2005). The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model. *Vision Research*, 45, 3145-3168.

Quick, R.F. (1974). A vector magnitude model of contrast detection. *Kybernetik*, 16, 65-67.

Robson, J.G. and Graham, N.V. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, 21, 409-418.

Rohaly, A.M., Ahumada, A.J. & Watson, A.B. Object detection in natural backgrounds predicted by discrimination performance and models. *Vision Research*, 37, 3225-3235 (1997).

Schafer, T.H. & Gales, R.S. Auditory masking of multiple tones by random noise. *The Journal of the Acoustical Society of America*, 21, 392-398 (1949).

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.

Stein, B.E. & Meredith, M.A. (1994). *The Merging of the Senses*. Cambridge MA: MIT Press.

Stevens, S.S. (1975). *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: Wiley.

Stromeyer, C.F. & Klein, S. (1975). Evidence against narrow-band spatial frequency channels in human vision: detectability of frequency modulated gratings. *Vision Research*, 15, 899-910.

To, M., Lovell, P.G., Troscianko, T. & Tolhurst, D.J. (2008). Summation of perceptual cues in natural visual scenes. *Proceedings of the Royal Society Series B*, 275, 2299-2308.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London Series B – Biological Sciences*, 353, 1295-1306.

von der Malsburg, C. Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5, 520-526 (1995).

Watson, A.B. & Ahumada, A.J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision*, 5, 717-740.

Watson, A.B. & Solomon, J.A. (1997). Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A*, 14, 2379-2391.

Yen, S.C., Baker, J. & Gray, C. M. (2007). Heterogeneity in the Responses of Adjacent Neurons to Natural Stimuli in Cat Striate Cortex. *Journal of Neurophysiology*, 97, 1326-1341

Zhaoping, L. & May, K.A. (2007). Psychophysical tests of the hypothesis of a bottom-up saliency map in the primary visual cortex. *PLoS Computational Biology*, 3, e62.