

Auditory scene analysis in humans: Implications for computational implementations.

Albert S. Bregman
McGill University

Introduction.

- The scene analysis problem.
- Two dimensions of grouping.
- Recognition errors if auditory scene analysis is done wrong.
- Speech perception suggests a strong role for schemas in auditory scene analysis.

Sequential organization

Cues that favor sequential grouping.

Idea that there is a competitive grouping of sounds according to the "distance" (D) between sounds, where D is some combination of distances along different dimensions.

- Frequency & temporal proximity
- Demo 3 from Bregman & Ahad disk: Galloping rhythm

Which feature of the temporal interval is important?

The interval from the end of one sound to the beginning of the next sound with which it might be grouped. Acts the temporal separation component of D. Rapid sequences enhance streaming because the size of the contribution of temporal distance to D is reduced relative to the contribution of other distances.

Demo: Tone duration. Increased segregation as alternating high and low tones get longer (increased duration) while keeping the stimulus onset interval (tempo) fixed.

- Fundamental frequency and spectral frequency. These are independent features of complex tones and each contributes independently to D.
- Spatial separation - either real, or signaled over headphones by time delays or amplitude difference - contributes to D.
- Differences among sounds in their spectral peaks contribute to D.
- Loudness differences make a weak contribution to D.
- Abruptness of property change in the acoustic sequence makes an independent contribution to segregation. Smoothly changing properties are marks of a single, changing sound.

Effects of sequential organization on perception

- Perception of melody: the melody is formed within an auditory stream,
- Rhythm emerges primarily within segregated streams;
- Judgments of timing are more precise within a perceived stream than across streams;
- Continuity of synthetic speech is lost if you suddenly change the pitch of the voice (Darwin & Bethell-Fox, 1977)
- Perceived spatial location can be affected by its sequential grouping. A sound that is occurring at one ear can be treated as a continuation of a previous monaural sound and prevented from combining with a contralateral sound of the same frequency and phase to create a spatially centered image (Research in progress with Y. Tagami and M. Rapaport).
- Perceived loudness can be affected if a set of components from a current spectrum is treated as merely a continuation of a previous sound (Richard Warren's "homophonic continuity" paradigm).

Simultaneous grouping (grouping of components that overlap in time):

Again we can conceive of a variable D (difference), but this time it applies to subsets of spectral bands or (spectral features) which can be segregated from one another.

Factors that contribute to D (increase segregation).

Different subsets of spectral features (e.g., spectral components):

- Belong to different harmonic series;
- Come from different places in space;
- Occupy distinct frequency regions;
- Show independent changes in loudness (synchronized dynamics);
- Are asynchronous in onset (and to a lesser degree, in offset);
- Undergo asynchronous changes in frequency (Maybe. There is still disagreement about this).

Effects of having grouped parts of the spectrum into separate “sounds”.

a) Each global sound has its own properties, including:

- Pitch;
- timbre;
- loudness;
- perceived location in space.

b) Organizing parts of the spectrum into separate groups reduces their perceptual interaction.

Competition between sequential and simultaneous grouping:

a) the “old-plus-new” strategy.

"When a spectrum gets more complex, try to interpret it as a continuing (old) spectrum to which a new spectrum has added some additional components."

This is the most powerful cue for segregating concurrent sounds. Chris Darwin at Sussex University has shown that it can even segregate a harmonic from a vowel, changing the identity of the vowel.

Demo 34 from Bregman & Ahad disk: Capturing a narrower band of noise out of a wider frequency band by alternating the narrower and wider bands

The old-plus-new phenomenon be viewed as a more pronounced case of asynchrony of onset.

Achievement of stability in the face of fluctuating acoustic evidence:

- by weighing different types of evidence against one another (as if the clues voted);
- by hanging on to an existing interpretation as long as possible.

Issues in the implementation of a system for computational auditory scene analysis (CASA).

Lessons from research on speech perception:

- Formants presented to opposite ears can be perceptually integrated to yield a speech sound, despite being also heard as separate sounds
- The same holds true for "sine-wave speech" (SWS) tones.

Demo 23 from Bregman & Ahad disk: "SWS"

- Vowels synthesized on the same fundamental and then mixed can in some cases be identified even though there is no acoustic cue to put together the right combinations of formants into two vowels.

These facts seem to negate the importance of auditory scene analysis in speech perception.

I would like to propose a set of principles for auditory scene analysis to handle these facts and to make predictions for non-speech signals

1. The default status of an input auditory array is the integration of all of it into a single sound. Unless there is evidence to partition the sound, it remains integrated. This explains why SWS is moderately intelligible, even if the component tones are divided between the ears. There is no competition. Grouping all of them can give the correct phonetic percept. We are currently doing research on mixed SWS signals, where grouping everything can't be effective and auditory scene analysis principles are very important.
2. Schema-based processes sometimes can select from the signal the components they need to satisfy their data requirements. The action of such schemas explains the perception of SWS and also how certain pairs of vowels, such as /ah/ and /ee/, that have been synthesized on the same fundamental and then mixed, can nonetheless be perceived separately.
3. Bottom-up ASA principles do not partition the signal into airtight packages and then hand them on to speech recognizers. Instead, they establish constraints that interact with constraints from top-down processes to converge on a description of sources that maximally satisfies both.

This description calls into question the current goal of most CASA research: to separate the acoustic data into distinct packages and hand each package separately to a speech-recognition program for recognition.

The fact that concurrent perception of both speech and non speech sounds can occur, as when formants are divided between the ears or when presented as tones in SWS, suggests:

- Allocation of input components to perceived sounds is not all-or-nothing.
- The same acoustic component or feature can contribute to the perception of both a speech and a non-speech sound.

This seems to be true of sounds other than speech. E.g., an unpublished experiment of Martine Turgeon, using the "rhythmic masking release" paradigm: a tone had its own rhythm, and at the same time was integrated with others to create another rhythm.

- Intelligibility of SWS suggests that one perceptual option, when bottom-up processes favor segregation, but not very strongly, is to allow integration to occur while some perception of the components is also available (duplex perception).

Lessons from the fact that cues seem to add up in determining perceptual grouping.

- Models should be designed so that there is a parallel activity of the systems that deal with different cues. In natural environments, sometimes a particular cue is useful and sometimes not. The most useful cues in any environment should be able to take over the burden of segregation and grouping in a seamless way.
- In a computational model, each cue-using mechanism should be able to be shut off without affecting the activity of the remaining ones (the system should degrade gracefully).
- Typically a CASA model attempts to test out the use of some particular cue for grouping. The modeler should be able to compare the success rates with and without the mechanism that uses this principle, so as to determine its “incremental” utility. This should be compared under different types of degradation of the signal. It may be that the cue in question has no incremental utility and also degrades just as quickly or more quickly than other cues in the same sets of environments.

Perhaps the most powerful cue to segregation of simultaneous events is the “old-plus-new heuristic”. Models should make more use of it.