Creating Mixtures: The Application of Auditory Scene Analysis (ASA) to Audio Recording

Albert S. Bregman and Wieslaw Woszczyk

1 The Art of Recording: The Creation of Mixtures

The focus of this article will be on the application of principles of auditory scene analysis (ASA) to the art of recording. In order to follow the present discussion, please also refer to "Controlling the Perceptual Organization of Sound: Guidelines from Principles of Auditory Scene Analysis" (see *Audio Anecdotes I*). To summarize, ASA describes how the auditory system sorts out the complex sense data received from a mixture of sounds into separate perceived sound sources, such as distinct voices or instruments. When many sounds are heard at the same time, the brain receives a whole array of sensory features of the signal. To hear the distinct sounds in the mixture, the brain has to create bundles of features that it treats as belonging to the same sound (the process of integration), and treats them as distinct from other bundles that are attributed to different sounds (the process of segregation). There are two types of organization involved: (a) sequential organization—those processes that integrate or segregate the sounds in a sequence, deciding whether they come from a single sound source; and (b) simultaneous organization—those processes that integrate or segregate frequency components or other sensory features that are present at the same time.

Many of the decisions made by the recording engineer are closely related to ASA because they are about the segregation or blending of parts,

where the parts are the different tracks that have been recorded of acoustic instruments and voices, synthetic and sampled natural sounds, effects (transformed sounds), room sounds, ambiances, etc. (see the articles by Dan Levitin, "Instrument (and Vocal) Recording Tips and Tricks" (see *Audio Anecdotes I*) and "How Music Recordings are Made" (page 1).

2 General Guideline

The general guideline from which all the detailed ones follow comes from the definitions of *integration* and *segregation*: to cause sounds to be distinct, strengthen the ASA cues that favor segregation, and to blend them, weaken those cues. We are not implying that ASA principles can tell recording engineers how to pursue their craft. Perhaps, however, by becoming aware of the general principles of perceptual grouping, as they have been uncovered in the laboratory, recording engineers can understand why their methods are successful. It is also possible that ASA principles might supply a framework upon which their craft could be systematized.

3 The Use of Space (Loudspeaker Separation)

An important technique that the audio engineer uses is the spatial placement of the numerous recorded tracks into two, three, five, or more spaced apart loudspeakers. One aspect of the art of mixing is to decide which of the original recorded tracks to mix into the same loudspeaker, which to separate into different speakers, and which to distribute across speakers.

As a crude first step, one could say that if you mix two sounds, A and B, into the same speaker, or into all speakers, they will blend, and if you mix them into separate speakers, they will be perceived as more distinct. However, research has shown that unless A and B differ in other ways, as well as in their spatial locations, their separation in space will do little to segregate them. In other words, if A and B are synchronous in onset, have the same amplitude envelopes, and the same kinds of attacks, spatial separation won't segregate them very well even though they have different pitches and timbres. Therefore, separating two steady-state sounds in space doesn't make them easier to segregate [1]. Spatial separation seems to work by accentuating the segregation that is based on other dissimilarities between the sounds. Happily, in any two sequences of sounds outside the psychophysicist's laboratory, there are many differences between them

from moment to moment: Their amplitude envelopes are rarely exactly correlated. Their attacks don't exactly overlap in time. Their pitches usually don't change in parallel. This is why spatial differences can be used effectively by the audio engineer to segregate signals, especially when they differ in their temporal characteristics, and when their difference in distance or angular separation is large. Similarly, mixing different instruments into the same loudspeaker, or spreading all of them out over a few speakers will contribute to the blending of their sounds.

4 Filtering and Equalizing

We know that the sequential grouping of sounds is affected by their timbres; this means that the ear will be more easily able to follow the same instrument or voice over time, if it has a unique timbre. There are two ways in which timbre might affect perception: (a) by providing features that the listener can track voluntarily over time; and (b) by influencing the automatic, bottom-up grouping of the sounds in a sequence. It is by no means certain that every feature that permits voluntary tracking is also a basis for automatic sequential grouping.

For both these forms of grouping, two of the most important features of timbre are the formant structure of a sound (layout of major peaks in its spectral envelope), and its onset and decay transients. Each musical instrument has a distinctive formant structure, which provides it with an "auditory signature" that can be tracked over time. However, transient structure (onset and offset) as well as dynamic envelope and fluctuations are also very important for the identification and tracking of a sound source. Playing any recorded sound backward preserves its spectral shape, but makes the source difficult to recognize particularly in the case of percussion, bells, piano, and plucked instruments. Audio Example 5 which appears on the accompanying CD-ROM presents four ten-second recordings of different solo instruments played backwards. Can you recognize the instruments? You can find the answers at the end of this article.

It has been demonstrated that when onsets of instrumental sounds are edited out, leaving only the sustained sounds, perception and classification of musical instruments is confused. A cornet, for example, can be mistaken for a violin, a cello for a bassoon, a French horn for a flute. Possibly, because onsets precede the sustained portions and are free of their own reverberation, transients can provide reliable cues in source identification. The nonstationary and resonant nature of musical

sounds makes them very robust carriers of redundant auditory cues—cues that allow them to be blended together or separated by the actions of the musicians.

Other identifying features are "brightness," "sharpness," and "roughness." Brightness and sharpness are qualities of experience that occur when the high partials of a sound are of greater intensity than its low partials (an intensity relation that raises the spectral "center of gravity"). Roughness is an experience that results from the beating of the partials of concurrent sounds at unrelated rates, either because the fundamentals of the two sounds are not in good harmonic relations with one another or because the sounds, in themselves, are inharmonic. Differences in any of these qualities will allow our voluntary attention to separate individual sounds, and their similarities over time will allow the sounds from a single source to be tracked. While not all these properties have been studied in the context of automatic sequential integration, it is known that "brightness" (spectral center of gravity) influences the d (combined difference in properties) that affects sequential grouping. Filters and equalizers can play a role in accentuating differences in brightness and can modify the spectral balance of formants and transients. A gradual high-pass filter will make the sound brighter whereas a gradual low-pass filter will make it duller. Segregating sounds by artificially induced differences in brightness may be undesirable from an esthetic point of view, but filtering a short phrase or two in the music in this way could clean up a muddy stretch of sound. Conversely, bandpass filtering of two sounds with the same filter settings will increase their tendency to blend.

Since each microphone is a filter, and each microphone placement captures a different spectrum of the source, engineers responsible for recording and balance use microphone selection and placement to shape the character and aural identity of the source that will make it distinct from others. Further equalization and filtering will be used to fit, isolate, match, and blend sounds with each other depending on the musical requirements of the mix. The entire process of composing, performing, recording, and mixing involves a careful consideration of which sounds should be fused together and which should be segregated so that the music can communicate its intended purpose clearly and fully.

Filtering can also be employed to bring out the pitch of a particular instrument, A. Since only a few harmonics, particularly the low ones, are needed to define a pitch, if there is a region in the spectrum where the lower harmonics of A are not mixed with those of other instruments, boosting the intensity of this spectral region will strengthen the pitch of A. Finding such a region is easier if A is a bass instrument, whose lower

harmonics are substantially lower than those of the other instruments in the ensemble.

5 Temporal Synchrony

By employing a rubato style, instrumentalists and vocalists take themselves out of exact synchrony with the accompanying instruments (assuming that the latter stay in synchrony with one another). This causes their sound to stand out from the rest of the ensemble. In the recording or post-production process, the recording engineer can achieve the same result by time-shifting the tracks that need to be emphasized relative to the other tracks. A delay of as little as 20 to 30 msec can be effective for an instrument whose tones have abrupt onsets that are well defined. Longer delays will be needed for slower onsets. This technique has to have musical and expressive justification and be used sparingly, perhaps causing the isolated instrument to be sometimes ahead of and sometimes behind the others. Otherwise, the time-shifted instruments will sound "out of time" and outside the context of the music.

Delaying or advancing a track with respect to others during post-production can also be used to synchronize tracks that were recorded out of synchronization with others; this can increase the perceptual integration of the group of instruments when this is desired.

A group of similar instruments (say electric guitars) can be blended into an ensemble when their individual envelopes are trimmed into synchrony using gates or keyed (synchronous) expanders. One of the envelopes is used as a master and is imposed dynamically on the other instruments to align their onsets with that of the master.

6 Spatial and Pitch Motion

The recording engineer can impose common motion in order to achieve greater sense of the ensemble, and create unity out of independent sources. Modern digital and analog processors allow group modulation of gain, frequency/pitch, time delay, and spatial position. For example, all reverberation sources in the mix can be modulated (or gated) by a single source signal, producing gated reverberation. Several sources or the entire mix can be compressed in amplitude by a compressor that imposes common dynamic changes. The result is always increased blend and interdependence of sounds subjected to the commonality of motion. Common

spatial panning of several instruments segregates them out of the mixture and groups them in the unity of motion. Pitch modulation or Doppler modulation, achieved when sources are reproduced via a rotating Leslie loudspeaker or its digital emulation, does the same to impose distinct aural characteristics blending the sounds together. The common spectral side-bands created by modulation are derived from the individual spectra and thus bind the individual sound together.

7 Reverberation

Even adding reverberation to an entire signal, or just a part of it, can affect perceptual organization. It can do so in two ways:

(1) An individual sound with reverberation added to it will stand out from a mixture (from other potentially masking sounds) because of the lengthening effect of reverberation. Reverberant decay sustains the spectral content of the source by delaying and recirculating this signal for as long as the reverberation time is set (on a digital room simulator, for example, the reverberation time indicates the time needed to achieve a 60 dB drop in reverberation level at mid-frequencies). The spectral content of the source is represented in the reverberant sound and is thus available for auditory evaluation because reverberation acts as a temporary (leaky) storage of that sound. This "running reverberation" (following the source closely in time) may help the perceived continuity of a stream of sound, by strengthening sequential integration. The lengthening can especially help the auditory system to more clearly register the pitch of short notes (since the pitch computation takes time) because these notes can still be heard in the reverberation even after a quick decay of the primary source. Each distinct reverberation pattern accompanying each different sound source will help to segregate these sources from one another by providing lengthening and differentiating characteristics (spatial or timbral) to these sounds.

(2) A mixture of sounds combined with a single reverberation derived from this mixture (by sending a number of tracks to the same reverberator) will act to blend and integrate these sounds. This is because the temporal, spectral, and spatial structure of the reverberator—either from the natural one (a room) or an artificial one (a digital room simulator)—will become the common attribute of all these sources. This common lengthening and spatializing of sounds

will enhance their similarity and thus promote their blending and integration. For example, background vocals should use the same reverberation if the intent is to provide a well-blended ensemble sound.

Adding the same type of reverberation to all tracks, or to the final mix, can reinforce the listener's sense that they are all in the same spatial context, lending a kind of esthetic unity to the mix. However, this effect is probably due to higher-level cognitive processes based on learning, and not on the low level, "bottom-up" processes that have been studied in ASA research, and which are believed to be innate.

It is also possible that by adding reverberation to one track (A) and not to another (B), this can help B to stand out against A, since B's attacks will be clearly heard against the smoother sound of A. This effect can be obtained in a weaker form by passing A and B through different reverberators that smooth A more than B.

For example, one reverberation (A_r) may give the impression of a small room while the other (B_r) of a large ballroom. The ASA will act to separate the two sounds, based on their reverberation difference including reverberation decay time and delay time of early reflections.

Here, we should perhaps point out that whenever the source produces sound in a large reverberant room, the auditory system subdivides the inputs into two streams. All direct sounds from the source plus the immediately following reflections that cannot be perceptually separated from them (e.g., floor reflection and that from the nearest wall or an object) are grouped to create the impression of the source. All indirect sounds created by the later arriving acoustic response from the room are grouped to produce an image of the room. Therefore, a listener is aware of the source and the surrounding enclosure as separate sounds, the source as a sequence of distinct sounds, and the enclosure as a continuous reverberation. In addition, a listener may identify and track other perceptually distinct sources such as that annoying flutter echo or a slap back from the rear or rumble of the ventilation system. All available sounds compete for membership in these perceptual structures on the basis of similarity and plausibility.

It is assumed that qualitatively similar sounds heard within the auditory system's integration limit will fuse together. The limit is generally considered to be between 5 ms and 40 ms, depending on the transient nature of the sounds, beyond which auditory fusion breaks down and the sounds are segregated [4]. Of course, strongly dissimilar sounds are able to maintain their perceptual independence and are not subject to the in-

tegration. Imprecise attacks in a group of instruments such as strings will be largely unnoticed, due to integration. Late attacks by instruments of different tonal characteristics (woodwinds, for example) will be more noticeable. It is considered that the permissible range of delayed starts of instruments in the orchestra is 50 ms.

Because the acoustic room response has similar tonal characteristics to those of the sound that caused it, perceptual integration of the source and early response of the room operates over a longer time span, perhaps as much as 80 ms, depending on the nature of the transients of the source. Beyond that delay, the "room sound" becomes separated from the source and is perceived independently from it. This is why we do not have a strong awareness of the acoustics in small rooms where room response decays quickly. Large rooms and concert halls provide a strong sense of a "separate" acoustic space having its distinct onset and decay pattern as if it were an independent musical instrument.

The classical recording engineer tries to capture and frame these two distinct images of source and enclosure using microphones. The pop recording engineer more often creates synthetic environments to enrich sources that have been captured in a dry studio. Both of them are fully aware that the right acoustic environment must be used to establish a unique mood and atmosphere able to enhance the intended musical illusion of time and place.

8 Transposing the Pitch

It is technically possible to transpose the pitch of a tone while keeping all its frequency components in the correct harmonic ratios. This processing can be used for improving the blend of tracks with one another. When a voice or instrument is out of tune with others, it has three distinct effects:

(1) It causes some beating that can be heard as roughness.

(2) It gives the impression of more voices (the "chorus" effect).

(3) Since consonant harmonic relations increase perceptual fusion, and being out of tune destroys these harmonic relations, it increases the segregation of concurrent sounds that would have been in good harmonic relations had they been in tune.

A more complex effect produced by a "chorus effects processor" (usually present in DSP multieffect devices) can blend a number of distinct

sounds (say, a group of instruments) into a softer "mass" of sound by imposing common pitch and phase modulation.

The simple chorus effect, which is created by adding the track to itself several times, each time with a small delay, tends to "fill out" the music so that the slightly asynchronous onsets of individual instruments in, say, the violin section are lost and the instruments blend together.

In other cases, the tendency to segregate may be undesirable. For example, in polyphonic classical music, segregation of the parts is usually desired in the middle of a phrase or section of the music. However, at the *ends* of phrases or sections, the parts usually come together into strongly fused chords. This fusion maintains the unity of the music despite its polyphony. Good harmonic relations at these points of fusion are important; if one instrument is slightly out of tune, it is important to use pitch transposition to correct this, in order to maintain the unity of the whole. At other places, the mistuning, as long as it is not sustained for long, can be tolerated, or even appreciated. For example, mistuning may lend a desirable human voice-like quality to electronic keyboard synthesizers. An interesting example of mistuning that is perceptually desirable is that applied by piano tuners who often tune (acoustic) pianos with the low end pushed just a little too low, and the high end just a little too high, in order to achieve proper perceived interval (melodic or harmonic distance) between extreme notes of the piano. Because very high and low pitches of the piano have many inharmonic components (due to the physically imperfect nature of metal strings), equally tempered melodic tuning would cause the pitch intervals to sound too close together in the absence of additional outward stretching of pitch.

A device called a harmonizer can be used to generate chords or additional notes transposed to a chosen pitch interval relative to the original track, all playing along with that track. The harmonic chorus effect created this way has a thicker, more immediate texture, but is usually used only sparingly to support the source itself or is fed to a reverberator to support the source through gentler ambient sound.

9 Interactions Between Modalities

Auditory and visual perception are not two independent processes functioning in isolation. Both modalities cooperate towards improving human efficiency and ability to track "objects" and "events" in a surrounding environment. When auditory information is supported by matching visual information, or when visual information is reinforced by a matching au-

ditory cue, the cooperative interaction between the modalities reinforces human awareness of the stimulus. The matching of auditory and visual data triggers perceptual synergy between modalities and promotes intermodal fusion. A powerful form of audio-visual interaction can be seen in a phenomenon called the "McGurk effect." When a video picture shows a person saying one consonant, and the audio has the person saying a different one—with the two signals appropriately synchronized—the observer hears (does not merely *decide upon*, but actually *hears*) a consonant that is maximally compatible with both sources of information, rather than hearing the sound that has been presented in the audio signal [3]. Later research has shown that the effect is very powerful. It occurs even when the auditory and visual stimuli are presented by different genders, or when the face is blurred. Even when the auditory stimuli lags behind the visual stimuli by as much as 180 ms, the McGurk effect is apparent.

This effect is taken by advocates of the "motor theory of speech perception" as showing that the speech recognition system does not use sound to recognize speech directly, but to infer the talker's vocal tract activity; then it hears the sound that this activity would have created. This is why the visual evidence can so strongly influence what is heard. From the point of view of ASA, the effect illustrates the potency of synchronizing picture and sound to achieve cross-modal integration. Movies with effective sound tracks also show the power of the same integrative force.

Another important intermodal phenomenon is the "ventriloquism" effect. When sounds are synchronized with a picture that comes from a different location, listeners hear the sound as coming from the location of the picture, or close to it [2]. This, too, appears to be automatic on the part of the listener. The pulling effect can be observed in delays of up to 200 ms, and spatial displacements of up to 30 degrees. In all cases of audiovisual integration, sound reduces the ambiguity of picture and helps to define it, while picture reduces the ambiguity of sound or its position and helps to define its purpose.

The scientific evidence of the interdependence of hearing and vision shows that this synergetic perceptual interaction depends on the matching between auditory and visual data displayed to the viewer [6].

The important matching factors are

(1) Temporal coincidence (synchrony),

(2) Spatial coincidence,

(3) Congruence of auditory and visual movement,

(4) Balance between picture size and the loudness of sound,

(5) Balance between picture quality and sound quality.

10 Conclusion

We hope that this exploration of the applications of ASA to music and recording will provide new insights into these rich arts and perhaps provide the craft of the recording engineer with a scientific foundation.

11 The Answers to Audio Example 5: Musical Instruments Heard Backwards

(1) Trumpet

(2) Guitar

(3) Cello

(4) Xylophone

These are anechoic sounds that are played backwards; so no room reverberation precedes the decay of the instrument.

The sounds were provided by Bang & Olufsen A/S and were prepared by Geoff Martin at Multichannel Audio Research Laboratory, at the Centre for Interdisciplinary Research in Music Media and Technology, and at McGill University, Faculty of Music.

Annotated Bibliography

[1] P. L. Divenyi and S. K. Oliver. "Resolution of Steady-State Sounds in Simulated Auditory Space." *Journal of the Acoustical Society of America* 85 (1989), 2042–2052.

This study created a simulated auditory space by using separate "transfer functions" for the two ears, simulating, over headphones, what the listener would have heard in real free-field listening. Listeners were asked for separate localizations of two sounds presented concurrently. Relatively untrained listeners required quite large separations (perhaps as large as 60 degrees). Complex sounds were easier to separate than pure tones.

[2] C. E. Jack and W. R. Thurlow. "Effects of Degree of Visual Association and Angle of Displacement on the Ventriloquism Effect." *Perceptual and Motor Skills* 37 (1973), 967–979.

Sounds that co-vary with visual stimulation originating at a different spatial location are localized at or closer to the visual display (the "ventriloquism" effect). This research used a display consisting of a videotape of a human speaker with the voice separated from the picture, and showed that the ventriloquism effect was greatly reduced when a lag of 200 ms was introduced between the visual and auditory stimulation. With co-variant auditory-visual stimulation, the effect operated over a separation as large as 30 degrees of visual angle.

[3] H. McGurk and J. MacDonald. "Hearing Lips and Seeing Voices." *Nature* 264 (1976), 746–748.

This brief paper reports the McGurk effect. It depends on the joint presentation of a video picture of a person saying one syllable and a sound track in which a different syllable is said. The resulting experience incorporates information from both sources to make a completely new percept. For example, when the picture of the face saying /ga/ is presented with the auditory syllable /ba/, the subject will perceive /da/. Like /ga/, /da/ doesn't involve a lip closure (which is absent in the picture), but it sounds more like /ba/ than /ga/.

[4] H. Wallach, E. B. Newman, and M. R. Rosenzweig. "The Precedence Effect in Sound Localization." *American Journal of Psychology* 62 (1949), 315–336.

Showed that the upper limit of the time interval over which fusion of separate sound stimuli (presented sequentially) takes place is about 40 ms for complex stimuli such as speech or music.

[5] F. Winckel. "Music, Sound and Sensation." New York: Dover Publications, 1967.

This small, yet excellent, book translated from German, combines physical acoustics in music with the evaluation of the art of music, the subjective character of musical hearing, and the analysis of musical structures and speech. The author who himself has a thorough understanding of music provides composers, musicians, and recording engineers/producers with a better understanding of physical acoustics and psychoacoustics by reviewing, in the context of music, the astonishing sensitivity of the ear to temporal, spectral, and dynamic properties of sound.

[6] W. Woszczyk, S. Bech, and V. Hansen. "Interactions Between Audio-Visual Factors in a Home Theater System: Definition of Subjective Attributes." *Proceedings of the 99th Convention of Audio Engineering Society*, Preprint No. 4133, October 1995, New York.

This paper reviews broad evidence of interactions between seeing and hearing, and explains an experimental design developed for measuring interactions in a home-theater viewing experience. The authors propose to test four dimensions of audio-visual experience, where cooperative interactions between vision and hearing can deliver convincing illusions: action, motion, mood, and space. The approach has been used successfully to assess the subjective quality of home-theater systems using Dolby Surround programs. The companion paper (Bech, Hansen, Woszczyk, "Interaction Between Audio-Visual Factors in a Home Theater System: Experimental Results," Proceedings of the 99th Convention of Audio Engineering Society, Preprint No. 4096, October 1995, New York) presents the results of these experiments.