

Discrete vs. Matrix: What Do the Words Mean?

Robert Reams, Chief Scientist
Neural Audio Inc.

There is great interest (again) in broadcasting surround content for consumption in an automotive environment. The marketing and branding machines are hard at work canonizing and demonizing various multichannel architectures. Touting subjective superiority while ignoring real world design constraints creates market expectations that are unmanageable, and in the long run, doom the technology to failure (again). Questions that should be asked are: a) what are the consumers expectations and b) how closely are they met?

Replica vs. Impression

There is presently no popular way to promote a technology that is designed to achieve less than perfection. That being said, it is generally accepted that perfection is virtually impossible to achieve. The author personally cringes every time he reads the words “perfect reproduction”, even when it refers to one of his products. It is difficult to get those who promote our technologies to adopt the philosophy of achieving “a reasonable and entertaining impression of the original” or “less annoying artifacts than the other guy” as an acceptable adjunct to “perfect transparency”. Let’s face it, nobody is going to win any medals for “best possible fit”, or “most practical compromise”. Those categories don’t exist.

However, it is apparent that the consumer is quite entertained with less than “a perfect replica of the original” if the impression is compelling enough to despite the shortcomings of the transmission medium. In fact, in terrestrial broadcast, satellite broadcast, cablecast or internet streaming, attributes of the content (dynamics, image and spectral balance) are often modified or discarded if it will reduce distortion or artifacts on the consumer end.

This is, of course, horrifying to the audio artist as he or she demands that an exact replica of their art is what the consumer will experience. The producer responsible for the value of the content has an expectation that the art will arrive as at least a reasonable facsimile of the original. The broadcaster has to manage the content to be compatible with the lowest common denominator of consumer expectation. Finally, the consumer wants to experience as much of the content as his or her personal budget will allow.

Exact replicas don’t happen. The impression of the original, however, can be extremely entertaining and hence, valuable, for a vast cross section of the intended audience.

What Does “Discreteness” Mean

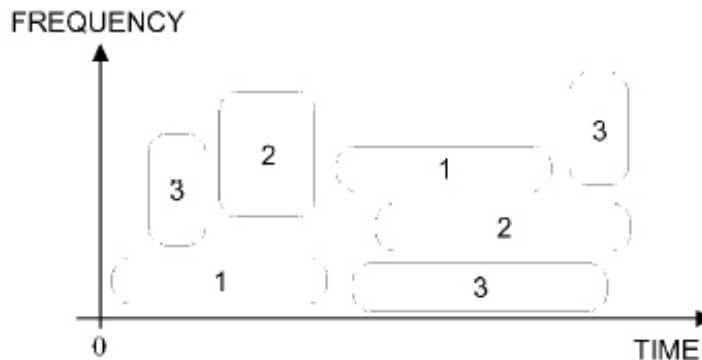
Separation or crosstalk between channels is a measurable and objective physical metric. How relevant is it when we include an average, naive human auditory system operating in a real world living room or car? The answer is not simple. It depends on the complexity and pervasiveness of

external sensory stimulus beyond that of the audible. If we assume a reasonable cognitive workload for the consumer in his or her environment we can predict the relevancy of inter-channel crosstalk as a function of relative loudness, duration and motion as a function of spectrum.

A better metric might be how many non-arbitrary “places” in frequency-space may be simultaneously achieved multiplied by the number of possible “fine structures”.

Frequency-Space Placement

The first factor would be a determining the available number of spectral parses or sub-bands in the frequency multiplexing portion of the image compression algorithm. For example, an advanced cochlear filter bank (CFB) would employ 98 parses or bands, a typical quadrature mirror filterbank (QMF) would employ 32 bands, SEE would employ between 1 and 1024 bands, a discrete cosine transform (DCT) may employ between 512 to 2048 bands and a typical “matrix” would typically enjoy 1 active band.



As an example, the time-frequency map (above) shows three sources which occupy non-overlapping regions in the time-frequency plane [after Faller] . This scenario would result in “zero crosstalk” as the sources do not have any conflict or overlap on either axis. It is easy to discern from the above map that as spectral granularity (the number of available bands) increases, so does the number of possible discrete “places”. As the number of bands decreases, likelihood of conflict along the frequency axis increases, and hence, so does crosstalk.

Fine Structures

A second factor to be considered would be determined by the number of simultaneously available image parses. In the ideal world of “5.1” this could be represented as five absolute, non-arbitrary “places” in the image (the “.1” or LFE portion is not included in the image). It can be argued that there is an infinite number of points in 2-D space that are simultaneously available.

It is also arguable, however, that these points in space are easily modified by external conditions (such as listener placement, room reflections, loudspeaker interferometry etc.) rendering them

arbitrary. For the sake of simplicity we will attempt to satisfy both arguments by counting the number of simultaneous “fine structures”. The maximum number of fine structures possible with a “5.1” medium will be a maximum of “5”. For example, an uncompressed “5.1” medium would yield 5 possible fine structures, SEE and certain other codec versions of would yield one or two fine structures and matrixes would yield two fine structures.

Multiplying the number of fine structures times the number of frequency-space places yields a metric for objectively defining the proportional “discreteness” capability of the image codec.

codec	fine structure	frequency places	“discreteness”
1)	2	1	2
2)	2	2	4
3)	2	32	64
4)	1	98	98
5)	2	512	1024
6)	5	256	1280
7)	2	1024	2048
8)	2	1024	2048

The above table depicts the range of “discreteness” available with seven different image compression algorithms. As is evident, there is tremendous variance. For the sake of clarity, the codecs at the bottom of the scale are of the “matrix” type while the codecs at the top of the scale are considered to be of the “sub-band” or “transform” type.

Discreteness is by no means a conclusive or absolute goodness metric for judging codecs as there are many other elements (psychoacoustic models, bugs, tuning, content etc.) that are factors in the efficacy of any lossy coding scheme. The discreteness metric, rather, is a measure of how much relative perceptual separation may be expected at the point of consumption.

After that, it is up to the system architect to discover the best possible implementation of the methods that achieve the best measure of discreteness.

SEE Image Data Transport Method

SEE imaging data may be described as a probable version of the original image envelope as a function of spectra, the lateral phantom source azimuth of the original image envelope can be represented by:

$$\alpha = \Delta L \left(\left[\frac{1}{S} \right]^t + \left| \frac{\Delta L}{\delta} \right|^t \right)^{-1/t}$$

α phantom source azimuth in degrees.

S slope parameter

t transition into saturation

δ loudspeaker azimuth

ΔL left and right channel intensity difference

The front-to-back phantom source “depth” of the original image envelope is represented by a running average of coherence (normalized cross-correlation)

$$\rho_{12}(j) = \frac{r_{12}(j)}{\frac{1}{N} \left[\sum_{n=0}^{N-1} x_1^2(n) \sum_{n=0}^{N-1} x_2^2(n) \right]^{1/2}}$$

where $\rho_{12}(j)$ is the cross-correlation coefficient, $-1 \leq \rho(j) \leq +1$.

$$r_{12}(j) = \frac{1}{N} \sum_{n=0}^{N-1} x_1(n)x_2(n+j)$$

$x_1(n)$ and $x_2(n)$ are the input blocks containing N data points each

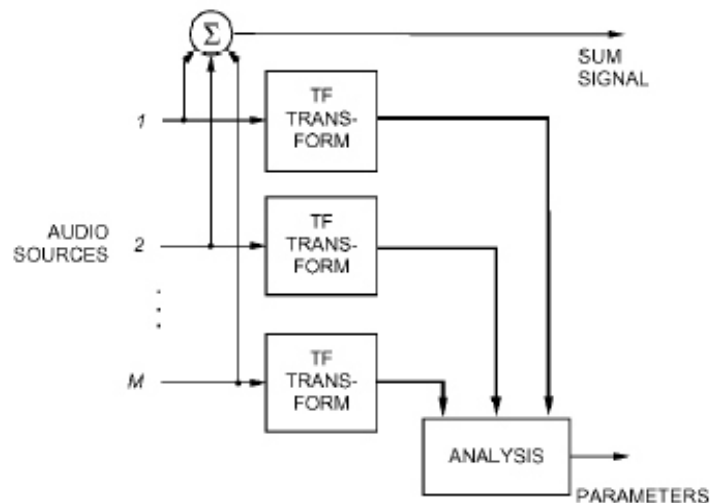
The acceptable level of smoothing is wholly dependent on the psychoacoustic model. Typical lossy compression algorithms use (frequency domain) smoothing as derived from the stationarity of the content to augment redundancy, hence, coding efficiency. In this case, smoothing is used to prevent departure from spectral pleonasticity caused by that which is perceptually irrelevant.

On the downmix side, the image envelope is imbedded in the audio in the form of watermarking. Intensity/coherence watermarking is an excellent choice because of its similarity to the image construct of naturally occurring 2-D stereo and compatibility with already prevalent Lt/Rt matrix content. This greatly simplifies the integration of 2.0 and 5.1 content on both the transmit and receive sides.

On the upmix side, the image envelope or the original 5.1 content may be re-synthesized based on the intensity/coherence information contained in the watermark. Using this methodology, an impression of the original source 5.1 content is reproduced with a high degree of merit. In fact, the use of intensity/coherence watermarking allows the codec to dedicate significantly higher bits/sec to the fine structures. At ultra low data rates (48kB/s) this would increase the number of bits available for the fine structure by 33% as there is no side data to transport. This also allows for the representation of legacy two source (stereo) content using the same image envelope approach used for original source encoded 5.1. The decoder segregates spatial elements naturally found in stereo based on the image envelope naturally residing in the content.

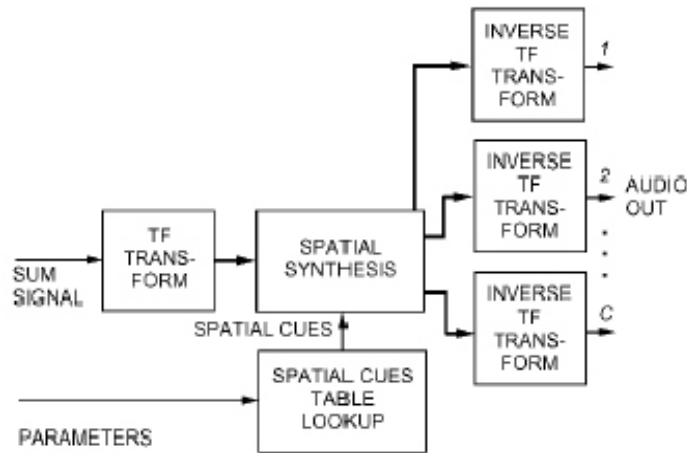
Other Image Data Transport Methods

With other codecs the side information approach is taken. On the encoder side, blocks of intensity and temporal information of the original source 5.1 are analyzed as a function of spectrum, quantized, compressed and multiplexed into the bitstream adjacent to the relevant block containing a two channel downmix of the 5.1 original source.



On the decoder side, the multiplexed block is unpacked and the quantized inverse of the analysis is performed on the one or two channel downmix, resulting in an upmixed 5.1 impression of the

source. In the absence of side information decoding, the downmix is reproduced in two channels.



Compatibility Issues

From a perceptual standpoint both methodologies work remarkably well with original source encoded 5.1 content. However, most image compression algorithms have no facilities for upmixing of legacy two channel content or compatibility matrix provisions for such, resulting in the same 2.0/5.1 content conundrum already experienced (and solved by SEE, for that matter) by the HDTV industry.

The following is quoted from an HDTV broadcast white paper addressing this conundrum:

“When dealing with HDTV audio there are additional challenges that crop up involving the interspersion of legacy (Lt/Rt, stereo or mono) content with original 5.1 content. The resulting image shift and collapse is startling and annoying. This phenomenon is easy to experience, just tune in to any HDTV broadcast and pay attention to the [stereo] commercials...no dialog anchor in the center channel, no envelopment and no bass. Typical listeners assume that HDTV is “unreliable” or the stations sponsors are “too cheesy to make 5.1 commercials.”

The core feature of the solution would be selective up mixing of mono, stereo, Lt-Rt (matrix encoded stereo) and 5.1 “discrete” content into a consistent and pleasing 5.1 spatial environment. “Upmixing” is not a new idea. What is needed is the intelligent detection of the content spatial “type” and appropriate smooth morphing (not switching) of one spatial scenario to the next into a consistent 5.1 spatial environment.”

The Solution

Continuing the quote from the HDTV white paper... “The solution would be “glue” processing (processing used to solve interface or incompatibility problems) that insures the “best possible fit” of content with highly variable spatial goals (mono, stereo, matrix stereo and 5.1 discrete) into a 5.1 spatial environment. This would solve the (HDTV) industry problem of transitioning content

from mono, stereo or matrix stereo to discrete 5.1 broadcasts. This would allow the station, affiliate or broadcast system to provide an uninterrupted stream of "5.1" content regardless of origin. The source of the content could be mono, stereo, matrix stereo or 5.1 Discrete. The problem of image transitioning could be solved using neural networks to categorize the content spatial "type" and take swift and appropriate action on the signal within the following modes:

- 1) Mono: Synthetic Aperture Mode
- 2) Stereo: 2-D Stereo
- 3) Lt, Rt: Matrix Up mix....."
- 4) 5.1 pass through

The downmixing challenges on the receiver side have been dealt with as well:

"...In the absence of a full multichannel loudspeaker complement, excellent results may be had by selective downmixing. The ever increasing ubiquity of DVD transports ameliorates any future cost issues and the inclusion of a DVD transport in even the most aggressively priced vehicle or consumer environment is an eventuality. Downmixing allows the designer to "future proof" the vehicle by allowing the playback of future as well as present content. The scarcity of 5.1 music material is temporary, so why not make it available to the owners of even modest vehicles."

Downmixers are not simple summing algorithms. An excellent downmixer design takes every possible content scenario into consideration and may include complex combining and adaptive networks to insure upmixing that is complementary to the intent of the original content. The downmixer may be additionally configured to produce two channel content compatible with the installed base of matrix style decoders.

Conclusion

2-D Downmix/N-channel Rendering image compression is a new methodology that meets the 2.0/5.1 transition challenges head-on. This novel encoding process allows the distributor or broadcaster the ability to capture original source 5.1 content and downmix it to a 2.0 channel format that survives aggressive lossy compression and conversion to analog. Encoded content may be broadcast or distributed through existing 2.0 infrastructures. To broadcasters, this could be a godsend as the existing infrastructure including production and storage are 2.0.

While this is one of the attractions of 5:2:5 matrixing, it is easy to discern from the above table that a discreteness metric of "2" or "4" falls far short of what is possible regarding discreteness (SEE N-Channel Rendering achieves a metric of "2048") [*]. Matrixes are an excellent solution in that they solve transition compatibility issues and allow, to a certain extent, 2.0 content co-exist with 5.1 content while extending the value of the ubiquitous 2.0 infrastructure. That being said, they fall short of the distributor and consumers expectation of what is now called "discrete" 5.1.

[*] Neural Audio has said all along “No, no, no it is *not* a matrix...it’s not, it’s not, it’s not...”.

The SEE decoding process allows the consumer to enjoy a consistent spatial environment with as many or few loudspeaker elements as is available with content ranging from 5.1 original source digital to mono analog. As with broadcast and/or distribution, the automotive audio infrastructure is also 2.0 and a codec that could utilize the existing 2.0 backbone could drastically reduce the cost of implementing 5.1, thus driving (pun intended) the ubiquity of the automotive 5.1 system.

It is obvious that the interspersions of legacy 2.0 and 5.1 content is a reality. Unless this is dealt with on a system basis, the result will be less than transition proof. In fact, inability to successfully integrate legacy content with “modern” content in such a way that meets the consumers expectation is unfortunately naive and slows the adoption to 5.1.