Recent Progress in Audio and Acoustic Communication:
What can video learn from it?

Karlheinz Brandenburg
Fraunhofer Institute for Digital Media Technology IDMT &
Technische Universität Ilmenau

December 7th, 2016
Overview

- Introduction
  - The quest for “High Fidelity” continues
- Where are we today
  - Two channel audio coding: A solved problem
  - There are many new topics in audio
- 3D Audio and object based audio: Old ideas come back in force
- The “inverse coder” and similar tools for forensics
- Psychoacoustics 2.0: Using cognitive effects
- Can video coding really learn from this?
- Final remarks
Early Audio Storage Media
The Dream of High Fidelity...
The Cello Blind Test Today

- Mono: o.k., might work
- Symphony orchestra: no, clearly audible
  - We need better spatial reproduction
  - Wave Field Synthesis can help (see later)
The Basic Paradigm of T/F Domain Audio Coding

Digital Audio Input → Filter Bank → Bit or Noise Allocation → Quantized Samples

Psychoacoustic Model

Bitstream Formatting → Signal to Mask Ratio → Encoded Bitstream
An Overview of Two Channel Audio Coding

- Once upon a time: MPEG-1 Layer 1,2,3
  - Not perfect, but good enough
- MPEG-2/4 Advanced Audio Coding
  - Good enough for golden ears -> Problem solved
  - Lower bit rates: HE-AAC and its derivatives
    - Widespread usage at 48 kbit/s for a two channel signal
- Other application areas:
  - Communications: LD-AAC, ELD-AAC, USAC, EVS
  - Compatible 5 channel coding: not wide spread
Immersive Sound/Object-based Audio

- A nice symbol (if you know “Star Trek”): The Holodeck (3D, all senses and interactive)
- At the movies there is already no difference between “realistic” and computer animation
- Computer games never had this difference
- More and more movies are 3D
- How to do this for audio?
  - Surround sound
  - Ambisonics
  - Wave Field Synthesis, VBAP etc.
An Early Idea: Wave Field Synthesis

Primary source

Secondary source

Adapt delay and amplitude

Diemer de Vries, Rinus Boone (TUD)
Redrawn by Frank Melchior (IDMT)
Some Mathematical Foundation

\[
P(r_R) = \frac{1}{4\pi} \oint_S \left[ j \omega \rho_0 V_n(r_s) \frac{\exp(-jk\Delta r)}{\Delta r} + P(r_s) \frac{1 + jk\Delta r}{\Delta r} \cos \varphi \frac{\exp(-jk\Delta r)}{\Delta r} \right] dS
\]

\(V_n\) – component of particle velocity in \(r_s\),

\(P\) – pressure,

\(\rho_0\)

– density function
Moving and Manipulating Audio Objects

- Single audio objects sound like moving on given trajectories
Object-based Sound Reproduction

- The dream of Gus Berkhout (TU Delft)
- A paradigm change:
  - Now: Channel-based formats (even for surround sound)
    2, 5.1, 7.1, 10.2, 22.2
  - The future: Object-based storage
    Each sound is rendered in a way, that whatever reproduction system is available, I always get the best possible auditory illusion
- Main advantages:
  - Complete flexibility on the recording side
  - Complete flexibility on the reproduction side
MPEG-H: A Universal Standard for 3D Audio

- In some respect, all the old dreams are realized (to some degree)
- Options for the recording side
  - HOA (Higher Order Ambisonics)
  - Audio-Objects (Track & Metadata)
  - SAOC (Extension to older MPEG-Audio standards) or
  - Channel-based formats (legacy-content)
- Reproduction
  - Object renderer via VBAP
  - Needs geometry data, metadata and one audio stream for each object
MPEG-H: Block Diagram Decoder

Source: MPEG-H Standard
SpatialSound Wave: 3D Object-based Audio System

- Object-based Audio

![Diagram showing the components of SpatialSound Wave: Object-based Audio System, including loudspeaker setup, audio objects, metadata, audio signals, audio rendering engine, and psychoacoustic model.]
Forensic Audio: Identifying Codecs

Inverse Decoding

Encoding

MP3, 64 kbit/s
AAC, 96 kbit/s

Decoding

wav
wav

Inverse Decoding

MP3, 64 kbit/s
AAC, 96 kbit/s
Inverse Decoding

Relation to Audio Forensics

- **Use case**
  - The attacker claims to have a continuous recording
  - The attacker *modified* the audio content
  - The content was *decoded* in order to be edited

- **General idea**
  - The attacker provides the court with a tampered evidence
  - The encoding traces will change
Inverse Decoding

Codec Footprints

- Many footprints are effected by lossy compression:
  - Number of inactive spectral coefficients
  - Spectral fluctuation
  - Difference of sorted spectral coefficients
  - Quantization step sizes

- Inverse decoding is possible
  - By detecting these footprints
  - On the correct framing grid

Inverse Decoding

Framing grid

- Lossy codecs
  - Divide the input file in frames
  - Using constant window length and hop size
  - Apply quantization to each frame
Inverse Decoding

Framing grid

- Lossy codecs
  - Divide the input file in frames
  - Using \textit{constant} window length and hop size
  - Apply quantization to each frame
  - Quantization traces are visible \textit{only} with the correct offset

Inverse Decoding

Framing grid

- Lossy codecs
  - Divide the input file in frames
  - Using constant window length and hop size
  - Apply quantization to each frame
  - Quantization traces are visible only with the correct offset

- This framing grid offset
  - For non-manipulated decoded file, is constant and fixed
  - Requires different methods, depending on the codec applied
Inverse Decoding

Algorithm

- For each known codec $C$
  - 1. Identify intervals encoded with $C$
  - 2. Determine offset $R$ for each interval
  - 3. Join contiguous intervals with the same offset

<table>
<thead>
<tr>
<th>Time [frames]</th>
<th># inactive coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(correct) offset $R$</td>
</tr>
<tr>
<td>50</td>
<td>(wrong) offset $R+1$</td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>250</td>
<td></td>
</tr>
</tbody>
</table>

- Mp3 is present with the offset under analysis in the first 120 frames
- Afterwards either the offset or the codec changes

Inverse Decoding

Algorithm

- For each known codec $C$
  - 1. Identify intervals encoded with $C$
  - 2. Determine offset $R$ for each interval
  - 3. Join contiguous intervals with the same offset

- After checking all known codecs
  - Mark the remaining intervals as never encoded
  - Output the final result

Inverse Decoder

Summary

Pros:
- Precision above 99%
- High applicability for content shared on Internet

Cons:
- Limited to the set of known codecs
- What if there was no encoding?

Psychoacoustics 2.0: What else do we need to know about ears and brain?

- What we hear is based on what we believe
  - New research on spatial hearing

- What about video?
  - Known: Better audio improves the perceived video quality
High End Audio Myth

- High frequency content up to 40 kHz
- Special cables, AC stabilization etc.
- Green ink
- Some frequent quotes from self-defined golden ears
  - “Blind testing does destroy the subjective pleasure of listening, should not be done”
  - “Anybody can hear this” for all the above
Some well known Facts about our Ears

- Absolute threshold of hearing
- Masking
- Binaural masking
- All stimuli go to the brain (CNS)
  - Which uses other modalities, too
  - Which bases our sensation on belief
The Audio Processing Chain in Ear and Brain

Loudness “integration”

Feature Analysis

Auditory Object Analysis

Cognitive and other Feedback

Mbits/sec

Mbit/sec

Kbit/sec

bit/sec
Do we really hear what we hear?

- The McGurk effect
  - Look at the picture and listen
  - What did you hear?
  - Now close your eyes and listen again
  - What did you hear?
How to create a perfect Audio Illusion

- Up to now, physics centered view
  - Wave equations
- Psychoacoustics
  - Spatial hearing: we just got two ears, but ...
    - Cone of confusion
  - Height perception: Blauert bands
  - Binaural cues
    - Interaural Time Differences (ITD)
    - Interaural Level Differences (ILD)
    - Head Related Transfer Function (HRTF)
The new Frontier: How to get real Illusion

This will be basic research for the next decade:
- Can we quantify cognitive effects?
- How accurately do we need to recreate waveforms?
- Is it possible to cancel room reflections enough to not spoil the illusion?
- How much does our brain adapt to less than perfect reconstruction?
- Can we quantify how our brain learns to adapt to acoustic features (like HRTF)?
Immersive Sound via Headphones

- We get great sound via loudspeakers
  - Immersive sound becomes widespread using Ambisonics, Wave Field Synthesis, VBAP or hybrid schemes

- Using headphones, usually
  - The sound comes from within the head
  - We get front/back confusion
  - It does not feel real

- There has been plenty of research in the last decades, but
  - It still does not feel real
When do we get a plausible spatial auditory illusion?

- We need a match between
  - Our expectation of a given sound in a given environment
  - A number of factors including
    - Cues for our own anatomy (HRTF including ear canal geometry)
    - Cues for the room (reflection patterns etc.)
    - Visual cues
    - Our experience

- None of these cues can override the rest for creating the convincing illusion, but
  - All of these can enhance the possibility for convincing illusion
  - Mono causal explanations do not help
  - One dimensional optimization will not lead to success
A Plausible Spatial Auditory Illusion

- **Visual influence** (sound source and room is visible or not)
  - ~16 % increase of externalization over investigated directions, personalization methods, and room conditions
  - No clear tendencies in respect to personalization and combination of listening room and synthesized room

- **Influence of personalization**
  - Externalization increases using individual binaural synthesis system
  - Higher effect size of personalization compared to room divergence effect
  - Higher effect size for congruent room conditions compared to divergence
What can video learn from all of this?

- Object based coding is now a major trend in audio coding
  - Will SNHC (Synthetic/Natural Hybrid Coding) of MPEG-4 come back?
  - What about a new generation of automatic segmentation algorithms?
- Coding may be connected to forensics
  - Is the world they show us real?
  - Can we prove manipulation of video scenes?
- Psychoacoustics 2.0: What is the video equivalent?
  - There are transcoders doing a smart bit allocation within the picture
  - Beware: Better audio quality results in better subjective video quality
    Too often, this is an afterthought
Final remarks

- Progress in audio coding and audio signal processing moved
  - From classic signal processing
  - To models including cognitive effects

- I do not believe in
  - Models using Euclidian distance
  - More generally: LTI models are mostly wrong in audio

- …there is plenty of research still to be done…