Auditory Scene Analysis

Albert S. Bregman
Department of Psychology
McGill University
1205 Docteur Penfield Avenue
Montreal, QC
Canada H3A 1B1
E-mail: bregman@hebb.psych.mcgill.ca

To appear in N.J. Smelzer & P.B. Bates (Eds.)
International Encyclopedia of the Social and Behavioral Sciences. Amsterdam:
Pergamon (Elsevier).

Abstract

Auditory scene analysis (ASA) is the process by which the auditory system separates the individual sounds in natural-world situations, in which these sounds are usually interleaved and overlapped in time and their components interleaved and overlapped in frequency. ASA is difficult because the ear has access only to the single pressure wave that is the sum of the pressure waves coming from all the individual sound sources (such as human voices, or foot steps). To construct a separate mental description for each source, ASA must analyze the incoming signal by the use of heuristic processes. These are based on regularities in the incoming sound that result from the fact that it is really the sum of some underlying sounds. For example, the incoming sound is built up of many frequency components. If a set, A, of these frequency components begin together at exactly the same time, whereas another set, B, all begin together, but at a different time from those of A, then the components of A will be considered as parts of one sound, and those of B as parts of a different sound. This grouping is justified by a regularity in the environment, namely that all the frequency components of a single sound tend to start at the same time. Other heuristic analyses are based on different regularities in how sound is produced. The grouping of these components can determine the perceived pitch, timbre, loudness, and spatial position of the resulting sounds.

130 Sensation and Perception. Auditory Scene Analysis

Sounds are created by acoustic sources (sound-producing activities) such as a horse galloping or a person talking. The typical source generates complex sounds, having many frequency components. [Its *spectrum* (pl. *spectra*) consists of the frequency and amplitude of every pure-tone component in it.] In a typical listening situation, different acoustic sources are active at the same time. Therefore, only the sum of their spectra will reach the listener's ear. For individual sound patterns to be recognized – such as those arriving from the human voice in a mixture – the incoming auditory information has to be partitioned, and the correct subset allocated to individual sounds, so that an accurate description may be formed for each. This process of grouping and segregating sensory data into separate mental representations, called *auditory streams,* has been named "auditory scene analysis" (ASA) by Bregman (1990).

The formation of auditory streams is the result of processes of sequential and simultaneous grouping. Sequential grouping connects sense data over time, whereas simultaneous grouping selects, from the data arriving at the same time, those components that are probably parts of the same sound. These two processes are not independent, but can be discussed separately for convenience.

1. Sequential Grouping

Sequential grouping is determined by similarities in the spectrum from one moment to the next (Bregman, 1990, Ch. 2). The *streaming phenomenon* (Fig. 1) provides a much-simplified example of sequential grouping. A repeating cycle of sounds is formed by alternating two pure tones, one of high frequency (H) and one of low frequency (L), of equal duration. The cycle begins slowly – say three tones per sec– and gradually speeds up to 12 tones per sec. At the slower speeds (Fig. 1a) listeners hear an up-and-down pitch pattern and a rhythm that contains all the tones. At the faster speed (Fig. 1b), they hear two streams of sound, one containing only the high sounds and a second containing only the low ones. It appears that there has been a perceptual grouping of the tones into two distinct streams (van Noorden, 1982). Intermediate speeds may lead to ambiguous organizations in which the listener can consciously control whether one or two streams are heard.

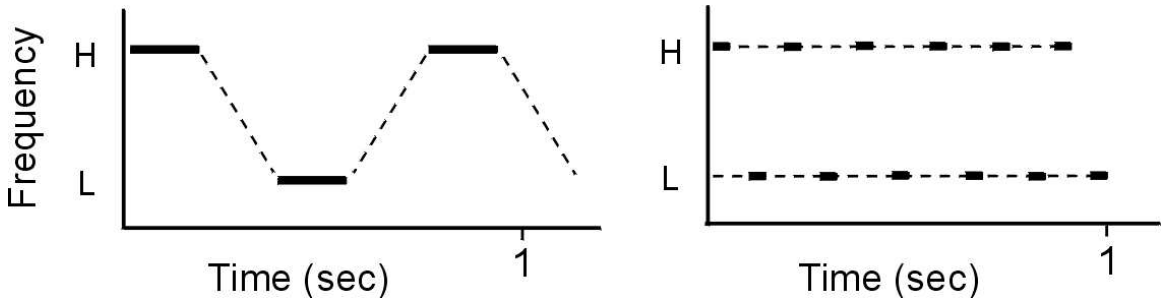

Figure 1

A cycle of alternating high (H) and low (L) pure tones. In Fig 1a (i.e., Panel a), the rate is 3 tones per sec, in Fig 1b, it is 12 tones per sec. Dashed lines show the perceptual grouping.

The Gestalt psychologists described analogous phenomena in vision: sensory components near one another are perceptually grouped into clusters. The separate high and low streams in the streaming phenomenon can be understood as the auditory version of such clusters. To show why, *d* is defined as the "perceptual distance" between any pair of successive auditory components, *A* and *B*. It is a weighted combination of the differences between *A* and *B* on a number of acoustic dimensions, including those of frequency and time. Sounds tend to group with their nearest neighbors, as defined by *d*. Gestalt psychologists hypothesized that perceptual grouping was competitive. An individual element, *A*, might perceptually group with *B* when all other elements were further away, but if a new element were introduced, which was very similar to *A*, then *A* might group with *it* rather than with *B*.

Streaming can be seen as the result of competitive grouping. First one must imagine that the time and frequency axes have been stretched or compressed until a unit on each of them represents the same perceptual distance. The graphs in Fig. 1 can be considered to be plotted on such axes. At slow speeds (Fig. 1a), the temporal separations are larger than the frequency separations. As a result, on the frequency-by-time surface, the combined separation, *d*, of each tone from the next one of the same frequency, two time steps away, is greater than its separation from the tone of the other frequency, only one time step away. Therefore each tone tends to group with its closest *temporal* neighbors, and a single sequence is perceived, containing all the tones. However, when the sequence is speeded up (Fig. 1b), this reduces the separation on the time dimension so that the *d* between two successive high tones, for example, is less than the *d* between each of these and the intervening low tone. Consequently, tones group best with their

nearest neighbors *in the same frequency range.* This results in the formation of two streams, one high and one low.

Similar contributions to *d* can be made by differences in timbre, in spatial direction from the listener, and in fundamental frequency (Bregman, 1990/1994). Sequential grouping is also affected by the nature of acoustic transitions. When sound A changes its properties *gradually* until it becomes B, the A-B sequence is likely to be heard as a single changing sound. However, when A changes into B abruptly, B tends to be treated as a newly arriving sound, this tendency increasing with the abruptness of the change. B can then be heard as accompanying A or replacing it, depending on whether the spectral components of A remain after B begins.

There are also cumulative effects in sequential grouping**.** If a sequence of the type in Fig. 1 is played at an intermediate speed, at which the grouping is ambiguous, the longer it is heard, the greater is the tendency to perceive separate H and L streams. It is as if the ASA system kept a record of "evidence" from the recent past, and strengthened the tendency to form a stream defined by a narrow range of acoustic properties, when newly arriving frequency components fell repeatedly within that range.

Although most existing research on ASA has been done on simplified examples of grouping in the laboratory, most ASA researchers believe that the same factors affect the perceptual organization of sounds in the natural environment.

The formation of streams has been shown to have powerful effects on perception:

(a) Fine judgements of timing and order are much more easily performed when they involve sounds that are part of the same perceptual stream.

(b) Rhythm and melody also seem to be judged using the set of tones of a single stream.

(c) In synthetic speech, if the fundamental of the voice changes abruptly in the middle of a syllable, a new stream is formed at the point of change. It sounds as if one talker has been replaced suddenly by another. Furthermore, it seems that any quality of a syllable that depends on information on both sides of the point of change will be lost (Darwin 1997).

Sequential integration is not only involved in the grouping of a sequence of discrete sounds as in Fig. 1, but also in the sequential integration of frequency components within a complex spectrum, for example the integration of the speech of a single voice in a mixture of voices.

2. Simultaneous grouping

When sounds are mixed, then if correct recognition is to occur, the auditory system must divide up the total set of acoustic components into subsets that come from different

sources. To achieve this, it uses properties of the incoming mixture that tend to be true whenever a subset of its components has come from a common source. For example, there is a broad class of sounds called "periodic", which includes the human voice, animal calls, and many musical instruments, in which all the component frequencies are integer multiples of a common fundamental. The auditory system takes advantage of this fact. If it detects, in the input, a subset of frequencies that are all multiples of a common fundamental, it strengthens its tendency to treat this subset as a single distinct sound. Furthermore, if the mixture contains two or more sets of frequencies related to different fundamentals, they tend to be segregated from one another and treated as separate sounds. This is an important cue for separation of a single voice from a mixture of voices, and is used by many computer systems for automatic speech separation (*see* Auditory Scene Analysis: Computational Models).

Other cues that tend to identify components that come from the same acoustic source are: (a) *synchrony of onsets and offsets* of components, a cue that is useful because parts of a single sound typically start at the same time (plus or minus 15-30 ms); (b) frequency components coming from the *same spatial location* (the spatial cue is weak by itself, but assists other cues in segregating components); (c) different frequency components having the *same pattern of amplitude fluctuation*; and (d) components that are *close together in frequency*. Cues tend to combine in their effects on grouping, as if they could vote for or against a particular grouping. Furthermore sequential and simultaneous grouping may be in competition, as when a spectral component B may either be interpreted as a continuation or reappearance of a previous sound (sequential grouping) or as a component of a concurrent sound (simultaneous grouping). This competition sometimes takes the form of the "old-plus-new heuristic": If there is a sudden increase in the complexity of the sensory input, the auditory system determines whether it can be interpreted as a new sound superimposed on an ongoing one. If done successfully, this aids in the partitioning of the total incoming sensory data.

The grouping of simultaneous components can affect many aspects of perception, including the number of sounds that are present and the pitch, timbre, loudness, location, and clarity of each. In music, it can effect the salience in perception of "vertical" relations, such as chord quality and dissonance.

There are both primitive ("bottom-up") and knowledge-based ("top-down") aspects of auditory scene analysis. Primitive processes, the subject of most ASA research, rely on cues provided by the acoustic structure of the sensory input. These processes are thought to be innate and are found in non-human animals (Wisniewski and Hulse, 1997). They have been shown to be present in the perception of speech (Darwin and Carlyon, 1995) and of music (Bregman, 1990, Ch 4 for music and Ch 5 for speech). The primitive processes take advantage of regularities in how sounds are produced in virtually all natural environments (e.g., unrelated sounds rarely start at precisely the same time). Top-down processes, on the other hand, are those involving conscious attention, or that are based on past experience with certain classes of sounds – for example the processes

employed by a listener in singling out one melody in mixture of two (Dowling, Lung, and Herbold, 1987).

Computational models of ASA have also been developed (see *Auditory Scene Analysis: Computational Models*). For a general view of the auditory system, see *Sensation and Perception: Auditory Models.*

Albert S. Bregman
Psychology Department,
McGill University

Bibliography

Bregman A S 1990 [1994 Paperback] *Auditory scene analysis: the perceptual organization of sound.* The MIT Press, Cambridge, MA

Bregman A S, Ahad P A 1996 *Demonstrations of auditory scene analysis: The perceptual organization of sound.* (Compact disk and booklet). The MIT Press, Cambridge, MA

Darwin C J 1997 Auditory grouping. *Trends in Cognitive Science* 1: 327-333

Darwin C J, Carlyon, R P 1995 Auditory grouping. In: B C J Moore, (ed.) *Handbook of perception and cognition: Hearing.* (2nd ed.), Academic Press, London: 387-424

Dowling W J, Lung K M-T, Herbold, S 1987 Aiming attention in pitch and time in the perception of interleaved melodies. *Perception and Psychophysics*, 41: 642-656

Van Noorden L P A S 1982 Two channel pitch perception. In: M Clynes (ed.) *Music, mind and brain.* Plenum Press, New York

Wisniewski A B, & Hulse S H 1997 Auditory scene analysis in European starlings (Sturnus vulgaris): Discrimination of song segments, their segregation from multiple and reversed conspecific songs, and evidence for conspecific song categorization. *Journal of Comparative Psychology,* 111(4): 337-350.