Coding Backward Compatible Audio Objects with Predictable Quality in a Very Spatial Way

Stanislaw Gorlow

To cite this version:

Stanislaw Gorlow. Coding Backward Compatible Audio Objects with Predictable Quality in a Very Spatial Way. 139th Audio Engineering Society (AES) Convention, Oct 2015, New York, NY, United States. pp.4. hal-01251648

HAL Id: hal-01251648
https://hal.science/hal-01251648

Submitted on 6 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Audio Engineering Society

Convention e-Brief

Presented at the 139th Convention
2015 Oct 29–Nov 1, New York, USA

This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Audio Engineering Society.

Coding Backward Compatible Audio Objects with Predictable Quality in a Very Spatial Way

Stanislaw Gorlow[1,2]

[1]*Gorlow Brainworks, 75014 Paris, Île-de-France, France*
[2]*Laboratoire Bordelais de Recherche en Informatique, 33400 Talence, Aquitaine, France*

Correspondence should be addressed to Stanislaw Gorlow (`stanislaw.gorlow@gmail.com`)

ABSTRACT

A gradual transition from channel-based to object-based audio can currently be observed throughout the film and the broadcast industries. One paramount example of this trend is the new MPEG-H 3D Audio standard, which is under development. Other object-based standards in the market place are DTS:X and Dolby Atmos. In this engineering brief, a newly developed prototype of an object-based audio coding system is introduced and discussed in terms of its technical characteristics. The codec can be of use everywhere where a given sound scene is to be re-rendered according to the listener's preference or environment in a backward compatible manner. The areas of application cover not only interactive music listening or remixing, but also location-dependent, immersive, and 3D audio rendering.

1 Introduction

Object-based audio coding is a more recent field of research that poses a challenge to the existing audio coding approaches. Its main objective is to represent a complex sound scene consisting of many different sound sources in a compact and reversible manner. A certain degree of quality loss is usually tolerated [2].

Not so long ago, a coding system for multitrack music recordings was presented in [3]. The output format in the case of music is typically stereo. This means that the mixture signal space is limited to two dimensions. As a consequence, the quality of the separated tracks has an upper bound that is determined by the spectral overlap and 2nd-order dynamics of the source signals, the spatial distribution of the sound sources, and their presence in the mixture in terms of loudness. In [4] it was shown that the system was sensitive to perceptual audio coding, such as MP3 or AAC. The two-channel mixing approach was extended to an arbitrary number of channels in [5]. There, it was shown that and how a set of single-channel source signals can be combined to form a multi-channel mixture and then separated or unmixed with the signal quality at the output being set in advance, i.e. prior to (down-)mixing. Accordingly, the compression ratio is given as the ratio between the number of sources and the number of channels.

In this brief, a revised coding system is presented that remedies the shortcomings of [3]. These are:

1. The intrinsic quality limitation that is due to the fixed number of mixture channels;

2. The audible degradation of sound quality that is due to perceptual audio coding.

In addition, the findings from [6] were also taken into account during the design of the system. As an extra, dynamic range compression/decompression [7] can be integrated in the new codec on demand [8]. The codec was developed in partnership with Aquitaine Science Transfert[1] and a C/C++ prototype is now available. It also features MP3 coding and decoding functionality through the LAME[2] and mpg123[3] libraries.

The novelty of the new coding system consists in the generation of extra data in addition to the regular, i.e. artistic, mixture, which allows the decoder to separate the tracks at the desired quality level. The extra data is generated adaptively on a block-by-block basis, which reduces the extra data rate significantly in comparison to, e.g., residual coding. It is important to understand in this context that the frequency content of a residual, as a rule, is equally complex as the frequency content of the source signal. Thus, to code the deviation of an unsatisfactory estimate from the original signal is just a clumsy way of coding the actual signal.

2 System Overview

In this section, the overall coding system is presented in the form of block diagrams (encoder and decoder).

2.1 Encoder

The encoder is depicted in Fig. 1.

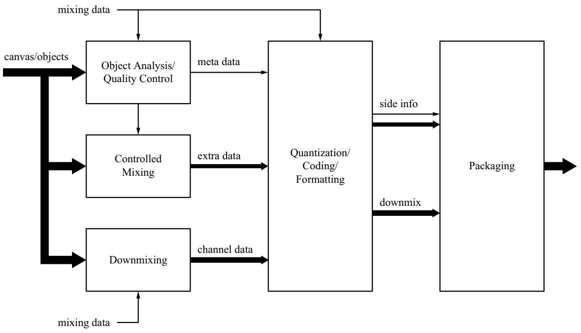

Fig. 1: Multichannel object-based audio encoder

2.2 Decoder

The decoder is depicted in Fig. 2.

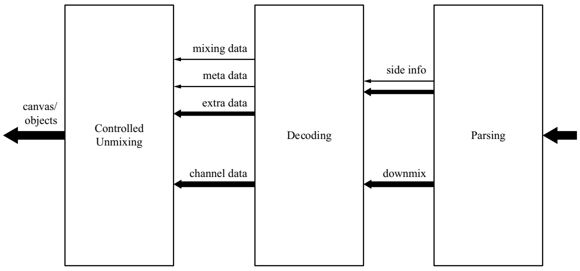

Fig. 2: Multichannel object-based audio decoder

3 System Properties

In this section, the most significant properties of the new coding system are elaborated in more detail. These are:

Flexible object format The audio objects can be single-channel *or* multi-channel. The static background of an acoustic scene, the sound canvas, would typically be represented as a multi-channel object, while the (active) objects (on top) would typically be single-channel.

Perception-aware quality metric The metric used for quality control is both spatial *and* perceptual. And so it gives a good estimate of the sound quality.

Predictable sound quality With the help of the quality metric it is possible to predict the attainable sound quality after decoding *prior to* downmixing.

Flexible mode of operation The codec can be operated in three different modes:

1. *Regular mode.* In this mode, only the artistic mixture is generated from the canvas and the objects. This is for (passive) listening.
2. *Virtual mode.* In this mode, a synthetic mix is generated, which maximizes the quality with a minimum number of virtual channels. This is for data reduction, but not for listening.
3. *Hybrid mode.* This is the default, or where the magic happens. This mode combines the two previous modes: the regular mixture channels are increased by the virtual channels in such a way that the desired quality is attained with a minimum number of virtual

[1] `http://ast-innovations.com/`
[2] `http://lame.sourceforge.net/`
[3] `http://www.mpg123.de/`

channels. This is the mode that enables both passive *and* active listening in high quality.

Coding efficiency The (spatial) coding is typically such that the total number of channels is at most equal to the number of single-channel objects. Accordingly, a stereo canvas would be counted as two objects. As an option, the number of virtual channels can be set to any number or determined a priori by the quality control mechanism and held fixed during the entire encoding process, which leads to a constant channel rate (CCR), or their number can be determined on a block-by-block basis and varied, thus resulting in a variable channel rate (VCR). This approach allows for the lowest average channel rate (ACR) w.r.t. the desired quality, as the number of channels and thus the spatial resolution is adapted locally to the signal constellation, which, in other words, translates into the number of active objects, the spectral dynamics, the spatial locations, and the relative volume levels. And since the content from the regular mix is used as well, the amount of virtual data is much smaller than what it would be, if all the objects were coded separately. To reduce the data rate even further, the (spatial) coder can be used in tandem with any type of (perceptual) audio coder. In that case, it acts as a spatial pre-processor for the encoder and as a post-processor for the respective decoder, accordingly.[4]

Digital rights management The regular and the virtual channels alike, both carry realizations of mixtures, which are generated from the audio objects. In the case of the virtual channels, these realizations may even be “sparse”. This accounts for the reduction of the overall data rate. This can be seen as some sort of *implicit* copy protection, because the objects are separately inaccessible without the decoder. Access can further be restricted in such a way that distinct objects cannot be soloed during playback, etc.

Audio objects over IP The spatial data that enables the system to meet the quality constraint and the meta-data are generated on a frame-by-frame basis. This is also the case for most if not all transform coders, such as MP3. Thus, the codec can also be used for audio *streaming*.

Mixing and mastering The mixing stage in the encoder may be followed by a mastering stage. This applies only to the artistic mixture. The mastering stage, in a typical scenario, would comprise a multi-band or single-band compressor and a limiter, possibly with look-ahead. These elements would then be undone on the decoder side through inverse modeling. This allows for the creation of mixes that are even *more realistic* and so, closer to the CD or the radio edited version of the same multi-track.

4 Possible Applications

In this section, some possible applications for the codec are mentioned. They mainly relate to:

Music publishing Consider the case where the mixing engineer exports the tracks (bass, guitar, vocal, etc.) that are to be mixed together and notes down the panning and the gains of the tracks. The mastering engineer does the mastering on a premixed version of the multi-track and notes down the compressor model and the respective parameters. Then, to generate a re-mixable version of the song, the encoder, which could also be a plugin for the DAW, would take the tracks as input together with the mixing and the mastering parameters. It would generate all the necessary side information and eventually generate the downmix. On the consumer side, the corresponding decoder would take the downmix and the side information as input. The side information would grant access to the composite tracks by separating them from each other in the mix. A re-rendering interface would allow the user to manipulate an apparently rendered piece of music without the direct access to the separate tracks.

The central idea is to have a backward compatible default version of a song for playback and some sort of a virtual interface that gives one access to the different objects (tracks) that compose the mix. Through the interface, one can e.g. re-spatialize the sound scene w.r.t. a new context or go back to the default at any time without “destroying” it. In short, one can *personalize* a generic rendition.

[4] Since the virtual channels contain (coded) audio, it might even be possible to use DTS’ Multi-Dimensional Audio (MDA) open format.

Imagine you download the radio version of your favorite song, but additionally you have the option to buy the remixing feature. If you have the corresponding decoder that support this feature, the re-rendering interface would enable you to remix the song in a very basic way that does not require any sound engineering knowledge. To realize this in high quality, the multi-track recording is required. This means that the content creator must work together with a studio or a record label.

3D sound The advent of new powerful VR consumer gadgets (VR headsets, VR gloves, etc.) brings not only new opportunities but also challenges, such as the need for the sound to keep pace with the visual and tactile stimuli. Hence, another area of application for the codec could be related to 3D sound rendering. More generally speaking, the codec could be useful in a context where the sound scene is to be matched to the topology of the room and/or the relative location of the viewer/listener in an enclosed space.

Imagine you enter a virtual 3D (sound) scene in which you are surrounded by sound objects. Using VR gloves you could displace the sound objects or move them closer of further away. The 3D sound rendering engine would adapt the aural impression to your relative position in that VR space and to the manipulations made. If you switch off 3D, the default sound scene, which could be in stereo, 5.1 or 7.1, is reproduced.

A concrete example is a vehicle with a 3D sound system. Imagine you buy a CD of your favorite artist that was "enriched" by the technology. You could choose to either listen to the stereo sound or you could switch on the AR mode and have the sound in 3D with the option to manipulate the spatial sound according to your personal taste.

Other applications would typically focus on new ways of delivering user-centered audio content with the aim of a greater audience engagement.

5 Summary

In this brief, a new multichannel object-based audio coding scheme was presented. Its key features are:

- Backward compatibility,
- Scalable data size and sound quality,
- Robustness against perceptual audio coding.

Its possible applications lie in the area of digital music publishing and 3D sound rendering related to AR/VR. More applications should emerge in the future.

Acknowledgments

This work was partially funded by "Aquitaine Science Transfert" through the Caumbot project (PJ 2014-089).

References

[1] Gorlow, S., *Reverse Audio Engineering for Active Listening and Other Applications*, Ph.D. thesis, Université Bordeaux 1, 2013.

[2] Gorlow, S. and Marchand, S., "Informed Audio Source Separation Using Linearly Constrained Spatial Filters," *IEEE Trans. on Audio, Speech and Language Proc.*, 21(1), pp. 3–13, 2013.

[3] Marchand, S., Badeau, R., Baras, C., Daudet, L., Fourer, D., Girin, L., Gorlow, S., Liutkus, A., Pinel, J., Richard, G., Sturmel, N., and Zhang, S., "DReaM: A Novel System for Joint Source Separation and Multi-Track Coding," in *133rd AES Convention*, pp. 1–10, 2012.

[4] Gorlow, S. and Marchand, S., "On the Informed Source Separation Approach for Interactive Remixing in Stereo," in *134th AES Convention*, pp. 1–10, 2013.

[5] Gorlow, S., Habets, E. A. P., and Marchand, S., "Multichannel Object-Based Audio Coding with Controllable Quality," in *IEEE ICASSP 2013*, pp. 561–565, 2013.

[6] Gorlow, S. and Marchand, S., "Informed Separation of Spatial Images of Stereo Music Recordings Using Second-Order Statistics," in *IEEE MLSP 2013*, pp. 1–6, 2013.

[7] Gorlow, S. and Reiss, J. D., "Model-Based Inversion of Dynamic Range Compression," *IEEE Trans. on Audio, Speech and Language Proc.*, 21(7), pp. 1434–1444, 2013.

[8] Gorlow, S. and Marchand, S., "Reverse Engineering Stereo Music Recordings Pursuing an Informed Two-Stage Approach," in *DAFx-13*, pp. 1–8, 2013.