

Improving Content “Codability” With Lossy Codecs

Robert Reams, Chief Scientist
Neural Audio Inc.

Any time that content is produced or mastered without regard to the realities of distribution, i.e. digitally broadcast, webcast, or sold as an MP3 file, the sound may be rendered needlessly flawed. Linear formats are forgiving in regards to sloppy spatial mapping, unnecessarily high spectral density, perceptually irrelevant intensity offsets, temporal errors and noise. Lossy coding is *not*. Common causes of avoidable coding artifacts are discussed. Real-world scenarios for the impact of poorly prepped content are examined. Solutions on the production and broadcast end are suggested.

History

“Excellent” content played back under “ideal” circumstances result in adequately masked “unnatural sounding” elements (USE’s) [1] that are a normal result of low bitrate coding based on perceptual principles. USE’s, historically, are adequately masked when optimum content is played back through a monitor quality two element stereo environment with the listener sitting in the “sweet spot” of an acoustically adequate environment.

The combination of “flawed” content and low bitrate coding based on perceptual principles may lead to annoying levels of artifacts if no precautions are taken in the mix/edit, mastering and processing stages of the distribution or broadcast chain. This paper discusses practical means to prepare content to be coded at low bitrates while keeping artifacts as manageable as possible.

Program content may be “planned” or “designed” to prevent the effect of quantization with insufficient precision due to an exhausted bit pool. It is assumed that we are not dealing with codecs utilizing a flawed psychoacoustic model. It is also assumed that we are dealing with “spatial codecs”, that is, codecs that have the capacity to deliver a stereo or 2-D impression of the original content. Parametric, Binaural Cue and Spatial Rendering types are included in this category.

Quantization Noise

Quantization noise is the most familiar of codec “artifacts”. Quant noise is often interpreted as the “rough” or “distorted” quality of most lossy codecs. Under extreme conditions, the quant noise will reside “just below” the masking threshold, rendering it “inaudible” under the conditions specified by the designers.

[1] The term “USE” has been coined to avoid writing or speaking the terms “quantization noise”, “zero quantizers” etc.

Zero Quantizers (birdies)

Zero quantizers cause the unnatural shifts in timbre during demanding content. Bit allocation algorithms are guided by a psychoacoustic model that selects the best quantizer to be currently used by predicting the conditions where quantization noise will be inaudible. This bit-allocation procedure varies from block to block. One of the possible quantizer scenarios during spectrally demanding content is the “zero-quantizer” which will quantize certain spectra to zero. As a result of the block to block bit assignment variances, spectral coefficients may temporarily appear and disappear. The resulting change in timbre and high frequency energy variations have a somewhat “chirpy” characteristic that, when unmasked, is unnatural sounding. “Birdies” are among the most frustrating of artifacts as they are so dependent on listening environment and position.

Problems in the Consumer Environment

There are conditions in the spatial environment which reinforce the masking properties of 2-D codecs. Any spatial scenario that strengthens temporal/spatial relationship between the quant noise, quant zero and the masker helps maintain the expected NMR predicted by the designer. Conditions of image commonality include: matched L/R intensity ratios, matched (-6dB) sum/diff intensity ratios and zero temporal offsets between correlated materials (excluding intentional artificial echos). These are conditions of the well known *cross masking* phenomena.

Environmental conditions exist on the consumer side that cause a phenomenon known as **acoustic unmasking**. The summing and differencing of signals occurs in the acoustical domain by means of interferometry. (Matrixing as a result of signal processing is of no relevance here.) Acoustical unmasking can be seen as the evil complement of *cross masking*. Scenarios exist where the signals attenuate one another through interference leaving quantization noise components unmasked. Environment acoustic transfer functions vary strongly with position and frequency. Acoustical unmasking is therefore unlikely to occur when the signals have any sizable bandwidth, however, possibilities for this kind of unmasking event are many.

The condition of acoustical unmasking is particularly noticeable when:

- a) listening to stationary signals within a standing wave node of a poorly conditioned listening room (typical home environment).
- b) listening to mono stationary signals through multiple sources with different distances or arrival times at the listener (typical automotive environment).
- c) the listening position defeats minimum conformance to expected intensity or temporal offsets within the environment (typical trade show demonstration environment).

Prevention in Production

There are already hundreds of thousands of hours of existing content that, on their own, don't code well. There are a few simple guidelines to follow to avoid becoming another source of content that codes badly. These suggestions are easy to follow, very general and for the most part, pretty much free. A lot of improvement may be had just by being careful about how we go

about the business of tracking and mixing. Following simple suggestions can turn bit-eating, code-hogging, sloppy, indeterminate content into one that codes and de-matrixes beautifully.

1) Be reasonable and accurate with your image mapping

- a) The obvious often isn't.
- b) If you are panning mono sources, make sure that when you are panning to center, that there is no left/right discrepancy.
- c) If you are using an analog console don't trust the center detent on the pan-pot...it's not close enough. L+R and L-R positioning must be exact. If you're using a pair of channels to create L-R spend the time to "null" the channels...make sure it's "just right".
- d) Never pan "just a little off center" or "just a little off right". Image mapping that doesn't respect the four cardinal points of Left, Sum, Right and Difference are doomed to code badly. An extra bonus...your stereo mix will also sound delightfully punchy and consistent. So keep your mix elements in the absolute left, center or right. Put room fill and ambiance in L-R.
- e) Always be wary of creating or using a stereo music mix and then adding additional mix elements like the music never existed. Two different mixes (with their own individual dominance events) may end up fighting each other perceptually and hence cause poor results through lossy codecs.

2) The ballistics of cognition, that is, how long it takes for you to recognize the existence of a perceptual event and where it is, should rule the expectations of your mix.

- a) Generalized direction (with non-repetitious aural objects) is on the order of 100ms with fairly precise hardening within 1second.
- b) 2D aural renderings consisting of one dominant and two recessive objects (with visual stimulus consisting of one active object) yielded generalized localization within 300ms with hardening within 1 second for the recessive objects.
- c) Don't assume that image complexity is proportional to success; quite the contrary, a complex mix may lose its articulation if several mix elements fight for dominance. Only one mix element may be dominant at a time.
- d) If there is going to be a dramatic pan of a mix element, it must (temporarily) become the dominant element. If it is too recessive in the level of the overall mix the pan change-of-position will become ambiguous.
- e) Use creative pauses in the content around the pan. If that is not possible try manually reducing the gain or ducking the recessive content during the pan.
- f) Psychoacoustically, a moving pan with start and stop points persisting less than 300 ms will be ineffective.
- g) "Arcing" pans should be executed at radial speeds of 6.0 ms/degree or slower. The more isolated a moving or "special effect" pan is, the less ambiguous and hence, more effective it will be.

3) Control the bandwidth of individual mix elements

Frequency response from DC to Channel 4 is impressive on the spec sheet, but when tracking, individual restriction of the spectral boundaries of mix elements will pay dividends after mixdown, mastering, distribution and right into an aggressive lossy codec.

- a) Bound the natural bandwidth of the individual mix element with high and low pass filters.
- b) Use transparent noise reduction wherever possible.

4) Use a clean, wide bandwidth signal path

Always use good, clean, filtered AC. Use high quality cables in good repair. Validate and maintain critical monitoring equipment. You already know this but it's even more critical now.

5) Get rid of all hum and noise inducing sources

Hum and noise elimination is equally important. They are uselessly bit-consuming non-content. In fact, for some content, hum and and/or noise is pushed into the perceptual envelope by (value adding) volume leveling and dynamic range processing. Even when they are is not perceivable, a function of the consumption environment, they drive down entropic opportunities of the codec making an already bit starved situation worse.

6) Stick to the rules (ITU-R BS.1423)

- a) that during surround program mixing, the resulting sound image should be checked by monitoring the signal after it has undergone the surround encoding/decoding process;
- b) that during mixing, the stereo and mono compatibility of the resulting signal be checked;
- c) that a surround sound decoder with characteristics of those used by consumer should be used for monitoring;
- d) that the program material produced in this manner should be clearly labeled as being surround encoded so that operational personnel will be aware of the signal format [2]
- e) that the primary multi-track source material (8 to 48 tracks, if available) used before the encoding should be preserved [3]. This will allow a discrete 5-channel mix to be produced.
- f) Use the ITU-R monitor loudspeaker placement standard

[2] Program providers may wish to label the program as being matrix surround encoded within the program content so that the audience can be aware of the signal format. (prior arrangement with licensee necessary).

[3] Recommendation ITU-R BR.1384 contains specifications on recording track assignments for multichannel programs.

Mastering or Broadcast

Digital mastering processes (not enhancement) can deliver an excellent experience into the consumers environment. Additional “care and feeding”, as discussed below, is necessary if the client is anticipating heavy digital air play or downloading.

The broadcaster has another challenge; there are already hundreds of thousands of hours of existing content that, on their own, don’t code well.

For the sake of presenting a consistent stream of entertaining content to the listener, the broadcaster must use processing to manage music and ad content. This is a tall order by itself. Although digital broadcast eliminates all multipath induced instabilities, we still have to deal with USE problems at the consumer end. Here are some general suggestions.

1) Image Management

Image management employs spatial transcoding. Spatial transcoding removes image impairing, bit consuming events from the content image prior to encoding. Image anomalies that are perceptually irrelevant (long-term L/R intensity offsets, short-term M/S displacements and long-term M/S imbalance) are reduced or removed by predicative (through forward-prop nets) and agile manipulation of the L/R and sum/diff axis pairs. Stereo music content may be subtly altered to remove inefficiency causing elements in the music using methods that have no perceptual impact on the content.

2) Limiting and Compression

Perceptual codecs do not respond linearly to overly aggressive level management. It is generally understood that perceptual codecs respond poorly to heavily clipped signals. What is generally not understood is that heavy dynamic range compression may deleteriously affect the codecs ability to operate efficiently. In fact, overly zealous, “in your face” compression generates intermodulation products that occupy more spectra than the original content. While this increase in spectral density produces a seemingly “magical” increase in perceived loudness, it also translates to coarser quantization at the same bit rate. Accepted psychoacoustic models predict that the additional spectra will mask the increased quantization noise. Anecdotally, many users disagree.

Aggravation of the codec’s temporal masking function occurs when “peaks” are removed for the sake of increasing the content “density” [sic]. Psychoacoustically, peaks are an important element in the “pre” and “post” masking of quantization noise. Removing the peak removes or reduces the duration of the window where the codec may “hide” additional bit-saving quantization.

This doesn’t mean that volume leveling and compression are not allowed for perceptual codecs. It does mean, however, that certain fundamental processing rules be followed. “Codec friendly” volume management employs smart leveling that bases gain on the perceptual “usefulness” of the content. In addition to this it is necessary to use variable ballistics that shape the changes in gain to be cooperative with the codec.

The volume management must define a window transparent to the temporal masking function (TMF) that would maintain maximum allowable volume with adequate peak control. Within the constraints of TMF transparency in conjunction with allowable (and adequate) peak control would exist a function to control the TMF transparent window to maintain constant perceived

volume.

a) In the absence of smart leveling, regular compressors may be used if ratios are kept below 3:1 and release or averaging times are kept above 500 ms.

3) Image Re-parsing

For low data rate codecs there is often a need for additional processing. Without assistance, undesirable artifacts may become “unmasked” in less than optimal spatial environments as mentioned above. Parametric image re-parsing allows for spatial “reassignment” of spectra to alternate portions of the image. This reduces coding load of the dominant image axis and reduces the occurrence of “zero-quantizing” without having to “throw away” content. Image re-parsing may evolve into various static or adaptive architectures depending on the codec being used, the codec’s expected application, data rate, performance criteria etc. Image re-parsing objects are often customized specifically for a particular codec. These objects are processor intensive and application specific, so they are best accomplished by processing outboard to the codec.

4) De-Noising

It can not be overemphasized how much attention to detail in producing or processing the content can influence the performance of aggressive lossy data compression. De-humming and de-noising can improve the perceived quality of uncompressed content in addition to making the content more pleasing to listened to through data compressed channels regardless of the playback environment. The codec will “spend bits” on noise as it does on content, get rid of it.

5) Signature Sound and Sweetening

Sometimes it’s necessary to alter the spectral balance of content in anticipation of the environment of consumption. This may be done on a per song or per album basis in the mastering process, or it there may be a spectral shape “target” that is maintained by a broadcast processor. In the case of broadcast, it’s *both*. The recent pop trend toward confusing “brightness” with “resolution” is resulting in content that is treble heavy. High frequency clipping has become a pronounced part of what the consumer has to listen to on a full time basis, rather than the rare occasion when a peak has to be controlled with little perceptual consequence. This is not an attack on broadcast processor design, rather an observation of how it is eventually used. The “reasoning” behind the pinna blistering hash resulting from this kind of use is beyond the understanding of the author (end of rant).

While digital broadcast doesn’t employ emphasis and HF clipping, lossy compression can still become overwhelmed by a spectrally dense, peakless content stream. Broadcast codecs deliver the *best sounding* performance when the content is presented with the *lowest* possible spectral density. Don’t equalize and limit too much.

6) Bandwidth Reduction

Under certain circumstance bandwidth reduction may be used to offset anticipated coding difficulty. Under certain conditions, simple low-pass filter may be enough to reduce the coding load to where the content becomes inoffensive. Tilting of the spectra is of no use, the filter must be somewhat sharp to have good effect.

Facilities for dealing with Gibb's phenomena must be employed to avoid unexpected overload due to "overshoot". Simple clipping or limiting won't work as both give rise to high frequency components above the low pass cutoff negating the bandwidth reducing benefit.

Broadcast processing designers have developed many clever methods of low passing and limiting for the sake of over-modulation and pilot protection. These same methodologies could migrate toward this new application. Missing four or five notes on the top octave may sometimes be a good trade-off to avoid audible artifacts in the rest of the audio at aggressive bit rates.

To stay on the safe side, set absolute limits 3 dB below FS, anticipating sample rate or analog conversion and reconstruction filter induced overshoot.

Conclusion

This note should shed some light on what to expect when (not if) lossy data compression and "problem content" meet. If specific preparative steps for both existing and future content are not taken, unpleasant surprises will occur at the consumer end. Of course, the consumer end requires additional discipline and sorting out, but that is the subject of another paper. On the other hand, perhaps data compression is just forcing upon us some disciplines we should never have put to the side. Whatever the motivation, excellent results are to be had by those who take heed to the above suggestions. They are by no means exhaustive or absolute, but should provide a truly excellent head start over those that don't.

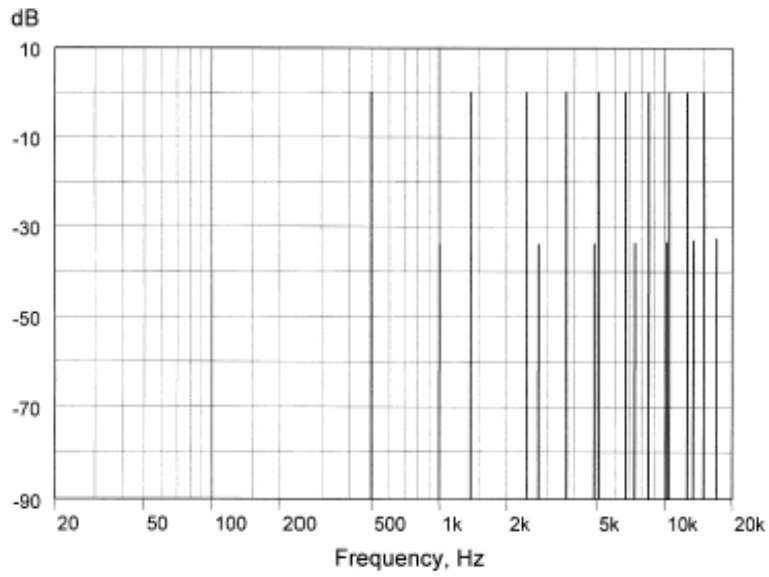


Illustration 1. Spectra generated by harmonic distortion (multi-tone stimulus) (After Cabot)

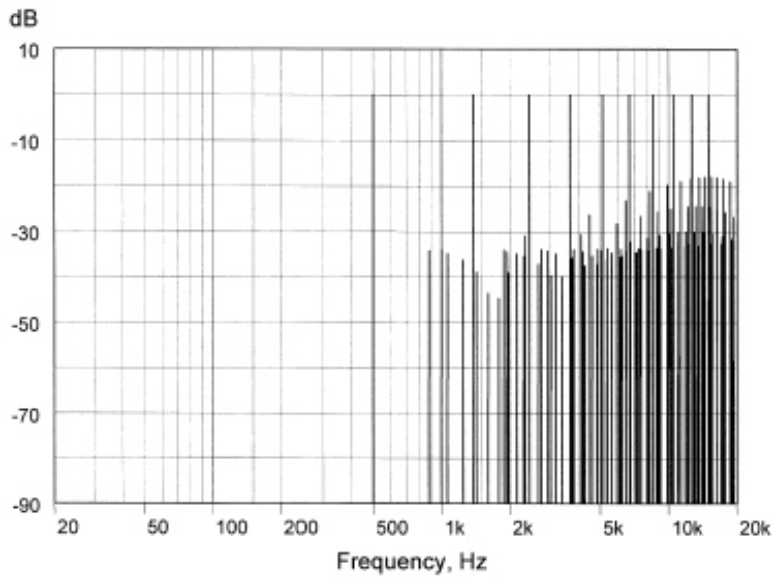


Illustration 2. Spectra generated by intermodulation distortion (multi-tone stimulus) (After Cabot)

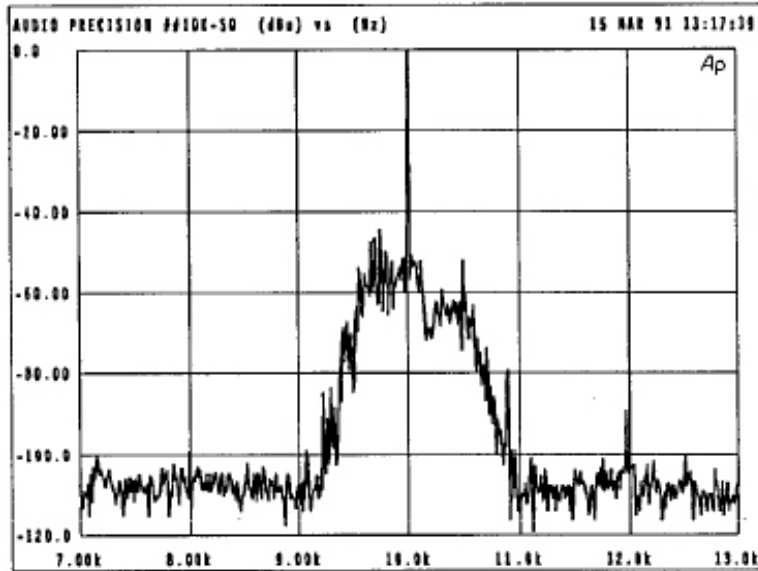


Illustration 3. Codec spectra output with 10 kHz and 10.5 kHz (-44 dB) tones. (After Cabot)

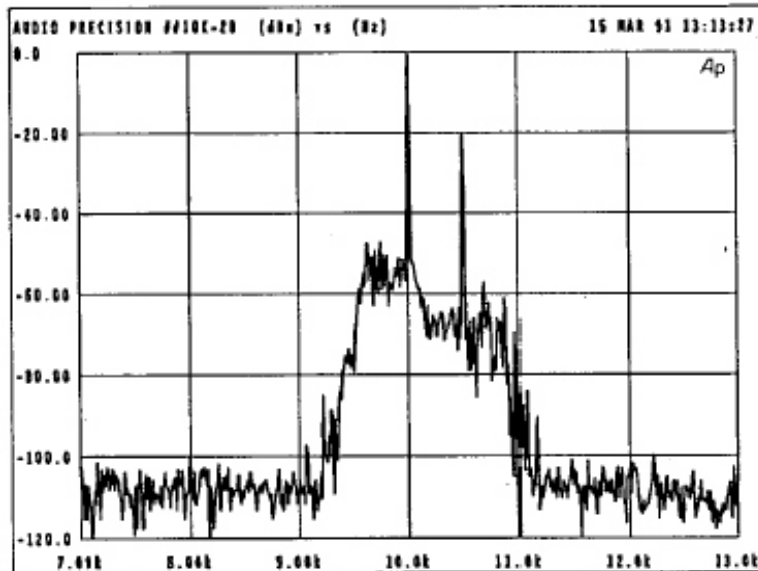


Illustration 4. Codec spectra output with 10 kHz and 10.5 kHz (-20 dB) tones. (After Cabot)

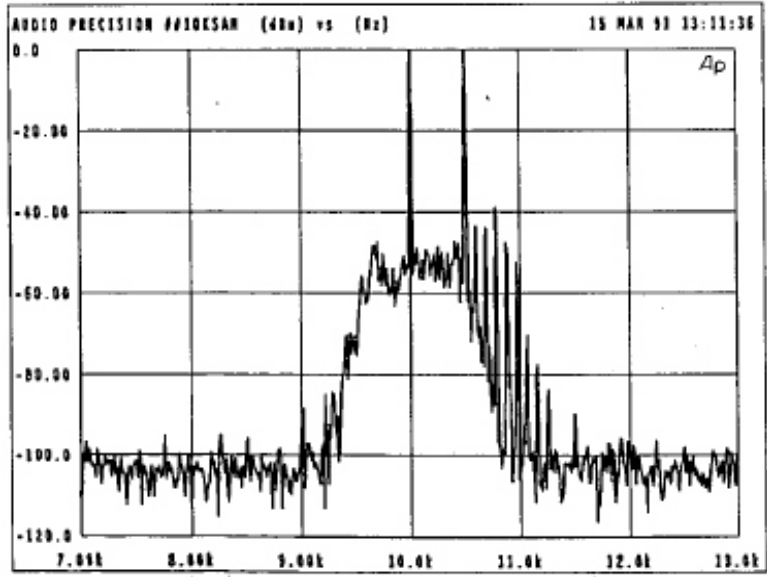


Illustration 5. Codec spectra output with 10 kHz and 10.5 kHz (-0 dB) tones. (After Cabot)