

Rendering: the Other Upmixer

Robert Reams, Neural Audio

Matrix Decoders 101

Before we get into the specifics of rendering, let's briefly examine how matrix encoders and decoders work.

All matrix codecs take advantage of a cognitive property known as *dominance*. A dominant event is one that has your attention. During a dominant event *all* other elements become *recessive*, meaning you are aware of them but you are not exercising analytical focus. As you become aware of additional elements, your cognitive workload is increased as you decide which element is dominant, all the while maintaining awareness on the elements you have deemed important but recessive. As this continuous stream of analysis occurs, many elements are included and discarded as the focus-sort-categorize-discard process of *scene analysis* is performed.

This means that you can only focus on one element at a time.

Matrix based systems are anchored in the premise that a spatial image may be represented within four *cardinal* points. These cardinal points are grouped in orthogonal pairs. The pairs are generally referred to as L+R (Sum)/L-R (Diff) and Left/Right. Each of the cardinal points is fed by a summer that receives complex and non-linear combinations of Left and Right to cause that particular point to behave as a point of delivery for Left, Center, Right or Surround. There are, of course, matrix decoders with more outputs, but the principle is the same.

The positioning of where image elements are "steered" is carried out within the decoder's *detector core*. The detector core typically consists of a pair of average log difference operations sourced from the Left and Right inputs for one and a Sum and Diff matrix for the other. These operations calculate the absolute average log ratio of Left:Right and the absolute average log ratio of Sum:Diff separately, applying the results via multipliers to modify or adapt the matrix.

As an example: in a Center Dominance condition, "L+R" multiplied by the averaged log difference of "L+R" and "L-R" is subtracted from both "Left" and "Right". This means that Left radiates Left and the $-(L+R)$ residual and Right radiates Right and the $-(L+R)$ residual. The residual is a dynamic error that is a function of the ballistics of the averaged log difference style detector core. After a short period of error, L+R is effectively removed from Left and Right individually...voila...no L+R comes out of the left or right speakers. If $L=R$ then L-R (the surrounds) output zero (no cancellation necessary).

If left or right become dominant, they are subtracted from L+R and L-R. This gives laterally panned dominant elements good separation from the center and surround speakers during dominance.

Surround dominance is the same as center dominance, except the log average L-R is greater than the log average L+R.

The detector core has (generally) variable or adaptive averaging that extends smoothing during pleonastic content and shortens it during transient content. This feature reduces the duration of the aforementioned error residual without making the matrix eternally “nervous”.

Matrix Realities

Loss of detail

Loss of detail is a common “complaint” about matrix based spatial compression. Image minutiae that make up the “detail” of the content image are often a casualty of the matrix. The presence of dominant content at a cardinal point may often reduce the “width” of recessive content. There is a subjective narrowing of image width along with loss of detail when a cardinal element is dominant. This often leads to subjective descriptions like (but definitely not limited to) “image collapse”, “center pile-up”, etc.

Cognitive *masking* is effective and useful *if* the listener is focusing on the element that *you* have deemed dominant. It only falls apart when the listener decides to focus on recessive elements (details) within the content. At this juncture the content may cease to be interesting because it has no depth beyond the dominant.

Dynamic Artifacts

“Pumping”, “breathing”, “twitchiness”, “drift” and “nervous sound field” are subjective descriptors of an ill-tuned or flawed detector core. Detector integration times that are not long enough to ignore irrelevant wiggles in the overall content image cause the “nervousness”. On the other hand, if the integration time is not short enough, the detector will be insufficiently agile to avoid perception of operations performed on the image, resulting in audible “pumping”. Most matrix decoders employ “gear shifting” or “adaptive” averaging to achieve program dependent agility. Some work better than others.

Leakage

Crosstalk is often the result of blind decoding. Content that is not created using complementary encoding usually results in content element placement with less precision than is necessary to achieve either perfect nulling of the element at the opposite end of the corresponding image axis. This results in “leakage” of “center to surrounds”, “left to right”, “right to left” or “surrounds to center”. Lack of precision is the cause of leakage. Leakage is often hidden by perceptual “cross masking” from adjunctive channels or temporally masked through the delay of the recessive end of the image axis. Advanced matrix decoders employ some types of energy or azimuth “servos” to compensate for content or medium errors. SWBTO.

Rendering

Rendering works in a completely different fashion. The loudspeaker number and placement is known by the upmixing algorithm. Each loudspeaker is associated with a “map” that corresponds

to its image sector in 2-D space. Correlation of the map with the content element image (as defined by the intensity/coherence descriptors associated with the element) results in reproduction of the element from the corresponding image sector. Since this process occurs over several hundred sub-bands, the discreteness is excellent.

There will be content elements that occupy common sub-bands in differing image sectors and there will be overlaps in the time-frequency plane, resulting in fusion or crosstalk. This is minimized by psychoacoustic methodologies in place that assist in the prediction and masking of the crosstalk. These methodologies rely on transform based architectures to spatially assign content elements, contentious or not, and present the appropriate spectral shape to the appropriate image sector at the appropriate time. There is contention, but these types of occurrences are perceptually managed and have little impact beyond “ghosts” perceivable only in the unnatural absence of the other channels or beyond the contours of a reasonable listening area.

There are no Bedrosian allpass pairs. There is no “matrix style” cross subtraction or detector core ballistics to deal with. The trappings associated with matrix style upmixer architectures is completely absent.

N-Channel Rendering (as applied in SEE) is a powerful methodology that allows the design engineer, at the consumer product end, to take full advantage of the 2-D Downmixed content. Through the use of Neural’s version of N-Chanel rendering (the Spatial Environmental Engine or SEE), the engineer may, through the specification of magnitude and coherence “maps”, define discrete image sectors, allowing the engineer to design systems that “render” the 2-D content to the few transducers that are available or to as many as they may wish.

This methodology is powerful. The number of reproduction channels is no longer tied to the number of source channels. If the consumer, for example, were to adopt to a “10.2” environment, SEE would enable “10.2” reproduction of legacy 2.0 and 5.1 content. The automotive system designer may easily solve dilemmas such as “5.1 DVD-A and stereo radio produced through four loudspeakers”.

Shameless Product Plug

The up-mixer of the Neural/Harris 5225 mix/edit appliance is based on a dedicated 5 channel 2-D rendering algorithm. That means that any stereo, Lt/Rt or 5.1 (via 5225) downmixed content will be rendered appropriately in 5.1. There are broadcast or distribution conditions where subsequent downmixing of an upmixed downmix are anticipated. The 5225 meets “near perfect reconstruction” criteria. This would insure minimum frequency domain and imaging artifacts in the downmix as well as lowest possible “tandem losses” on concurrent upmixes.

Production and Editing (and more plugging)

Radio stations have lots of stereo “music bed” content in their archive. A desirable scenario is upmixing the stereo music to 5.1 while creating layers of discrete mix elements. With rendering methodologies, there is no contention between the music and discrete mix elements for dominance. This results in a focused, stable mix that doesn’t wobble, swim or collapse.

The 5225 is a “transform” based technology that intelligently “shapes” the image over several

hundred spectra according to a psychoacoustic model similar to the ones used in modern codecs. The result is a perceptually compelling rendering of the mix, just as the mixer intended. The mixer may add as many layers of mix elements as they like. The mix may be ruled by artistic decisions, not the shortcomings of the downmix/upmix tools.

The 5225 downmix does not use “side information”, that is, data multiplexed into a bitstream to aid or enable spatial reconstruction of discrete 5.1 content. Instead, the 5225 embeds the spatial reconstruction descriptors directly into the waveform of the audio. This allows the mixer to treat the downmixed audio as (time domain) “audio”. The mixer can cut and paste the downmixed 5.1 just like any other stereo content without the fear of “ticks”, “pops” or image instabilities.

Benefits

Additional benefits of N-Channel Rendering are:

- 1) Completely scalable upmixing (mono to 256 channels)
- 2) Mixing of 5.1 and stereo elements in a common monitoring environment
- 3) easy and intuitive to use
- 4) Allows use with favorite “old school” equalization and dynamics processors
- 5) Natural sounding upmixes
- 6) Cross-fading between stereo and 5.1 in a common spatial environment

Summary

The concept of N-Channel Rendering is powerful. Creators of content can produce lively, dynamic and spatially involving content without the added burden of monitoring all possible upmixes. Rendering enables the user to monitor content mixes quickly and confidently while providing facilities for testing matrix, stereo or mono compatibility.

Rendering, as provided in the 5225 upmix, provides perceptual confidence monitoring that insures that the art will be heard just the way it was designed.