Efficient Representation of Sound Images: Recent Developments in Parametric Coding of Spatial Audio

Jürgen Herre
Fraunhofer Institut für Integrierte Schaltungen (IIS)
Erlangen, Germany

Introduction: Sound Images … ?

- Humans live in a world of sound
- … continuously listen to the sound surrounding us
- … perceive the acoustic waves reaching the ears and interpret them
- … (re)construct an acoustic scene
- In analogy to visual perception, we form a "sound image" . . .

This talk: How to efficiently represent and reproduce sound images!

Overview

Part I

- What constitutes a "sound image"?

Part II

- Efficient coding of spatial sound images
  - Perceptual audio coding
  - Coding of multi-channel / surround sound
  - "Spatial Audio Coding" & MPEG Surround

Part III

- Next-generation interactive coding / rendering of sound images: "Spatial Audio Object Coding"
  - Demonstration

Part I:

What Constitutes A “Sound Image”?

Fraunhofer Institut
Integrierte Schaltungen

Spatial Sound Perception

The Physics Part

- We are permanently listening to sound through our two sensors (ears)
- Ears receive a complex sound mixture of
  – direct sound that is radiated from several sources (frequently concurrently), and
  – many reflections from our acoustic environment ("ambient sound")

Fraunhofer Institut Integrierte Schaltungen

Spatial Sound Perception (2)

The Perceptual Part

- **Humans interpret sound to make sense of it**
  - Map it to an internal sound scene which is a *perceptual correlate* of the actual physical scene (not necessarily identical!)
  - Very complex & sophisticated process - by far not fully understood!

Some Key Words

- **Psychoacoustics** [Zwicker, Moore, Fletcher, Blauert, …]
- **Perceptual streaming, auditory scene analysis** [Bregman, …]
  - How to form a consistent interpretation of the acoustic world from the stream of received information fragments

Fraunhofer Institut Integrierte Schaltungen

Spatial Sound Perception (3)

A Note of Caution

- In the following, we will look at the phenomenon of spatial audio perception in a grossly simplified, yet illustrative and very useful way …
- Allows immediate technical application to efficient coding of spatial sound
- Highlight some analogies between “sound images” and “visual images”

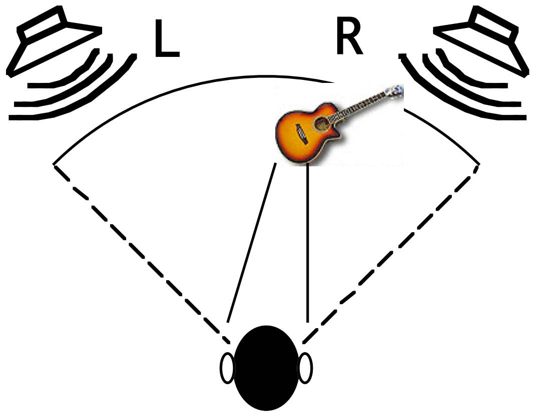
Frequency Selectivity

- The Human Auditory System processes sound in a frequency selective way (roughly logarithmic, ca. 1/3 octave → BARK and ERB scales)

Fraunhofer Institut
Integrierte Schaltungen

Spatial Sound Perception (4)

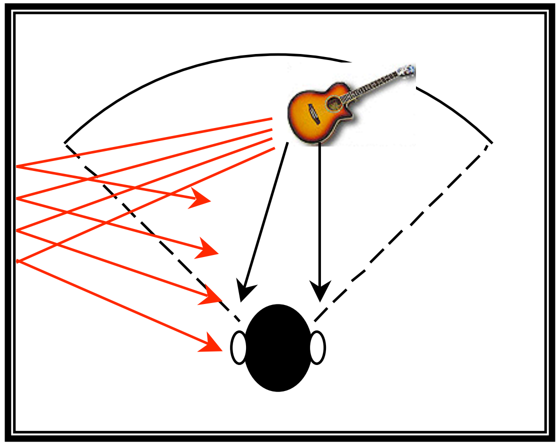
Sound Localization



- Complex process, mostly driven by interaural signal differences
- Depending on its position, a sound source produces differences in the ear signals
- Most important perceptual cues contained in the ear signals: e.g. [Blauert 1984]
  - Interaural Level Differences (ILD)
  - Interaural Phase/Time Differences (IPD/ITD) (relevant up to ca. 4 kHz frequency)
  - Temporal signal envelope (high frequencies)

⇒ ILD, IPD/ITD determine perceived ***lateral position*** of sound sources!

Spatial Sound Perception (5)



- In realistic closed rooms, reflections from the walls (and other objects) occur
- These blur the simple ILD, IPD/ITD relations by *decorrelating* the ear signals

⇒ Decorrelated sound at the two ears represents reflections and thus provides information on the *acoustic environment*

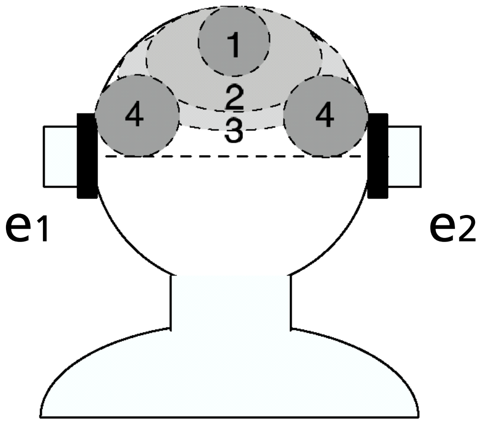
  – Distinct early reflections → nearby walls

  – Late reverberation → statistical summary of acoustic environment / room properties

⇒ Auditory system needs to “parse” sound into direct & ambient sound!

Fraunhofer Institut Integrierte Schaltungen

Spatial Sound Perception (6)



[Faller 2004]

- Important role of Interaural Coherence (IC): Determines spatial extent (=width) of sound event [Schuijers03] [Faller03]
- Perceived width of auditory event increases as coherence decreases (1 → 4); eventually separates into 2 distinct events
- Max. of normalized cross-correlation:

$$IC = \max_d \frac{\left| \sum_{n=-\infty}^{\infty} e_1(n) \cdot e_2(n+d) \right|}{\sqrt{\sum_{n=-\infty}^{\infty} e_1^2(n) \cdot \sum_{n=-\infty}^{\infty} e_2^2(n+d)}}$$

⇒ Key to source width and source/ambience perception!

Some A↔V Analogies in Scene Attributes

Visual Domain	**Auditory Domain (→ auditory cues)**
Foreground object	Sound sources (→ high IC) - directional
Object position	- Object position (→ ILD/ITD/IPD for lat. position)
Object size	- Object size (→ IC)
Background	Ambience (→ low IC) - non-directional

Part II:

Efficient Coding Of Spatial Sound Images

Fraunhofer Institut
Integrierte Schaltungen

Basics of Audio Coding

Goal

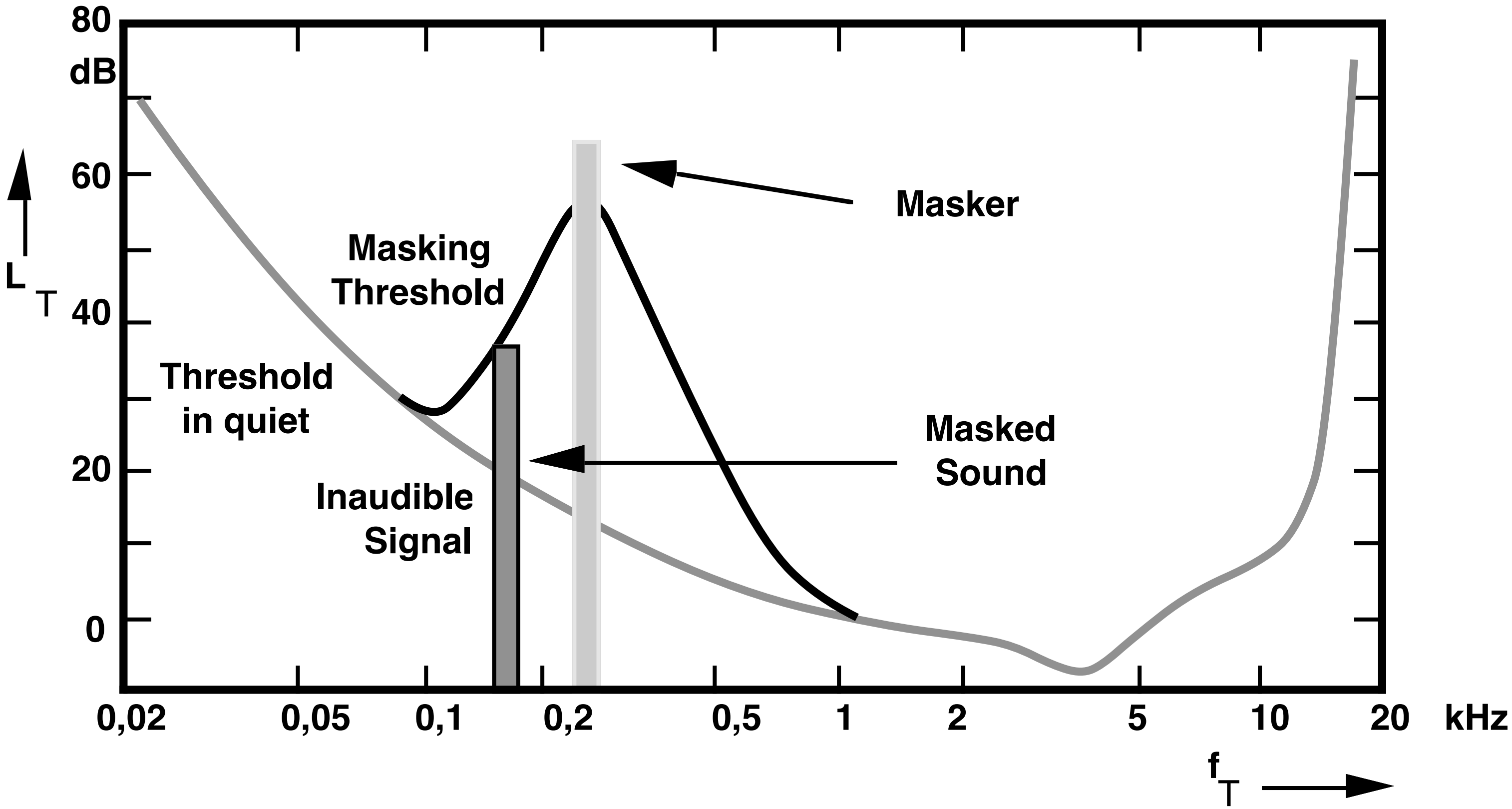
- Represent audio data as compactly as possible while maintaining sound quality (ideally: "transparent" coding)

Predominant Approach

- Concept of *perceptual audio coding*
  - Optimize *subjective quality* rather than objective distortion metrics (e.g. MSE/SNR)
  - Use knowledge about signal receiver: Psychoacoustics gives limits of perception
  - Keep coding distortion below limits!
  - No universal source model available (unlike in speech coding)

IIS
Fraunhofer Institut Integrierte Schaltungen

Psychoacoustics



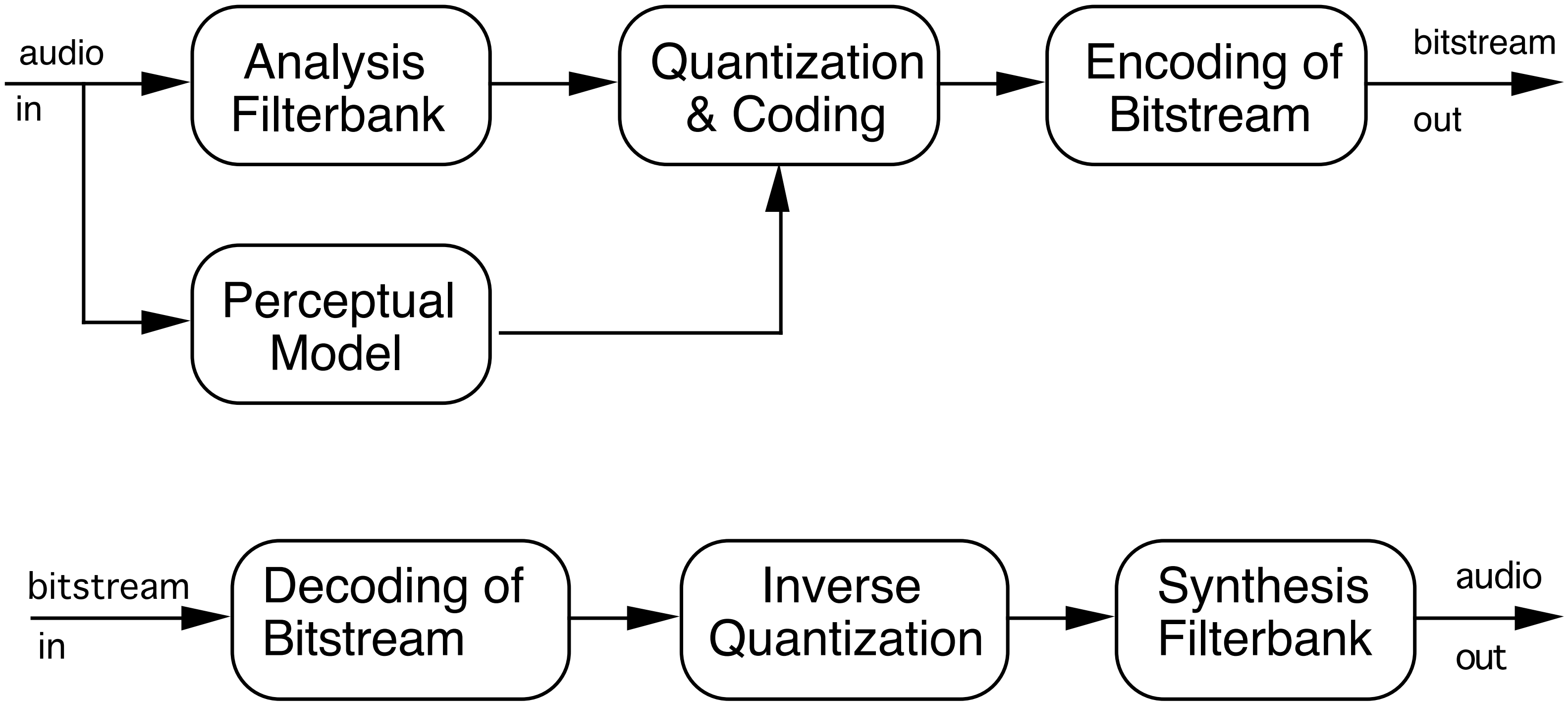
Demonstration: The "13 dB Miracle"

Historic demonstration by James D. Johnston and Karlheinz Brandenburg at AT&T Bell Laboratories in 1990 using the best psychoacoustic model available at that time ...

- Original signal
- Original + white noise, SNR = 13,6 dB
- Original + noise at threshold, SNR = 13,6 dB
- Difference signal: White noise
- Difference signal: Noise at threshold

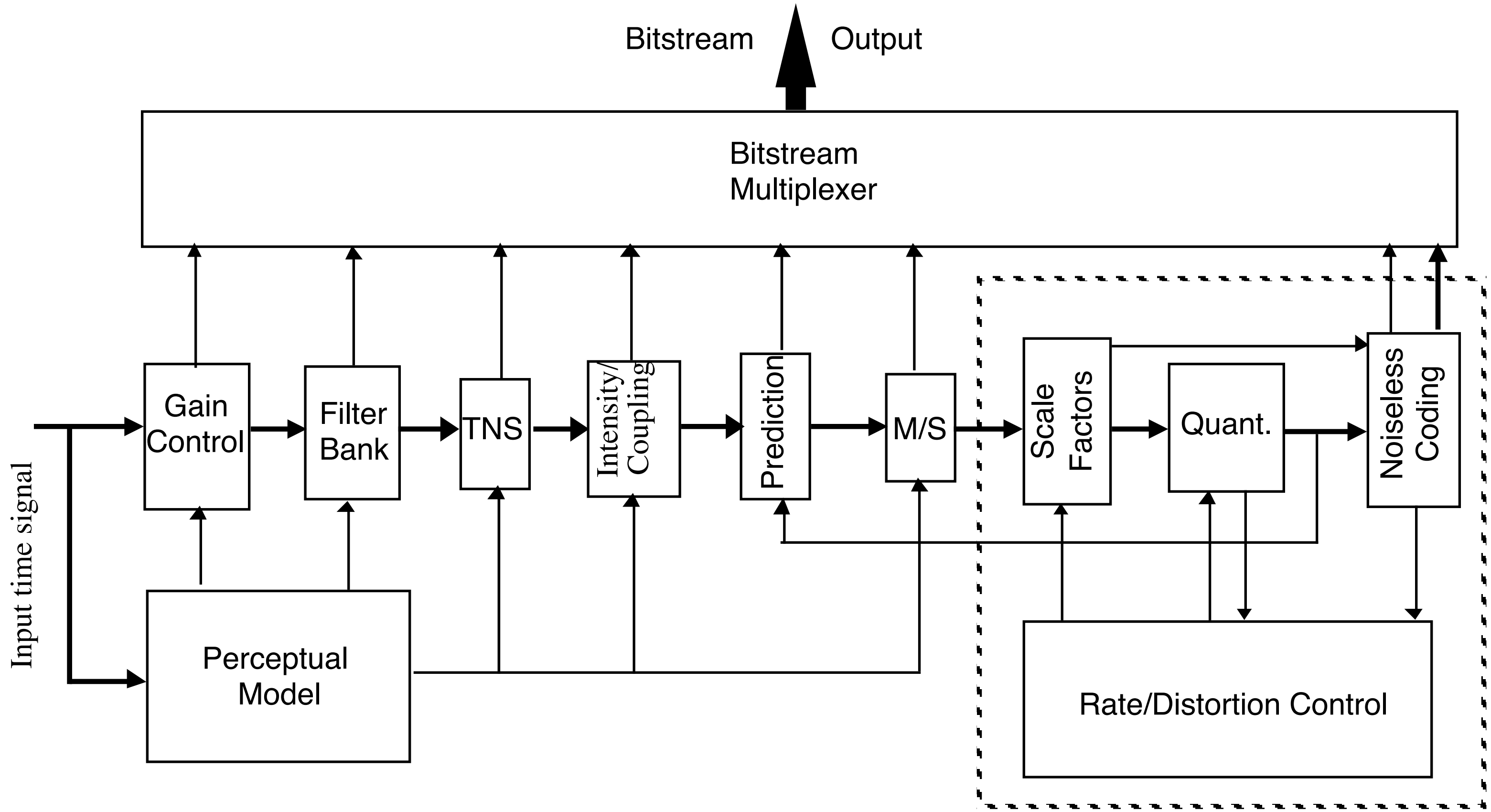
⇒ SNR *does not* adequately describe subjective sound quality!
⇒ Putting psychoacoustics to work makes a huge difference!

Basic Paradigm of (Monophonic) Perceptual Audio Coding



Fraunhofer Institut
Integrierte Schaltungen

A Real Audio Coder (MPEG-2 AAC, 1997)



Fraunhofer Institut
Integrierte Schaltungen

The Better Spatial Sound Image:
Surround Sound / Multi-Channel Audio

- Significantly increased spatial realism over stereo, envelopment
- Origins in movie sound (5.1); now also for music, broadcasting
- Increasingly adopted in consumers' homes

Traditional Delivery Formats For Surround Sound

Matrixed Surround (Prologic, Neo6, …)

- Downmix of 5.1 sound into stereo signal, upmix at the receiver side
  - Efficient in terms of transmission bandwidth (same bitrate as stereo)
  - Backward compatible to stereo delivery
  - Limited computation necessary
  - *Significant loss in subjective audio quality*

Discrete Surround (AAC, AC-3, …)

- Separate transmission of each channel
  - *Significantly higher bitrate than stereo*
  - Moderate amount of computation
  - High subjective audio quality possible

Fraunhofer Institut Integrierte Schaltungen

A Major Step Ahead: “Spatial Audio Coding”

- Rather recent development

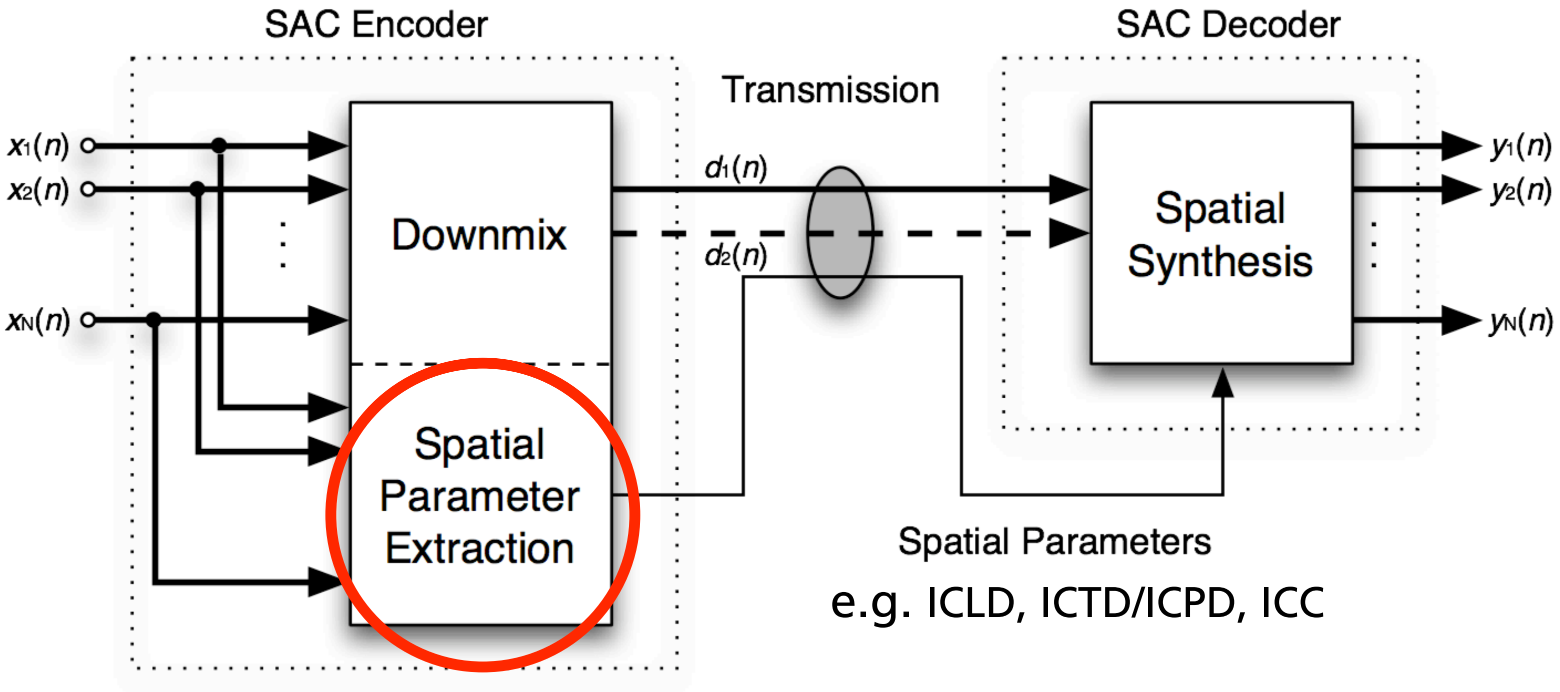
Main Characteristics

- Compression efficiency:
  Transmits multi-channel audio at bitrates used for 2-channel stereo (or even mono)
- Backward compatibility:
  SAC multi-channel audio is coded in a backward compatible way
  ⇒ existing infrastructures can be seamlessly upgraded to multi-channel / surround!
- High subjective audio quality

Heavily based on exploiting perception rather than waveform coding!

The Spatial Audio Coding (SAC) Concept

"Spatial Audio Coding" = Downmix + Parametric Spatial Synthesis



Extracts & reproduces *Inter-Channel* Counterparts of *Inter-Aural* Parameters

Related Technology

Generalization / Extension of

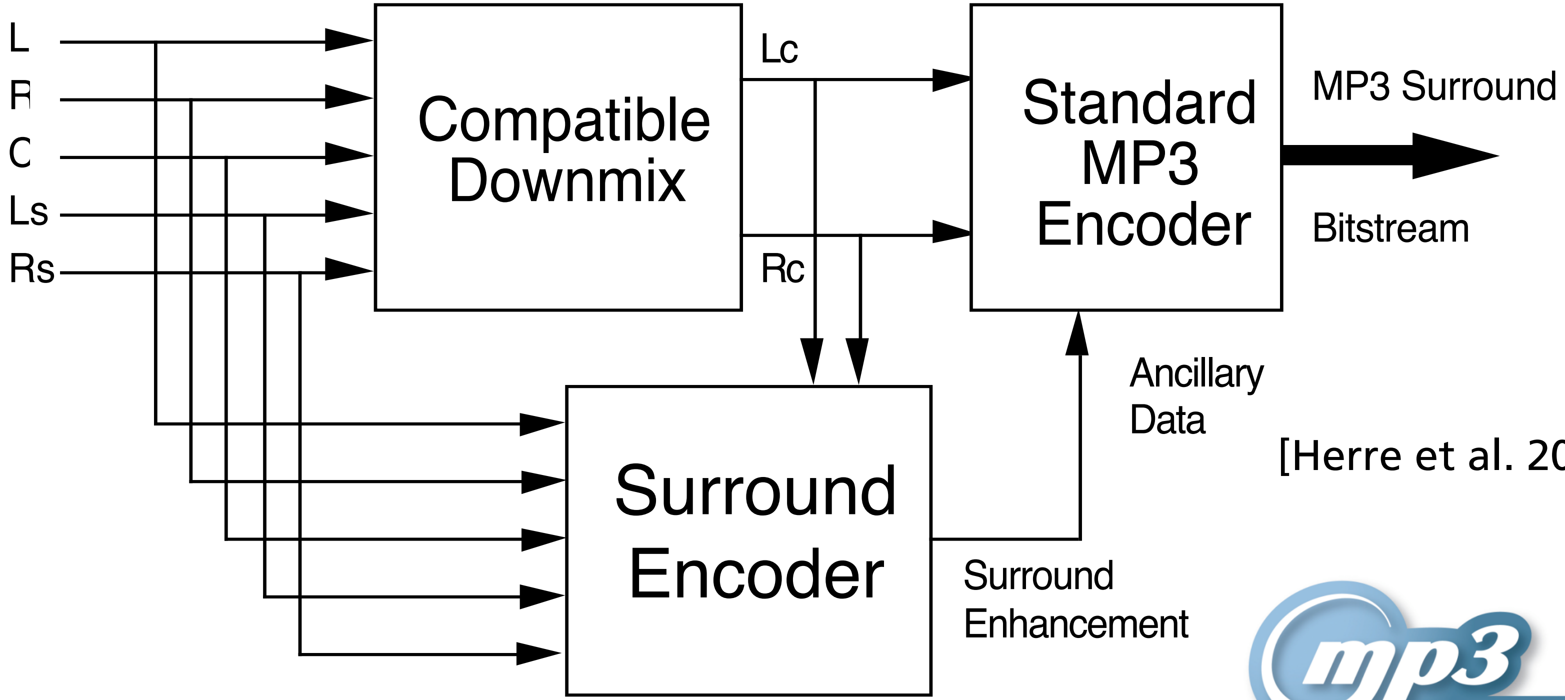
- Binaural Cue Coding
  Parametric Coding of Multi-Channel into a Mono Channel
- Parametric Stereo
  Parametric Coding of 2-Channel Stereo into a Mono Channel
- Matrix Surround
  (Dolby Prologic, Logic 7, Circle Surround …)

First commercial Application (2003)

- MP3 Surround
  Backward compatible extension of stereo MP3 towards 5.1 surround sound at 128 - 192kbit/s

Example: MP3 Surround Encoding

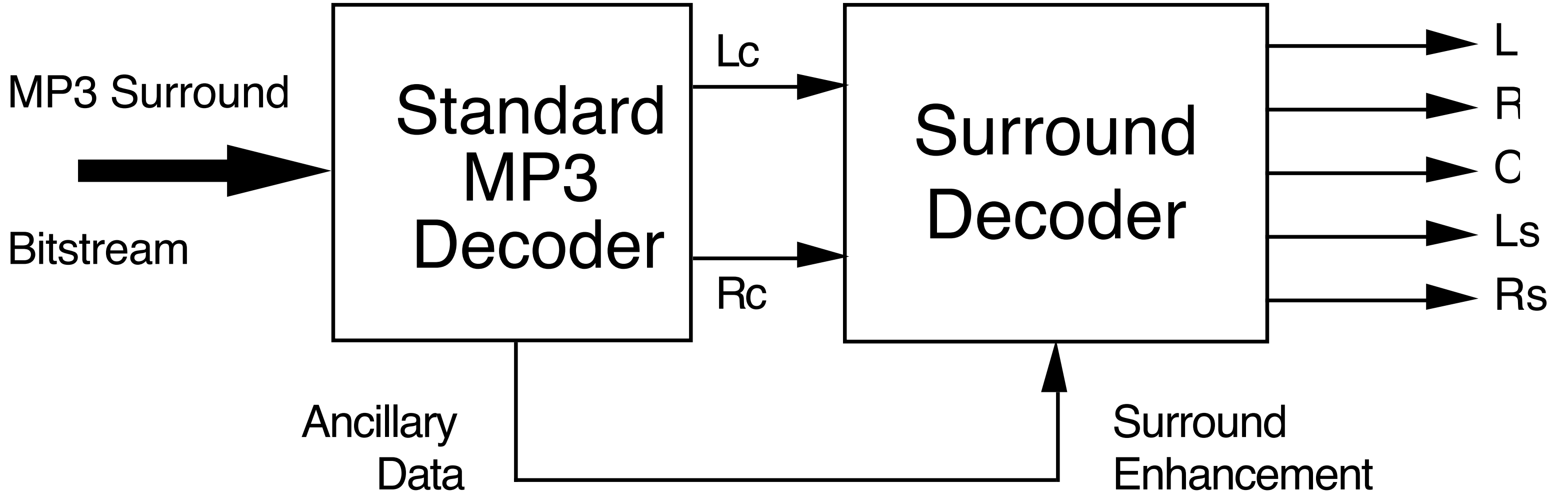
Principle: Surround → Stereo + Side Information



[Herre et al. 2004]

Example: MP3 Surround Decoding

Principle: Stereo + Side Information → Surround



[Herre et al. 2004]

Fraunhofer Institut
Integrierte Schaltungen

MPEG Spatial Audio Coding / MPEG Surround

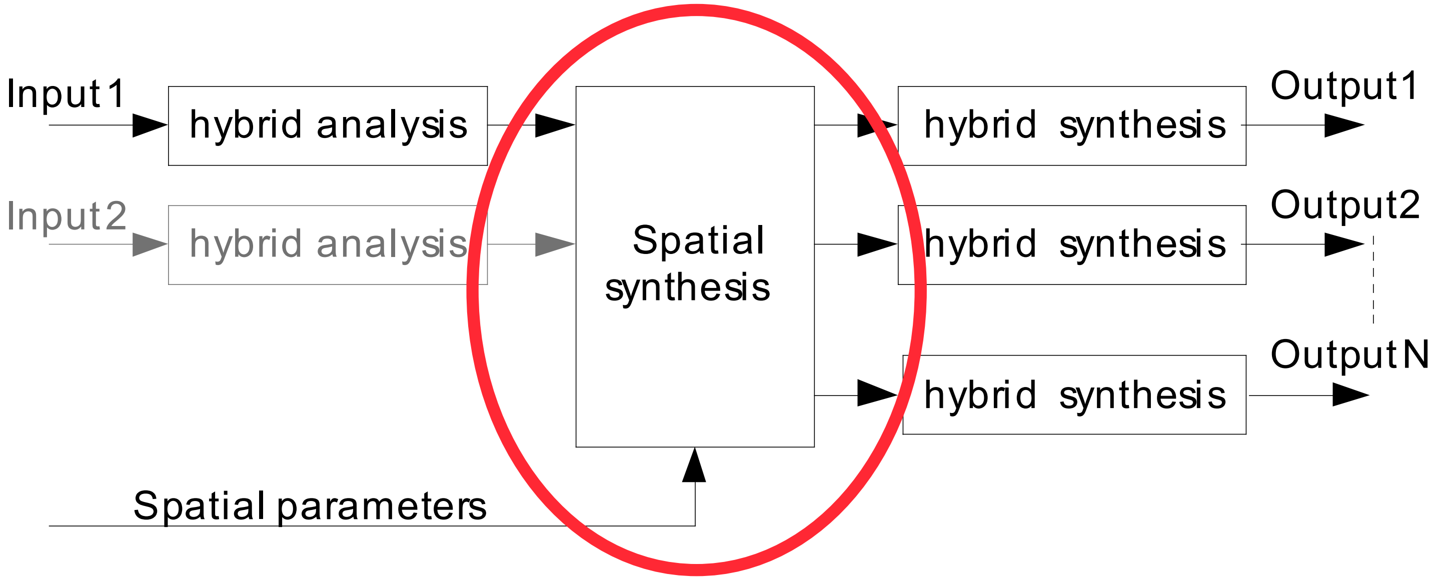
- Work item “Spatial Audio Coding” (SAC)
- Technical development 3/2004 - 7/2006
- Main contributors: Fraunhofer IIS, Agere Systems, Coding Technologies and Philips
- Renamed into MPEG Surround (MPEG-D)

Applications:

- Efficient & backward compatible upgrade of audio distribution to multi-channel, e.g.:
  – Music download service
  – Multi-channel streaming / Internet radio
  – Digital audio broadcasting

MPEG Surround Synthesis: General Concept

Decoder Structure

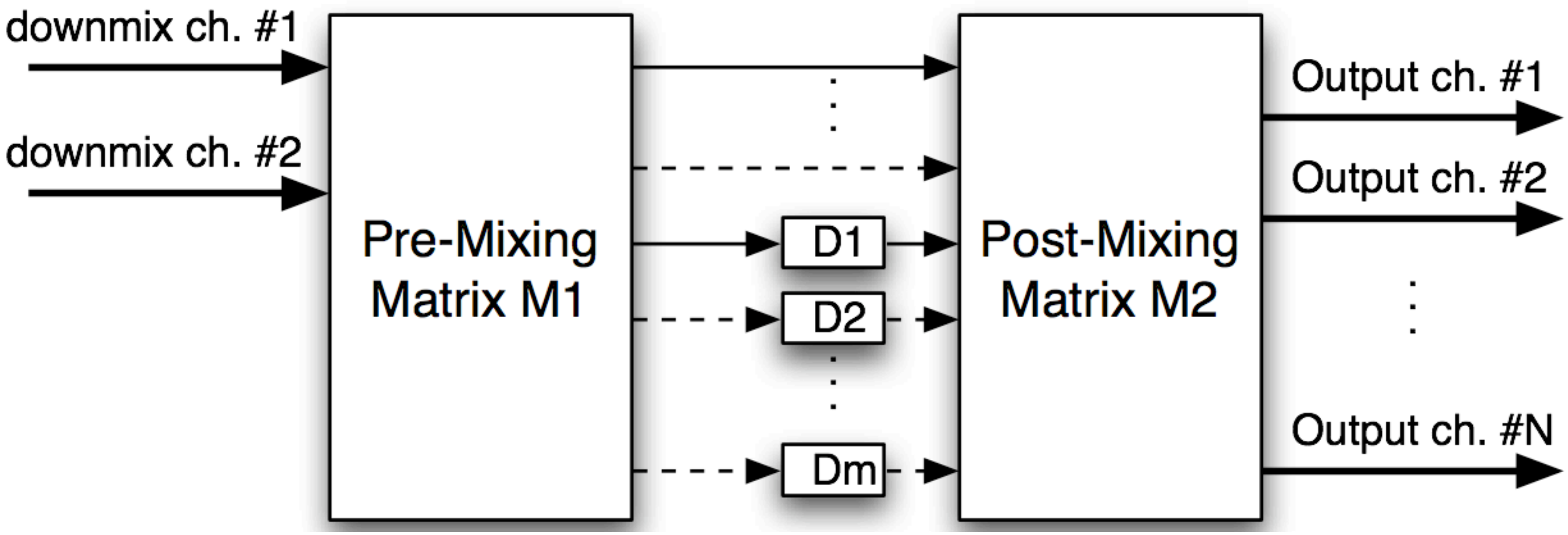


- "hybrid" = QMF filterbank + 2nd stage → non-uniform frequency resolution relating to frequ. resolution of human auditory system
- Same QMF filterbank as in MPEG-4 HE-AAC (AAC + SBR)

Fraunhofer Institut Integrierte Schaltungen

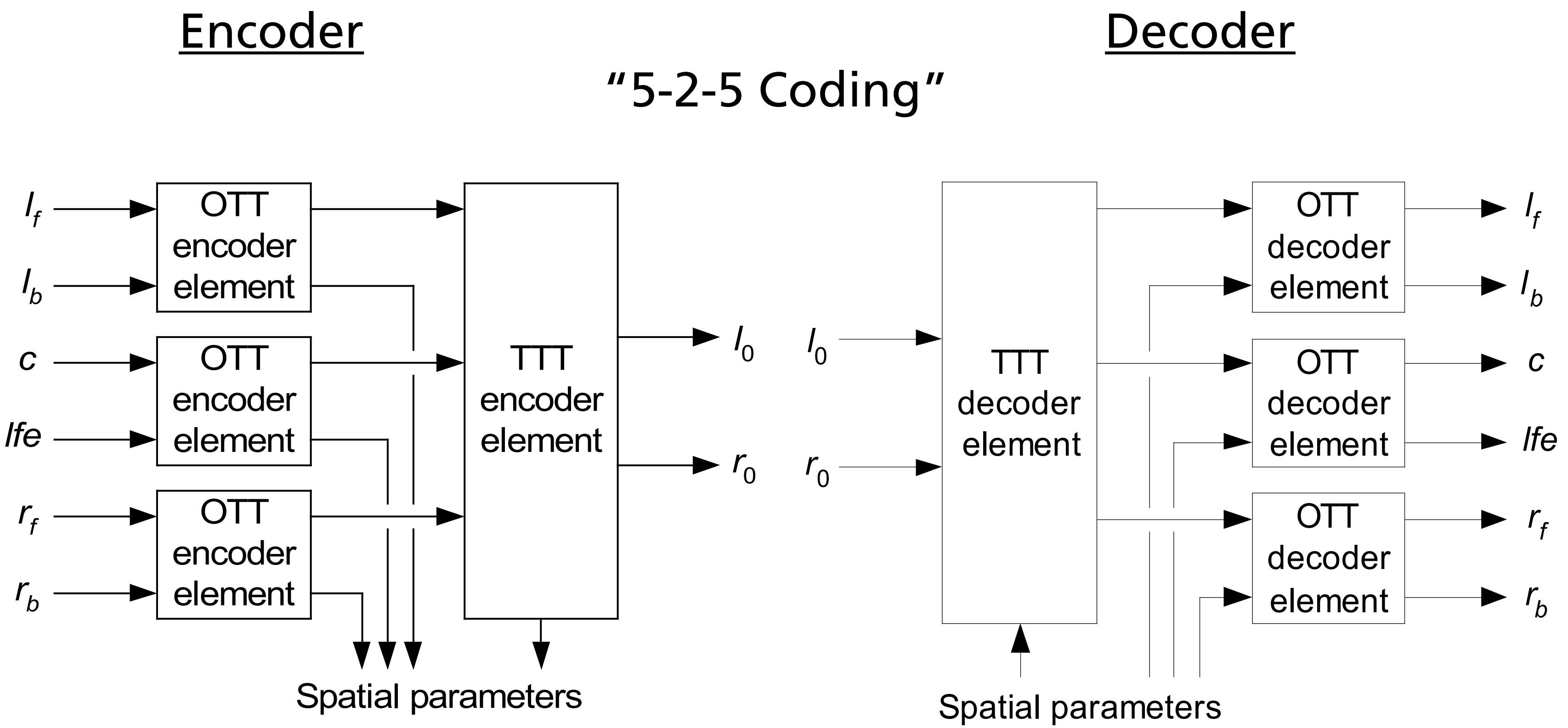
MPEG Surround Synthesis: General Concept (2)

- Frequency selective processing: Dynamic matrix + decorrelation



- Many more details, please see dedicated publications!
- Side information rate typ. 3 - 32 kbit/s

Underlying Idea: Hierarchical En/decoding



MPEG Surround Concepts

Generalization

- Similar trees for mono-based operation ("5-1-5"), other modes ("7-2-7", "7-5-7") …

Spatial Parameters

- Channel Level Differences (CLDs)
- Inter-Channel Correlations (ICCs)
- Channel Prediction Coefficients (CPCs)
- Prediction errors (residuals)

Other Aspects

- Decorrelation by QMF-domain all-pass filters
- Several tools for handling fine temporal envelope structure (both without and with additional side information)

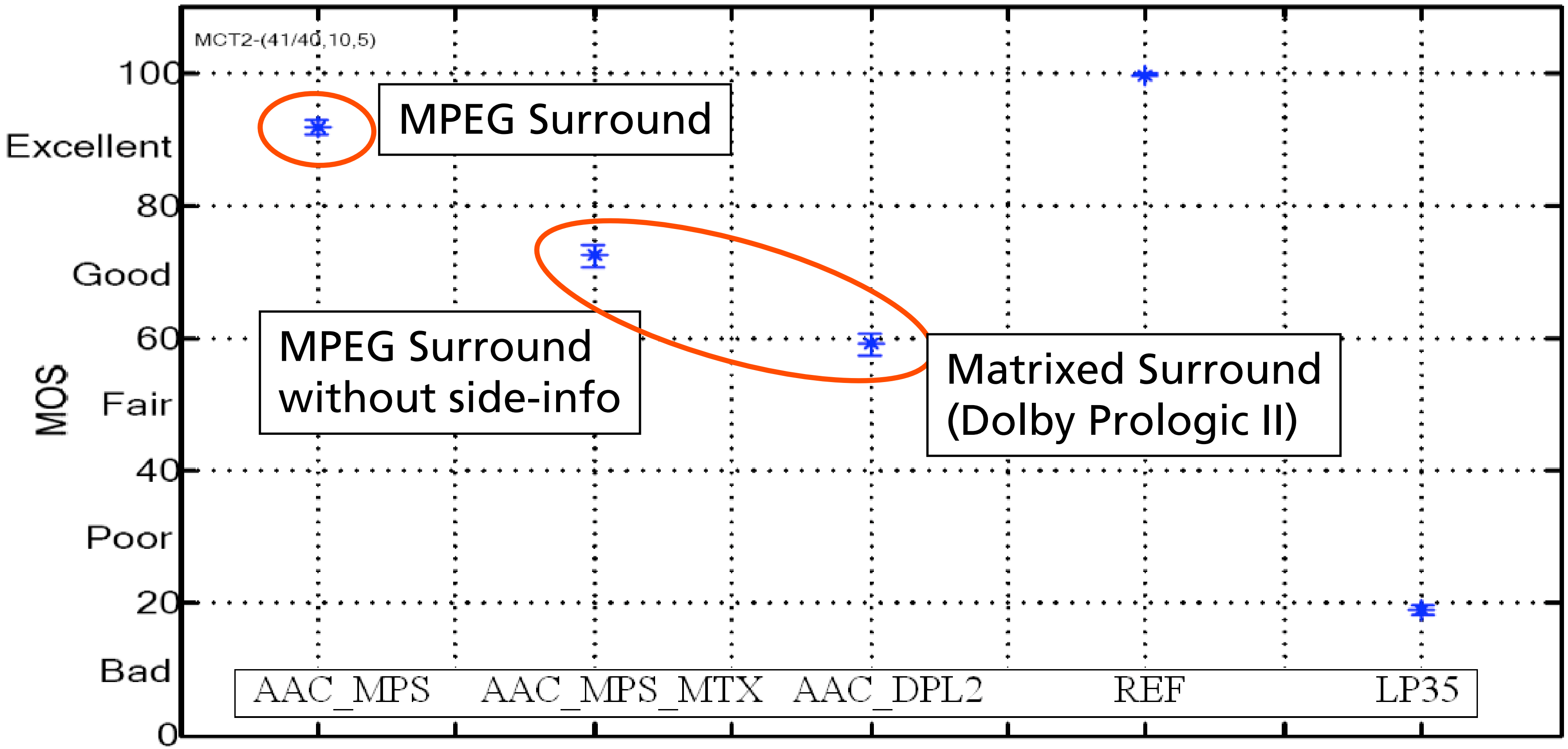
MPEG Surround: Additional Functionalities

Artistic Downmix	• Externally created downmixes can be used
Matrix Surround Compatibility	• Stereo downmix can be made compatible with common matrix surround decoders
Enhanced Matrix Mode	• MPEG Surround decoder can decode matrix surround signal (i.e. work without side info)
Binaural Rendering for headphone	• Downmix can be generated as virtual surround, or MPEG Surround can be decoded directly into virtual surround very efficiently

$\Rightarrow$ Rich set of attractive features for practical application

Fraunhofer Institut Integrierte Schaltungen

MPEG Surround: Recent Verification Test



“Music-Store” test scenario: Stereo downmix coded using AAC@160kbit

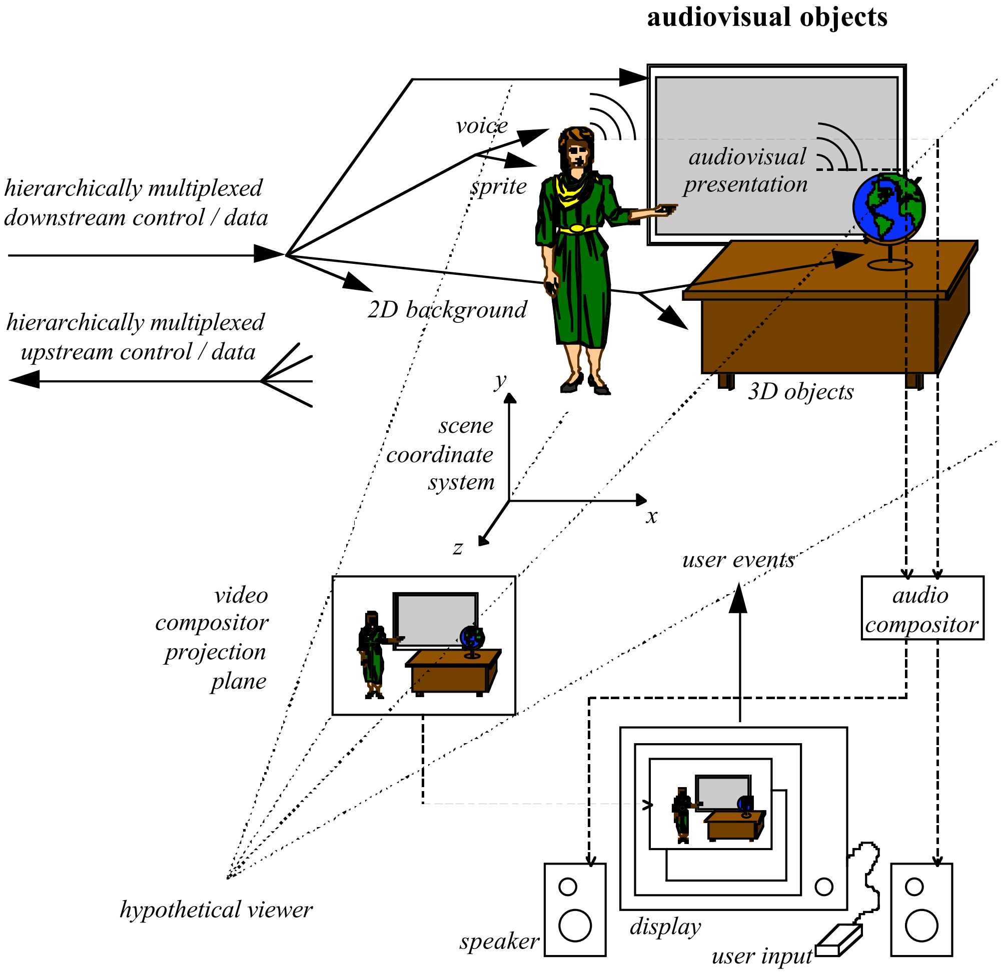
Part III:

Next Generation Interactive Coding / Rendering of Sound Images

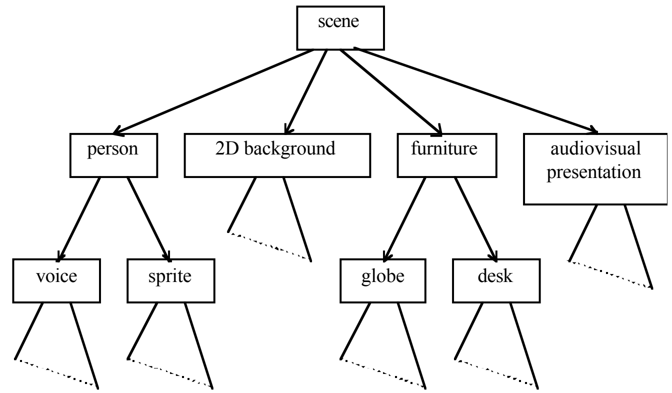
From *Spatial Audio Coding* to
Spatial Audio <u>Object</u> Coding

Fraunhofer Institut
Integrierte Schaltungen

Classic MPEG-4 Interactive Scene Composition (1996ff)

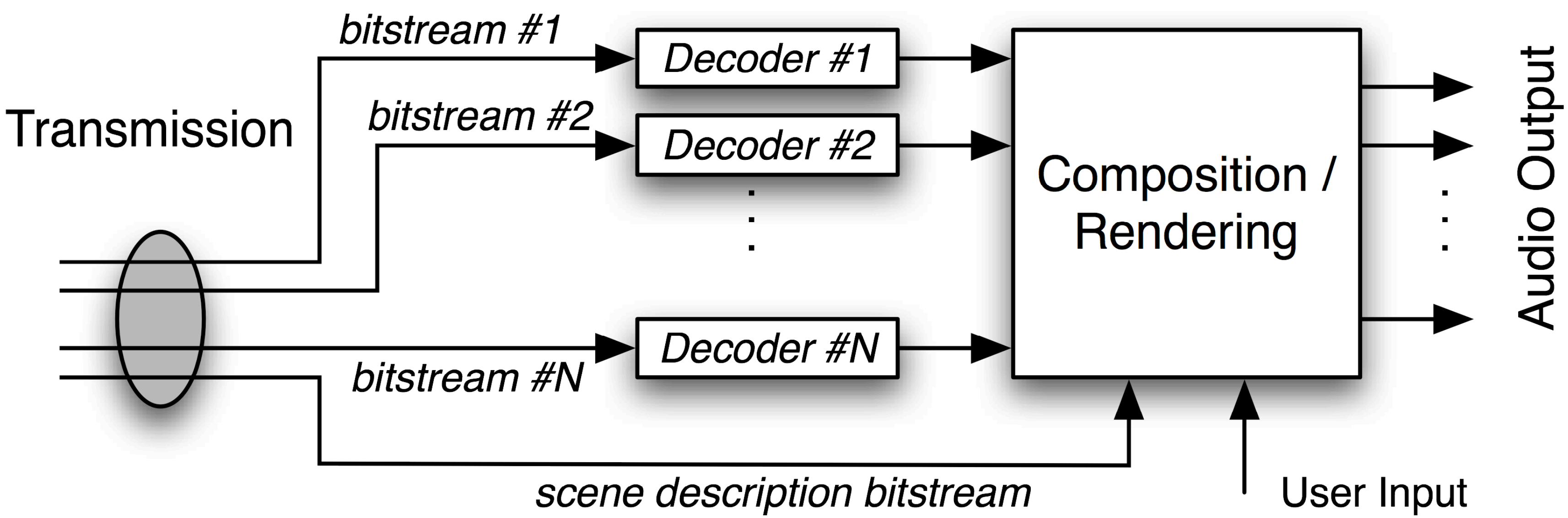


Hierarchy of objects



Scene is composed of multiple A/V objects and can be rendered interactively

MPEG-4 Object Based (De)Coding

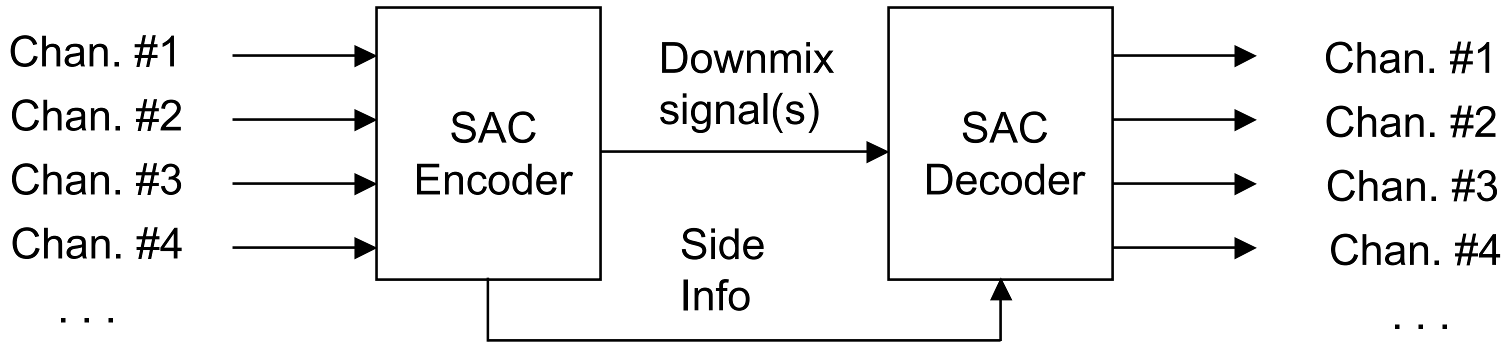


Discrete approach comes at a rather high price:

- Bitrate and decoding complexity grow with number of objects
- Structural complexity

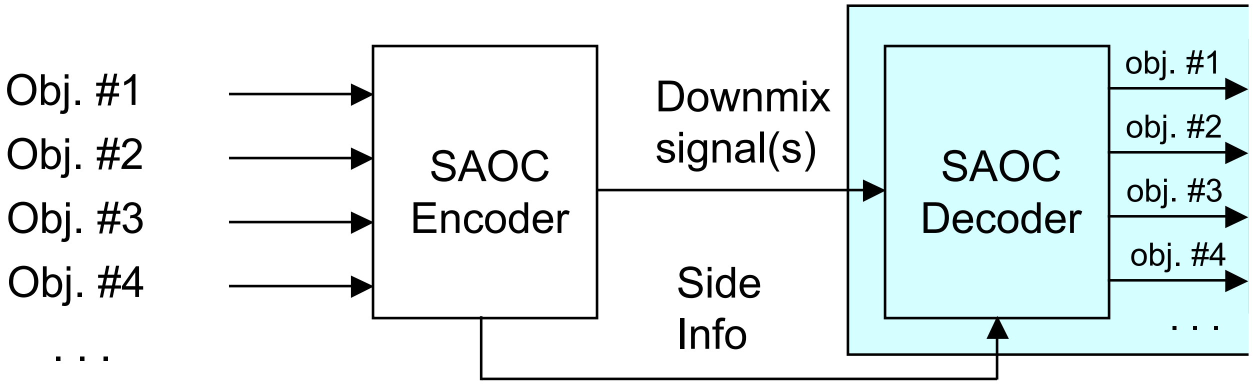
From Spatial Audio Coding (SAC) to SAOC

Regular Spatial Audio Coding: *Channel-oriented* scheme (MPEG Surround)



From SAC to SAOC (2)

Alternative: *Object-oriented* Spatial Audio Coding



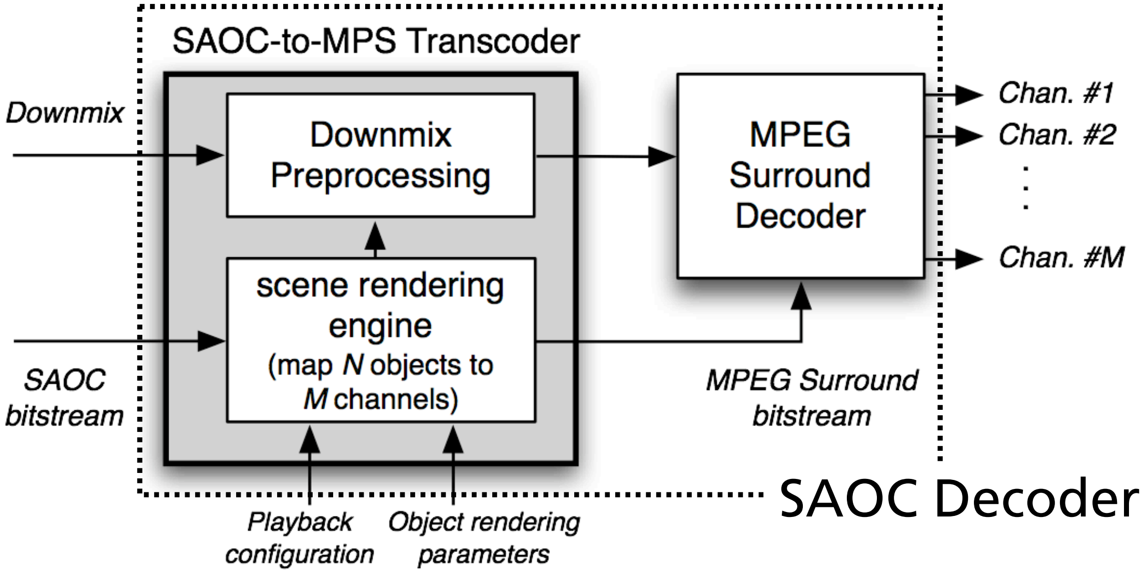
- Processes object signals instead of channel signals
- “Mixing”/rendering parameters vary according to user interaction
- Combined obj. decoding & rendering ⇒ computationally efficient!
- Previous work by Faller & Baumgarte [2001ff] and Faller [2006]

Demonstration ...

Real-time interactive rendering of audio objects from a mono audio downmix + SAOC side information

New MPEG Standardization Activities

- Work on "Spatial Audio Object Coding" (SAOC) started
- Transcoding approach: "SAOC" + rendering info → MPEG Surround



- Reference model and working draft recently established (10/2007)

Conclusions

- "Sound Images" carry some analogy to images in the visual world
- *Spatial Audio Coding* schemes code surround sound based on perception (rather than on waveform match)
  - "Object positions" are represented by perceptual spatial parameters
  - "Audio Object Texture" is coded using regular mono/stereo coder
- Such schemes can bring surround sound into existing infrastructures
  - High compression factor (surround sound at 64kbps and below!)
  - Stereo / mono backward compatibility
- Extension towards efficient interactive, object-based scene coding / rendering (Spatial Audio Object Coding) is currently on its way …

Thank You For Your Attention!

Fraunhofer Institut Integrierte Schaltungen