# Acoustic cues to visual detection: A classification image study

**David Pascucci**
Department of Psychology, University of Florence, Florence, Italy, &
Center for Mind/Brain Sciences, Department of Cognitive Sciences and Education, University of Trento, Trento, Italy

**Nicola Megna**
INO-CNR, Florence, Italy

**Michela Panichi**
Department of Psychology, University of Florence, Florence, Italy, &
Institute of Neuroscience, CNR, Pisa, Italy

**Stefano Baldassi**
Department of Psychology, University of Florence, Florence, Italy

A non-informative sound is known to improve contrast detection thresholds for a synchronous visual target (M. Lippert, N. K. Logothetis, & C. Kayser, 2007). We investigated the spatio-temporal characteristics of the mechanisms underlying this crossmodal effect by using a classification image paradigm specifically suited to investigate perceptual templates across both space and time (P. Neri & D. J. Heeger, 2002). A bright bar was embedded in 2D (space–time) dynamic noise and observers were asked to detect its presence in both unimodal (only visual) and bimodal (audio–visual) conditions. Classification image analysis was performed and the 1st and 2nd order kernels were derived. Our results show that the cross-modal facilitation of detection consists in a reduction of activity of the early mechanisms elicited by the onset of the stimulation and not directly involved in the identification of the target. In fact, the sound sharpens the 2nd order kernels (involved in target detection) by suppressing the activation preceding the target, whereas it does not influence the 1st order kernels. These data suggest that the sound affects some non-linear process involved with the detection of a visual stimulus by, decreasing the activity of contrast energy filters temporally uncorrelated with the target, hence reducing temporal uncertainty.

Keywords: classification images, cross-modal perception, visual detection, reverse correlation, multisensory, contrast threshold

Citation: Pascucci, D., Megna, N., Panichi, M., & Baldassi, S. (2011). Acoustic cues to visual detection: A classification image study. *Journal of Vision, 11*(6):7, 1–11, http://www.journalofvision.org/content/11/6/7, doi:10.1167/11.6.7.

## Introduction

Events in the real world constitute an overwhelming source of sensory signals, thus the ability to flexibly integrate or combine different sources of information plays a fundamental role in our perception. The integration of acoustic and visual information is one of the most important issues in the cross-modal studies of perception and attention (Burr & Alais, 2006; Driver & Spence, 2004; Ernst & Bulthoff, 2004; Vroomen & de Gelder, 2000).

Cross-modal stimulation affects performance in visual detection and spatial discrimination (Driver & Spence, 2004; McDonald, Teder-Salejarvi, & Hillyard, 2000) as well as in covert attention tasks (Driver & Spence, 1998; McDonald & Ward, 2000) and may generate misrepresentations of some visual stimuli features, leading to perceptual illusion (Alais & Burr, 2003; McGurk & MacDonald, 1976; Shams, Kamitani, & Shimojo, 2000, 2002).

Moreover it has been shown that a sound presented in synchrony with a visual stimulus in tasks requiring the detection of visual targets enhances the sensitivity to specific visual features like contrast (Lippert, Logothetis, & Kayser, 2007), intensity (Stein, London, Wilkonson, & Price, 1996), and pattern configuration (Vroomen & de Gelder, 2000).

Some recent contributions have focused on the level at which the audio–visual interaction occurs (Mishra, Martinez, Sejnowski, & Hillyard, 2007; Shams et al., 2002; Wallace, Carriere, Perrault, Vaughan, & Stein, 2006). In particular, Lippert et al. (2007), using vertical Gabor gratings at variable contrast, compared the effect of a synchronous sound presented alone ("sound informative" condition) or combined with a visual cue (a gray

doi: 10.1167/11.6.7 Received September 29, 2010; published May 11, 2011 ISSN 1534-7362 

frame surrounding the target, "sound uninformative" condition) in a contrast detection task. They found that the cross-modal facilitation of visual contrast detection disappeared when the sound was redundant with the visual display. The authors interpreted this finding in terms of a cognitive, high level interaction, ruling out the possibility of a low level interaction as suggested by previous studies (Marks, Ben-Artzi, & Lakatos, 2003; Odgaard, Arieh, & Marks, 2003).

On the other hand, Mishra et al. (2007) and Shams et al. (2002) interpreted the robustness of the "sound induced flash illusion", consisting in the perception of multiple flashes when a single flash is presented with multiple beeps, as an evidence of the action of a mainstream circuitry, providing an interpretation in favor of a low-level neural integration of audio–visual signals, which is supported by the multisensory activation observed in both visual and auditory primary cortices (Kayser & Logothetis, 2007; Martuzzi et al., 2007) and by the presence of multisensory neurons in the superior colliculus (Stein, Meredith, & Wallace, 1993; Stein, Stanford, Ramachandran, Perrault, & Rowland, 2009).

The aim of the present study is to probe the mechanisms of acoustic facilitation of visual detection through the use of the Classification Images technique (Ahumada, 2002; Ahumada & Lovell, 1971), which is also referred to as Psychophysical Reverse Correlation. This method is based on the analysis of the visual noise characteristics leading to specific observers' responses and has been very useful in revealing the characteristics of the perceptual templates exploited by an observer in visual tasks such as Vernier acuity (Beard & Ahumada, 1999), disparity discrimination (Neri, Parker, & Blakemore, 1999), illusory-contour perception (Gold, Murray, Bennett, & Sekuler, 2000), orientation discrimination (Solomon, 2002) as well as spatially cued detection (Eckstein, Shimozaki, & Abbey, 2002). In particular, a recent spatio-temporal version of this technique (Neri & Heeger, 2002) has provided a powerful tool to probe behaviorally the characteristics of the mechanisms involved in visual detection and discrimination across both space and time. Using spatio-temporally modulated white noise and classification images analysis of both the 1st and 2nd order kernels, constituted by the mean and variance template respectively, the authors were able to dissociate two processing stages: an early 'detection' stage, in which initial and strong noise variations are able to engage automatic and exogenous mechanisms of attentional capture, and a later 'identification' stage that follows detection by about 100 ms and is characterized by the use of image intensities to identify the luminance polarity of the signal (a bright or dark bar).

In the present study, we use this paradigm during a visual detection task in a Unimodal (only-visual) and a Bimodal (audio–visual) condition, in order to investigate the nature of the interaction between the auditory and the visual system in response to cross-modal stimulation. We hypothesized that, if the facilitation of detection induced by a sound synchronous to the signal depends on the same low-level mechanisms of visual detection *per se,* then the improvement should be reflected on the pattern of activation of the 2nd order kernels. The results confirmed our predictions showing that the effect of the sound is reflected by the non-linear stage probed by the noise variance, providing novel insights to explain how a sound interacts with a visual stimulus to make it more detectable.

# Methods

## Stimuli and procedure

Visual stimuli (Figure 1) were generated on MatLab, using a CRS VSG 2.5 graphic card, and presented on a gamma calibrated CRT monitor (Barco Calibrator) with a mean luminance of 35 cd/m$^2$ at a frame rate of 100 Hz.

Each trial consisted in the rapid presentation of 9 frames of unidimensional noise centered at fixation, each containing 11 bars of random luminance displayed for 28 ms (35 Hz), hence the stimulus lasted 252 ms. Each bar was $0.1° \times 1.1°$ and its luminance was randomly determined from a uniform discrete distribution of $35 \pm 4$ cd/m$^2$.

The target was a bright vertical bar whose exact luminance was set at different values to match the accuracy criterion of 75% for the two conditions (Unimodal and Bimodal) and was added on the central bar of the fifth frame in 50% of the trials. The entire stimulus could be described as a $9 \times 11$ bidimensional spatio-temporal matrix containing all the luminance values of each bar in time and space.

In the Unimodal condition, subjects had to detect the presence/absence of the target on any given trial by pressing one of two keys (see Experiment 1 in Neri & Heeger, 2002). In the Bimodal condition, an acoustic cue (1000 Hz square wave played for 28 ms at 70 db of intensity) was presented contemporaneously with the target in both the 'target present' and the 'target absent' trials. Observes were informed both about the target-sound synchrony and the non-predictability of the target appearance on the base of the sound presence on any given trial.

Audio–visual synchrony was ensured by using an analog to digital converter (ADC) interfaced with the VSG graphic card. A red fixation cross of $0.1° \times 0.1°$ remained visible throughout the experimental session. The inter-trial interval was variable within a 200 to 500 ms range.

The Unimodal condition was identical to the first experiment of Neri and Heeger (2002), with a bright target only.

The Bimodal condition differed only for the presence of the sound, which was presented in each trial including

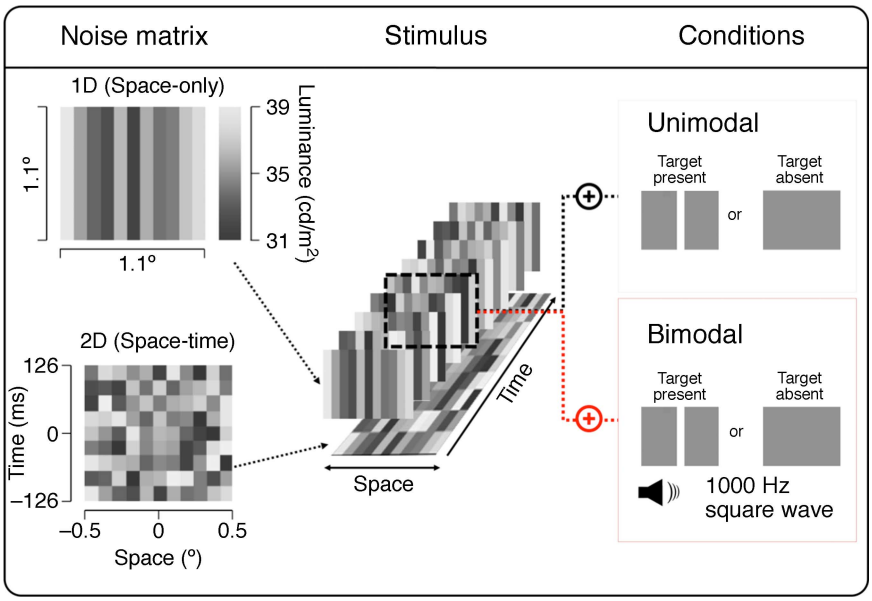


Figure 1. Stimuli and conditions. Each frame (28 ms) contained a "barcode" noise matrix with dimension 1.1° × 1.1° of visual angle (top left image). The succession in time of nine frames generated spatio-temporal dynamic noise lasting 252 ms (bottom left and central images); the target (a bright bar in the central position) was an increment of luminance of the central bar of the fifth frame of the stimulus and, in the bimodal condition, a sound (1000 Hz, 28 ms) was played in physical synchrony with the target frame on both target absent and target present trials (right images). The entire stimulus can be described as a 2-Dimensional (space (x) and time (t)) matrix containing the seeds generator of the random gray levels that each bar assumed in time.

those in which the target was absent. Observers were informed that, in the Bimodal condition, the beep was perfectly synchronous with the frame potentially containing the visual stimulus and that its presence was uninformative in predicting the actual presence or absence of the target. Noise free 'reminder' trials containing the target without noise were shown every 20 trials, in both conditions, to recall the representation of the signal; they were not included in the analysis.

## Observers

Two authors (DP & MP) and two naïve observer (DO & GC) participated to the experiment, all of them with normal or corrected-to-normal vision. They collected 4000 trials per conditions by alternating Bimodal and Unimodal stimulation every block, each lasting 100 trials.

# Reverse correlation data analysis

Depending on the observers' responses, we classified the noise matrix of each trial as a Hit or a False Alarm when a signal present answer followed a 'noise plus signal' or a 'noise only' image, respectively. Noise matrices classified as Correct Rejections or Misses followed signal absent responses in the absence or the presence of the signal superimposed on the noise, respectively.

The noise distributions of each response class were analyzed separately by calculating their mean (1st order statistics) and variance (2nd order statistics) and then combined to compute the perceptual templates by using the following formulae:

$$MeanCL = \mu_{S(1),R(1)} + \mu_{S(0),R(1)} - \mu_{S(1),R(0)} - \mu_{S(0),R(0)}, \quad (1)$$

$$VarianceCL = \sigma^2_{S(1),R(1)} + \sigma^2_{S(0),R(1)} - \sigma^2_{S(1),R(0)} - \sigma^2_{S(0),R(0)}, \quad (2)$$

Where $\mu$ and $\sigma^2$ are the mean and variance of the spatio-temporal matrices of each category, respectively, defined by $S$, that is the target condition (0 = absent, 1 = present), and $R$, that is the subject's answer (0 = absent, 1 = present). By summing the noise information leading to a "Signal Present" response and subtracting it to the noise samples leading to "Signal Absent" response, we computed the templates representing the noise pattern that led the observer to a "yes" response. We analyzed mean and

variance of these templates (hereafter named 1st and 2nd order kernel, respectively) in order to probe the stage where the facilitation introduced by the sound occurs and how this is accomplished.

The reliability of the kernels and the derived measures was tested with a Bootstrap procedure (N = 2000) in which at any bootstrap sample we measured a template derived from a subset of noise matrices. In particular, the procedure created small samples of 800 noise matrices randomly selected (with replacement) out of the 4000 available for each subject, with an internal distribution per response category (Hits, Correct Rejections, Misses and False Alarms) that matched the empiric distributions of our data set. 1st and 2nd order kernels were calculated and a new iteration started. Each pixel of the final 1st and 2nd order templates represent the average across the 2000 bootstrap samples and its value is set to 0 (mid-gray in Figures 3 and 4) when a one sample *t*-test comparing the pixel values to 0 was not significant based on the criterion of $\alpha$ 4 0.01. Therefore, the templates shown in the figures represent only significant activations. This is a conservative procedure that leads to more solid results by relying on small samples, i.e. on lower Signal-to-Noise Ratios.

The final templates for Unimodal and Bimodal conditions (Figures 3 and 4) are plotted in Z scores (as previously done by other authors, e.g., Neri & Heeger, 2002). The use of Z units and the consistency of the data across observers legitimated us to pool the noise matrices of the four observers in order to work out the templates for a 'Super Subject' (elsewhere named 'aggregate observer', e.g. Neri, 2009). We reasoned that in the presence of relatively consistent data across individual subjects, the Super Subject would led to a cleaner general representation of the mechanism probed by our task than any other central measure of tendency to the processed data.

# Results

## Detection thresholds in noise

We first measured visual detection thresholds in Unimodal and Bimodal conditions in order to confirm with our noisy stimuli the basic effect of improvement of visual sensitivity in the presence of synchronous uninformative sounds. We used a Yes/No procedure and a stimulus set that matches the main Classification Image experiment, with the only difference that the target intensity was varied according to the adaptive procedure QUEST (Watson & Pelli, 1983) whose parameters were set to obtain an entire psychometric function. At each trial, the stimulus was a spatio-temporal matrix identical to those described in the Methods section (see Figure 1), and observers were asked to report the presence or absence of the target, a bright bar added in the spatio-temporal center of one half of the noise patches presented, in random order. Figure 2 reports the psychometric functions of three observers in the two conditions (filled red, Bimodal; empty black, Unimodal) fitted by a cumulative Gaussian function (dashed red, Bimodal; straight black, Unimodal) and the resulting thresholds, marked by the two arrows along the abscissae, for the Bimodal and the Unimodal conditions.

In the Bimodal condition, contrast thresholds for both subjects improve by about one octave for all observers

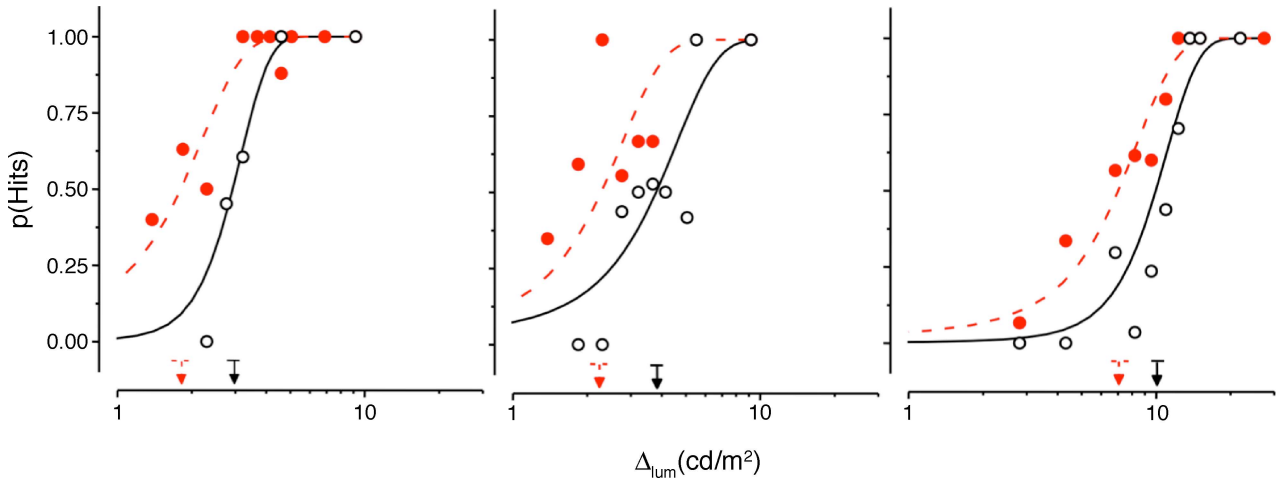


Figure 2. Psychometric functions of the YES/NO detection task for the Unimodal (black empty symbols and straight line) and Bimodal (red filled symbols and dashed line) conditions. Thresholds, marked as arrows in the abscissae, were obtained with a bootstrap procedure that recalculated and refitted the cumulative Gaussian 200 times, giving raise to the confidence intervals defined by the arrows' caps.

and, importantly, the entire psychometric function is shifted to the left in the presence of a sound. This confirms that the reported sound-induced facilitation (Lippert et al., 2007) occurs not only within the spatio-temporal dynamic noise paradigm used in our experiment, but also while using a stimulus that effectively must be detected against a pedestal (the noise).

## Classification image analysis: First order kernels

By applying the formula Equation 1 to the noise matrices stored according to the response classes reported in the Methods section, we calculated the first order kernels for the detection task and reported the results as linearly interpolated, Gaussian-filtered ($\sigma$ = 0.2 noise pixels) data in Figure 3. The matrices reported to the left side of the figure represent the 1st order "information" (the distribution of luminance in the spatio-temporal matrix) leading to a "yes" response in the detection task. Each line of plots reports the data of individual observers, with the Super Subject at the bottom (see Methods). The two columns of graphs reported to the right-hand side of Figure 3 show a bi-dimensional representation of the kernel activation in space (left), obtained averaging across time, and in time (right), obtained considering only the central position of space rather than by averaging as it is the location providing most information about the temporal pattern of the kernels. In all cases the Unimodal and the Bimodal condition do not differ significantly from each other, showing a peak of activation in correspondence of the actual spatio-temporal location of the signal (i.e. the center of the matrices on the left of Figure 3 and the middle of the 2d graphs on its right). The temporal dimension reveals a positive activation preceding and following the physical appearance of the signal, revealing a form of temporal blur around the target, which is coherent with the well known impulse response function properties of neuron involved in detection tasks (e.g., Watson & Nachmias, 1977). The spatial dimension show a well localized peak of positive activation at the location of the stimulus and two lobes of negative activation (i.e. activation in anti-correlation) at its flanks, confirming what found by Neri and Heeger (2002).These results suggest that the improvement observed adding a sound as a temporal cue cannot be explained by modifications of the 1st order template: the patterns of noise leading to a yes response in the Unimodal and in the Bimodal condition share the same spatial and temporal features.

## Classification image analysis: Second order kernels

By applying the formula Equation 2 to the noise matrices classified according to the observers' responses we calculated the 2nd order, or variance, kernels and reported the results in Figure 4. The structure of the figure matches that of Figure 3. These matrices (one for each observer and condition) represent the 2nd order "information" leading to an increase of probability of signal present responses in the detection task. According to Neri and Heeger (2002) the information carried by the 2nd order kernel represents the luminance variability of the noise bars against the mean luminance of the screen, which can be considered the contrast energy of the noise. In other words, the variance kernel informs us on the structure of the variability of noise associated to the detection of the target.

As explained in Methods section, the data plotted in Figure 4 report only the significant activation as calculated by a *t*-test to the bootstrapped data (see Methods). The variance kernels plotted in the color graphs on the left are relatively less structured in space and time than the mean kernels, but separating the spatial and the temporal dimension reveal interesting differences between the two experimental conditions, especially in the temporal dimension.

More specifically, the kernels for the Unimodal condition share the major positive peak of activation being placed very close to the spatiotemporal position of the target in spite of a slight anticipation. Data show a trend for negative activation at locations adjacent to the target, that is also visible from the 2d plot reporting the activation in space (third column of plots from left). The profile of the 2nd order negative activity is much noisier and less spatially localized than in the 1st order kernel and, because variance is by definition positive, it implies higher variance associated to 'Signal Absent' than to 'Signal Present' responses in the bluish locations rather than simple anti-correlation of the relevant pixels' luminance polarity. However, as with the 1st order kernel analysis, the spatial profile of the 2nd order kernels does not reveal any substantial difference between the Unimodal and the Bimodal condition.

The temporal profile shows instead consistent differences of activation between conditions. The rightmost column of plots of Figure 4 plots a family of curves resulting from the linear interpolation of the spatial slice containing the target for the Unimodal (black) and the Bimodal (red) condition. These data reveal a specific, consistent trend that resists to the inter-observer variability visible in the data and is confirmed by the Super Subject analysis. First, detection without a sound is characterized by a more sustained pattern of activation of the variance kernel at and around the time of appearance of the target (gray vertical band), with weaker and more scattered peaks of activation under bimodal stimulation. In the Unimodal condition in fact, the 2nd order activation ramps to form a relatively broad pole of positive activation that peaks for all observers just before the target frame and fades out at the time of the target for all observers except DP (for whom there is still a decay,

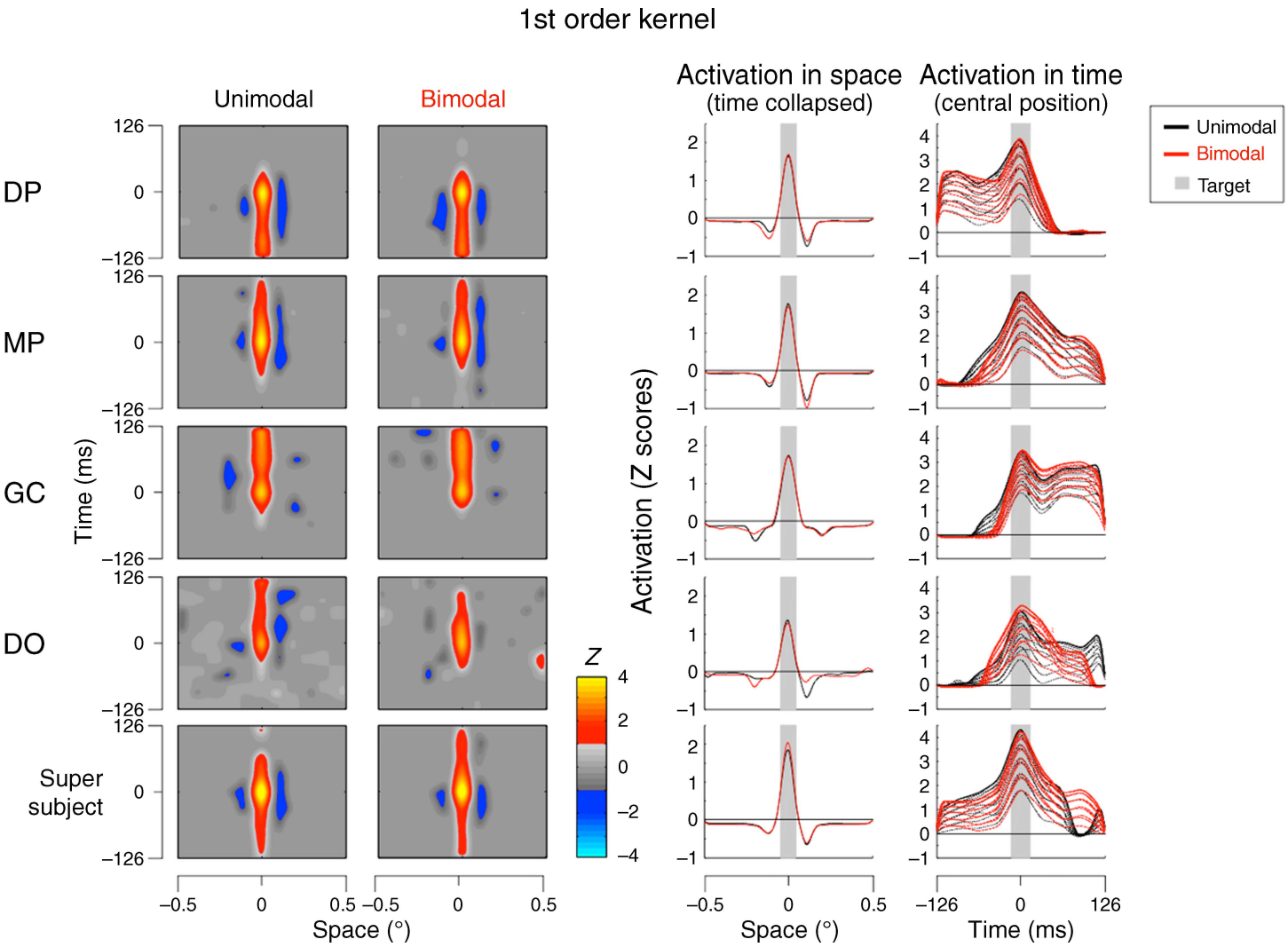


Figure 3. First order kernels analysis for the four observers and the Super Subject, shown in the bottom line of plots. The two columns of graphs to the left-hand side map the pattern of activation-inhibition of the 1st order, mean kernel. The leftmost panels report the kernels of the Unimodal condition, whereas the column to their right are the kernels for the Bimodal condition. The colors are linearly interpolated Z scores of the calculated templates. Positive and negative activations are highlighted by the red and blue side of the colormap scale reported in the legend, respectively. The two columns of graphs to the right-hand side summarize the spatial and temporal activation of the kernels. The plots to the left reports the spatial profiles of the templates, obtained averaging across time the 1st order kernels for the Unimodal (black line) and the Bimodal (red line) conditions; the central, vertical band represents the spatial position of the target. The rightmost plots of the figure represents the 1st order kernel as a function of time in the central spatial position, with the different traces representing different linearly interpolated points of space. The kernels show a well localized spatial profile of activation with two lobes of inhibition at its sides, featuring the typical shape of the early filters for detecting lines. The temporal profile is slightly different for each observer, but they all show a clear peak of activation at the time of the target.

but weaker). In the Bimodal condition the overall 2nd order activity is lower and follows a different temporal pattern that is mainly characterized by a strong depression of activation before and during the presentation of the target. This is true for all observers except for DP, who shows anyway a statistically significant reduction of the Bimodal kernel activation at the target time. Another significant difference between the two conditions is the burst of activity shown by all subjects at the beginning of the stimulus (or slightly delayed for observer DO) and a clear return to high variance after the stimulus frame. In other words, the Bimodal stimulation introduces a depression of the noise variance leading to positive responses and changes the overall temporal pattern of the 2nd order activation, with high variance at the beginning and at the end of the temporal array of noise in our stimulus.

In order to provide a more meaningful representation of the effect of an uninformative sound on the temporal activation of the 2nd order kernel, which seems to explain

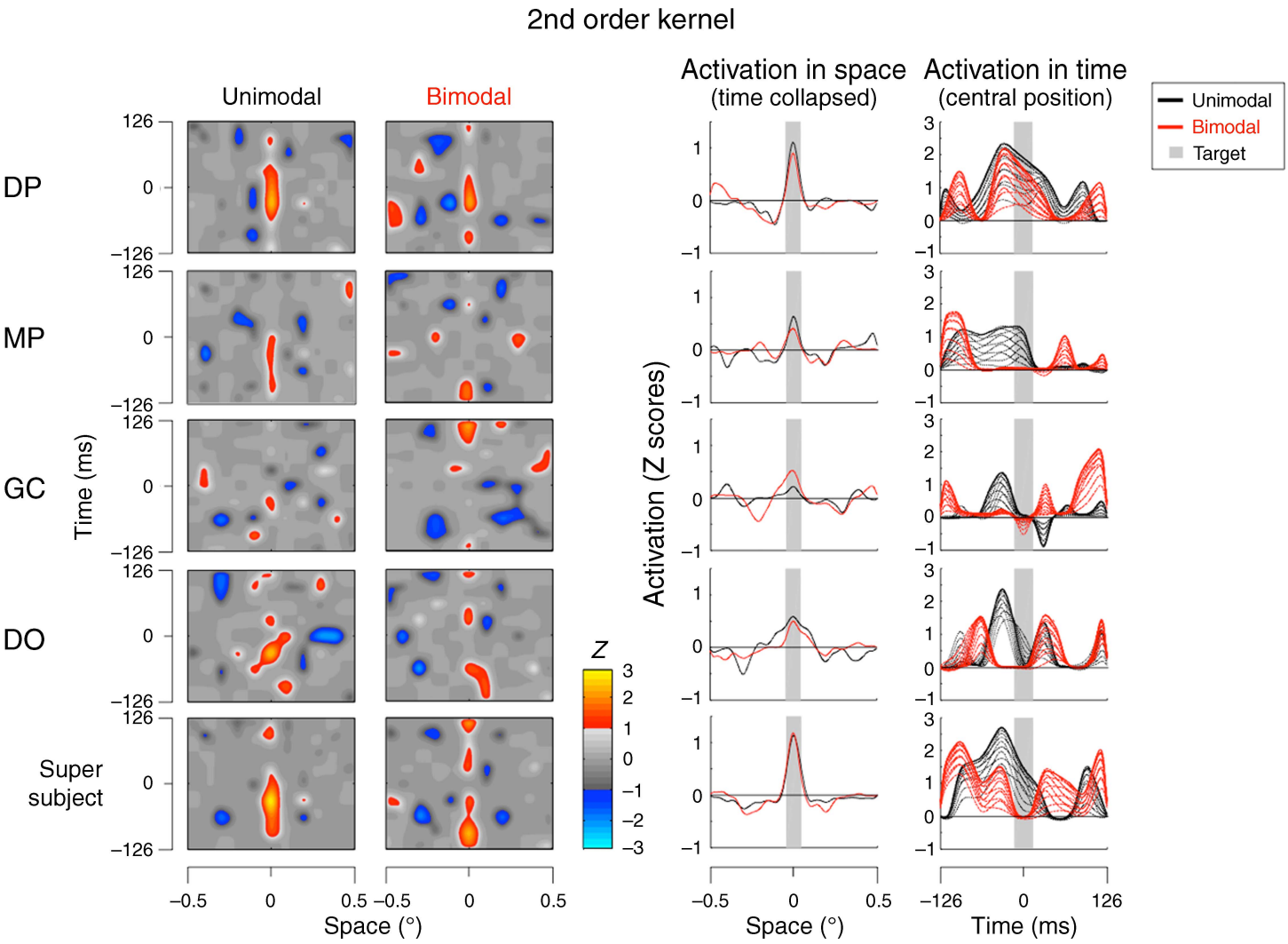


Figure 4. Second order kernels analysis for the four observers and the Super Subject, shown in the bottom line of plots. Arrangement of panels, colors and conventions follow the organization of Figure 3. The variance kernels show a generally localized activation at the site of the target (i.e. the central bar) and a temporal pattern that differs in the two conditions. In the Unimodal condition the main focus of activation is sustained and peaks at the time of the stimulus, or immediately before its onset. In the bimodal condition the same intervals show a suppression of activation and there are two bursts of activity at the beginning and at the end of the temporal array of noise frames.

the detection threshold improvement in our study, we have subtracted the 2nd order activation of the central stripe of the stimulus under Unimodal stimulation from that of the Bimodal condition. Figure 5 plots the family of interpolated curves representing this difference (blue waveforms) and the actual differences without interpolation (light blue stairsteps) for the Super Subject and, in the small panels underneath, for the individual observers. Positive values imply stronger Bimodal activation, negative values stronger Unimodal activation. Importantly, we have calculated the statistical reliability of the difference by comparing the bootstrap samples of the two experimental conditions with a two-tailed *t*-test, and the intervals yielding non-significant differences (i.e., $p > 0.01$) are highlighted by a black marker on top of each panels (present only in the 6th frame for DP and the 3rd frame for GC). What emerges clearly from this analysis is that the sound suppresses the 2nd order activation at the time of the target onset and for the preceding 30 to 60 ms. This pattern, though it is weaker in DP holds for all observers. Outside this window, all observers show a boost of Bimodal activation at the very first frames, while the pattern is relatively inconsistent across subjects, with mono- and bi-phasic sound-induced activations of the last frames that are hard to interpret.

## Discussion

Perceptual performance, in particular visual detection, improves in the presence of multisensory input. In the present study we investigated this effect using a psycho-

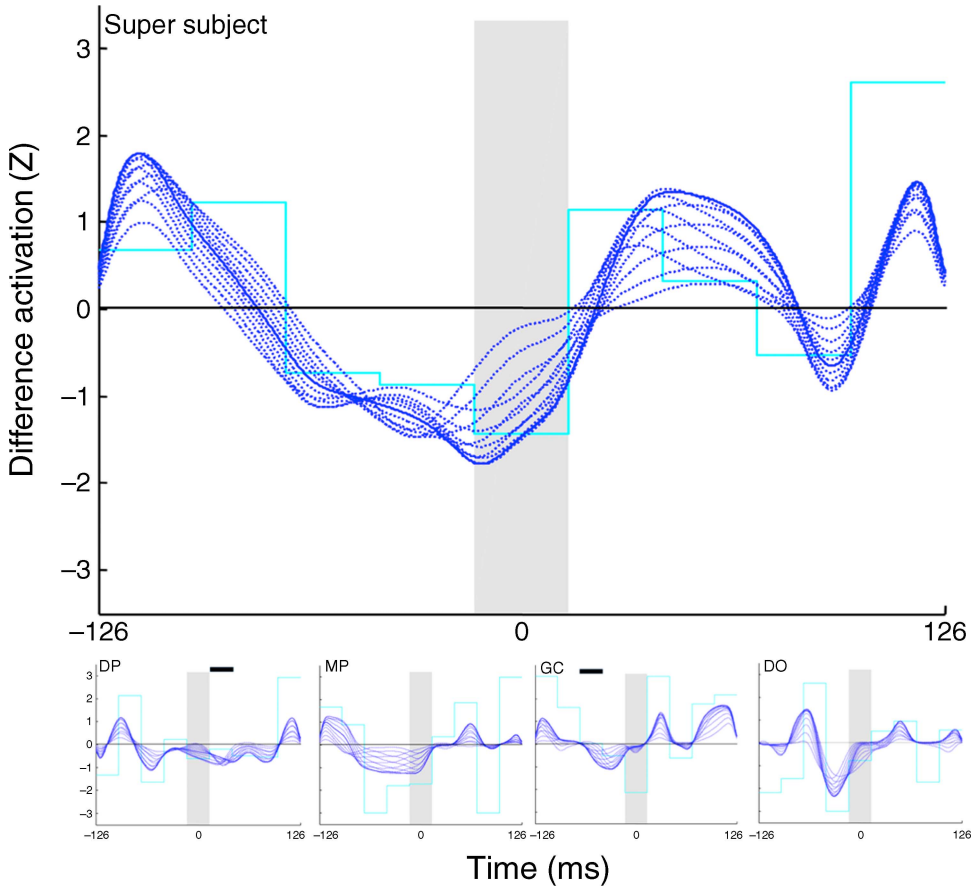


Figure 5. Difference activation (Bimodal–Unimodal) for the 2nd order kernels in the temporal dimension. The top, large graph reports the data of the Super Subject, while the four panels on the bottom of the figure are for individual observers. The family of interpolated curves (continuous straight and dashed lines) is plotted in dark blue, whereas the non-interpolated functions (stairsteps lines), are overlapped in light blue. The frame containing the target is marked by the gray vertical band in the middle of each graph, while non-significant differences ($p > 0.01$) are marked by the small black rectangles at the top of the panels (only in DP and GC). The difference highlights in all cases an initial burst of bimodal activity at the first 1–2 frames, a sound-induced suppression of the noise variance in the frame of the target and the preceding 1–2 frames (~50 ms), and a larger, irregular activity in the frames that follow the target.

physical reverse correlation paradigm requiring detection of a visual target embedded in dynamic visual noise (Unimodal condition) and in the presence of a sound played synchronously with the visual target (Bimodal condition). First, we found a reduction of contrast detection thresholds during cross-modal stimulation, confirming that the effect is solid and occurs even in the presence of spatio-temporally modulated noise. Then we observed that the presence of the auditory signal changes the pattern of activation of the 2nd order kernels while leaving unaffected that of the 1st order kernels, revealing an effect on the non-linear processes involved in detection of a luminance increment in noise.

The analysis of contrast energy obtained computing the noise variance revealed an interesting difference between Unimodal and Bimodal condition: the typically found strong variability of noise that may facilitate target identification in the 100 ms window preceding the target narrows significantly in conditions of cross-modal stimulation. This is the main finding of the present study that led us to interpret the cross-modal interaction in terms of increased gain of the visual signal when it is accompanied by a sound, reducing the intrinsic uncertainty about the channel detecting a signal across the temporal dimension. In particular, we found an interaction between auditory signal and the dynamics of the visual mechanisms that is coherent with an effect at the level of energy extraction and not explainable with the high-level, cognitive interpretations of cross-modal perception (Lippert et al., 2007; Mishra et al., 2007; Shams et al., 2002). The sound did not modify the visual activity related to the identification stage (1st order kernel) that has been suggested to reflect the behavior of mechanisms involved in simple identification, i.e. the simple cells (Neri & Heeger, 2002). This can be explained taking into account the temporal uncertainty characterizing our task: even if a sound is

played with the target frame (the 5th of 9) the stimulus elicits activity throughout the 9 frames (252 ms), and this activity cannot be completely suppressed or reduced.

Indeed, the relevant findings of this study resides in the Bimodal modulation of the templates obtained by estimating the 2nd order kernels. Within the theoretical context introduced by Neri and Heeger (2002), the spatiotemporal configuration of the kernel activations in our data fits very well with the activity of the early stages of visual processing, that is separated from the features identification stage and could reflect an active role for attentional capture mechanisms engaged by abrupt onsets of stimulation. The activity of these mechanisms displays an evident reduction in Bimodal condition and a distribution shifted on the very first frames of the stimulation, implying that the temporal information provided by the sound diminished the role of attentional capture.

Our data are coherent with the idea that the sound would introduce a gain factor to the visual information processing. This weight would amplify the power of synchronous visual signals, by reducing temporal uncertainty. Because of the different temporal resolution of visual and auditory system (with acoustic information being processed faster than visual one), the physically synchronous sound would act as a temporal pre-cue for the detection of a target embedded in dynamic noise. Following this, the 'knowing when' facilitation (that is the gain factor) would reduce the 2nd order kernel activity which is normally engaged to analyze the noise variability in a way to prepare the system to detect any abrupt variation of luminance (i.e. the stimulus onset). In this sense, a gain amplifying the power of the stimulus by pre-cueing it, would reduce the need for contrast energy extraction and so the activity of the 2nd order kernel if compared to a Unimodal condition. Therefore, when in a Unimodal, visual-only task some form of temporal precision is required, the system would benefit from the ability to capture attention by stimuli with abrupt and strong onsets (Jonides & Yantis, 1988; Yantis & Hillstrom, 1994) that operate as a cue (Nakayama, 1989; Nothdurft, 2002), but when a sound signal is presented along with the target, the system no longer needs the contribution of attentional capture mechanisms, and this is possibly due to the higher temporal resolution provided by the acoustic stimulation and to its integration with the target signal.

Such a general interpretation is in agreement with low-level theories of multisensory integration (Kayser & Logothetis, 2007; Martuzzi et al., 2007; Stein et al., 1993, 2009), since we observed an "amplification" of visual signal (contrast) when it was synchronized with sound, rather than a different decisional behavior of subjects or a criterion shift. Speculating about the neural site of this interaction, we could hypothesize a role of the superior colliculus, where neurons in different layers are known to be responsive in tasks that involved covert shifts of attention and attentional capture (Ignashchenkova, Dicke, Haarmeier, & Thier, 2004; Posner & Petersen, 1990) as well as during multisensory stimulation (bimodal neurons).

Our results are indeed in line with the possibility of a very low level interaction of cross-modal signals, coherently with evidences about the existence of neuronal pools in V1 showing strong activation during cross-modal stimulation (Martuzzi et al., 2007; Watkins, Shams, Tanaka, Haynes, & Rees, 2006). However, future studies and the development of detailed models tailored to explain the exact nature of 2nd order kernels (e.g., Neri, 2009) are needed to understand in details the specific non-linear mechanism underlying this puzzling cross-modal effect.

## Acknowledgments

This research was supported by the ERC grant STANIB.

Commercial relationships: none.
Corresponding author: Stefano Baldassi.
Email: stefano.baldassi@unifi.it.
Address: Via di San Salvi, 12 pad. 26 Firenze 50135, Italy.

## References

Ahumada, A. J., Jr. (2002). Classification image weights and internal noise level estimation. *Journal of Vision, 2*(1):8, 121–131, http://www.journalofvision.org/content/2/1/8, doi:10.1167/2.1.8. [PubMed] [Article]

Ahumada, A. J., Jr., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America, 49,* 1751–1756.

Alais, D., & Burr, D. (2003). The "Flash-Lag" effect occurs in audition and cross-modally. *Current Biology, 13,* 59–63.

Beard, B. L., & Ahumada, A. J., Jr. (1999). Detection in fixed and random noise in foveal and parafoveal vision explained by template learning. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 16,* 755–763.

Burr, D., & Alais, D. (2006). Combining visual and auditory information. *Progress in Brain Research, 155,* 243–258.

Driver, J., & Spence, C. (1998). Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 353,* 1319–1331.

Driver, J., & Spence, C. (2004). *Crossmodal space and crossmodal attention*. Oxford: Oxford Univ. Press.

Eckstein, M. P., Shimozaki, S. S., & Abbey, C. K. (2002). The footprints of visual attention in the Posner cueing

paradigm revealed by classification images. *Journal of Vision, 2*(1):3, 25–45, http://www.journalofvision.org/content/2/1/3, doi:10.1167/2.1.3. [PubMed] [Article]

Ernst, M. O., & Bulthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8,* 162–169.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology, 10,* 663–666.

Ignashchenkova, A., Dicke, P. W., Haarmeier, T., & Thier, P. (2004). Neuron specific contribution of the superior colliculus to overt and covert shifts of attention. *Nature Neuroscience, 7,* 56–64.

Jonides, J., & Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception & Psychophysics, 43,* 346–354.

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure & Function, 212,* 121–132.

Lippert, M., Logothetis, N. K., & Kayser, C. (2007). Improvement of visual contrast detection by a simultaneous sound. *Brain Research, 1173,* 102–109.

Marks, L. E., Ben-Artzi, E., & Lakatos, S., (2003). Cross-modal interactions in auditory and visual discrimination. *International Journal of Psychophysiology, 50,* 125–145.

Martuzzi, R., Murray, M. M., Michel, C. M., Thiran, J. P., Maeder, P. P., Clarke, S., et al. (2007). Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cerebral Cortex, 17,* 1672–1679.

McDonald, J. J., Teder-Salejarvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature, 407,* 906–908.

McDonald, J. J., & Ward, L. M. (2000). Involuntary listening aids seeing: Evidence from human electrophysiology. *Psychological Science, 11,* 167–171.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Mishra, J., Martinez, A., Sejnowski, T. J., & Hillyard, S. A. (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *Journal of Neuroscience, 27,* 4120–4131.

Nakayama, K. M., M. (1989). Sustained and transient components of focal visual attention. *Vision Research, 29,* 1631–1647.

Neri, P. (2009). Nonlinear characterization of a simple process in human vision. *Journal of Vision, 9*(12):1, 1–29, http://www.journalofvision.org/content/9/12/1, doi:10.1167/9.12.1. [PubMed] [Article]

Neri, P., & Heeger, D. J. (2002). Spatiotemporal mechanisms for detecting and identifying image features in human vision. *Nature Neuroscience, 5,* 812–816.

Neri, P., Parker, A. J., & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature, 401,* 695–698.

Nothdurft, H.-C. (2002). Attention shifts to salient targets. *Vision Research, 42,* 1287–1306.

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2003). Cross-modal enhancement of perceived brightness: Sensory interaction versus response bias. *Perception & Psychophysics, 65,* 123–132.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Reviews in Neuroscience, 13,* 25–42.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature, 408,* 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Brain Research. Cognitive Brain Research, 14,* 147–152.

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision, 2*(1):7, 105–120, http://www.journalofvision.org/content/2/1/7, doi:10.1167/2.1.7. [PubMed] [Article]

Stein, B. E., London, N., Wilkonson, L. K., & Price, D. D., (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience, 8,* 497–506.

Stein, B. E., Meredith, M. A., & Wallace, M. T. (1993). The visually responsive neuron and beyond: Multisensory integration in cat and monkey. *Progress in Brain Research, 95,* 79–90.

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J., Jr., & Rowland, B. A. (2009). Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research, 198,* 113–126.

Vroomen, J., & de Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1583–1590.

Wallace, M. T., Carriere, B. N., Perrault, T. J., Jr., Vaughan, J. W., & Stein, B. E. (2006). The development of cortical multisensory integration. *Journal of Neuroscience, 26,* 11844–11849.

Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage, 31,* 1247–1256.

Watson, A. B., & Nachmias, J. (1977). Patterns of temporal interaction in the detection of gratings. *Vision Research, 17,* 893–902.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33,* 113–20.

Yantis, S., & Hillstrom, A. P. (1994). Stimulus-driven attentional capture: Evidence from equiluminant visual objects. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 95–107.