<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Dynamics of perceptual grouping: similarities between the organization of visual and auditory groups</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Elliott, Mark</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2001-06</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>Taylor &amp; Francis</td>
</tr>
<tr>
<td><strong>Link to publisher's version</strong></td>
<td><a href="http://dx.doi.org/10.1080/13506280143000043">http://dx.doi.org/10.1080/13506280143000043</a></td>
</tr>
<tr>
<td><strong>Item record</strong></td>
<td><a href="http://hdl.handle.net/10379/1574">http://hdl.handle.net/10379/1574</a></td>
</tr>
</tbody>
</table>
Dynamics of perceptual grouping: similarities between the organization of visual and auditory groups

1Aleksandar Aksentijević, 2Mark A. Elliott & 1Paul J. Barber

1School of Psychology, Birkbeck College, University of London, London, England

2Institut für Allgemeine Psychologie, Universität Leipzig, Leipzig, Germany

DRAFT (Please do not cite)

Correspondence to:
Aleksandar Aksentijević
School of Psychology,
Birkbeck College, University of London,
Malet St
London, WC1E 7HX
England

Tel: ++44 (0)20 7631 6207 (secretary)
Fax: ++44 (0)20 7631 6312
Email: a.aksentijevic@psyc.bbk.ac.uk
Abstract

In vision, the Gestalt principles of perceptual organization are generally well understood and remain a subject of detailed analysis. However, the possibility for a unified theory of grouping across visual and auditory modalities remains largely unexplored. Here we present examples of Gestalt grouping in audition and vision, that share many similar organizational properties, in particular, similarities are revealed between stimulus grouping in apparent motion, visual and auditory streaming. Further, given evidence that visual and auditory grouping may share a common neural algorithm (the frequency-specific synchronization of neural activity), a further possibility is that supramodal principles of Gestalt organization may find a common algorithm at the physiological level. Taking the available evidence into consideration, it is suggested that some basis exists for the development of cross-modal principles for Gestalt organization that are strongly supported by algorithmic correspondence at the level of neural action.
Introduction

In the last four decades, a substantial effort has been devoted to the investigation of neural mechanisms of (visual) perceptual grouping proceeding from the perspective that neural feature analyzers represent the primary mechanisms from which visual grouping proceeds. This approach has lead to many significant advances in our understanding of the microstructure of visual information processing and visuo-cortical function, and has indeed produced evidence that Gestalt grouping can correlate with activity at the level of the single neuron (see, e.g. Gray, König, Engel, & Singer, 1989; von der Heydt & Peterhans, 1989). This evidence has lead to detailed theories concerning the algorithms responsible for information processing in the brain, which can trace their origins, in part, to the principles of perceptual organization originating in early Twentieth Century Gestalt psychology.

This paper is concerned with general principles of perceptual organization that may be evident from similarities between the Gestalt organization of visual and auditory stimuli. Although striking similarities have often been noted between certain classes of auditory and visual grouping contexts (Bregman, 1990; Koffka, 1935), the possibility that these similarities represent general, modality-independent invariants of perceptual organization remains an open issue. Further, the contribution of this notion, that perception may be fundamentally based upon modality-independent invariants to the current debates concerning ‘stimulus binding’ (that is the search for a common algorithm or principle underlying the perception of complex stimuli), remains for the most part unexplored. Rather than speculate directly on the neural algorithm that may underlie supramodal principles of perceptual organization, in this paper we aim to examine some of the evidence, and ideas in support of a common framework for visual- and auditory-perceptual organization based upon Gestalt principles of stimulus
Perceptual organization in vision and audition

grouping. In order to achieve this we have divided this paper into four sections: in the
first section, a simple grouping situation is described and discussed with reference to
both auditory and visual contexts; the second section addresses the similarity between
three well-known instances of perceptual grouping, namely, visual and auditory
streaming (Bregman, 1990) and apparent motion (Wertheimer, 1912); in the third
part, a brief overview is given of the current attempts to understand the grouping
processes. Most of these focus on the co-operative behavior of neurons and neuronal
assemblies; finally, in the discussion section, some of the unanswered questions are
addressed with a view to highlighting the importance of supramodal invariants of
perceptual organization.

Instances of grouping in vision and audition

Consider the display in Figure 1. The line segments are mutually equidistant
along the abscissa. When the gaps between the segments are uniformly shortened, the
segments will tend to be perceived as a single zigzag line. As the vertical distance
between the segments increases, the percept gradually changes into two separate
streams (Figure 2). Shifting the upper ‘stream’ horizontally relative to the lower one
so that they completely coincide will result in a sequence of line-segment pairs
(Figure 3). In addition, there is the intermediate case where the relationship between
the vertical and horizontal distance produces a bistable percept. The display can be
interpreted either as a zigzag line or as two horizontal lines (Figure 4).

Figures 1, 2, 3 and 4 about here
In all four instances, the perceptual outcome is mediated by the horizontal/vertical relationships. Changing the horizontal distance between the segments will, depending on the other conditions, either increase or decrease the grouping stability of the percept. It is important to note that the grouping properties of this simple configuration are affected by its absolute magnitude, expressed in terms of visual angle. The smaller the overall display, the more closely grouped it will appear. Thus, other things being constant (e.g. the contrast between the lines and the ground, the length of the segments), it can be said that in the above example, the organization of the elements is governed by three factors. These include the vertical and horizontal distance between the segments and the absolute magnitude of the display.

The above example serves a dual purpose. First, it demonstrates that the interaction of the three factors generally produces three stable grouping solutions: a single interrupted line, two separate ‘lines’, or a sequence of line-segment pairs. In addition, a specific configuration of the elements results in ‘unstable’ perceptual solutions. The important point is that these three factors provide the context (the number of degrees of freedom) within which the grouping occurs. The example also represents a visual illustration of a well-known auditory phenomenon - the streaming effect. According to Bregman (1990), the streaming effect illustrates the general principle of auditory stream segregation. How does it relate to the above visual example? To all intents and purposes, the visual and auditory versions of the streaming effect are governed by identical dynamics. Substituting frequency for the vertical, and time for the horizontal distance, we can interpret the above observations in terms of auditory grouping. The stable solution illustrated in Figure 1 corresponds to the perception of a single auditory stream. Increasing the frequency differential between the alternating pure-tone segments and/or decreasing their temporal
Perceptual organization in vision and audition

separation leads to the splitting of the sequence into two perceptually separate streams. When the two streams have common onset (phase angle 0 degrees), the listener hears a sequence of harmonic complexes (Figure 3). In a demonstration of the effects of temporal coincidence on streaming, Bregman observed the transition from the ‘harmonic’ to the ‘one-stream’ stable solution from a sequence of two alternating pure-tone segments each lasting 250 milliseconds (Bregman, 1990). Bregman found that when the inter-stream temporal overlap was 50%, the components were perceptually assigned to different streams. Once the overlap reached 88%, the segments belonging to different streams were fused into unified ‘vertical’ percepts, which sounded like complex tones. Although other parameters were not mentioned, the example demonstrates that we can know, at least at the probabilistic level, when things are ‘perceived as belonging together’. Finally, under the right circumstances, both stable solutions can become equally available. As Bregman noted: “In the streaming effect, when the sequence is played with an intermediate speed and frequency separation, the sequence can often be heard as either a single stream or two, first one than the other organization taking charge.” (p. 199).

In the case of audition, the vertical grouping solution corresponds to the perception of the pitch of complex tones (harmonic complexes). While several spectral frequencies coexist within a harmonic series of a complex tone, the product of their grouping is a distinctive auditory figure defined by a single predominant pitch, which is generally unaffected by the absence of the fundamental frequency from the stimulus spectrum (see, Schouten, 1940). The separate spectral pitches can be ‘heard out’ under certain conditions, namely, when a conscious effort is made to analyze the complex percept (see, Helmholtz, 1863). This suggests that the spectral components of complex tones represent the fundamental descriptors, or building
blocks of music and speech, which, according to some investigators (e.g. Terhardt, 1986) may operate in a remarkably similar way to the function of edges and contours in early visual feature analysis.

The phenomenon of auditory stream segregation represents an important manifestation of the limited organizing potential of the auditory system. For a given number of stimuli, only a certain number of stable grouping solutions is available. At the same time, these grouping outcomes are a function of the interaction between stimulus frequency and temporal factors – stimulus duration and the interstimulus interval (ISI). In other words, the closer together the tones are in terms of frequency or time, the more easily they will be assigned to a single stream. Further, stream formation is facilitated by continuous/smooth transition. Continuous glides and ramped tones tend to be grouped into single streams (Bregman & Dannenbring, 1973). Similarly, tones that undergo common frequency or amplitude modulation will be perceived as belonging to a single stream. In addition, auditory grouping is facilitated by spectral similarity. Harmonically related tones that have common onset will also tend to form a coherent percept. In order for the partials (pure-tone components of harmonic complex tones) to fuse into a single percept, the harmonic ratio has to be maintained within a relatively narrow margin. Simultaneously stretching (or compressing) the spacing between the harmonics by as little as 7% destroys the sensation of harmonic pitch (Cohen, 1980; Slaymaker, 1970). Similarly, mistuning a low harmonic by between 3 -6% will cause the shifted harmonic to be perceived as separate from the (somewhat weakened) harmonic complex (Moore, Glasberg & Peters, 1985).
Common factors influencing grouping in vision and audition

Another interesting parallel between auditory and visual grouping is provided by the phenomenon of apparent motion (Wertheimer, 1912). Briefly, the perception of two (or more) successive light flashes changes as a function of the distance between the flashes, the size of the interstimulus interval (ISI), and the intensity of the flashes. As the ISI is increased, the perception changes from simultaneity (less than 30 ms) and partial movement (30 – 60 ms), to optimum (about 60 ms) and phi movement (60 – 200 ms). Finally, at ISI of above 200 – 400 ms, no movement is perceived (Graham, 1965). According to the Körte’s third law (Körte, 1915) as the speed of presentation increases, distance between flashes must be reduced in order for the smooth motion to be preserved. This interdependence of vertical and horizontal factors characterizes not only apparent motion, but as already mentioned, also the auditory streaming effect. Specifically, for a given pure-tone segment duration, the speed of presentation must be reduced as the frequency difference between the segments increases, if the alternating tones are to be perceived as a single stream. Bregman (1990) generalizes this observation to music and refers to melodic motion that is preserved through a trade-off between frequency separation and speed.

More intriguing is the apparent parallel between the manifestations of Körte’s law in apparent motion and in static visual displays of the kind described earlier (Figures 1 to 3). Let us look again at Figure 3. If we uniformly compress the display along the abscissa by simultaneously shortening both the segments and gaps while preserving the vertical distance, the resulting figure will almost certainly not be grouped as a single visual stream represented in Figure 1. In fact, depending on the other two factors, the stream will have fragmented either into a set of weakly grouped segments or into two streams. This simple demonstration suggests that the same
dynamics govern grouping in three different contexts: apparent motion, auditory and
visual stream segregation (although the list could be longer). Discrete auditory and
visual stimuli are connected in some way and the patterns of connection are a function
of the spatial or spatio-temporal relationships obtaining between the stimuli. The
question then arises as to the apparent correspondence between the ‘horizontal’ spatial
dimension in the visual streaming example, and the temporal dimension of the
apparent motion and auditory streaming effects. Can this correspondence be ignored
or explained away?

A common neural code for auditory/visual binding?

For the last twenty years, the search for the factors underlying organization in
vision and audition has been focused on the neural correlates of perceptual grouping.
A substantial body of (primarily visual) research has been based on the premise that
the process of percept formation consists of at least two stages: an early stage of
stimulus extraction and analysis, followed by the subsequent recombination and
integration of these features into perceptual wholes (or Gestalten). Implicit in such a
model is the notion of a mechanism responsible for integrating or ‘binding’ the
features belonging to an object while segmenting those that belong to different
objects. Among a number of competing hypotheses put forward in response to the
binding problem, the temporal-correlation hypothesis (von der Malsburg, 1981) has
been particularly prominent. Generally, analyzed object-features are recombined into
coherent percepts via the synchronization of topographically separate feature-coding
neuronal assemblies in the cortex (Singer, 1993). Numerous studies have indicated
that synchronized assemblies tend to oscillate within the gamma band (30-80 Hz; see
Basar-Eroglu, this issue). The discovery of transient, stimulus-locked gamma-band
oscillations in visual (Gray, et al, 1989) and auditory cortices (Galambos, Makeig &
Talmachoff, 1981), has engendered the ‘gamma-band’ hypothesis, according to which, the binding of stimulus features is related to the ‘synchronized’ activity of relevant neuronal assemblies at a subset of gamma-band frequencies (not always, but typically referred to in terms of synchronized neuronal oscillations at around ‘40 Hz’).

How does the synchrony-binding hypothesis relate to Gestalt grouping phenomena? A number of researchers have addressed the relationship between the grouping properties of simple stimuli and synchronized neuronal oscillations. To illustrate, 40-Hz evoked response, phase-locked to the stimulus onset has been recorded with static (Tallon, Bertrand, Bouchet & Pernier, 1995) and coherently moving visual stimuli (Gray et al., 1989). Similarly, Pantev et al. (1991) have hypothesized that auditory gamma-band evoked potential could be related to the synchronization of separate tonotopic cortical loci coding pure-tone components of harmonic complex tones. Notwithstanding the ongoing debate on the functional role of different stimulus-dependent responses (see e.g. Tallon-Baudry & Bertrand, 1999), gamma-band synchronization represents a potential neuronal correlate of certain grouping phenomena. To illustrate, recent psychophysical evidence indicates that external and specific gamma-band entrainment engenders a matching frequency code that facilitates detection of illusory figures (Elliott & Müller, 2000; Elliott, Herrmann, Mecklinger & Müller, 2000). Yet, neuronal synchronization per se might not be entirely sufficient to account for the dynamical aspects of grouping. Here, a more comprehensive model would be necessary in order to explain the interaction of spatial and temporal factors on the perceptual impact of changing stimulus relationships. One general approach (Geissler, Scherbera & Kompass, 1999) posits a hierarchical temporal structure, based on a psycho-neural time quantum, which provides a dynamical framework for perceptual organization (the cited study deals with apparent
motion). Specifically, the dynamics of binding can be described in terms of shifting patterns of synchronization, constrained by the value of the time quantum. At the core of the ‘temporal’ accounts (see also Dehaene, 1993) lies the conviction that the fine temporal structure, arguably detectable in psychophysical measurements (e.g. Lisman & Idiart, 1995; Pöppel, 1968), could represent the ‘missing link’ between the psychophysical and neural levels of explanation. This view receives support from models such as the ‘shifting correlation’ model of auditory streaming by Wang (1996), based on Terman & Wang’s (1995) global-competition-local-cooperation network of oscillators. The model accounts for the relational aspects of auditory streaming in terms of fast synchronization of locally connected oscillators constrained by the activity of a global inhibitor. The degree of synchronization is determined by the mutual proximity (along pitch- and time axes) of the stimuli. A different approach to the problem of visual grouping has been proposed by Grossberg (see e.g. Grossberg & Grunewald, 1996). The perceptual constancies characteristic of vision have to be maintained in a changing environment. A mechanism labeled ‘perceptual framing’ ensures that the Gestalt properties of a stimulus are preserved, notwithstanding the fact that its parts are processed at different rates. According to Grossberg, fast resynchronization of spatially distributed object parts, based on a global-cooperation-local-competition model, simulates the effects of global context on the processing of the parts. The perceptual-framing concept can, among other things, explain apparent motion. It is of interest, if not altogether surprising, that the above mentioned grouping phenomena are described by models based on different dynamical principles. Wang’s model operates on two degrees of freedom – vertical and horizontal distance between stimuli, while Grossberg’s model takes into account the possibility that synchronized assemblies require an additional level of control in order
Perceptual organization in vision and audition

to maintain object coherence – a requirement not crucial in the simple streaming context. However, the examples presented at the beginning suggest that a successful model of visual grouping should at least account for the spatio-temporal correspondence between the static and moving grouping contexts.

This brief overview only hints at the immense effort invested in understanding temporal factors that underpin the binding problem. It is possible that a better understanding of the neuronal algorithm, will throw light on the still insufficiently understood invariances of perceptual organization. Particularly promising in this respect is the dynamical analysis and modeling of neural oscillatory behavior. Yet, the compelling similarities outlined in the previous sections also represent a challenge for experimental psychology.

Discussion

One advantage of the above approaches lies in the possibility of establishing a link between psychophysics and neurophysiology. Starting with the physiological properties of the neural substrate on one side, and the psychophysical measurement on the other, the hope is that the study of the relationship between the domains can bring about a satisfactory explanation. Yet, certain questions remain and importantly, these questions relate specifically to the evidence that visual and auditory grouping may be similar at the Gestalt level of organization. For example, are the above-described supramodal constancies important? The apparent correspondence between the static and moving visual- and auditory contexts raises several interesting possibilities. First, the organizing potential of the perceptual system is limited, for otherwise the distinction between stable and unstable grouping solutions would not exist. Second, the perceptual impact of a particular arrangement of stimuli (salience of individual stimuli and their grouping) depends on the context – illusory contours cease to be
contours if a single element of the figure is displaced. A harmonic complex is weakened if a harmonic is mistuned. The salience of individual objects (auditory and visual) is inversely proportional to the number of objects present. Spatial or spatio-temporal proximity in vision is in a sense equivalent to proximity in audition, defined in terms of time and pitch. Further, proximity and similarity are dynamically related in both modalities, despite the fact that the embedding dimensions are different. Importantly, the apparent similarity between the static visual and auditory examples suggests that the dynamics of grouping are not a function of selective attention. If anything, the opposite is true, namely, attention is guided by the stable grouping solutions, or to quote Wang (1996), “directed to streams” (p. 445). As can be seen in Figures 1, 2, 3 and 4, attention keeps switching between the competing organizations long after the system has fully encoded the display. It cannot do otherwise. This, of course, brings us back to the beginning, to the possibility that a percept might be more than the sum of its parts. Although the above remarks restate the well-known Gestalt observations, perhaps the time is right for a rethink. Can the dynamics of grouping be investigated from within a single framework, as a problem in its own right? In other words, should Gestalt phenomena be studied systematically, without a direct recourse to neurophysiology? Although neuronal synchronization, together with other brain-oriented approaches, remains a promising avenue for research, the gap between the perceptual transparency of the grouping phenomena outlined earlier and the opacity of the underlying neural mechanisms is still vast. While some important aspects of neural dynamics are still insufficiently understood (for instance, the relationship between synchronization and oscillations, the functional role of various brain responses, not to mention the relationship between the behavior of single cells and the behavior of cell assemblies), the grouping phenomena are stable and relatively easily
studied in their own right. Further, as our understanding of the role of the brain in perception improves by the day, psychology is finding it difficult to reassess the legacy of Gestalt and move forwards. As Chen (this issue) notes, the main obstacle to a systematic psychophysical investigation of the dynamics of grouping has been the lack of a theory. In order to explain the sensitivity of the visual system to topological properties of objects (see also Chen, 1982), Chen proposes that these topological properties (connectedness, number of holes) could represent the invariants, which govern perceptual organization in vision. Further, according to Chen, the increasing selectivity of the visual system could be represented by a formal system akin to the hierarchy of geometries (topology, projective, affine, Euclidean) originally put forward by Klein (1872).

These considerations prompt the question as to what kind of mathematical model could describe the supramodal grouping invariants described earlier. Mainstream psychology never accepted the original attempt by Köhler (1920) to incorporate the Minimum principle into perception, partly because of the unsustainable psychophysical parallelism of the ‘physical Gestalten’ model. Further, there is the problem of quantifying grouping. Without venturing into a detailed exposition, we propose that a reformulation of the Minimum principle in the framework of surface topology/analytical geometry might provide a link between the primitive relational properties of auditory and visual grouping phenomena described in this paper, and psychophysics. What would be the starting assumptions for such a model? First, the finite organizing potential of the perceptual system could be represented as a perceptual ‘space’, conceptualized as a surface of uniform elasticity. In the first instance, the visual ‘sheet’ is defined by two spatial dimensions and (in the context of apparent motion) time. Frequency and time would then define auditory
‘space’. Consequently, perceptual objects could be conceived of as patterns of continuous distortion of perceptual space. The second assumption is that the perceptual system is sensitive to these patterns and that the more homogeneous the distorted region, the more easily the objects will be grouped. Analysis of the general geometrical properties of the patterns of distortion could provide a means for relating the perceptually transparent dynamics of grouping to psychophysical measurement, and eventually, neurophysiological data.

**Conclusions**

Understanding the principles of perceptual organization remains a major task for psychological science. Three examples (two visual and one auditory) described here demonstrate that at least some of these principles operate across sensory modalities. Further, the apparent similarity between static and moving visual and auditory contexts suggests a form of spatio-temporal equivalence that needs to be addressed by any general model of perceptual grouping. Many accounts of auditory and visual binding have focused on the cooperative behavior of neurons and neuronal assemblies, with the result that the dynamical interactions observed in grouping are often interpreted in terms of shifting patterns of synchronized neuronal activity (see e.g. Singer, 1994). Although systematic investigation of grouping-related neural dynamics holds considerable promise, the current state of knowledge leaves room for complementary approaches. A theory of grouping based on a topological interpretation of the Minimum principle might represent such an approach.
Perceptual organization in vision and audition

References


Acknowledgements

Mark A. Elliott is supported by Deutsche Forschungsgemeinschaft project grant SCHR 375/8-1. Correspondence concerning this article should be addressed to Aleksandar Aksentijević, School of Psychology, Birkbeck College, University of London, Malet St, London, WC1E 7HX, England (email: a.aksentijevic@psyc.bbk.ac.uk).
**Figure Legends**

**Figure 1**: A stable (auditory and visual) streaming solution. The sequence is perceived as a zigzag line or a single auditory stream. In the latter case, the abscissa represents time (in milliseconds), and the ordinate represents frequency (in Hertz).

**Figure 2**: The second stable solution. The sequence has split into two horizontal lines (or auditory streams). For auditory streaming, the abscissa represents time (in milliseconds), and the ordinate represents frequency (in Hertz).

**Figure 3**: The third stable solution. In the visual context, the display is interpreted as a sequence of line-segment pairs. In the auditory grouping context, the figure corresponds to a sequence of harmonic pitches. In this case, the abscissa represents time (in milliseconds), and the ordinate represents frequency (in Hertz).

**Figure 4**: An unstable solution. The sequence is perceived either as a zigzag line, or two horizontal lines. In the auditory context, the figure corresponds to the situation where the single-stream and two-stream solutions are equally available. In this case, the abscissa represents time (in milliseconds), and the ordinate represents frequency (in Hertz).
Frequency (Hz)

Time (ms)

Figure 1

Perceptual organization in vision and audition
Perceptual organization in vision and audition

Figure 2

![Diagram showing two time-frequency axes with ellipses depicting frequency and time intervals.](image-url)

Figure 2
Figure 3
Perceptual organization in vision and audition

Figure 4

Frequency (Hz)

Time (ms)