

# Codec Pre-Conditioning for Parametric Stereo

Robert Reams, Chief Scientist  
Neural Audio Inc.

Use of ultra-low bit rate audio coding like Parametric Stereo brings new sensitivity to production and mastering attributes within the stereo content. Modern audio processing techniques must solve these issues to maintain consistent and pleasing results in less than perfect listening environments.

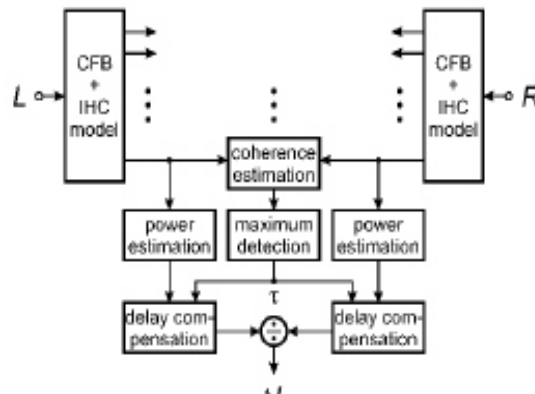
What it is:

The simplest definition of parametric stereo coding is as follows: a mono signal, down-mixed from the stereo source and compressed by powerful modern coding algorithms, is converted back into stereophonic sound by means of a coded parametric description of the spatial properties of the original signal.

Parametric Stereo offers a compact parametric representation of auditory spatial information such as localization cues inherent in multi-channel audio signals. Parametric Stereo as well as BCC and other like technologies, allows reconstruction of the spatial image given a mono signal and spatial cues that require a very low rate of a few kbit/s. This is a brief review of relevant auditory perception phenomena exploited by these technologies.

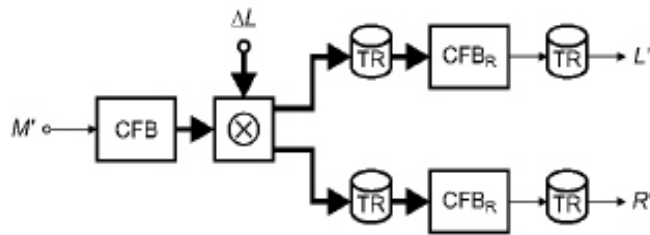
The Encoding Process:

An encoder (using BCC as an example) generally consist of two cochlear filterbanks (left and right) with simple inner hair cell models as a method of spectral parsing and detection. The power of each parse is estimated and a coherence weighted L/R intensity ratio is calculated as a transmittable coefficient. The coefficient group representing a spectral/spatial “map” of an encoded block is transmitted along with a downmix of the left and right channels.



The Decoding Process:

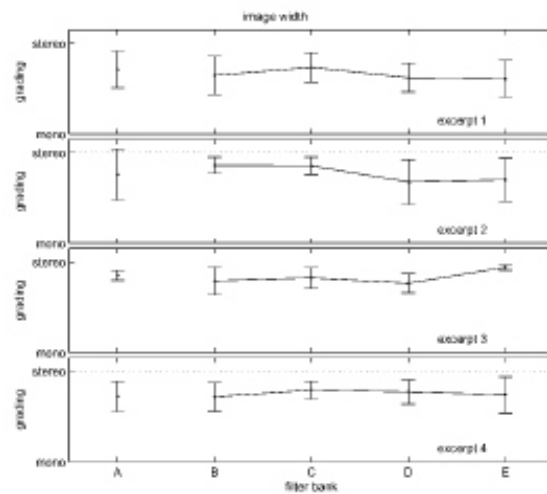
At the receiver end the coefficient group representing the spectral/spatial block “map” is then



decomposed into a series of ratios representing L/R intensity balance for each spectral parse of the cochlear filterbank. The ratio is then “split” into a pair of values; one for left and one for right. Each value is then convolved with it’s respective spectral parse . The parses are then summed to reconstruct the individual left and right channels.

| label | FB  | size                 |
|-------|-----|----------------------|
| A     | CFB | 98 non-uniform bands |
| B     | FFT | 2048                 |
| C     | FFT | 1024                 |
| D     | FFT | 512                  |
| E     | FFT | 256                  |

| excerpt | category    | left               | right   |
|---------|-------------|--------------------|---------|
| 1       | speech      | male               | female  |
| 2       | singing     | tenor              | soprano |
| 3       | percussions | castanets          | drums   |
| 4       | applause    | (stereo recording) |         |



This is a clever tradeoff between image distortion/stability and coding efficiency. Most experienced listeners note a perceived “image narrowing” that is a result of sacrificing the image fine structure. This is because of the statistical reduction (and sometimes elimination) of the “L-R” component of the stereo image during predominantly “L+R” passages. Perhaps the coherence estimation method suffers a common trade-off between “smoothness” and “agility” and as such sacrifices the image fine structure, giving an impression of a “collapsed image” under certain playback conditions. Again, this is clever trade-off.

As an example, the upper left chart illustrates the efficiency of a cochlear filter bank (CFB) vs. that of a Fast Fourier Transform (FFT). The CFB may be described with far fewer coefficients (98), making it a good choice for describing the spectral/spatial block “map”. The lower left chart depicts the test content. Note that the left and right elements of the first three excerpts are dissimilar or “hostile” and the fourth excerpt (stereo applause) is both transient and incoherent, this is most difficult to code. The right chart illustrates the “perceived image width and 95% confidence windows”. As can be seen, the CFB is an excellent choice. Validated performance is comparable with a 512 band FFT at a coding gain of better than 5:1. Again, this is a good tradeoff. However, the content must be preconditioned to prevent any “gotchas” as we anticipate which environments the content is going to be experienced in.

## USE’s

When played back through headphones, any “unnatural sounding” elements (USE’s) that are simply a result of these aggressive codecs are *adequately masked*.

When played back through a stereo “two speaker” environment with the listener sitting in the “sweet spot”, the USE’s are *adequately masked* with only the aforementioned “image narrowing” being noticeable only when compared with the original content.

When played back blindly through common matrix decoders (like Dolby’s “Pro-Logic II” and DTS’s “Neo-6”, Lexicon’s “Logic 7” and SRS’s “Circle Surround”) the surrounds channels completely unmask the resultant USE’s that are normally masked when it is reproduced in relatively ideal two channel stereo.

## Overcoding Conditions

Overcoding conditions exist in many forms. It usually starts with ill-conceived (random mixing of intensity and M/S image tracks and samples) content that was poorly mastered (usually with “mastering plug-ins” and other various do it yourself mastering in a box devices...not real mastering) sometimes made worse by seemingly unavoidable tandem coding of the content. This results in content with temporal, intensity or image mis-alignments (very common) which, in turn, causes drastic reductions in a codec’s efficiency as it must always overcode the opposing end of either image axis (L/R and S/D being the image axis pairs). The resultant overcoding causes more

violation occurrences of the codec's psychoacoustic model resulting in **more noticeable artifacts**.

Solutions must be designed to accomplish the following tasks while remaining psychoacoustically "invisible":

A) Spectrally re-parse the content image to reduce the audibility of USE's in:

- 1) less than optimal stereo environments
- 2) playback through "blind matrix" decoders
- 3) playback into a "synthetic spatial environment" (SEE)

B) Detect temporal centroid of L/R image axis and reduce mean deviation to zero

C) Detect intensity centroid of L/R image axis and reduce mean deviation to zero

D) Detect image centroid of S/D image axis and reduce mean deviation to zero

E) Reduce bit-consuming low and high frequency content noise

F) Reduce bit-consuming 60Hz hum (and harmonics) in content

These powerful architectures may be "tuned" to any lossy compression algorithm including AACplus and HDC. Addition of volume/peak management and spectral/image processing satisfy "signature sound" needs .

The results are excellent. Anecdotal results from 40 "expert" listeners yield a better than 59% (from 22% to 35%) averaged improvement in listener preference.

## Summary

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, de-noising, image management and image re-parsing 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.