Auditory: Auditory Scene Analysis

In Squire, L. (Ed.) *New Encyclopedia of Neuroscience*, Oxford, UK: Elsevier

Albert S. Bregman
Psychology Department
McGill University
Montreal, Quebec,
Canada H3A 1B1

Key words

acoustics
audition,
auditory organization,
auditory perception,
auditory scene analysis,
electrophysiology,
hearing,
mismatch negativity,
perception
perceptual organization,
sound,
stream segregation,
streaming,

Synopsis

A mixture of sounds, though distinct in the environment, arrives in the form of a single pressure wave at each ear. From it, listeners must extract the signals coming from individual sources of sound, a process called auditory scene analysis (ASA). ASA first characterizes the incoming waveform by its frequency components and other features, then creates subsets (auditory streams) that extend over time, each representing a single environmental sound source. ASA exploits regularities in the signal to determine how to parse it. ASA is present at birth in humans and has been found in other animals.

The scene analysis problem

In the natural world that surrounds us, and in which we evolved, it is rarely the case that only one sound at a time is present. What reaches our ears is a mixture of all the sounds present at a given moment: a voice speaking, the recorded music in the background, a car passing by, a bird singing. If you ask people how they can hear an individual sound in this mixture, they say that they just focus their attention on whatever they want to listen to. This answer, however, ignores a fundamental difficulty. The mixture that reaches each of our ears is a single pressure waveform that is the moment-by-moment arithmetic sum of the pressure patterns that arise from individual events. This summed wave does not have written on it how many sounds contributed to it or how each sound is buried in it. Yet our auditory systems, and those of other animals as well, have the capacity to find the individual sounds in the mixture—a capacity called auditory scene analysis (ASA).

Every sound of finite duration can be thought of as the sum of a set of frequency components of different amplitudes and phases (a spectrum). An example of a mixture is shown in Figure 1 in the form of a spectrogram, (showing time on the x-axis, frequency on the y-axis, and the intensity at any time-by-frequency position as darkness). This representation is relevant to ASA because there is evidence that the first stage in the neural processing of an incoming acoustic signal involves analyzing its frequency composition. The ASA problem is equivalent to finding a set of spectrograms, which when superimposed and summed, gives us the observed spectrogram. This decomposition would be made easier if natural sounds were compact in frequency or in time, but generally, they are not. Think of two voices heard at the same time. Far from being compact in frequency or time, the frequency components of the two are intermingled on both dimensions.

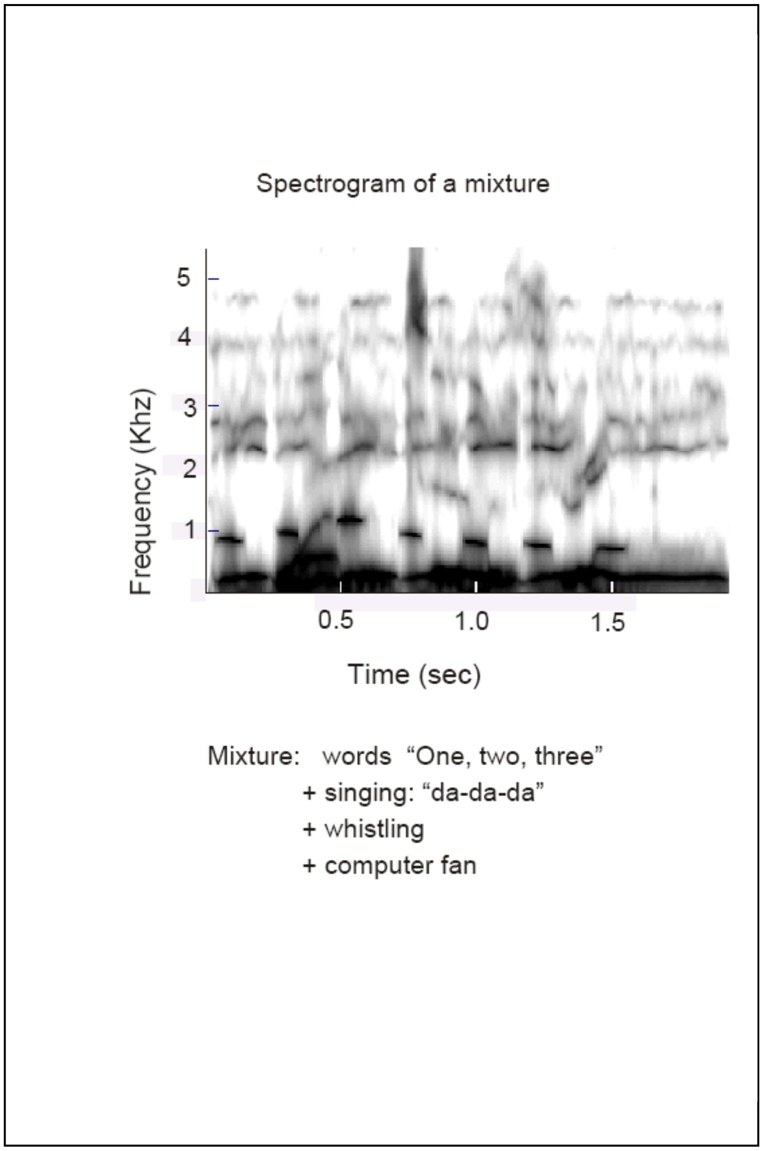

Figure 1. Spectrogram of a mixture of four sound sources: the words, "one, two, three"; a voice singing "da-da-da", a person whistling, and a computer fan.

It should be evident that there are a virtually infinite number of ways in which this a spectrogram can be decomposed into two or more component spectrograms, if the decomposition is not guided by principles. But what sort of principles? Engineers have used mathematical procedures, such as principle components analysis, that solve the ASA problem decisively, but only in very restricted circumstances. Since no single computation will always solve the problem in a broad range of environments, animals are forced to work with sets of principles that are called “heuristics” because they are very helpful, but are not guaranteed to always work. These heuristics exploit regularities in the world such as the following: since it is unlikely that two unrelated sounds in the world will start at exactly the same time, if the auditory system registers a set of frequency components whose onsets are approximately simultaneous, it is highly probable that they are parts of the same sound and therefore should be grouped together.

One cannot be sure that every type of animal solves the problem in the same way. Some undoubtedly have specialized mechanisms to extract critical sounds (for detecting prey, predators, or mates m the mixture). The specialized mechanisms may well be neural circuits that act as detectors for the required information, responsive to certain time and frequency relations, and unresponsive to sounds that lack these relations. However, larger-brained animals deal with sound in much more complicated ways, and can learn about the dangers and affordances of new sounds, but only if they can extract them from mixtures. For these animals, including ourselves, it is of great value to have general methods for decomposing a mixture into its component sounds, regardless of whether the latter are familiar or not.

Primitive and schema-based processes.

There seem to be at least two types of brain mechanisms involved in the grouping of information. The first is a set of primitive (unlearned), bottom-up processes, that is probably shared with non-human animals. As described by Albert Bregman, in his book, Auditory Scene Analysis, these processes group the information by using similarities and discontinuities in the signal, similar to those described by Gestalt psychology in their study of vision.

The second type of mechanism consists of a set of brain processes (called schemas) for dealing with frequent and important patterns in the environment. They may be, completely, or in part, innate, but large-brained animals such as humans can modify them, and develop new ones through learning. These are the mechanisms that permit recognition of conspecific animal sounds, familiar words, melodies, and so on, and probably assist in segregating these patterns from their backgrounds. Schemas are extremely numerous in humans, numbering at least in the hundreds of thousands for a particular adult (consider just the schemas for the tens of thousands of words that an adult can recognize). They can operate in conjunction with attention, as when we are trying to listen to a familiar voice in a crowded room, or prior to attention, as when, in that same room, our own name pops out of the background sound, attracting our attention. Despite the probable importance of schema-based perceptual processes in isolating familiar patterns from their contexts, the present discussion focuses on the primitive processes of ASA.

ASA as grouping

The decomposition of the information in the mixed spectrogram of Figure 1 can be viewed as a problem of grouping. This grouping has two aspects, simultaneous and sequential. Simultaneous grouping determines which parts of the complex information presented simultaneously to the senses should be allocated to the same description of an environmental event. One can think of this as sharing out the energy of the spectrum that is present at a given moment. In the spectrogram, this represents the grouping of energy on the y dimension, and even on the z (intensity) dimension, since the energy at a single frequency-by-time location may not all have come from a single environmental sound.

Sequential grouping is responsible for answering the question of which spectral components should be connected up over time (the horizontal dimension of the spectrogram) to yield a representation of a distinct train of events in the environment. The two types of grouping interact, but it is convenient, for exposition, to describe them one at a time.

Sequential integration & segregation

A laboratory phenomenon called stream segregation, or streaming, illustrates sequential integration. The reader can view an illustration of a pattern of sounds in Figure 2 and listen to it in (Demonstration 3 from Bregman & Ahad, 1996). The stimulus is a rapid cycle composed of two tones, a higher-frequency one (H) and a lower-frequency one (L), formed into triplets, separated by short silences: HLH–HLH–HLH–(repeated). The H and L tones are far apart in frequency. This cycle gradually speeds up. At slow rates, the HLH triplets are heard as repeating units with a galloping rhythm; but as the sequence gets faster, the high tones seem to segregate from the low ones and we hear two parallel sequences, one consisting of the a sequence of high tones and the other, a slower sequence of the low tones. These two streams of sound seem to be going on at the same time, yet are perceived independently. Attention seems to be able focus on one stream or the other, but not on both at the same time. When the high and low tones are close to one another in frequency, the perception of the gallop persists even at higher speeds, and the sequence remains as a single coherent stream. This shows the importance of frequency separation, as well as speed, in causing the H and L tones to form separate streams. (These effects are illustrated in Figure 2 and Demonstration 3 from Bregman & Ahad, 1996) This stream formation is a form of ASA. When the streams are segregated, the system is betting that there are two sources of sound in the environment, not one. While sequential grouping of the sensory data in a complex environment may not operate in as simple a way as in the streaming phenomenon of the laboratory, it is thought that this phenomenon is a glimpse of the sequential process of ASA in a pure form. Consequently, this artificial version of ASA has been used extensively to study sequential grouping.

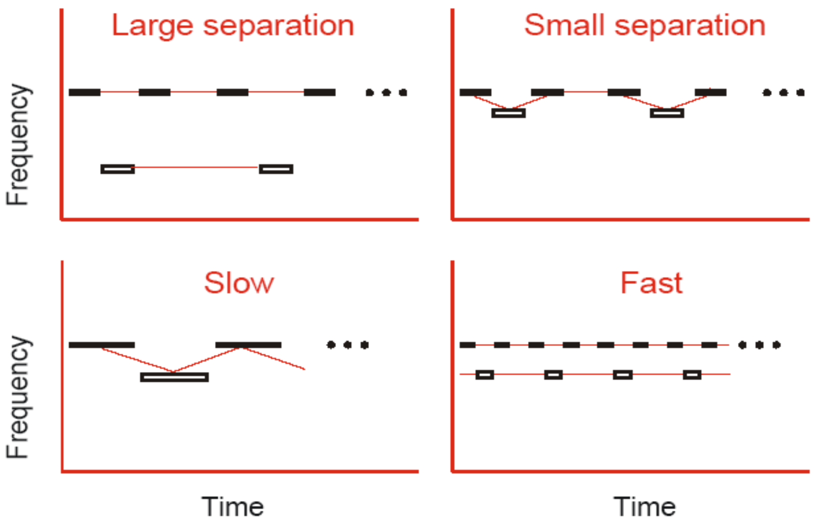

Figure 2. Diagram of four condition in the galloping pattern of a higher tone (H) and a lower one (L) played repetitively in a galloping pattern: HLH–HLH–HLH–…. The connecting lines indicate the perceived streams. The top two panels show the effects of frequency separation; the bottom panels show the effects of speed.

The streams need not be as simple as those of Figure 2. One can interleave the notes of a melody with distractor tones, destroying our ability to recognize it ((Demonstration 5 from Bregman & Ahad, 1996)) However, if the melody and the distractors are separated in pitch range, one can hear each sequence in a separate auditory stream. Incidentally, the number of concurrent streams is not restricted to two. The upper limit is unknown, If one uses only musical notes, varying only their pitch, the limit seems to be three or four. However, if one were to add quite different sounds to the mixture, such as the clicking of a clock, the ringing of a telephone, spoken digits, and so on, the number would undoubtedly be higher.

Factors contributing to sequential segregation

We can think of the tones in each panel of Figure 2 as laid out on a two dimensional surface (time by frequency) and can imagine that grouping occurs as a result of relative proximity on this surface. When the frequency separations are small and the time separations large (at low speeds), the time dimension—having the greater range of values—will dominate the grouping, and tones will group with their nearest temporal neighbors, regardless of the small frequency differences. However, when the temporal separations are small (high speeds) and the frequency separations large, then frequency differences will be dominant and tones will group with tones that are close in frequency. It is as if the auditory system forms clusters in the frequency-by-time space that minimize the distances within clusters and maximize the distances between them.

However, two dimensions, e.g., the frequency and time separations of pure tones, are not enough to represent all possible differences between sounds. Other differences also contribute. The segregation of complex (as opposed to pure) tones can be based on: (a) differences in fundamental frequency, $F_{0,}$ (even when the harmonics are restricted to the same frequency range), (b) differences in spectral shape, with F_0 held constant, perceived as differences in timbre (Demonstrations 9 & 10 from Bregman & Ahad, 1996) (c) differences in spatial location (Demonstration 38 from Bregman & Ahad, 1996), (d) in intensity and (e) in amplitude envelope (e.g., rise and fall times).

The effects of these differences combine. For example, if, in a rapid sequence of alternating A and B tones, A and B are different in two ways, say in timbre and in spatial location, they will segregate more readily than if different in only one of these ways. It is as if the best clusters are formed in a multidimensional similarity space, including time as one of its dimensions.

A separate factor that contributes to grouping is continuity. If changes from A to B in the repeating sequence ABAB… are abrupt, the A’s and B’s are more likely to form separate streams than if the properties of A smoothly transform into those of B. For example, if A and B differ in frequency, then frequency glides joining A and B tend to hold the tones together as a single stream of sounds.

Cumulative effects

There is a gradual increase in the tendency of A and B to form their own streams with increased numbers of repetition of the A-B alternation. However, the alternation of two tones in a regular pattern is not a requirement for the buildup of such streams. Two sets of tones all different, but separated into two distinct frequency bands, can also segregate into separate streams. The segregation tendency dies away gradually during a silent gap and starts building up again after the silence.

It has been proposed by Stuart Anstis and Shinya Saida that the effects of repetition can be explained by the existence of frequency-transition detectors, whose function it is to integrate successive tones into a single stream. Repetition of ABAB… transitions lead to the habituation of these detectors, so that they can no longer perform this integrative function. A problem for this theory is that it appears that any perceptible difference at all between two tones may promote their segregation into two streams; so the habituation theory would require the auditory system to have a very large number of types of transition detectors.

An alternative theory, functional, rather than physiological, argues that the default condition of grouping is to assign all incoming sounds to a single stream, but the repeated occurrence of tones in different frequency regions builds up evidence that the sounds are coming from two different sources and should be assigned to separate perceptual streams.

Van Noorden's two boundaries

Stream segregation seems to take two forms, shown by the two curves in Figure 3, based on the data of Leo van Noorden who studied segregation of the HLH–HLH–… pattern. The x-axis represents time between adjacent tones, while the y-axis represents the frequency separation of the H and L tones. Each of the two curves represents the time-by-frequency threshold between hearing one versus two streams. At frequency-time values above the curve, one hears two streams and below the curve, one stream. However, there are two different curves. The upper one, the temporal coherence boundary (TCB), was obtained when the listeners were trying to hold onto a single-stream percept, and the lower one, labeled fission boundary (FB), when they were trying to hear two streams. Two facts are evident: The first is that when trying to hold onto a single stream they could do so at higher frequency separations and speeds than when they were not. This is not surprising. More remarkably, when they were trying to segregate the streams, the rate of presentation had a very small effect.

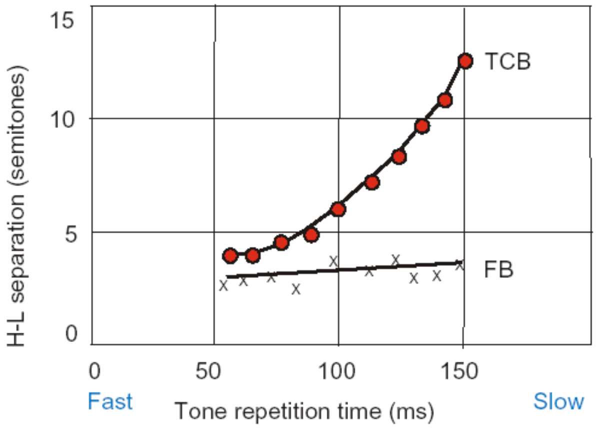

Figure 3. Curves by van Noorden showing the thresholds for stream segregation. For each curve, the area above it shows the region (in frequency x time) in which two streams are heard; in the area below it, only one stream is heard. The temporal coherence boundary (TCB) is obtained when the listener is trying to hear all the tones as a single stream. The fission boundary (FB) is found when the listener is trying to hear the tones in two separate streams. The tonal pattern used was a galloping sequence of the form LHL–LHL–…, where L and H are lower and higher frequency tones respectively.

The different shapes of the two curves reveal the activity of two different mechanisms of segregation. When trying to hold onto all the tones as a single stream, the process that interferes with this goal is probably the primitive, bottom-up process of grouping. When trying to hear separate high and low streams, the process employed is one of selective attention, trying to hear either the high or the low tones – a top-down process, which is limited only by the listener’s ability to discriminate the higher and lower tones at high speeds..

Effects of sequential grouping

When streams are strongly segregated from one another, the effects are numerous: (a) melodies and rhythms seem to include only within-stream sounds; the temporal relations between concurrent streams become uncertain. Indeed, it appears that for any pattern of sound to be clearly perceptible, it must involve only the elements of a single stream.

Simultaneous integration

Cues favoring the grouping of simultaneous components

One of the most effects on the allocation of spectral energy to simultaneous sounds “harmonicity”. Many natural sounds including the vowel-like segments of human speech, and the pitch-possessing portions of the calls of other animals, have repetitive waveforms. So also do certain manufactured sounds, such as the sound of a violin. Repetitive waves have a pitch, and are composed of harmonics—frequencies that are integer multiples of the lowest frequency (the fundamental). The auditory system has mechanisms for detecting and grouping a subset of frequency components that are multiples of the same fundamental (a harmonic series). Furthermore, it can find more than one harmonic series at a time, allowing the listener to hear concurrent sounds that have different pitches. Both humans and computers depend strongly on harmonicity to segregate simultaneous components. For example, it is easier for us to segregate a man’s voice from a woman’s that from another man’s, and computer models of the segregation of speech from other interfering sounds typically use the harmonic structure of the vocalic sounds as the main grounds for finding a stream of speech in the mixture.

Harmonicity is not the only basis for segregating concurrent sounds. Another important one is synchrony or asynchrony of onset. It is very likely that all the frequency components from a single environmental sound will start together, and it is very unlikely that the components of unrelated sounds will start at the same moment. So synchrony and asynchrony are used by the ASA system to determine whether or not frequency components should be allocated to the same sound. A third factor is the frequency separation of concurrent components. Components that are further apart are less likely to be treated as part of the same sound.

Another difference between components that is used by ASA is their difference in spatial location. This difference, by itself, does not powerfully group and segregate sets of concurrent components (although it is very powerful in sequential stream segregation). However, when other principles, such as asynchrony of onset, act to segregate concurrent components, spatial differences seem to multiply the strength of this segregation. It is surprising that spatial differences have such a limited effect in human ASA. Engineers use a technique called blind separation, in which the separation of voices is primarily or even exclusively based on spatial separation. However, spatial information is not always reliable in natural environments (for example when the sounds are coming around a corner). Perhaps this is why spatial cues are not used by humans as a primary basis for grouping components.

<u>Achievement of stability in the face of unreliable acoustic evidence</u>

It is not only spatial cues that are unreliable. While harmonicity is a good cue, many sounds are not harmonic. There is no pitch involved in footsteps, accelerating cars, scratches, or bumps—yet these sounds are informative and we need to know how to group their components across the spectrum and over time. Because even the best cues are not always reliable, the ASA system assesses many cues and allows them to reinforce, or compete with, one another in controlling the decisions about grouping (as if the various cues could vote for their own preferred organizations).

Another method that the ASA system uses because cues are less than perfectly reliable is the conservative strategy of maintaining an existing interpretation until evidence piles up that it is wrong. The system seems to start off with the hypothesis that all acoustic input is part of a single sound. As evidence builds up, an organization in terms of a number of distinct sounds emerges. Maintaining a stable percept, despite the fact that cues can suffer interference, means not altering the grouping of the sounds upon encountering a brief drop in intensity, or a momentary change in interaural spatial cues, as the listener passes behind an obstruction. If each brief glitch caused a reorganization, our perceptions would be highly unstable. Hanging on to an existing organization is a valuable strategy in a world that is more stable than the cues about it are.

What, then, are the perceptual effects of simultaneous grouping—sometimes called fusion? Its most obvious one is on the number of perceived sounds and the distinctness of their qualities. When segregated, we hear more distinct sounds, each with its own qualities of pitch, timbre, loudness, location and so on, whereas when fused, the full set of frequency components create the qualities of a single global unit, which, at any given moment, has only one pitch, one loudness, one spatial location, and one timbre.

Competition between sequential and simultaneous grouping

Although we have been looking separately at the two major dimensions of grouping—sequential and simultaneous (the horizontal and vertical dimensions, respectively, of a spectrogram)—it is important to recognize that each of these affects the other. Often they compete. This competition is illustrated in Figure 4, a simplified spectrogram whose horizontal and vertical dimensions show time and frequency respectively. A, B and C are three pure-tone components. First A occurs alone, followed by B and C together, and this pattern is repeated cyclically. The BC spectrum can be interpreted as a complex tone with two frequency components or else as two simple tones, B and C, that happen to occur at about the same time. One can hear the total cycle as a two-tone stream formed of A and B, repeating over and over, accompanied by a one-tone stream, consisting of repetitions of C. Alternatively, one can hear a pure tone, A, alternating with a rich tone, BC.

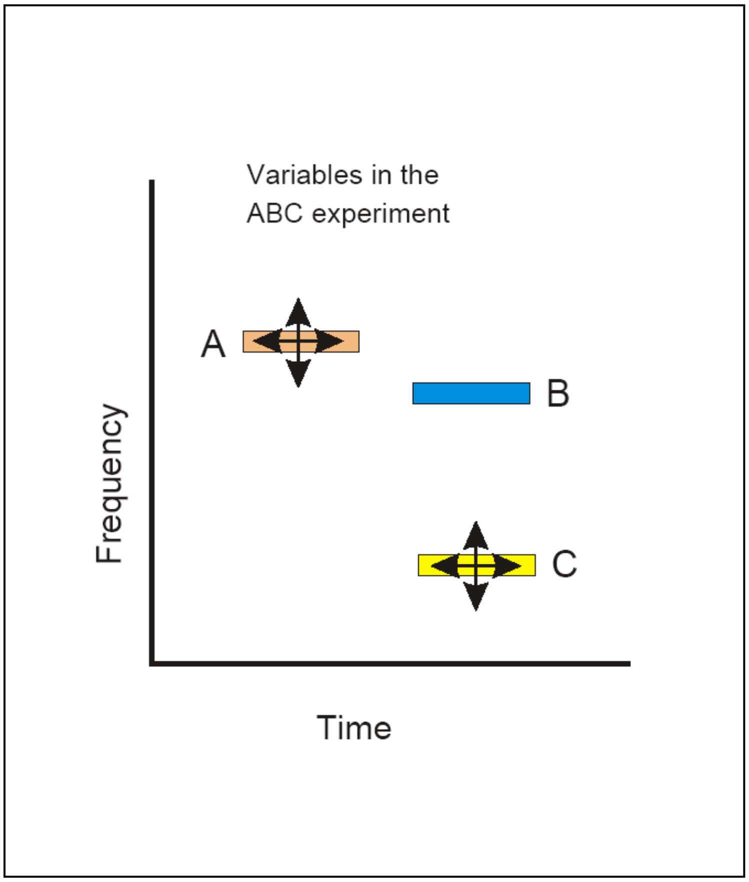

Figure 4. Diagram of an stimulus pattern in which a pure tone, A, alternates repeatedly with a pair of pure tones, B and C. The arrows show that the temporal positions and frequencies of both A and C may be varied. As A comes closer to B in frequency or time, it captures the latter more strongly into a sequential stream. As C comes closer to B in frequency or becomes more synchronous with it, it captures B more strongly into a fused perceptual unit.

This is a good stimulus for demonstrating the competition of sequential and simultaneous organizations. First we can increase the sequential grouping of A with B by moving A closer to B in frequency as in (Demonstration 25 from Bregman & Ahad, 1996) This not only increases the AB sequential grouping, this grouping weakens the BC simultaneous grouping, so that the BC spectrum is heard as less rich. Conversely, if the simultaneous grouping of B with C is manipulated by altering the BC asynchrony (Demonstration 26 from Bregman & Ahad, 1996), not only is the fusion of B and C affected, but also the sequential grouping of A with B. As the BC fusion becomes weaker due to greater BC asynchrony, B becomes more available to form a sequential pure-tone stream with A. It is as if A and C were competing to make a connection with B.

The old-plus-new heuristic

Another example of an interaction between simultaneous and successive organization is the old-plus-new heuristic. In natural environments, sounds do not strictly follow one another in time, nor are they exactly superimposed on one another. Instead they are typically overlapped. The old-plus new heuristic uses the moment when a new sound enters a mixture to derive a very good description of the newly entering sound. It works as follows: When a spectrum suddenly becomes more intense or more complex , the auditory system carries out an analysis of this spectrum to determine whether the spectrum just before the change—or one very similar to it—is still present after the change. If so, it treats the changed spectrum as the sum of a continuing old spectrum plus a new sound (hence the name old-plus-new). Since it can compute the difference between the earlier and later spectra, it can derive, with some accuracy, the properties of the new sound. Note that it is not necessary that the earlier spectrum contain only a single sound, as long as no sound within it drops out at exactly the instant that the new sound begins. So the moment of onset of a new sound plays a critical role in identifying it and it would be surprising if the nervous system did not have mechanisms that were specialized for the analysis of sudden increases in spectral intensity or complexity

Explanations

Explanations of stream segregation and other manifestations of ASA have fallen into two categories: functional and neurophysiological. The discussion up to this point has been about the job of the ASA system and which of its properties allows it to carry out its job. What cues does it use? How does it put them together? How does evidence accumulate? How do top-down and bottom-up processes interact? The concept of an ASA system is not actually a theory but a set of concepts and observations that can act as constraints on the form of such a theory.

An example of a physiological theory is one offered by Leo van Noorden and a closely related one proposed by William Hartmann and Douglas Johnson. The basic argument is that for stream segregation to occur, the members of the two streams must activate non-overlapping populations of hair cells in the cochlea of the listener (van Noorden), or, different frequency channels in the cochlea (Hartmann and Johnson). It is argued that the reason that streams can form on the basis of frequency differences is because each frequency has its strongest effect on a different population of hair cells. The reason that streams can form on the basis of differences in ear-of-arrival is that there are different populations of hair cells (or frequency channels) in the cochleas of the two ears.

This theory is attractive because it is simple and makes the peripheral sensing apparatus responsible for stream segregation . If true, it would open streaming to investigation by well-understood physiological techniques. Unfortunately, a number of findings fail to support this theory. The most clear-cut evidence comes from an experiment in which complex tones containing exactly the same spectral components were made to sound different in timbre and pitch by manipulating the phases of their components. When two tones that had different phase relations among their components were rapidly alternated in the galloping pattern described earlier (ABA–ABA–…) each tone formed its own stream, despite the fact that it activated the same frequency channels within the cochlea as the other tone did.

In some trivial sense, the external sense organ is indeed responsible for stream segregation. After all, unless sounds differ in some way at the early receptive levels, the rest of the brain will never hear about it. However, this is far from saying that the neural computation that establishes the separate streams is close to the sensory periphery. One reason to believe that stream segregation, and in general, ASA, is computed higher up in the brain is that the instructions given to listeners, which activate top-down processes, can affect grouping in ambiguous cases. Another is that various auditory differences among the set of incoming sounds (e.g., pitch, spatial location, onset asynchronies) have an combined effect on the segregation of streams. Yet these cues may be provided by different sorts of feature analyzers. Bringing together the “votes” of the different feature analyses has to be done at a level beyond the one at which the features are first detected, and at a level at which the effects of prior learning can play a role (probably in auditory cortex).

While adequate physiological explanations of ASA have not yet been forthcoming, physiological methods have been able to attack questions that are not easily answered using behavioral methods. One such question is whether the creation of auditory streams (the brain’s representation of distinct environmental sounds) out of a sequence of sounds can precede the involvement of attention or requires the participation of attention. It is very hard to solve this problem using behavioral methods, because whenever human listeners have to carry out a task, such as reporting whether a particular set of sounds forms a separate stream, this task focuses their attention on the sounds. Elyse Sussman and her colleagues have studied the involvement of attention in stream formation by recording event-related potentials (ERP) from scalp-mounted electrodes when people are exposed to sequences of sounds while their attention is distracted by a visual task. She has used a component of the ERP called mismatch negativity (MMN), to detect whether the sounds are forming separate streams. The results suggest that at least three separate streams can be formed while the listener is not paying attention to the sounds, but as soon as attention is focused on one stream, the MMN evidence for the existence of the other streams disappears. So at least in some circumstances, stream segregation may be pre-attentive.

The ERP method is particularly well suited for studying auditory organization in persons who are not capable of reporting on their experiences, such as young infants. Do newborns organize sound the way adults do? If they do not, it suggests that the ASA methods may be have to be learned. István Winkler and his colleagues played sequences of sounds to sleeping newborn infants at 2 to 5 days of age. There was clear evidence for auditory stream segregation. This supports the idea that the mechanisms of ASA are primitive, in the sense of being inborn, and can give an initial boost to the infants’ early learning about the important sounds in their environments, preventing them from memorizing the accidental combinations of properties exhibited by fortuitous combinations of sounds.

Auditory scene analysis in other animals

Non-human, just like human animals live in a world where sounds come mixed with others. Yet they must respond only to certain sounds in the mixture. An important motivation for studying ASA in non-human animals is that its physiological basis could be investigated. Accordingly, ASA has been studied in a number of species, including birds, bats, frogs, fish, pinnapeds and insects. However, a word of caution is in order. ASA is an accomplishment, not a mechanism, Even if different animals succeed in partitioning the sound mixture to the extent needed to detect their particular predators, prey, or potential mates, this does not mean that each of them does it by the same mechanisms that the others do or that humans do. It is probable that many species have mechanisms to extract specific features and patterns that are important to the particular type of animal involved, and probably prepare certain actions that are appropriate responses to the detected patterns. Even humans may have them: some theorists have claimed that we possess specific mechanisms for extracting speech patterns from mixtures. Another example is that bats use specific bands of acoustic energy in the echolocation of their prey and have neural mechanisms specialized in the extraction of specific patterns from the echo data.

Macaque monkeys are large-brained animals and might be expected to share the general ASA mechanisms with humans. In one study with macaques, a rapid alternation of tones of two different frequencies, A and B, was delivered to them while recordings were made from their auditory cortexes. In humans, as the rate of alternation of A and B is increased, stream segregation increases. In the macaque auditory cortex, cells that respond best to A also respond a little to B. But as the rate of A-B alternation increases, these A-sensitive cells stop responding to B. It is as if the rate increase had caused stream segregation to occur, so that the cells that responded to A no longer "saw" B. The same finding has also been reported in bats and starlings. Of course, even if this effect really is part of the stream-segregation mechanism, it does not automatically imply that the A-sensitive cell is, by itself, responsible for stream segregation. It may merely be a point at which one of the effects of stream segregation can be detected by an outside observer.

An interesting physiologically motivated neural network model, the ARTSTREAM model of Stephen Grossberg and his colleagues, has attempted to give a foundation for ASA in terms of neural computation. The process is much more complex than the activity of feature-sensitive cells.

It is possible to conceptualize ASA as the binding together of acoustic information of various types, in order to create acoustic objects, be they sequences or single sounds. Remember that more than one acoustic object (or stream) is being formed at the same time. So the brain has to register, for example, that it is the loud sound that has the rich timbre, and the soft sound that has the purer timbre, and not the reverse. It is possible that this binding of the right features to individual sounds could be carried out by temporarily recruiting some cells, activated by all the to-be-bound features, to represent the object as a whole (sort of a Hebbian phase sequence).

However the existence of binding need not imply the convergence of all the information about the acoustic object to a common pool of neurons. This may not be the way "objectness" is encoded in the brain. It has been proposed by DeLiang Wang, following the approach of Christoph von der Malsberg, that individual auditory features activate oscillatory processes in the brain and that their binding occurs when the oscillations representing particular features are driven into synchrony.

Computational auditory scene analysis (CASA)

Wang's theory, implemented in the form of a computational model, is one of many such models, inspired by the findings about human ASA. A new field of research, known as computational auditory scene analysis (CASA), attempts to create computational models of sound segregation. Most such models focus on the strongest cues for the grouping of sensory input: spatial location and harmonicity. However, a beginning has been made by Darryl Godsmark and Guy J. Brown, on a more open system that allows a multiplicity of features to influence the final organization, using a blackboard architecture.

There are important practical benefits in creating effective computational systems for ASA: Computer systems have great difficulty in recognizing speech mixed with other sounds; so a system that segregated sounds in the course of recognizing them would have a greater chance of successful recognition.

Conclusions

Auditory scene analysis (ASA) affects all perceptible features of sound. The perceived loudness, position in space, pitch, rhythm and timbre of sounds all depend on how the sensory input is organized. As we have become increasingly aware of this fact, the topic of ASA has stimulated research in psychophysics, cognitive science, biology, neuroscience, mathematical and computational modeling, speech, hearing science, audio engineering, and the psychology of music.

Further Reading

Alain, C., Reinke, K., He, Y., Wang, C., & Lobaugh, N. (2005). Hearing two things at once: neurophysiological indices of speech segregation and identification. *Journal of Cognitive Neuroscience, 17(5),* 811-818.

Bee, M.A., & Klump, G.M. (2004). Primitive auditory stream segregation: a neurophysiological study in the songbird forebrain. *Journal of Neurophysiology 92_*, 1088-1104.

Bregman, A. S. (1990) *Auditory scene analysis: the perceptual organization of sound*. Cambridge, Massachusetts: MIT Press. (Paperback 1994)

Bregman, A.S., & Ahad, P. (1996) *Demonstrations of auditory scene analysis: the perceptual organization of sound.* Audio compact disk. Montreal: Authors (Distributed by MIT Press)

Cooke, M. P., & Brown, G. J. (1993). Computational auditory scene analysis: Exploiting principles of perceived continuity. *Speech Communication, 13(3-4)*, 391-399.

Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In Moore, Brian C. J (Ed). *Hearing*. (pp. 387-424). xxi, 468 pp. San Diego, CA, US: Academic Press.

Ellis, D. P. W. (1999) Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication, 27(3-4),* 281-298.

Fishman, Y.I., Reser, D.H., Arezzo, J.C., & Steinschneider, M. (2001). Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. *Hearing Research, 151*, 167-187.

Grossberg, S., Govindarajan, K. K., Wyse, L. L, Cohen, M. A. (2004). ARTSTREAM: a neural network model of auditory scene analysis and source segregation. *Neural Networks. 17(4),* 511-536.

Hartmann, W.M., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, 9(2), 155-184.

Hulse, S. (2002). Auditory scene analysis in animal communication. In Slater, P.J.B. & Rosenblatt, J.S. Snowdon, C.T., & Roper, I.J. (eds.). *Advances in the study of behavior, Vol. 31* (pp. 163-200). San Diego, CA, USA: Academic Press.

Klump, G. (2005) How does the hearing system perform auditory scene analysis? In Van Hemmen, J.L., & Sejnowski, T. Jr. (eds). *23 problems in systems neuroscience.* ch. 15, Oxford: Oxford University Press..

Winkler, I., Kushnerenko, E., Horváth, J., Čeponienė, R., Fellman, V., Huotilainen, M., Näätänen, R. & Sussman, E. (2003). Newborn infants can organize the auditory world. *Proceedings of the National Academy of Sciences of the United States of America, 100(20),* 11812–11815.

Moss, C. F & Surlykke, A. Auditory scene analysis by echolocation in bats. *Journal of the Acoustical Society of America, 110(4),* 2207-2226.

Rosenthal, D. F., & Okuno, H.G. (eds.) (1998). *Computational Auditory Scene Analysis*. Mahwah, NJ. USA: Erlbaum.

Sussman, E. (2005). Integration and segregation in auditory scene analysis. *Journal of the Acoustical Society of America, 117(3, Pt 1),* 1285-1298.

Wang, D. (1996) Primitive auditory segregation based on oscillatory correlation. *Cognitive Science*, 20, 409-456.

Wang, D., & Guy J. Brown, G.J. (2006). *Computational Auditory Scene Analysis : Principles, Algorithms and Applications.* Piscataway, NJ, USA: Wiley-IEEE Press.

Yost, W.A. (2004). Determining an auditory scene. In Gazzaniga, M.S (ed.). *The cognitive neurosciences,* 3rd edn, pp. 385-396. Cambridge, MA, USA: MIT Press.

Spectrogram of a mixture of four sound sources: the words, "one, two, three"; a voice singing "da-da-da", a person whistling, and a computer fan.

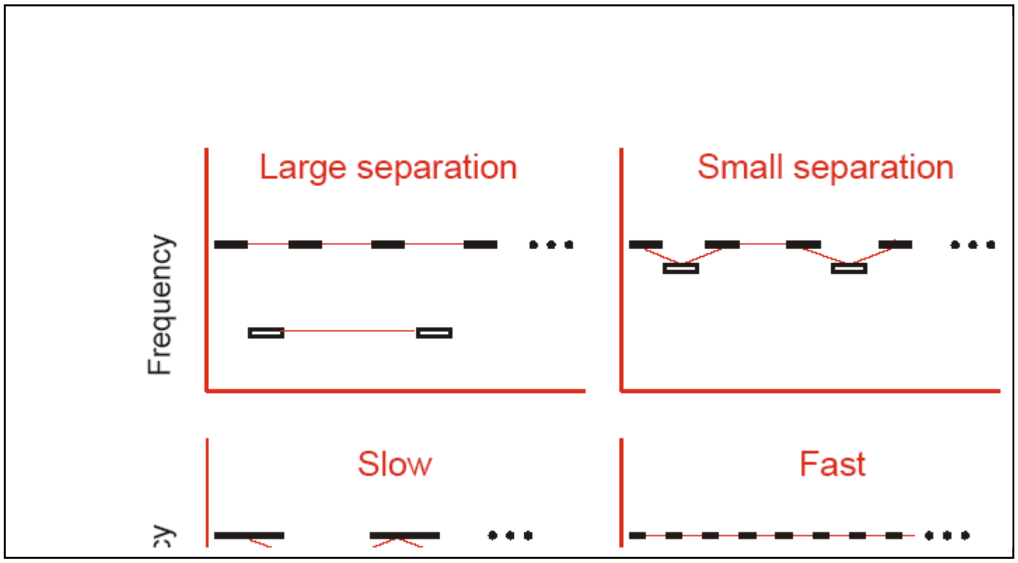

Figure 2. Curves by van Noorden showing the thresholds for stream segregation. For each curve, the area above it shows the region (in frequency x time) in which two streams are heard; in the area below it, only one stream is heard. The temporal coherence boundary (TCB) is obtained when the listener is trying to hear all the tones as a single stream. The fission boundary (FB) is found when the listener is trying to hear the tones in two separate streams. The tonal pattern used was a galloping sequence of the form LHL–LHL–…, where L and H are lower and higher frequency tones respectively.

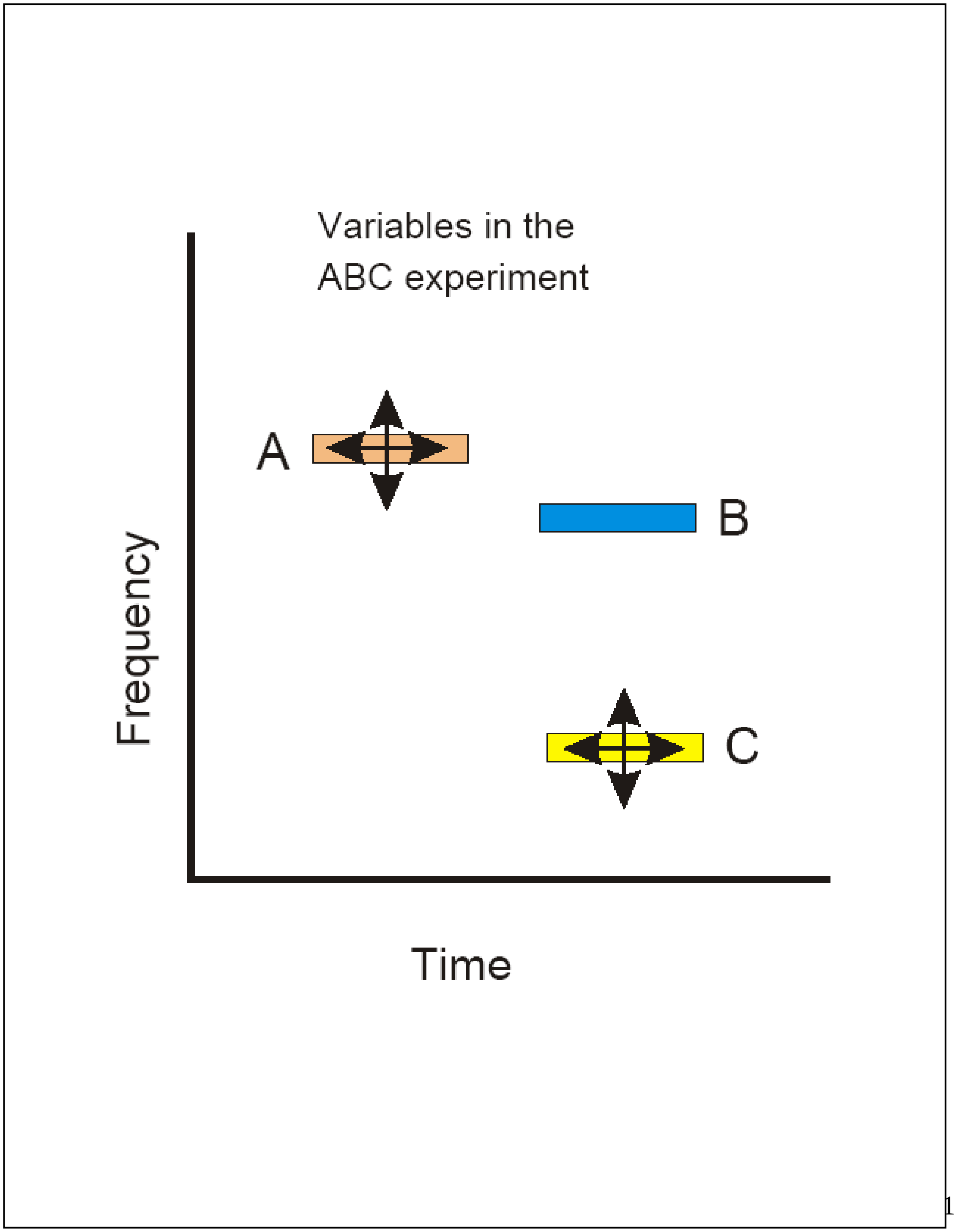

Figure 3. Diagram of an stimulus pattern in which a pure tone, A, alternates repeatedly with a pair of pure tones, B and C. The arrows show that the temporal positions and frequencies of both A and C may be varied. As A comes closer to B in frequency or time, it captures the latter more strongly into a sequential stream. As C comes closer to B in frequency or becomes more synchronous with it, it captures B more strongly into a fused perceptual unit.

Audio Examples

1. A rapid cycle of higher (H) and lower (L) tones in a repeating pattern of HLH–HLH–…(repeated). In the first part, the tones are well separated in frequency. As the sequence speeds up, it appears to split into two separate perceptual streams, one containing repetitions of H, and the other, repetitions of L. In the next part, the tones are close together in frequency. The sequence resists splitting into two streams, even when speeded up.

2. A familiar melody is interleaved with distractor tones from the same pitch range and is completely camouflaged, On successive repetitions the melody is transposed upward in pitch until it forms a separate pitch-based stream and can be identified.

3. A galloping pattern, ABA–ABA–… (repeated), of complex tones of the same pitch but differing in timbre (position of spectral peaks). As it accelerates it splits into separate streams, one for each timbre

4. A galloping of identical noise bursts, in a triplet pattern, NNN–NNN–…(repeated), is segregated by spatial position. First all the bursts come from the center and then the first and third one of each triplet migrate to one side of the head while the middle burst migrates to the other side, leading to strong stream segregation.

5. A pure tone A is followed by two simultaneous tones, B and C. In successive examples, A moves closer in frequency to B, more effectively capturing it into an AB stream, leaving C in its own stream.

6. Similar to the previous example except it is the synchrony of B and C that is varied. As the asynchrony grows, A more effectively captures B into a stream that is distinct from the one containing the repetitions of C.