

2-D Downmixing and Rendering as an Alternative Method of Transporting 5.1

Prepared for: "Spatial Coding of Surround Sound: A Progress Report"
Robert Reams, Neural Audio Inc. *AES 2004*

There has been significant movement toward what is known as "discrete 5.1" by both the content provider and the consumer. This is both good news and bad news for the broadcaster. On the upside, this renews the interest of the content hungry consumer in broadcast content, raising viewership and/or listenership. On the downside, integration of modern and legacy content *without the proper tools* can be a nightmare. The key driver of this challenge is the consumer expectation of what is considered "discrete 5.1". The consumer "knows" what 5.1 is supposed to "sound" like and expects all content to live up to this new expectation.

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 award 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. In fact, in terrestrial broadcast, satellite broadcast, cablecast or internet streaming, portions of the content (dynamics, image and spectral balance) will be modified or discarded if it will reduce distortions or artifacts on the consumer end.

This is, of course, horrifying to the audio artist.

As the content makes it's way down the production and distribution chain priorities may change. As an example:

- 1) The artist demands that an exact replica of their art is what the consumer will experience.
- 2) 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; packaged for real world distribution.
- 3) The broadcaster has to manage the content to be compatible with the lowest common

denominator of content transport to the consumer.

4) Finally, the consumer wants to experience as much of the content as his or her personal budget will allow.

The fact is: 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.

With modern spatial coding techniques, new 5.1 music content may be distributed over legacy broadcast infrastructures. A most recent issue to surface is the question of what constitutes sufficient discreteness. This is a brief discussion addressing this issue.

Achieving “Discreteness”

Separation or crosstalk between channels is a measurable and objective physical metric. It is possible to achieve “perfect” discreteness or “zero” crosstalk if the number of source elements never exceeds the number of available fine structures and/or there are no overlaps in the time/frequency plane of the content transport system. That being said, the result of eventual spatial and spectral “collisions” result in phenomenon akin to the effects of crosstalk. If we assume a reasonable playback environment, we can predict the relevancy of these “collisions” using established psychoacoustic models. From these we may construct a framework for predicting the contour of crosstalk irrelevancy to achieve perceptually acceptable discreteness in a predictable listening area.

Crosstalk Prediction

Cross-masking contours that define the “sweet spot” of the crosstalk mask may be predicted from available parameters. With these parameters we may predict loudspeaker and listener placement that would insure adequate performance.

Prediction of the crosstalk contour requires:

- 1) knowledge of the number of “places” (sub-bands) on the time-frequency plane
- 2) knowledge of the available number of simultaneously available discrete fine structures
- 3) a psychoacoustic model (including the loudspeaker and listener placement) predicting the crosstalk mask

1) Frequency-Space Placement

As the number of spectral bands decreases, *likelihood* of conflict along the frequency axis increases, and hence, so does the *probability* of crosstalk. A source which occupies non-overlapping regions in the time-frequency plane could be rendered with “zero crosstalk” regardless of how many “channels” it was “rendered” to if the sources had no conflict or overlap due to sufficient spectral granularity. [after Faller]

2) 2-D Placement and Downmixing

As the number of simultaneously available fine structures approaches the number of original channel sources the probability of crosstalk approaches zero. This merits (brief) discussion of the origin of the 2-D content as well as spatial rendering.

3) Psychoacoustic Model

Modern psychoacoustic models are applied to lossy compression algorithms to predict how much quantization (and the resulting noise) is allowed per critical band before audibility. A similar principle may be applied to predicting where crosstalk is perceptually irrelevant. The model is expanded to include listener placement contours within a fixed array of loudspeakers. With this model we can predict where crosstalk can exist with little perceptual impact and construct maps of the effective listening environment. We can also predict where concentrations of unnatural sounding elements will occur.

2-D Downmix Image Data Transport Method

On the downmix side, 2-D downmix imaging information may be described as an averaged version of the original image envelope as a function of spectra. The lateral axis, or azimuth of the original image envelope is conveyed as averaged left and right channel intensity differences as a function of spectra.

The front-to-back or “depth” axis of the original image envelope is conveyed as coherence between the left and right fine structure as a function of spectra.

Features worthy of note are: 1) there is no bitrate cost, 2) the content remains completely editable as the image info is embedded in the content waveform, 3) the image reconstruction info survives concatenation with unspecified lossy compression algorithms and analog and 4) 2-D downmixes are compatible with existing matrix style decoders.

On the upmix side, the lateral portion of the image envelope is reconstructed via the L/R intensity difference descriptor while the depth portion of the image envelope is reconstructed from the L/R coherence descriptor via a running average of normalized cross-correlation. As with any perceptual coding algorithm, acceptable (or even required) envelope averaging is wholly dependent on a psychoacoustic model resident within the decoder.

This also allows for the multichannel 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. This is an important feature as it reduces the consumer objection to co-mingling of content with different spatial goals and is preferred to an “all left and all right” approach of representing stereo in a multi-channel equipped environment.

In the absence of multi channel rendering the downmix is reproduced in two channels with no

spectral artifacts and ***no bit rate reduction***. These particular features are very important to broadcasters.

The image envelope, therefore, is imbedded in the audio. Intensity/coherence embedment is an excellent choice because of its similarity to the image construct of naturally occurring stereo and compatibility with already prevalent “legacy” content. This greatly simplifies the integration of 2.0 and 5.1 content within the transmission plant and allows both legacy 2.0 and 5.1 downmixes to be broadcast at an unimpaired bit rate. At ultra low bit rates (64kB/s) this would increase the number of bits available for the fine structures by 20% as there is no side data to transport.

Compatibility Discussion

From a perceptual standpoint both side data and embedment methodologies work remarkably well with original source encoded 5.1 content. However, within the transmission plant, there are challenges for side information based methods.

- 1) **Editing** becomes an issue because bitstreams can't be mixed, cut or cross-faded.
- 2) **Servers** or archives are either linear or compressed, usually with the content format already decided (MPEG-1 Layer 2, linear etc.) with the legacy content already in place. This means that any system that requires a change in format also requires the server to be “re-tooled” to the new side information based compression format.
- 3) **Switchers, mixers, sweetening**, the very wire the stations is made up of must be replaced.
- 4) **The entire infrastructure** of the radio station must be changed.
- 5) **Board-ops must be re-trained**

On the other hand, the 2-D downmix is **100% compatible** with existing servers, switchers, wiring (analog or not) and may be edited with existing systems. The server may be operated “as is”, compressed or not and all consoles and processing equipment may be used as always.

The proposed future linear 5.1 plant doesn't solve all of the challenges:

- 1) **Editing** is still an issue because eventually 5.1 and 2.0 will still have to be mixed/crossfaded (where in the cross-fade do you cease to be 5.1 and become 2.0? Meta-data approaches don't work in television and they don't work here)
- 2) **Servers** (including RAID, tape backup and troubleshooting) will still drive up both time and expense (does 2.0 take up the same bit rate as 5.1?)
- 3) **Switchers, mixers, sweetening**, the very wire the stations is made up of will still need to be

replaced.

4) The entire infrastructure of the radio station must still be changed.

5) Board-ops must still be re-trained

Do you really get significant benefit for all of this cost???

Conclusion

2-D downmixing and multichannel rendering are new methodologies that meet 2.0/5.1 interoperability challenges head-on. 2-D downmixing allows the distributor or broadcaster the ability to capture original source 5.1 content and downmix it to a 2.0 channel spatial format that survives aggressive lossy compression and conversion to analog. Downmixed content may be broadcast or distributed through existing 2.0 infrastructures. To broadcasters, this could yield new profit centers as web-casting, pod-casting are 2.0.

This was one of the attractions of 5:2:5 matrixing. Matrixes were an excellent solution in that they solved transition compatibility issues and allowed, to a certain extent, 2.0 content to co-exist with 5.1 content while extending the value of the ubiquitous 2.0 infrastructure. Although they have fallen short of the distributor and consumers expectation of what is now called “discrete” 5.1, they are still a format to which the broadcaster must remain compatible. To do otherwise is to risk incompatibility with some 120,000,000 decoders in consumerland.

The SEE n-channel rendering 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 complexity of implementing 5.1, thus driving the ubiquity of the automotive 5.1 system.

It is obvious that the interspersion of legacy 2.0 and 5.1 content is a reality. Unless these are made interoperable 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 broadcaster’s and consumer’s expectations slows the adoption of 5.1.

The proposed methodology allows content providers to promote their content to the consumer without bankrupting the radio station. It allows the radio station to conduct “business as usual”, with editing for play-out, voice-overs, creation of promotional materials and station ID tags while archiving in 5.1 and stereo content on the existing servers. To quote Mike Pappas, Chief Engineer of KUVU in Denver and noted 5.1 production expert... “It’s simple and effective”.