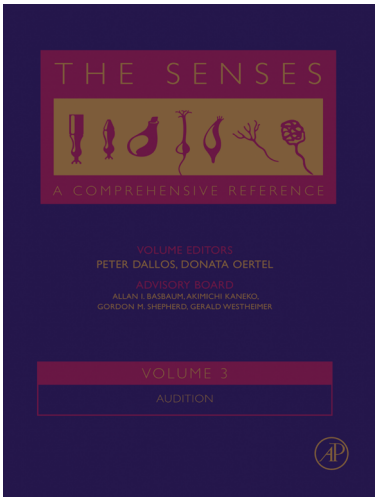

A S Bregman, Auditory Scene Analysis. In: Allan I. Basbaum, Akimichi Kaneko, Gordon M. Shepherd and Gerald Westheimer, editors The Senses: A Comprehensive Reference, Vol 3, Audition, Peter Dallos and Donata Oertel. San Diego:
Academic Press; 2008. p. 861-870.

3.49 Auditory Scene Analysis

A S Bregman, McGill University, Montreal, QC, Canada

Glossary

auditory scene analysis (ASA) Auditory scene analysis is the name for both a problem and a perceptual process. The problem is how to form mental representations of individual sounds from the summed waveform that reaches the ear of the listener. It is also the name of the brain process that accomplishes this result (with a greater or lesser degree of accuracy).

auditory stream A perceptual interpretation, produced by auditory scene analysis, of a mixed auditory input. Parts of the input are heard as separate coherent patterns evolving over time (auditory streams), and interpreted as individual acoustic sources present in the auditory environment (such as a voice, a violin, or a machine). More than one stream can be present at the same time. A stream (in ordinary conversation, we call it a sound or a sequence of sounds) is a perceptual entity, analogous to a visual object, capable of binding a cluster of properties (e.g., one sound is high in pitch, loud, and on the left, whereas the other sound is low in pitch, soft and intermittent, and straight ahead).

event-related potential (ERP) The electrical activity of the brain is measured on the scalp while external events (such as sounds) are presented to the subject. A large number of samples of the electrical recording that correspond temporally with some particular class of external event (such as an acoustic signal of a certain type) are extracted from the record of the brain activity and then aligned temporally using the start of the external event as a zero point, and averaged together. This averaging is thought to eliminate all other influences on the electrical wave except the occurrence of the event. This average wave is the ERP.

mismatch negativity (MMN) The MMN is a negative deflection in the human event-related potential recorded from frontal scalp electrodes, and is thought to originate in auditory cortex. It occurs when a change in a regularly repeating stimulus is detected by the brain, first a series of identical presentations of an auditory stimulus is presented. The repeating stimulus can be as simple as a single sound or as complex as a short pattern of sounds. Then, without a break, a change is introduced in the stimulus. The introduced change can be as simple as a single feature of the stimulus or as complex as an abstract relation between the stimuli of the repeating pattern. MMN is thought to represent the activity of an automatic process for

The Senses: A Comprehensive Reference, vol. 3, pp. 861-870

detecting change, since it occurs whether or not attention is directed toward the repeating stimuli.

objectrelated negativity (ORN) ORN is a component of ERPs that has a negative peak at around 150 ms relative to the onset of an acoustic signal. It occurs when the incoming spectrum is interpreted as two sounds rather than one, and is not sensitive to the involvement of attention.

stream segregation The formation of more than one auditory stream from a mixed input.

streaming A phenomenon in which a rapid sequence of tones or other acoustic units is interpreted as the co-occurrence of two or more auditory streams. The perceptual decomposition of the sequence of sounds into its component streams takes time to evolve.

3.49.1 The Problem of Mixtures

Most studies on perception, either by humans or other animals, have been done in very simplified setups, especially in the area of auditory perception. In a typical experiment in this area, listeners have been presented with an auditory signal in a background of silence, or perhaps noise, and required to judge (or, in the case of animals, to respond to) something about it: whether the signal is present, or whether it is different from a second signal in some quality such as loudness, pitch, or timbre. In the most complex case, speech signals, they may be asked to recognize the speech. This approach has given us what might be called the psychophysics of the isolated stimulus.

However, the acoustic signals we encounter in the natural world do not occur in isolation. The waves emanating from different events co-occur and overlap in time. A recording of a busy household might register a conversation, a child playing a game in an adjacent room, the clamor from a television set, and the mumble of an air conditioner, all overlapped in time. In the world of nonhuman animals in their natural environments, most of the signals are quieter than in a human household. However, these low-level signals are many: the wind, the flutter of birds' wings, insects buzzing, the chatter of small mammals, and the footsteps of larger ones. The symphony of the field is as complicated as that of the household, but with the volume turned down. In both cases, in order for listeners to recognize the signals individually, they must be able to undo the acoustic mixture. The brain process that accomplishes this has been labeled auditory scene analysis (ASA) (Bregman, A. S., 1984; 1990/1994). The critical importance of ASA can be understood by considering what would happen if the process made an error. Syllables from two different voices might be joined to make a single word. A cough in the audience might be heard as part of a piece of music. For a small animal, the noises made by the approach of a predator might simply contribute its qualities to the sound of a waterfall.

For clarity, in this discussion, we will use different terms to refer to the physical acoustic signals from those used for the mental structures and qualities to which they give rise. In the physical category, we will refer to acoustic sources and acoustic signals, or just signals, and, in the context of an experiment, stimuli. These will have properties or features, such as fundamental frequency, amplitude, spectral shape, etc. By an acoustic source, we mean any physical object or process that generates an acoustic signal by creating audible vibrations. By properties, or features, we mean physical properties (to be contrasted with perceptual qualities). Examples of acoustic sources include a person speaking, a violin playing, the wind howling, and a bird singing.

In the mental category, we will refer to sounds (not signals). These will have qualities (not properties), such as pitch, loudness, timbre, etc. So in our terminology, a signal has no pitch until the brain makes one or more sounds out of it, and assigns pitches to some of them. Timbre is a quality of a sound, not a property of a spectrum. Are these distinctions simply the product of an academic obsession with definitions? They seem to be so when we consider only the psychophysics of the isolated stimulus. However, they become essential when considering mixtures, where the qualities that emerge from the analysis depend on the particular parsing that the auditory system gives to the incoming spectrum, rather than on the way that experimenters describe the features of the physical signals that they have created. The perceived qualities of timbre, pitch, loudness, rhythm, etc., emerge from the way ASA allocates energy to individual auditory streams. A spectrum, in itself, does not

have a timbre, only a spectral composition; it may, for example, furnish the energy for two concurrent sounds, each of which has its own timbre.

Because the extracting of separate sounds usually goes on quite automatically in our own auditory systems, it is hard for us to realize how profound an accomplishment it really is. But consider the problem in physical terms. The time-varying acoustic pressure wave that reaches each of our ears is the arithmetic sum of the pressure waves emanating from the individual events that produced them. The individual events overlap in time, starting and stopping asynchronously, shaping the summed pressure wave. Yet what is needed for an appropriate response to the events around us is not a perceptual description of this happenstance mixture, but separate descriptions of the individual acoustic sources. Unfortunately, examining a short stretch of the waveform, or even its spectrum cannot tell us how the mixture was formed, since there are an enormous number of ways of decomposing it. The problem can only be solved by examining a longer window of time, analyzing the changes that occur over time, making use of some *a priori* constraints on the plausibility of particular decompositions.

3.49.1.1 How the ASA Problem Can Be Solved

We can see some of these constraints by looking at the spectrogram shown in Figure 1, of a mixture of the words, 'one, two, three', the sung syllables, 'da-da-da', a man whistling, and a computer fan. The spectrogram of the mixture is roughly what you would see if you made a spectrogram of each of these acoustic sources on separate pieces of clear plastic and superimposed them. In the audio recording from which this spectrogram was made, the individual signals are not hard for the ear to pick out.

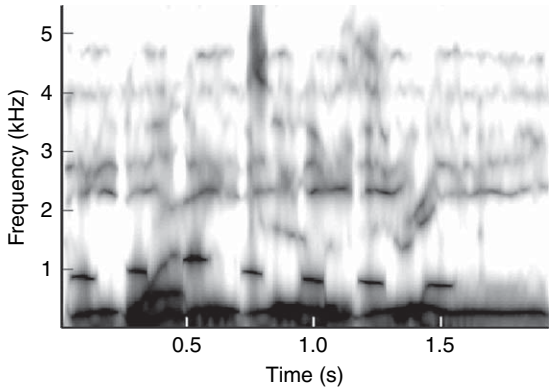

Figure 1 A spectrogram of a mixture, including the spoken words, 'one, two, three', a person singing 'da-da-da', a person whistling, and a computer fan. Adapted from Bregman, A. S. and Woszczyk, W. 2004. Controlling the Perceptual Organization of Sound: Guidelines Derived From Principles of ASA. In: Audio Anecdotes: Tools, Tips and Techniques for Digital Audio (*eds.* K. Greenebaum and R. Barzel), Vol.1, pp. 35–64. A K Peters.

The problem of using the summed spectrum to derive perceptions of the component sounds can be viewed as a problem in perceptual organization. Various parts of the energy present in this spectrum have to be allocated to form one percept and segregated from other groupings of energy that will be used to create other percepts. The difficulty is that the energy from the individual sound sources overlaps in frequency and time, and an individual frequency component may result from the summing of components from two sources. In Figure 1, we can see two dimensions of the problem, represented by the horizontal and vertical dimensions of the spectrogram.

In the horizontal dimension, listeners must group together the energy that came from the same acoustic source over time, a process called sequential organization (Bregman, A. S., 1990/1994, chapter 2). The organization across time does two things: (1) it forms each individual 'da' syllable from the moments of energy that compose it, giving it a beginning, a duration, and an end, thus making it into a temporal unit, and (2) it groups the individual 'da' syllables into a 'da-da-da' melody.

In the vertical dimension, the ASA process must allocate all the energy present at the same time to one or more concurrent sounds, each of which has emanated from a distinct acoustic source. This is called spectral organization or simultaneous organization (Bregman, A. S., 1990/1994, chapter 3).

3.49.1.1.1 The auditory stream

The mental representation of a distinct acoustic source, created by ASA, is called an auditory stream. Any acoustic input may yield a number of concurrent streams. Examples of streams include the mental representations of: (1) an individual voice, (2) a sequence of footsteps, (3) an approaching car, and so on. A stream in audition, like an object in vision, is a mental entity that is intended to correspond to an individual entity in the world, although its activities may unfold over time. Its role in perception is to act as a structure to which perceptual qualities can be bound. We say that a certain sound (the name we typically give to a stream) is loud, or nearby, or high in pitch, or voice-like, etc. There can be a number of

concurrent sounds (streams), each with its own qualities.

3.49.1.1.2 Regularities in acoustic mixtures

ASA takes advantage of regularities that cut across environments regardless of the provenance or meaning of the signal. For example, the various frequency components arising from an individual acoustic source (such as a voice or instrument) tend to start and stop at the same time. While not true without exception, this statistical regularity holds in deserts, jungles, households, in music, speech, and on Mars. The use of such general regularities by the auditory system is always useful. It is plausible that over evolutionary time, the auditory system has incorporated processes into its hard wiring that detect the general cues that signal the presence of an independent acoustic source within a mixture. These can be called primitive or built-in processes of ASA. One would expect them to be found even in young infants and in nonhuman animals.

However, the human listener is also capable of learning many regularities that apply in narrower environments. These probably include the phonetic components of one's own language and the articulations that give rise to them, the vocabulary of one's language, musical keys and scales, the sound qualities of various machines, the voice of one's own child, and so on. Humans operate in so many different environments that we possess a robust system for learning acoustic patterns, which can provide help in decomposing mixtures.

On the other hand, some nonhuman species, such as bats, operating in very specialized sonic environments and having a vital need to respond quickly to special types of acoustic signals, undoubtedly possess innate mechanisms in their perceptual systems that detect these signals even in mixtures, using the specific structure of these signals, rather than just the general regularities found in most acoustic mixtures.

3.49.1.2 Grouping of Acoustic Energy Based on Acoustic Regularities

Let us consider how the auditory system uses the general regularities to achieve sequential grouping. Going back to Figure 1, if we observe a sequence formed of spectral patterns that resemble one another, it is reasonable to assume that they may emanate from the same source. This assumption is based on the statistical regularity that most sounds change slowly if the proper timescale is chosen. The strategy derived from this regularity is: 'if patterns of spectral energy, not too far apart in time, resemble one another, they should be assigned to the same auditory stream'. This resemblance can be in terms of fundamental frequency, spectral center of gravity, spectral shape, intensity, temporal dynamics, spatial position, and perhaps other acoustic features.

Spectral or simultaneous organization also makes use of general regularities. Again we can use Figure 1 to understand them. We can see many examples in the spectrogram where energy at different frequencies starts and stops at the same time. This reflects the general fact that the components of a single sonic event tend to start and stop together. Accordingly, one of the principles used by ASA to organize simultaneous components is: 'if energy in different frequency regions or in different parts of space change in amplitude in the same direction at the same time, assign these to the same auditory stream'.

The justification for these principles and others like them derives from statistical regularities in our acoustic world. It is probable that successive wave patterns arising from the same source will resemble one another, though their properties may change slowly over time. This is not true without exception, but it is generally true, especially if you consider sufficiently short periods of time. This regularity justifies the principle of grouping a pair of signals A and B, which are close together in time and similar in properties into the same auditory stream and rejecting another dissimilar signal, C, that may occur between A and B into a different stream.

Here is another example. Many natural vibrations, including the human voice, animal calls, and insects flying, are harmonic. That is, they are formed of a fundamental frequency and many harmonics of that fundamental. The ASA system can use this as a constraint which favors the grouping of simultaneous components as parts of the same perceived sound if they appear to be harmonics of the same fundamental. Again, a regularity in the world becomes a constraint on the process of decomposing the acoustic input.

A final example has been called the old-plus-new heuristic (Bregman, A. S., 1990/1994, pp. 14, 222–227). It derives from the fact that unrelated signals rarely stop and start at the same time. Instead they overlap in time. As a signal starts, its spectrum is added to the ongoing spectrum resulting from already-present signals; so the spectrum becomes more complex and has a greater intensity.

This fact justifies the following rule in decomposing spectra: 'if a spectrum quickly changes so that it becomes more intense and more complex, do an analysis to determine whether the spectral components that were there just prior to the change are still present after the change. If they are, treat them as a continuation of the prechange spectrum and subtract them from the mixture. Treat whatever is left over as a new sound that has entered the mixture and ascertain its properties'. The moment of entry of the new sound into the mixture is an unparalleled opportunity to discover its properties since we can compare the spectrum just before and just after its entry.

3.49.1.2.1 Cooperation and competition in ASA

Despite the fact that these strategies capture important regularities in the world, and can serve as constraints in decomposing the mixture, each of them can be unreliable. For example: (1) occasionally the properties of the signal coming from a single acoustic source can change rapidly instead of slowly – think of yodeling; (2) the separate spatial origins of parts of the acoustic mixture can be unavailable if the signals are coming around a corner; or if we are listening to a monophonic broadcast or recording; and (3) the synchrony or asynchrony of onset of energy in different parts of the spectrum may be blurred in a highly reverberant environment. However, this unreliability can be dealt with by not depending too much on any individual regularity, but allowing them to collaborate and compete in determining the organization that emerges. If a process analogous to a voting system is used, so that the perceptual grouping is controlled by the greatest number of votes supplied by different strategies, the reliability of the system as a whole can be improved.

It is not just that different regularities (such as spatial proximity and fundamental frequency) compete with one another to control the parsing. Even the use of individual regularities such as separation in a frequency–time space may involve competition. Consider Figure 2. Tone B follows tone A at a certain distance in frequency and time. The question is whether A and B, as two tones in a four-tone repeating cycle, will be placed in the same stream or in different ones by ASA processes (Bregman, A. S., 1978). The answer depends on the properties of the other tones that are present. In the left panel, A and B are far from tones X and Y in frequency; so ASA isolates A and B in their own stream and X and Y in a second. The dashed lines show the resulting

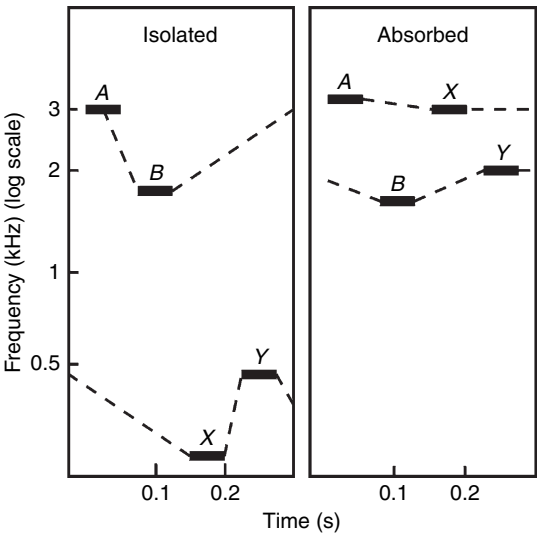

Figure 2 The stream membership of tones A and B, separated by a fixed frequency and time, depends on the context of other tones. The dashed lines show the streams that are formed. In the left panel, A and B are isolated in frequency from X and Y; so each set forms its own stream. In the right panel, X absorbs A into one stream and Y absorbs B into a second. Adapted from figure 2.27 in Bregman, A. S. 1990/1994. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press (paperback 1994), with permission of MIT Press.

streams. The listener hears a repeating sequence AB – – AB– – ···, where the dashes represent the durations of X and Y, which appear as silences in the AB stream since they have been allocated to a stream of their own. In a lower frequency range, the listener hears – – XY – – XY···. The right panel shows how A and B can be absorbed into separate streams. Here X and Y are in the same frequency region as A and B, with X close to A, and Y close to B. Under these conditions, A and X form one stream, with the rhythm A – X – A – X ···, and B and Y form a second, with the rhythm – B – Y – B – Y···. The separation between A and B in frequency and in time are the same in the two panels. If there were a fixed separation that determined whether two sounds were assigned to the same stream, the grouping of A with B should be the same in both conditions. We can see, instead, that the frequency proximities compete, with the ASA system finding the tightest clusters that it can in the frequency–time space. Our ability to draw a picture that predicts the auditory grouping from the visual grouping encourages us to think that there are strong similarities between the grouping processes in vision and audition (Handel, S., 2006), many of which

seem to follow the principles set forth by the Gestalt psychologists (Koffka, K., 1935).

3.49.1.2.2 Validity of principles in the natural world

Simplified laboratory paradigms have been set up to study ASA. On the whole, involve the setting up of simplified stimulus patterns that can be perceived in two different ways depending on the perceptual organization, and then varying the features of parts of the stimulus pattern to see the effects on the perceived sounds. However, since the goal is to come to conclusions that are valid in the natural world, the relevance of the laboratory simplification has to be established by verifying whether the principles discovered in the laboratory are also found in the real world. To some extent this has been done by determining whether the discovered principles can be applied to music (Huron, D., 2001), to speech (Darwin, C. J., 1997, 2001), and to auditory displays (Shilling, R. and Shinn-Cunningham, B., 2002; see also Bregman, A. S., 1990/1994, chapters 5 and 6).

3.49.2 Sequential and Simultaneous Aspects of ASA

3.49.2.1 Sequential Organization

One paradigm for the experimental simplification of sequential organization is the phenomenon of stream segregation, sometimes called streaming (Bregman, A. S., 1990/1994, chapter 1) or fission (Van Noorden, L. P. A. S., 1975). Two tones of different frequencies (A and B) are played in a repeating cycle, either in the pattern ABABAB. . . , or the pattern ABA – ABA– · · · , where the gap between successive triplets is the same duration as the B tone. The latter is sometimes referred to as a galloping pattern because of its distinctive triplet rhythm. If the sequence is played rapidly, say 100 ms per tone, and if the frequencies of A and B are well separated, the sequence seems to split into two perceptual streams, a high one and a low one, each with its own timing pattern. In the case of the simple alternating ABAB. . . sequence, one hears an A stream, A – A – · · · , and a B stream, – B – B · · · , each with a rhythm at half the rate of the original integrated sequence. (Bregman, A.S., and Ahad, P. A., 1996; Van Noorden, L. P. A. S., 1975). In the case of the ABA – ABA– · · · pattern, the galloping rhythm disappears from perception and is replaced by two streams, one with the rhythm – A – A – · · · , and a slower one with the rhythm – – B – – B – – · · · In each case, the dashes represent the tone or tones lost to the other stream.

3.49.2.1.1 Main findings

The probability of segregation into two streams increases both with the time interval between the offset of one tone and the onset of the next tone of the same frequency, for example, the A-A interstimulus interval (Bregman, A. S. *et al.*, 2000) and with the frequency separation between the A's and the B's. One would think that these two facts would allow us to create a simple law relating seperation in frequency and time to stream segregation. However, segregation as a function of frequency and time is not the same for different patterns. For example, it is different for the two patterns described above, and may be different for different listeners (Van Noorden, L. P. A. S., 1975). However, the qualitative findings show that when acoustically different sounds occur in a rapid sequence, if some of the sounds are similar or identical, these will form their own stream.

The acoustic difference in the preceding example was in the frequency of pure tones. However, other differences have been shown in the laboratory to cause separate streams to emerge in rapid sequences of tones. These include differences in rise times, point of spatial origin, intensity, and the abruptness of transition between successive tones (smooth transitions favor integration of all tones into a single stream). Among complex tones, important differences include the repetition rates of their waveforms (perceived as pitch) and their spectral compositions (perceived as timbre). In sequences of bandlimited noise bursts, differences in their spectral centers on a logarithmic frequency scale (perceived as height) are important.

When substreams are created by the grouping of subsets of tones or noises, there are many effects on perception. For instance: (1) it is easier to detect the presence of a pattern of tones that is formed by the elements of one stream than one that crosses streams, (2) melodies and rhythms are formed within auditory streams, (3) judgments of timing are more precise when the time to be judged relates two sounds in the same stream then when the sounds are in different streams, (4) the continuity of synthetic speech is lost if you suddenly change the pitch of the voice, or (5) the perceived loudness of a signal can be reduced if part of its energy is interpreted by the old-plus-new heuristic as belonging to a previous sound (even if the signal is a pure tone).

3.49.2.2 Simultaneous Organization

3.49.2.2.1 The ABC paradigm

One of the simplest experimental paradigms for studying the grouping and segregation of simultaneous signals has been the ABC paradigm (Bregman, A. S. and Pinker 1978). The stimulus pattern is shown in Figure 3. We see a pure tone, A, followed by a pure tone, B, not far in frequency from A. Tone B is accompanied by another tone, C. The pattern is repeated over and over as a cycle. Its interpretation is ambiguous. B is sometimes heard as a second pure tone, grouping with A to form a sequential stream of pure tones, so that with the cyclic repetition, we hear A – B – A – · · ·. At the same time we hear a second stream, derived from C alone: – – – C – – – C· · ·. The arrows in the figure show the experimental variables. In different conditions, A may be closer or further from B in either frequency or time (Bregman, A.S. and Pinker, S., 1978; Rappold, P. W. *et al.*, 1993). As A gets closer to B, it is much more likely to capture it into an AB stream, leaving C in its own stream. The frequency and temporal position of C can also be changed in different conditions. The greater the asynchrony of onset or offset B of C, the more likely they will be allocated to separate streams. In the case of B, this means entering into an AB stream.

When B and C are both pure tones, the harmonic relation between them (e.g., whether B is exactly an octave away from C or not) has little effect. However, some experimental stimuli have been constructed in which C is not a pure tone like B but a tone with many harmonics. B and C may be aligned so that B serves as one of the harmonics in the complex tone C (that harmonic being otherwise missing from C's spectrum). If A and B are of the same frequency, B can be captured out of the C spectrum into an AB stream (Van Noorden, L. P. A. S., 1975, section 3.3). However, B is easier to capture when it is a little higher or lower in frequency than its ideal harmonic position (Hartmann, W. M., *et al.*, 1990). When it is a true harmonic of C, it seems to better able to resist the tendency to be captured by A into an AB stream. This illustrates the role played by harmonic relations in holding a set of spectral components together to form one coherent sound. Strangely enough, it is not only harmonic relations that integrate B with the components of C. Even if C is a nonharmonic set of components, but has a regularity in the spacing of its components (e.g., frequencies of 230, 430, 630, 830, . . . Hz), if B fits into this regular pattern, it seems to make it harder for it to be captured by A (Roberts, B. and Brunstrom, J. M., 1998; 2001). Why should ASA respond to this type of regularity? It is hard to think of a spectrum in nature whose components are not multiples of a common fundamental, but are nonetheless spaced regularly in some other way. The strongest possibility is that a regular, but unnatural, spectrum may trick a mechanism that has been evolved to integrate natural, harmonic components.

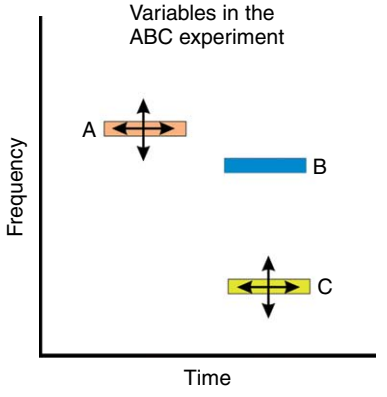

Figure 3 Stimuli for the ABC experiment. A, B, and C are pure tones. The pattern is repeated over and over as a cycle. The arrows show that both A and C may be varied in frequency or time. Adapted from figure 1.16 in Bregman, A. S. 1990/1994. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, (paperback 1994), with permission of MIT Press.

The ABC experiment shows that proximity in frequency and time tend to integrate a sequence of tones as parts of the same stream, just as they did in the galloping pattern of the stream segregation paradigm. It also shows that concurrent tones are better integrated into a single coherent sound when their onsets and offsets are synchronized., and when they fit into a set of frequency components that form a regular pattern in the frequency domain (the most important being the harmonic series).

3.49.2.2.2 Other findings

Other research has shown that simultaneous frequency components are somewhat more likely to be audible as separate sounds if they come from the different locations in space (Divenyi, P. L. and Oliver, S. K., 1989). This is, however, a weak effect when compared to those of asynchronous amplitude changes or the violation of harmonic relations.

Even in the integration of simultaneous components, the frequency separation of the components plays a role. This can be seen using narrow band noise bursts. The greater their frequency separation, the less likely they are to be integrated and heard as a single sound (Turgeon, M. *et al.*, 2002).

The perceptual effects of simultaneous organization are many. Whereas sequential integration forms auditory streams, the integration and segregation of simultaneous components determines which sounds are heard and their perceptual qualities. The number of sounds, their pitches, timbres, musical chord qualities, and even spatial locations depend on the organization of concurrent parts of the spectrum.

3.49.3 Brain Recording and the Role of Attention in ASA

There have been two major projects in the study of ASA using a biological approach. One has had the goal of clarifying the involvement of attention in ASA by recording the electrical activity of the brain that can be measured on the scalp. Is ASA, as claimed by (Bregman, A. S., 1990/1994) a preattentive process, or is attention involved? The advantage of brain recordings over purely behavioral testing is that it is easier to study the involvement of attention in ASA. The difficulty with trying to answer this question with behavioral research is that as soon as subjects have a task to perform with the acoustic material, they pay attention to it. With electrical recording, the subjects can be ignoring the acoustic input.

One approach has been to employ the mismatch negativity (MMN) component of auditory eventrelated potentials (ERPs) as a tool. After a series of repetitions of an unchanging short tonal pattern at a fixed rate, if some property of the pattern changes (e.g., the order of the tones in it) the ERP includes a negative component (the MMN), evoked by the detection of the change (Sussman, E. *et al.*, 1998; 1999). It is believed that this component reflects preattentive processing of the signal by the brain (Näätänen, R., 1992). MMN is found even if the subject's attention is engaged by a visual task (Näätänen, R., *et al.*, 2001). It has been possible to set up a paradigm in which the presence (or absence) of the MMN indicates that a sequence of tones has been segregated (or not segregated) into two or more auditory streams (Sussman, E. *et al.*, 1999; Sussman, E. S. *et al.*, 2005). Recent research on this topic has shown that when subjects are performing a difficult visual task and ignoring a sequence that has been made by interleaving three subsequences of tones, each occupying a distinct frequency range, three auditory streams can be formed concurrently. However, when the subjects are asked to focus their attention on one subsequence (e.g., the high tones), the MMN shows that the high-pitched stream is strengthened as a distinct perceptual entity, while the other two become weak or nonexistent (Sussman, E. *et al.*, 2005). There could be two possible explanations of this effect: (1) the ASA process requires auditory attentional resources that are being used to focus on one of the streams; therefore, it can only structure that particular stream; and (2) the absence of MMN is not due to the failure of the ASA system but rather of the MMN generating process: attention to one stream may suppress the MMN in the other streams. It is not yet clear whether either (or neither) of these explanations is correct.

A method has also been designed to use brain recordings to study the coherence of the sound resulting from multiple simultaneous components (Alain, C. *et al.*, 2001; 2002). Human subjects heard complex sounds composed of multiple harmonics, one of which could be mistuned so that it was no longer an integer multiple of the fundamental. With enough mistuning, the mistuned component stands out as a separate sound (Moore, B. C. J. *et al.*, 1986). In some conditions the subjects watched a silent movie at the same time to determine whether the mistuned component could be automatically detected. ERPs were recorded. Whenever a subject heard the mistuned component as a separate sound, a characteristic wave was obtained, first negative, peaking at about 150 ms after the onset of the sound, then positive, peaking at about 350 ms. The early negative wave, called object-related negativity (ORN) by Alain C. *et al.* (2001) was present both when the subjects were attending to the sound or watching the movie. The researchers interpreted the ORN as an index of a process that, without the involvement of attention, automatically decomposes the incoming signal into perceptual groups that can later be identified by higher processes in the brain.

These ERP studies are complemented by the work of Carlyon and his co-researchers, who studied the issue of whether ASA can go on without the involvement of attention, using behavioral tests on normal subjects and on patients with a unilateral attentional deficit (Carlyon, R. P. *et al.*, 2001). The research on normal subjects showed that the buildup of auditory streaming over time was greatly reduced or absent when they attended to a competing auditory task in the contralateral ear. Patients with an attentional neglect of the left side of space showed less stream segregation of tone sequences presented to their left than to their right ears. The researchers concluded that attention was needed for the

streaming phenomenon to occur. Presumably this research would lead one to suspect that ASA, in general, required attention.

Another question is whether attention is an indivisible mental process. The research by Sussman E. S. *et al.* (2005) cited above, suggests that auditory and visual tasks do not draw on the same pool of those particular resources that we label as attention. Performance in an engaging visual task permits three concurrent auditory streams to form; but attending to one of these streams (an exercise of auditory attention) may prevent the others from forming. Auditory attention itself may be divisible; even if ASA involves attentional resources, we still do not know whether unattended streams can be structured (perhaps by allocating a minority of attentional resources to them), in the circumstances of everyday life. It would be very useful to the animal if they could be. Then when attention was switched, it would be switched to a new partly organized auditory object, one for which an object node already existed in the nervous system, and not just to a smear of energy somewhere in the spectrum.

3.49.4 Animal Studies and Innate Processes of ASA

The study of nonhuman animals represents another approach to the biological study of ASA. If there are general regularities in the spectrum of a mixture that are present in all environments, and these regularities tell the listener how to parse it, one would expect that operations that used these regularities would be preprogrammed in the nervous systems of many species. However, one cannot assume that ASA works in the same way in all animals. For example, ASA has been studied in the echolocation of bats (e.g., Moss, C. F. and Surlykke, A., 2002). Some bats emit ultrasonic vocalizations containing frequency modulation (FM) at fairly stable rates (within 5%). They operate in a fairly complex environment in which both they and their prey are moving, and in which other bats are hunting too. Unlike humans, bats control the sounds they use for hunting; so that in order to segregate their own echoes from those of other bats, they can use strategies that go beyond the methods used by humans in ASA. It is argued that an individual bat adjusts the spectrum and sweep rate of its ultrasonic vocalizations so its own echoes will differ from those of its fellow hunters. Moss, C. F. and Surlykke, A. (2002) postulate that "the bat's perceptual system organizes acoustic information from a complex and dynamic environment into echo streams, allowing it to track spatially distributed auditory objects (sonar targets) as it flies." These streams serve the same function as those of humans – packaging a meaningful grouping of sounds (the reflections coming from a single prey animal) – but operate in an entirely different sonic environment. While the differences in how bats and humans achieve ASA is extreme because of bats' use of echolocation, even animals that do not use this method may live in acoustic worlds that are unlike the one humans inhabit. While the outcome of ASA has to be similar to that of humans, the particular cues and grouping principles may have to be different. Nevertheless, if we find a species whose ASA processes resemble those of humans, it should be possible to use physiological methods to investigate the brain structures and processes that humans may use.

3.49.5 Summary

This chapter describes some of the central issues in ASA. Considerable difficulty is involved in converting the incoming mixture of acoustic signals into separate mental representations of distinct sounds, each with its own properties. To parse the incoming spectrum, the human auditory system uses heuristic methods, based on general regularities in the world of acoustic mixtures. A set of such methods is used but no single one is decisive. They collaborate and compete to determine the final perceptual organization into distinct sounds and streams of sounds. There are two aspects of the ASA problem: how to group incoming energy over time, and how to do so over the set of components (or perhaps features) that are present concurrently. These two forms of grouping involve different regularities, but a competition for the simultaneous and successive grouping of a bit of spectral energy can occur until a final interpretation is resolved. Apart from the general regularities that are the basis for these methods, there are particular regularities in limited situations, such as in a particular language, or style of music, or type of machine, which may be learned and employed in addition to general regularities to parse the spectrum into separate overlapping auditory streams.

The role of attention in ASA is still in question and there have been some recent attempts to resolve this issue by means of brain recordings as well as behavioral studies. However, a decisive picture has

yet to emerge. Finally, it is hoped that the ASA concepts may clarify the perceptual activities of some nonhuman animals, and that having a good animal model of ASA may allow physiological explorations of the underlying processes in the central nervous system.

References

Alain, C., Arnott, S. R., and Picton, T. W. 2001. Bottom-up and top-down influences on ASA: evidence from event-related brain potentials. J. Exp. Psychology: Hum. Percept. Perf. 27, 1072–1089.

Alain, C., Schuler, B. M., and McDonald, K. L. 2002. Neural activity associated with distinguishing concurrent auditory objects. J. Acoust. Soc. Am. 111(2), 990–995.

Bregman, A. S. 1978. Auditory streaming: competition among alternative organizations. Percept. Psychophys. 23, 391–398.

Bregman, A. S. 1984. ASA. In: proceedings of the Seventh International Conference on Pattern Recognition, pp. 168–175. Silver Spring, Md. IEEE Computer Society Press (Library of Congress No. 84-80909).

Bregman, A. S. 1990/1994. Auditory Scene Analysis. The Perceptual Organization of Sound. MIT Press (paperback 1994).

Bregman, A. S. and Ahad, P. A. 1996. Demonstrations of ASA. The Perceptual Organization of Sound. (Compact disk and booklet.) Auditory Perception Laboratory, Psychology Department, McGill University. Distributed by MIT Press.

Bregman, A. S. and Pinker, S. 1978. Auditory streaming and the building of timbre. Can. J. Psychol. 32, 19–31.

Bregman, A. S. and Woszczyk, W. 2004. Controlling the Perceptual Organization of Sound: Guidelines Derived From Principles of ASA. In: Audio Anecdotes: Tools, Tips and Techniques for Digital Audio (*eds*. K. Greenebaum and R. Barzel), Vol. 1, pp. 35–64. A K Peters.

Bregman, A. S., Ahad, P. A., Crum, P. A. C., and O'Reilly, J. 2000. Effects of time intervals and tone durations on auditory stream segregation. Percept. Psychophys. 62(3), 626–636.

Carlyon, R. P., Cusack, R., Foxton, J. M., and Robertson, I. H. 2001. Effects of attention and unilateral neglect on auditory stream segregation. J. Exp. Psych. Hum. Percept. Perform. 27(1), 115–127.

Darwin, C. J. 1997. Auditory grouping. Trends Cogn. Sci. 1, 327–333.

Darwin, C. J. 2001. Auditory grouping and attention to speech (keynote paper). Proc. Inst. Acoust. 23, 165–172.

Divenyi, P. L. and Oliver, S. K. 1989. Resolution of steady-state sounds in simulated auditory space. J. Acoust. Soc. Am. 85, 2042–2052.

Handel, S. 2006. Perceptual Coherence: Hearing and Seeing, Oxford University Press.

Hartmann, W. M., McAdams, S., and Smith, B. K. 1990. Hearing a mistuned harmonic in an otherwise periodic complex tone. J. Acoust. Soc. Am. 88, 1712–1724.

Huron, D. 2001. Tone and voice: a derivation of the rules of voice-leading from perceptual principles. Mus. Percep. 19(1), 1–64.

Koffka, K. 1935. Principles of Gestalt Psychology. Harcourt, Brace and World.

Moore, B. C. J., Glasberg, B. R., and Peters, R. W. 1986. Thresholds for hearing mistuned partials as separate tones in harmonic complexes. J. Acoust. Soc. Am. 80, 479–483.

Moss, C. F. and Surlykke, A. 2002. ASA by echolocation in bats. J. Acoust. Soc. Am. 110(4), 2207–2226.

Näätänen, R. 1992. Attention and Brain Function. Erlbaum.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, and Winkler, I. 2001. Pre-attentive cognitive processing (primitive intelligence) in the auditory cortex as revealed by the MMN (MMN). Trends Neurosci. 24, 283–288.

Rappold, P. W., Mendoza, L., and Collins, M. J. 1993. Measuring the strength of auditory fusion for synchronous and nonsynchronous amplitude-fluctuating, spectrally disparate narrow-bands of noise. J. Acoust. Soc. Am. 93(2), 1196–1199.

Roberts, B. and Brunstrom, J. M. 1998. Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes. J. Acoust. Soc. Am. 104, 2326–2338.

Roberts, B. and Brunstrom, J. M. 2001. Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch. J. Acoust. Soc. Am. 110, 2479–2490.

Shilling, R. and Shinn-Cunningham, B. 2002. Virtual Auditory Displays. In: Handbook of Virtual Environments: Design, Implementation, and Applications (*ed*. K. Stanney), pp. 65–92. Lawrence Erlbaum.

Sussman, E. S., Bregman, A. S., Wang, W. J., and Khan, F. J. 2005. Attentional modulation of electrophysiological activity in auditory cortex for unattended sounds within multistream auditory environments. Cogn. Affect. Behav. Neurosci. 5(1), 93–110.

Sussman, E., Ritter, W., and Vaughan, H. G. 1998. Attention affects the organization of auditory input associated with the MMN system. Brain Res. 789(1), 130–38.

Sussman, E., Ritter, W., and Vaughan, H. G. 1999. An investigation of auditory stream segregation using event-related brain potentials. Psychophysiology. 36, 22–34.

Turgeon, M., Bregman, A. S., and Ahad, P. A. 2002. Rhythmic masking release: Contribution of cues for perceptual organization to the cross-spectral fusion of concurrent narrow-band noises. J. Acoust. Soc. Am. 111(4), 1819–1831.

Van Noorden, L. P. A. S. 1975. Temporal coherence in the perception of tone sequences. Doctoral dissertation, Eindhoven University of Technology, Eindhoven, The Netherlands.

Further Reading

Divenyi, P. 2004. Speech Separation by Humans and Machines. Springer.

Relevant Websites

http://www.psych.mcgill.ca – Department of Psychology, McGill University. ASA

http://www.ebire.org – East Bay Institute for Research and Education.

http://www.aecom.yu.edu – Dr. E. Sussman, Albert Einstein College of Medicine.