W19 Hybrid Audio Coding

Parametric Spatial Coding of Immersive Audio

HEIKO PURNHAGEN

DOLBY SWEDEN AB

142ND AES CONVENTION, 2017 MAY 20-23, BERLIN

IMMERSIVE AUDIO FORMATS (3D AUDIO)

Object-based Immersive Audio

- Intuitive content creation
- Optimal reproduction over a large range of playback configurations using suitable rendering systems
- Example: Dolby Atmos

Channel-based Immersive Audio

- Evolution of established 5.1 surround format
- Content creation and workflow similar to 5.1
- Examples: 5.1.2, 5.1.4, 7.1.4

DELIVERY OF IMMERSIVE AUDIO TO CONSUMER ENTERTAINMENT SYSTEMS

Broadcast or streaming of immersive audio at low bit rates requires efficient representation

- Approach: Parametric spatial coding

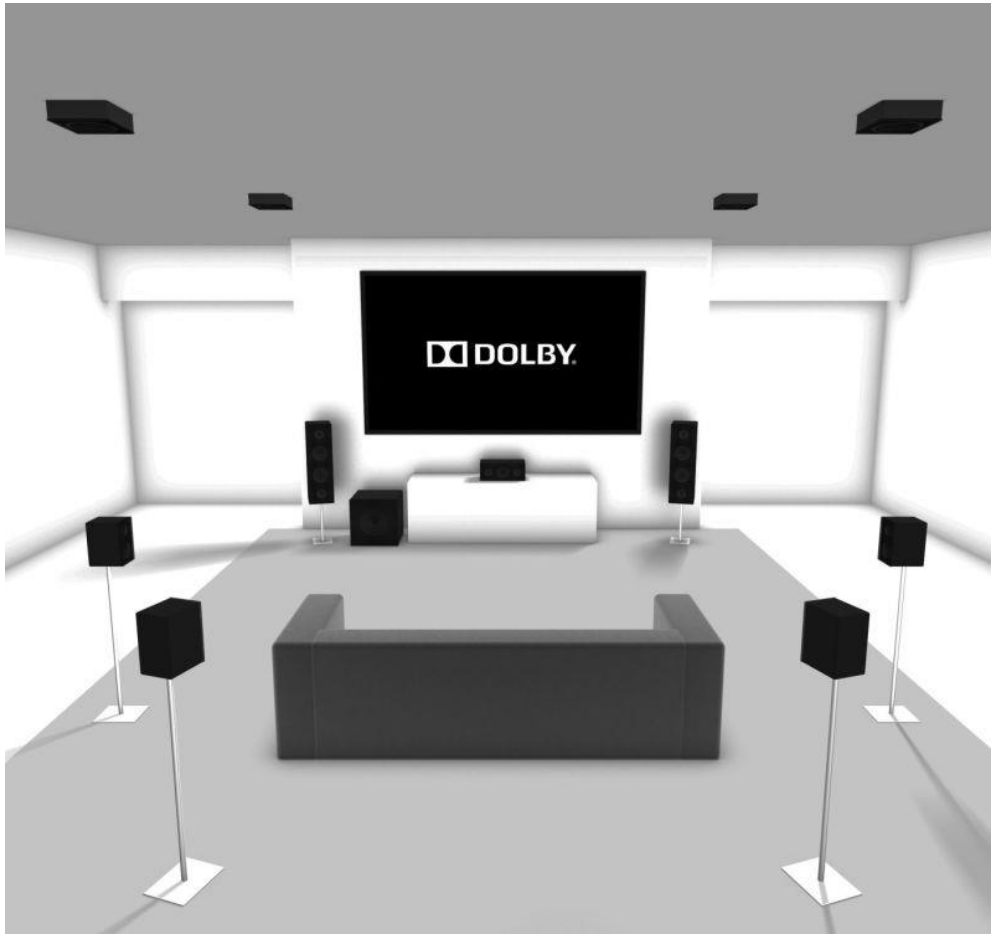
Example: Dolby AC-4 System

Channel-based immersive audio in AC-4

- Advanced Coupling (A-CPL)
- Advanced Joint Channel Coding (A-JCC)

Object-based immersive audio in AC-4

- Advanced Joint Object Coding (A-JOC)



PARAMETRIC SPATIAL CODING FOR CHANNEL-BASED CONTENT

Parametric spatial coding

- Convey N-channel signal using M downmix channels and parametric side information
- N-M-N system

Decoding

- Time- and frequency-varying upmix, controlled by parametric side information
- Decorrelators help restoring perceptual cues such as ambience or width
- Goal: Re-instate time- and frequency-varying covariance matrix of N-channel signal

Examples

- Parametric Stereo: 2-1-2 system
- MPEG Surround: 5-2-5 system (cascade of one 3-2-3 and two 2-1-2 modules)

Problem: How to handle channel-based immersive audio with many channels?

- Approach: Arrange channels in groups, use one downmix channel per group

PARAMETRIC SPATIAL CODING USING SINGLE DOWNMIX CHANNEL

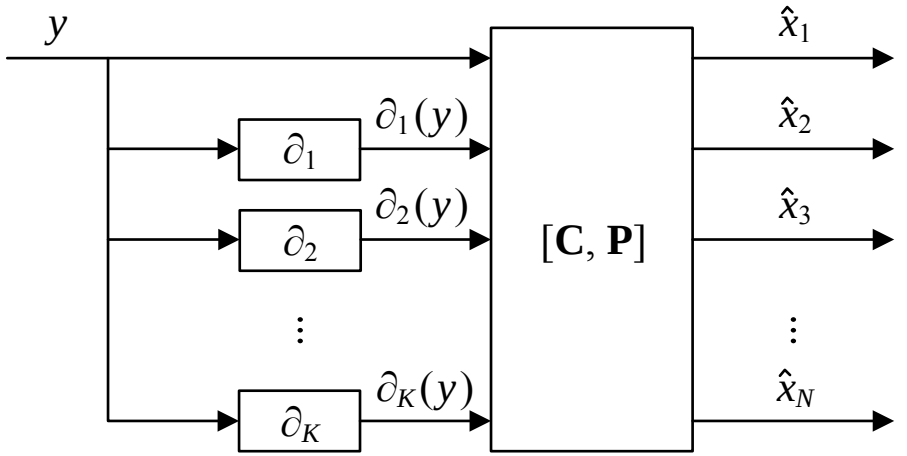
Convey group of N channels using N-1-N system

Upmix with K = N-1 decorrelators

- Dry upmix coefficients: $\mathbf{C}$ (size N x 1)
- Wet upmix coefficients: $\mathbf{P}$ (size N x K)

Goal: Re-instate covariance in each time/frequency tile

- Dry coefficients: best waveform approximation (least squares)
- Wet coefficients: compensate for missing covariance in dry-only upmix



Compact parametrization (assuming downmix is sum)

- 2-1-2 system: 2 coefficients: α, β
- 3-1-3 system: 5 coefficients: $c_1, c_2, h_{11}, h_{12}, h_{22}$

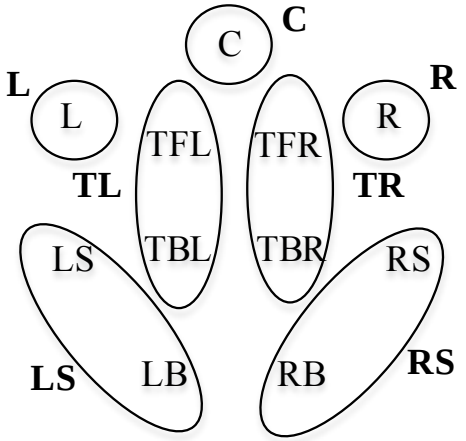
$$\mathbf{C} = \frac{1}{2}\begin{bmatrix} 1+\alpha \\ 1-\alpha \end{bmatrix} \qquad \mathbf{P} = \frac{1}{2}\begin{bmatrix} \beta \\ -\beta \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} c_1 \\ c_2 \\ 1-c_1-c_2 \end{bmatrix} \qquad \mathbf{P} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix}$$

ADVANCED COUPLING (A-CPL)

Convey 7.1.4 content using 5.1.2 downmix

- Four groups with 2 channels (2-1-2)
- Three “groups” with 1 channel
- LFE
- Typically 7.5 kb/s side information



ADVANCED JOINT CHANNEL CODING (A-JCC)

Convey 7.1.4 content using 5.1 downmix

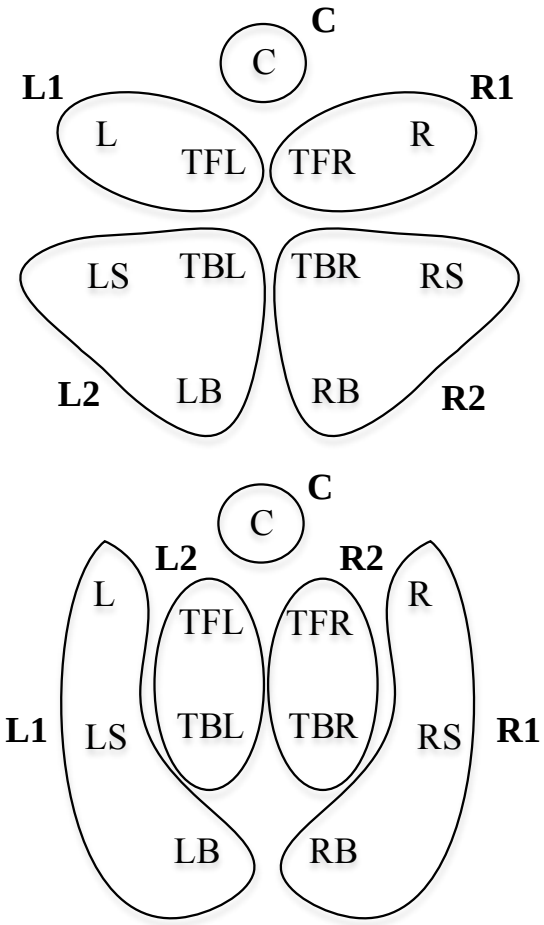
- Two groups with 3 channels (3-1-3)
- Two groups with 2 channels (2-1-2)
- One “group” with 1 channel
- LFE
- Typically 7.5 kb/s side information

Promising downmix configurations

- 5.1.0
- 3.1.2

Content-adaptive downmix

- Select downmix that requires less “wet” contributions
- Interpolation of full upmix matrix (11 channels) ensures smooth transitions



EXPERIMENTAL RESULTS: QUALITY-RATE CURVE

MUSHRA test with 7.1.4 content

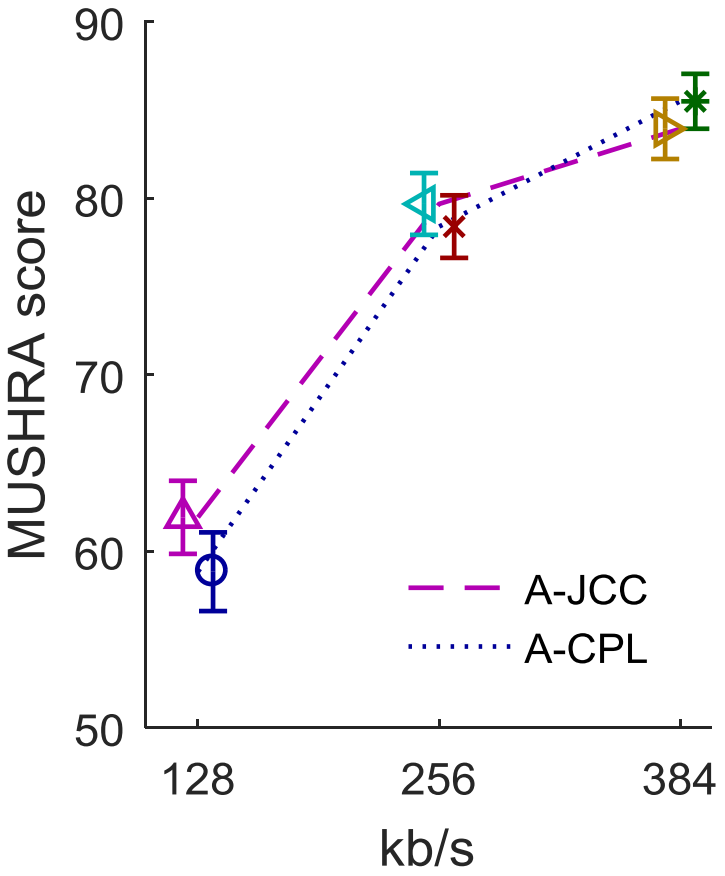
- A-JCC and A-CPL
- 128 kb/s, 256 kb/s, 384 kb/s

Comparison: A-JCC vs. A-CPL at 128 kb/s total bitrate

- A-JCC: 5 full-band downmix channels, typically 24 kb/s per channel
- A-CPL: 7 full-band downmix channels, typically 17 kb/s per channel

A-JCC uses more extensive parametric spatial coding

- Beneficial at lower rates:
  Fewer downmix channels coded at better quality
- Quality saturates earlier towards higher rates:
  Limitations of the underlying parametric model



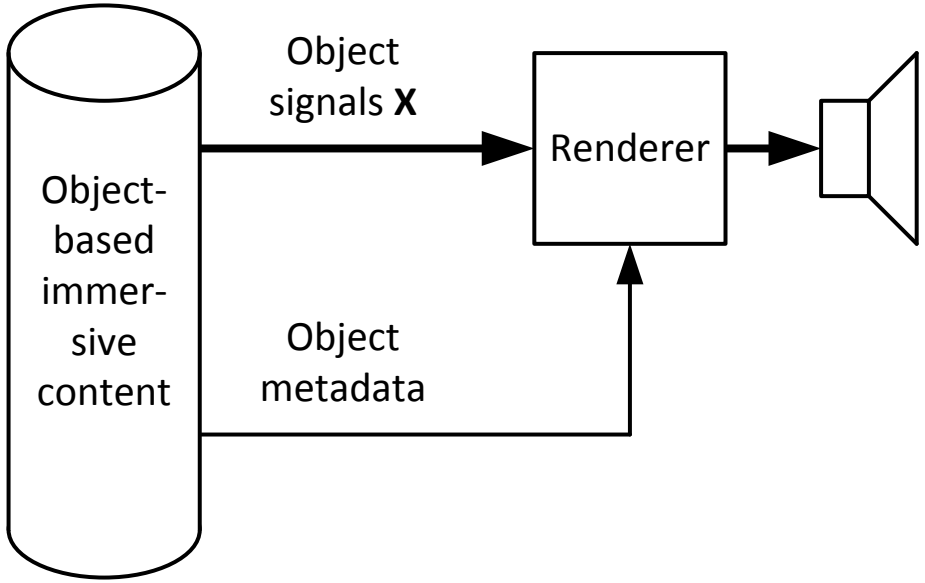
PARAMETRIC SPATIAL CODING FOR OBJECT-BASED CONTENT

MPEG Spatial Audio Object Coding (SAOC)

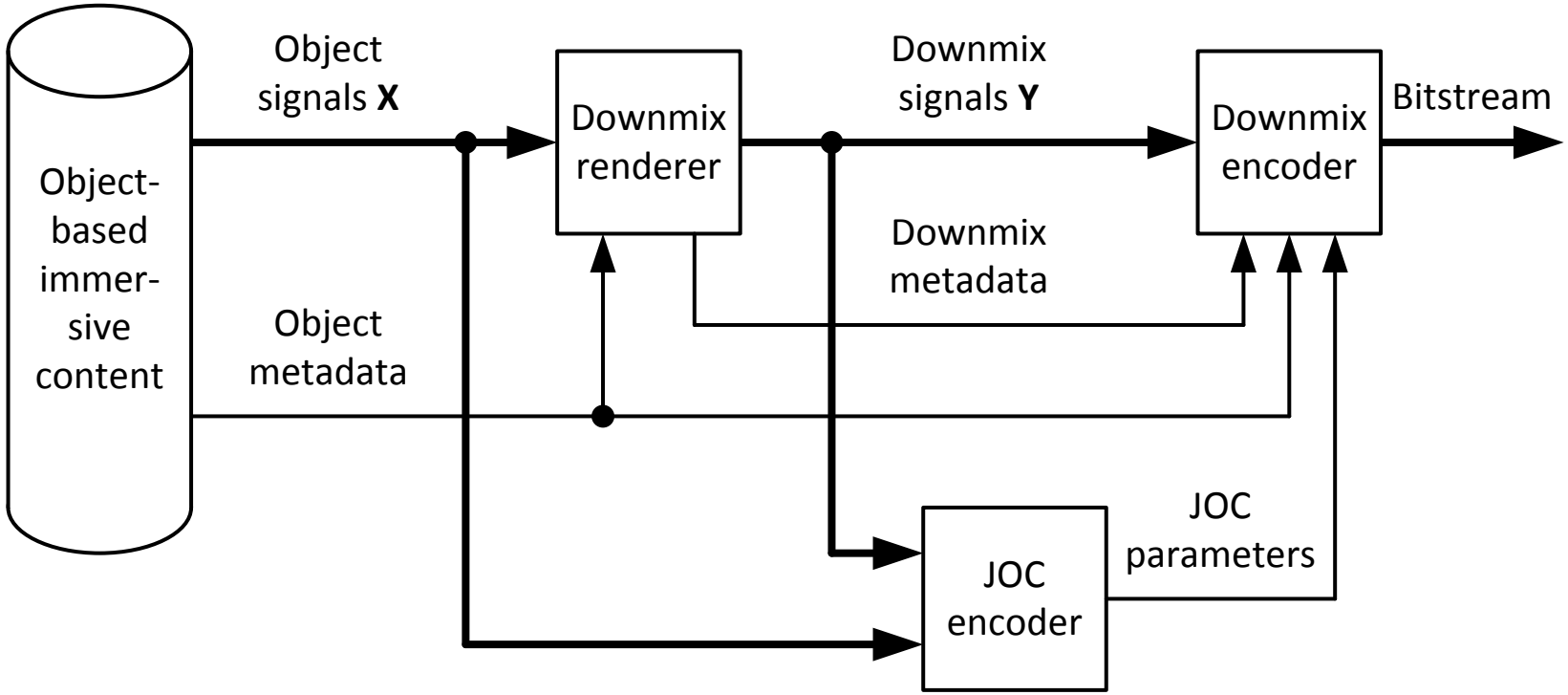
- Combines object reconstruction and rendering

Joint Object Coding (JOC)

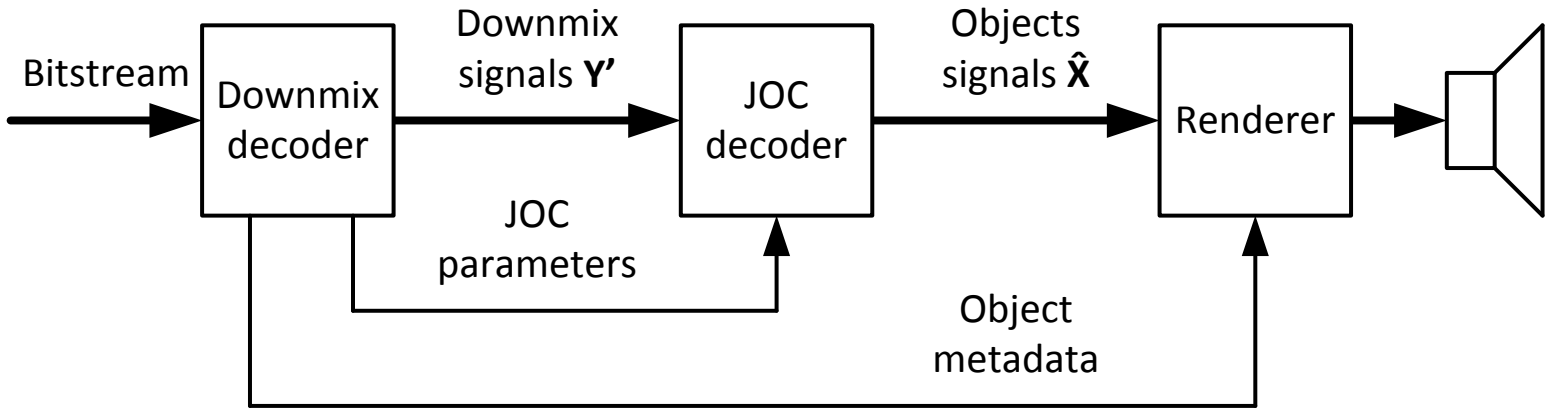
- Reconstruct object signals
- Enables rendering for various playback systems:
  - Immersive speaker playback (e.g. 7.1.4)
  - Immersive soundbar playback
  - Binaural headphone playback



THE JOINT OBJECT CODING PARADIGM – ENCODER

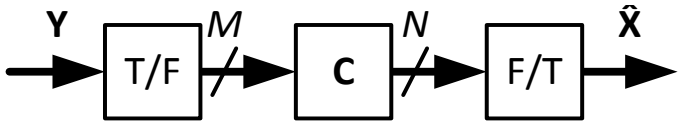


THE JOINT OBJECT CODING PARADIGM – DECODER



BASIC APPROACH

Reconstruct N object signals **X**
from M downmix signals **Y**
using dry upmix matrix **C** size N x M



JOC parameters (transmitted as side information)

- Matrix elements of **C** per T/F tile (time- and frequency-variant)
- Typically 7 to 12 frequency bands (grouped complex-valued QMF bands)
- Typically 32 to 43 ms temporal stride (with interpolation)

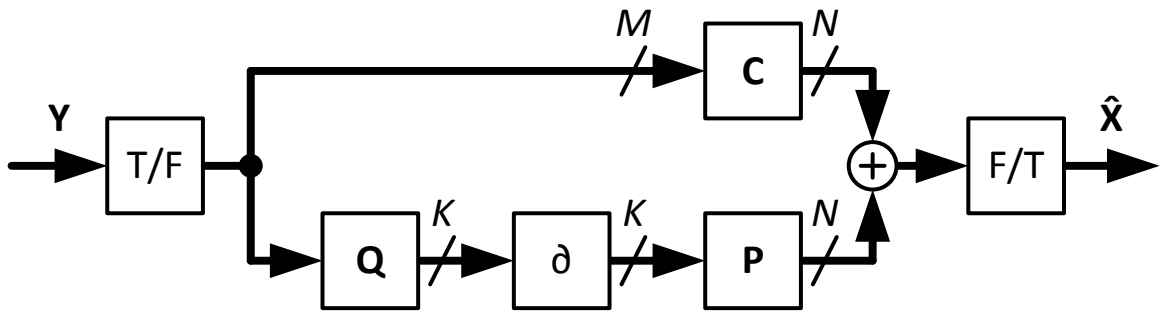
JOC encoding strategies

- Least squares reconstruction of each object in each T/F tile
- Compensate for prediction loss by “gaining” elements of **C** (variance reinstatement)

OBJECT DECORRELATION

Problem

- Dry upmix can't ensure object covariance reinstatement



Approach

- Introduce K decorrelators
  (thumb rule: K = N – M, but fewer decorrelators are often sufficient)
- Transmit also wet upmix matrix **P** size N x K (per T/F tile)
- Pre-matrix **Q** is computed from **C** and **P** in decoder

Encoding strategy

- Target object covariance reinstatement

DOWNMIX STRATEGIES IN JOC ENCODER

5.1 channel-based downmix **Y**

- Compatible with legacy decoders without JOC decoding

Adaptive downmix **Y**

- Each downmix channel is a dynamic spatial group of neighboring objects
- Example:
  - Some downmix channels carry objects in listener plane
  - Other downmix channels carry objects in the ceiling
- Downmix metadata enables low-complexity core decoding of adaptive downmix

CONCLUSIONS

Parametric spatial coding enables efficient representation of immersive audio for delivery to consumer entertainment systems at low bit rates

Channel-based immersive content

- Joint Channel Coding
- More extensive parametric spatial coding advantageous towards lower target bit rates

Object-based immersive content

- Joint Object Coding
- Reconstruction of object signals enables rendering for various playback systems