

Audio Processing for Compressed Digital Television

Robert Reams, NAB Member

The audio portion of television broadcast is often variable in format, spectral balance, loudness and image width. This not only complicates matters for the HDTV broadcaster, who has to merge stereo, Lt Rt and 5.1 content but it may also wreak havoc at the consumer end with either “discrete” or “matrix” style decoders at home. In this paper we will discuss present and future challenges as well as the transitional technologies in ingest, mixing and editing and the “B-chain” that will allow us to “stay in business” as we make the long, arduous and hopefully profitable transition from SDTV to HDTV.

1.0 Loudness

“There has recently been renewed interest in loudness in the broadcast industry, and this has prompted the development of a new Loudness Indicator...Because of continuing complaints from the public about loudness, and especially about loud commercials on television, the Federal Communications Commission (FCC) initiated its own study. It reported on a series of measurements on the loudness of commercial material and concluded that the problem of "excessively loud" commercials still exists in television broadcasting. The FCC conclusions were based on judgments of a listening panel and readings... In FCC tests...[the] panel....led to the conclusion that approximately one-third of commercial material was considered loud.” Jones and Torick inventors of the CBS Audimax. *The time was 1981*

This is still a real issue within the broadcaster community. The viewer really does “change channels” if they become annoyed enough. The integration of “modern” high dynamic range content with (lower dynamic range) legacy content and loud blaring (high density) commercials is effectively “viewer repellant”.

There is working metadata technology that takes this problem into consideration, however, there are metadata integration challenges between the content and the consumer as well as legacy content issues (pre-existing content that has no associated metadata).

Let’s look at some history. At one time SMPTE standardized -20 dBFS as the "operating level" for digital audio systems and established VU zero as -20dBFS to produce typical PPM peaks of about -10 dBFS for VU peaks of 0. There appeared to be difficulty maintaining this as a consensus so dialog normalization was made variable within a range from -31 dBFS to -1 dBFS. Although a dialnorm meter has become commercially available, proper dialnorm measurement requires choosing a suitable portion of dialog within the program and relies on the discretion of

the operator while monitoring in a highly controlled environment. These measurements require a skilled operator with the time to perform a complete level assessment of every show (not possible in a broadcast environment). If all goes well and all of these conditions are met (all broadcast engineers may stop laughing now), dialnorm must pass intact to all destination decoders.

Does this mean that metadata is “bad”? Of course not, but the problem of “everybody’s gotta have one before it works” might cause a few pragmatists to hesitate in spending rarely available cash on something that works “sometimes”. So...what to do in the mean time?

Generally, the overall shape of the loudness control transfer function is where a problem could develop. A default “target map” of the program dynamics could be defined and maintained in the absence of metadata. In the presence of valid metadata the target map could “morph” into the compression profile described by the metadata. Should the metadata vanish or become corrupt the compression profile would simply morph back into the default target map.

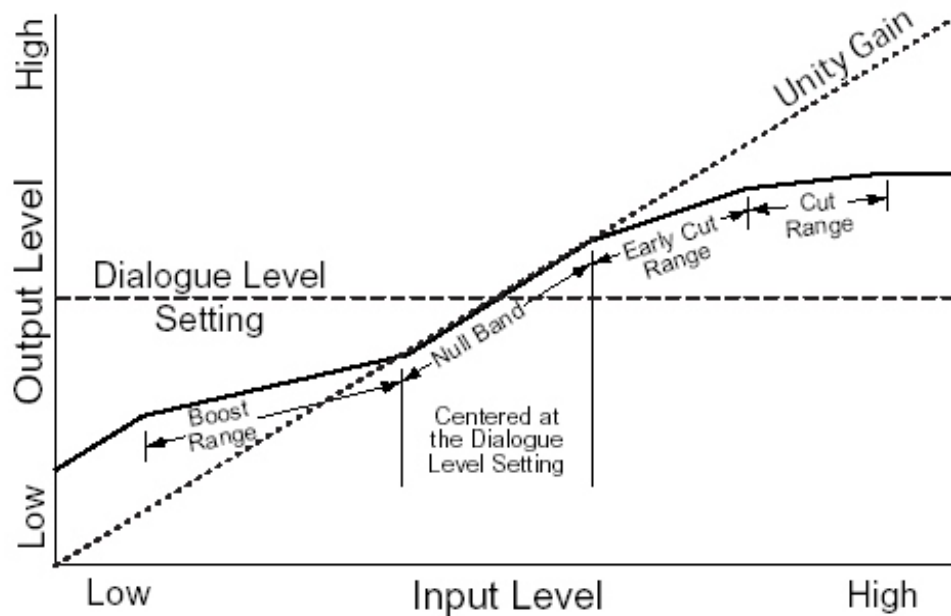


Fig. 1 The Compression Profile

Maintaining the long term perceived loudness (the center of the “null band” within the compression profile) of the overall program under all conditions is a desirable feature. Although instantaneous correction is not possible, satisfactory (local) null band gain normalization is achievable if the recovery/reduction ballistics are shaped according to psychoacoustic principles.

The broadcast engineer could then have a choice to override the local norm in the presence of valid metadata. This feature would allow the station to “back out” the “local norm” and the

default target map features as metadata became better understood and more reliable. If all goes well, maintenance of a local compression profile target map and null band gain normalization will become unnecessary with the exception of the stations who have their own dynamics preferences.

2.0 Spectral Balance

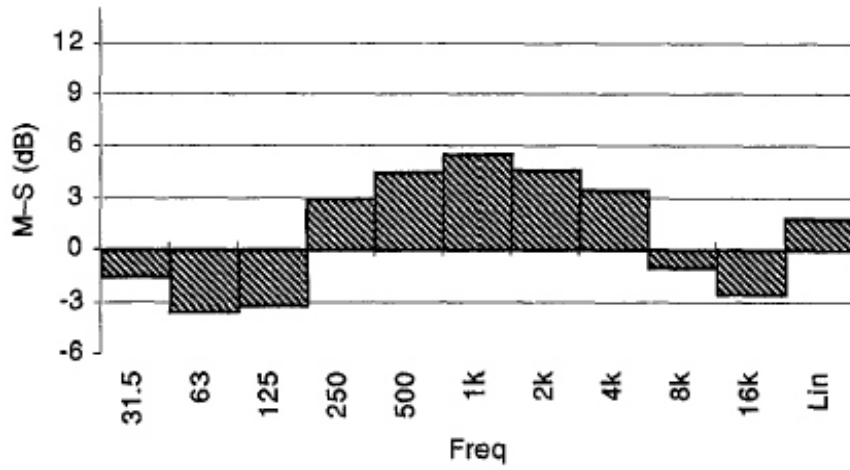
What makes content pleasing? What sets winning programming apart from poor programming? Content that is statistically popular interspersed with ad content specifically tailored to the anticipated lifestyle of the listener demography is shown historically to be most important. Maintaining consistency and polish within the boundaries of what is considered most pleasing (within the listener demography) is not only sensible but imperative if the television station is going to hang onto the viewer once he or she tunes in. Any perceptual discontinuities in the content flow has been shown to (consciously or sub-consciously) impact the *attention level* and the *attention span* of the listener, and hence, affect ratings value of the content stream. (Per Werrbach and Stuart) Additionally, these two features determine the effectiveness of ad content delivery. The overall flavor or signature sound of the content must generally match the chosen genre and hence, the audience's perceptual expectations. To assure the quality and consistency of the content flow certain perceptual aspect targets must be determined and maintained. Different genres of program draw viewers with different perceptual expectations. Maintaining an overall loudness target and (dynamic) range is assumed to be an accepted and generally understood perceptual aspect and is addressed above in "Loudness".

Two other important perceptual aspects are the spectral and image balance of the broadcast. Terms historically used to describe spectral balance would be bright, punchy, warm, subterranean, etc. Language used to describe image balance has been, "in your face", focused, clear, immersive, etc. Some of these terms have been used to describe the complex interaction of both spectra and image (more immersive means less clear). There lies the rub. Perceptually, these two aspects (spectral balance and image) are not parametric; that is, they interact with each other.

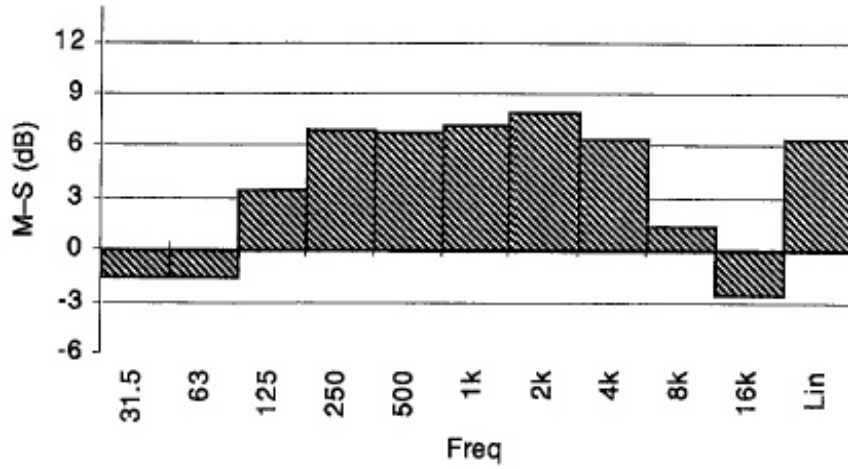
Rather than trying to parameterize these two aspects separately it is possible to create a map that adequately describes both simultaneously. The spectral image map (SIM) describes the complex relationship of spectra and image over a fixable time interval. The spectral image map serves as the programmable target for an adaptive spectral and image management process.

In the following figures note the variance (ah...art!) in the ratio of sum (M) to difference (S) energy as a function of spectrum. (Per Rumsey) These figures describe image width as a function of frequency-not the aggregate spectral balance.

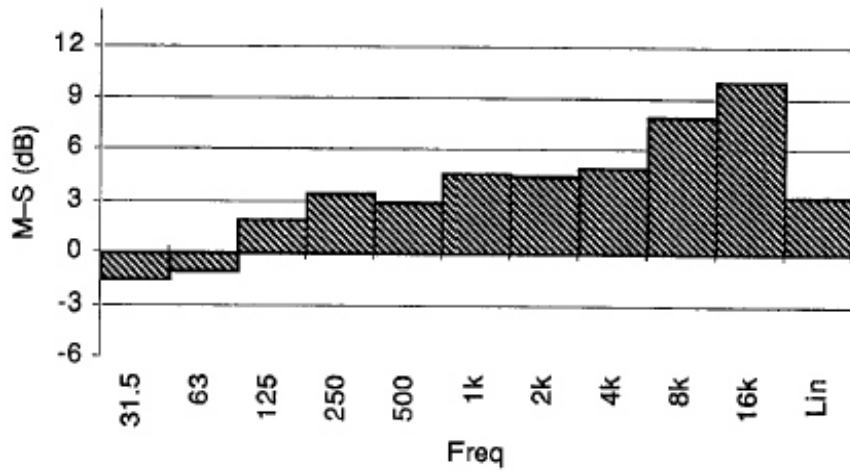
Orchestral (coincident pair)



Sci-fi radio drama



Rock (Parsons)



Aggregate spectral balance must be asserted simultaneously if the spectral image map is to be successfully maintained.

Maintaining a spectral image map target is accomplished by stabilizing spectral energy variances between the mono and side channels of stereo element *pairs*. This is accomplished by matrixing the stereo pair into an M/S pair, spectrally parsing the channels and shaping the medium-term gain of the spectral parses with low gain back-prop nets. The mono net sum and side net sum are then de-matrixed back into the stereo element pair. This maintains a “spectral image map” consistent with the pre-determined genre of the content. Conformance to the map must fall within ballistic restrictions derived from cochlear mechanics. These restrictions minimize the audibility of envelope changes. Although it is possible to reduce long term (and even short term) variances to zero, this behavior is inappropriate in anticipation of use with lossy data compression. Lossy codecs rely on variances in relative energy of differing spectra in a common temporal envelope to effect masking in the frequency domain.

The spectral image map may be extended to a 5.1 spatial environment by establishing independent maps of C (side channel=0), L/R and Ls/Rs, parameterizing (establishing as a target) the mean balance between L/R and Ls/Rs and then parameterizing the mean balance between C and L/R, Ls/Rs as a group.

3.0 Lossy Compression

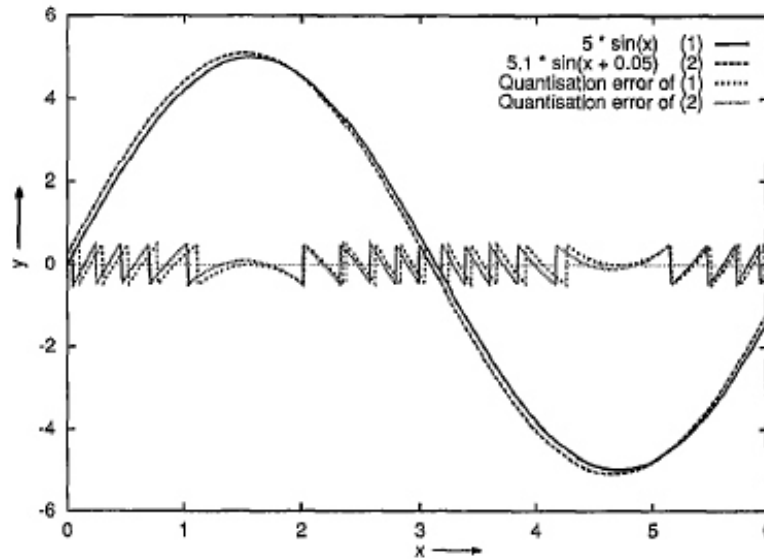
Perceptual coding that is so prevalent in the broadcast path relies on the masking properties of the human auditory system. Using time-to-frequency mapping, the encoder may apply either transform or sub-band filters to achieve data reduction by quantizing parts of the audio signal in narrow frequency bands with less bits than necessary for quantisation of the broad band signal. Typically, 1.5-3 bits (and less) per sample are needed to achieve the same audio quality as a PCM audio signal with 16 bits of resolution. The power of the quantization noise in a perceptually coded signal is therefore much greater than in a broad band quantized signal with the same audio quality.

In most of the cases matrixing (downmixing) alone does not cause problems with the perceptual based low bit-rate coding. In the digital domain, the relative phase and magnitude of two signals is constant along the whole transmission path. Therefore, the correlation of the quantization noise with the content signal is constant.

Upon de-matrixing (upmixing), the resulting quantization noise may no longer correlate with the coded signal. They (the masking signal and noise) no longer coexist in the same acoustic space and hence, the noise becomes audible along with the resulting signal. When the signal is coded without having this in mind, two excellent signals may form a bad signal when upmixed. The signal which was supposed to mask the quantization noise may be spatially displaced, resulting in ineffective masking of the quantization noise.

Only in a very few cases the two signals forming the sum or difference are actually identical or identical with opposite signs. Many mono signals only seem to be mono. The digitized samples

however are not identical. The fact that the signals are *different* means that the quantization noise is uncorrelated. Forming the difference between two *almost* identical signals means that the signal disappears and the quantization noise remains. In fact noise from the two channels is added, causing a 3 dB rise in the noise level. (Per Nielsen)



Uncorrelated quantisation noise of two "mono-like" quantised sine tones with a stepsize of 1

In a matrix listening set-up, a mono-like signal does not really disappear, it is emitted from the center loudspeaker, or in the case of a "phantom center loudspeaker" set-up, from the L and R loudspeakers. The fact that the masked noise comes from the L, R and surround loudspeakers means that only a partial masking takes place, depending on the listening position. The situation where the matrixed signals are out of polarity (only surround present) rarely occurs. In normal stereo it may of course happen, but in matrix surround coded material it is unlikely that unmasking will occur as the surround signal is often mixed at a lower level into the front channels. In most cases the difference signal is causing problems as the mono-like signals are much more common than pure out of phase signals.

The combination of matrixed audio signals and low bitrate coding based on perceptual principles may lead to annoying levels of artifacts due to an unmasking effect of the quantization noise and/or artificial shifts in timbre if no precautions are taken in the mix/edit and processing stages of the distribution or broadcast chain.

On the content consumer's side, circumstances consisting of blind playback through common matrix decoders (like Dolby's "Pro-Logic II" and DTS's "Neo-6", Lexicon's "Logic 7" and SRS's "Circle Surround") result in the unmasking of resultant unnatural sounding elements (USE's) that are normally masked when it is reproduced in relatively ideal two channel stereo.

De-matrixing bit-rate-reduced content under these conditions may result in a perceptual compromise in audio quality. The audibility depends on both the correlation between the input signals and the noise/zero-quantizer components associated with these signals. In the case of de-matrixing where there is a clear correlation between the signals, the effect is more apparent. This effect may be referred to as **unmasking**. The general basis of this concept is that quantization images from the other channels reside in the matrixed channel. This should not be confused with the effect of quantization with insufficient precision (due to either a flawed psychoacoustic model or an exhausted bit pool).

An example of unmasking is reproducing the difference portion of a bit-rate-reduced signal and the original; only the quantization image is heard. In more practical terms the noise components in the compatible signal may become unmasked as the signal that was masking them is removed or at least displaced to a different time/space by de-matrixing. It should be noted that the masker is still present in the multichannel reproduction, it just doesn't coexist in the same temporal/spatial location as the quantization noise.

The prevention of uncorrelated quantization noise image is not a trivial task. As conditions that cause this phenomenon are impossible to predict or eliminate, it is still possible to calculate a "best possible fit" for each possible error scenario and thus, statistically reduce the occurrence. This process is most easily described using stereo element pairs and is extendable to a 5.1 spatial environment. (See Spectral Balance)

The three stages of processing or "image packing" as it becoming to be known are described as follows:

- a) Detect the temporal centroid of L/R image axis and reduce mean deviation to zero
- b) Detect the intensity centroid of L/R image axis and reduce mean deviation to zero
- c) Detect the image centroid of S/D image axis and reduce mean deviation to zero

There are several permutations and methodologies of image packing, however it is beyond the scope of this paper.

4.0 Image Management

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. (another excellent "viewer repellent"). This phenomenon is easy to experience, just tune in to any HDTV broadcast and pay attention to the 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 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: Expanded Sweet Spot Stereo
- 3) Lt, Rt: Matrix Up mix
- 4) 5.1: Pass-through

The inputs and outputs would be primarily 5.1, that is, three AES pairs grouped in a defacto TV standard:

- 1st pair: L, R
- 2nd pair: C, LFE
- 3rd pair: Ls, Rs

A secondary function would be to provide a (fourth AES) watermarked “Lw, Rw” output of the content of the processing input. This two channel output may be considered a “down-mix” or “up-mix” depending on the original content’s spatial goal. This “Lw, Rw” output would be:

- 1) Substantially compatible with stereo
- 2) Substantially compatible with all known consumer matrix decoders
- 3) 100% compatible with the new SEE rendering engine.

This would insure perceived spatial consistency within the consumer’s “consumption environment” and would be fundamental in easing the transition from contemporary stereo or matrix presentation to the discrete 5.1 presentation of the future.

Conclusion

We have addressed some real world problems that the broadcast engineer faces every day. Our fondest wish is for the engineer “in the trenches” to continue to identify these challenges so that we may affect a few simple tools that will allow the continuous and uninterrupted flow of entertainment from the broadcaster to the consumer. This must occur during the process of major transitions of several technologies. The following, however, is irrefutable: the industry conversion

to “digital”, with its fits and starts, is the culmination of the work of giants. *We* get to be a part of its *completion*.