August 28, 2003

Transform coding of audio impulse responses

M. Sc. Thesis

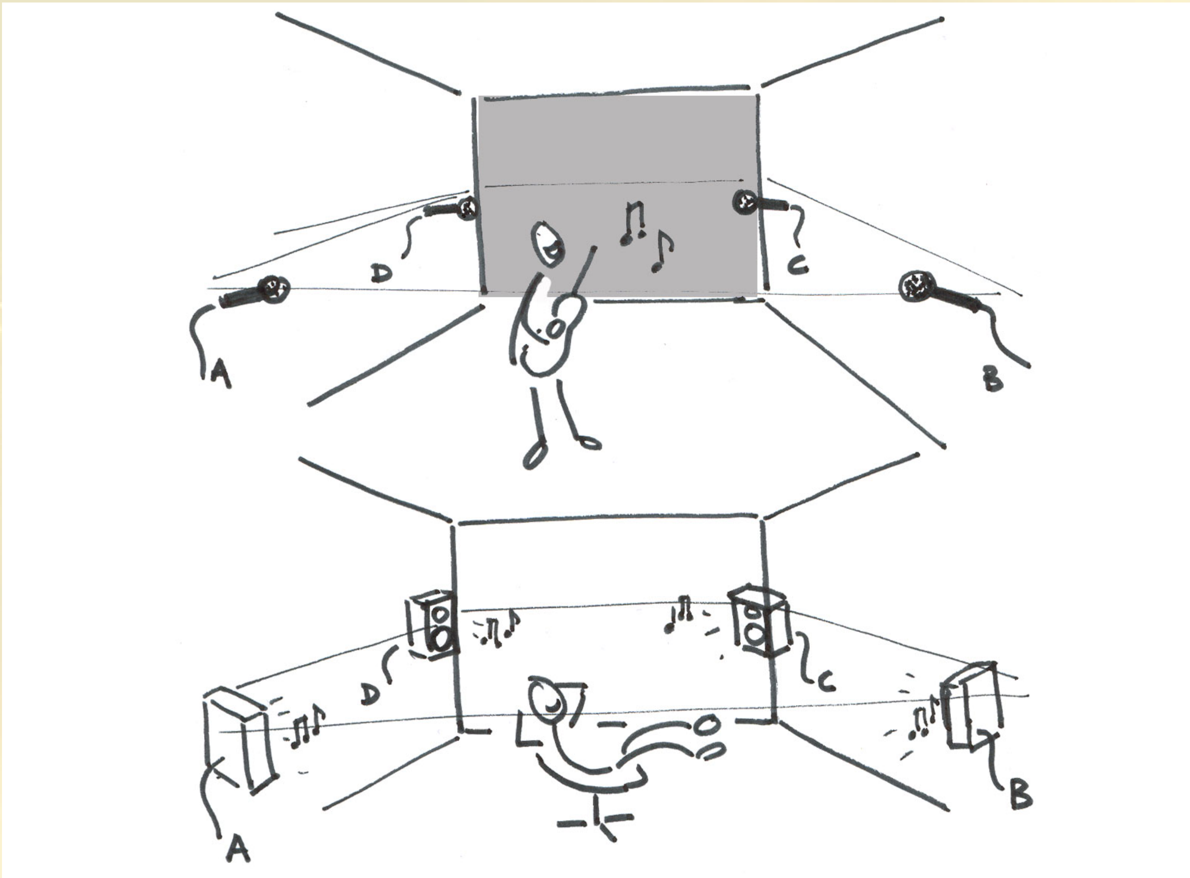
Jochem van der Vorm

Supervisors: prof. dr. ir. A. Gisolf, dr. ir. D. de Vries

Introduction - Context

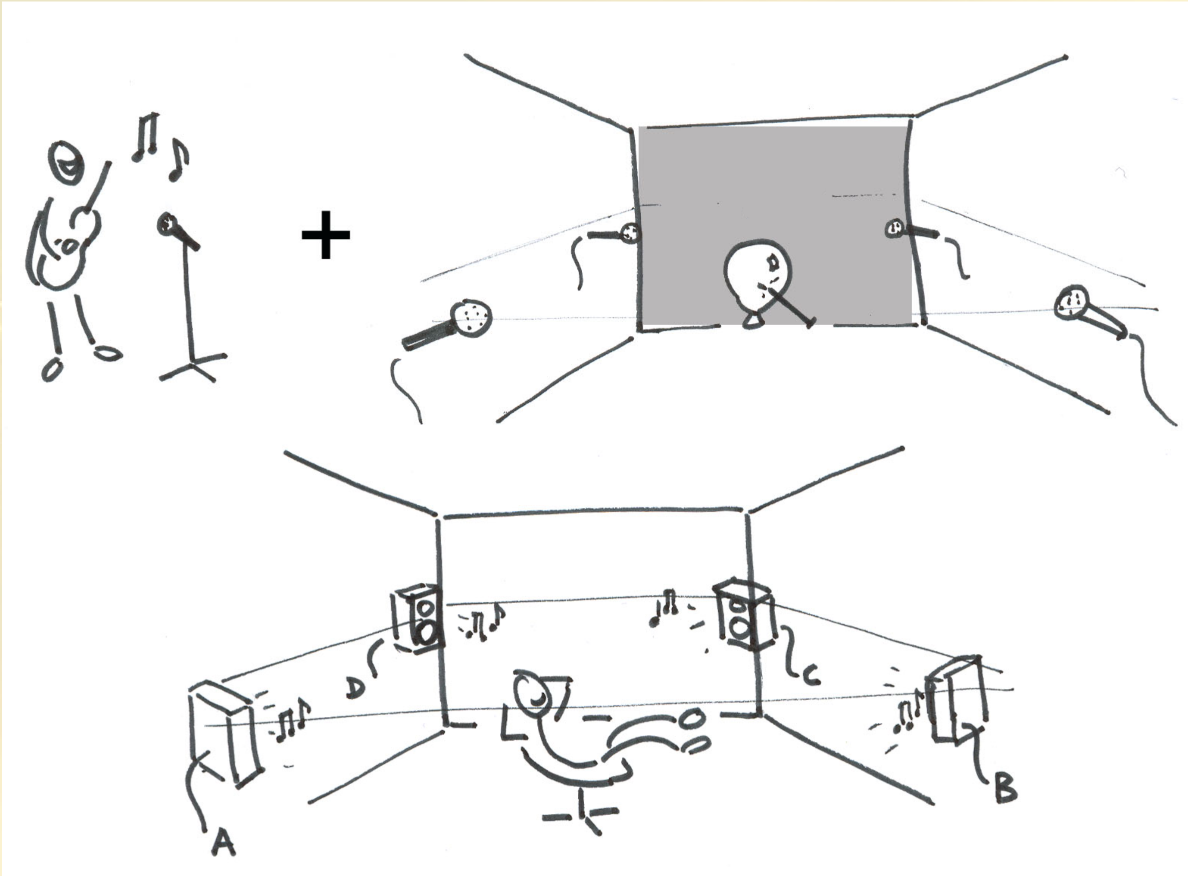
- **Carrouso project** $=>$ Creative, assessing, and rendering in real-time of high-quality audiovisual environment in MPEG-4 context.
- **Wave Field Synthesis** $=>$ Developed at TU Delft, a method for spatial and temporal reproduction of a sound field.
- **Compression** $=>$ Since 70's speech and later music is compressed to save bandwidth, using a wide collection of methods, most well-known nowadays is MP3.

Introduction - Context



Traditional multi-channel audio (4 audio channels transmitted)

Introduction - Context



Possible WFS approach (1 audio channel + acoustics transmitted)

Introduction - Research goals

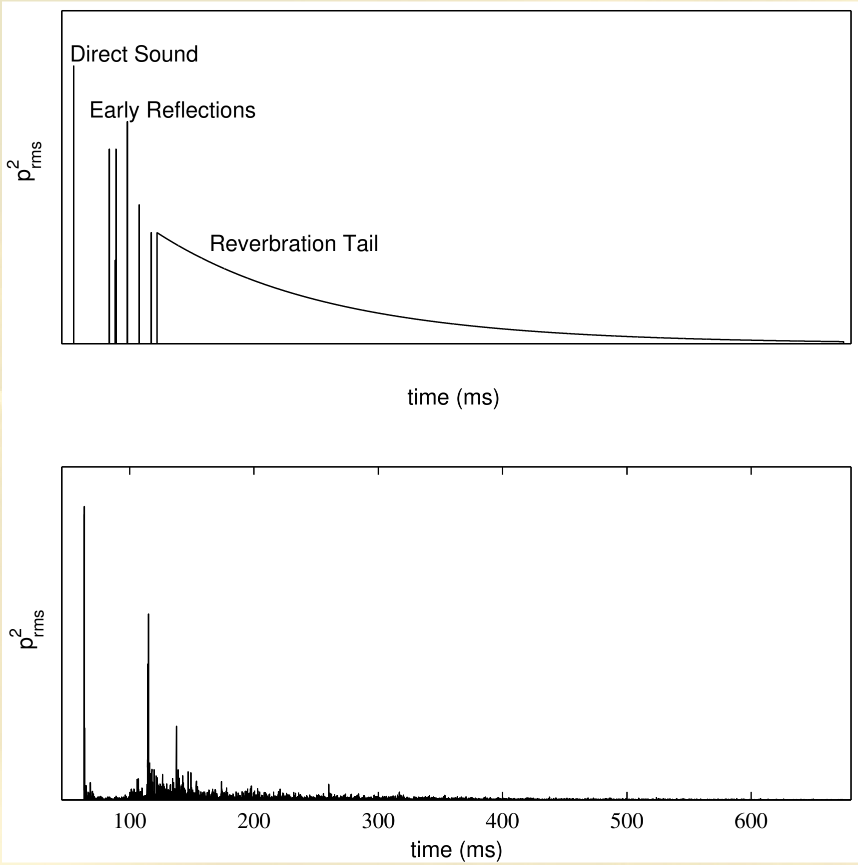
- Develop coding structure for audio impulse responses
- Reconstruction indistinguishable from original (when 'used')
- Compression factor must be (much) higher than music coders
- Model must apply to a wide range of inputs (different 'acoustical environments')

1. **Introduction**
   (a) Context
   (b) Research goals
2. **Theory**
   (a) Audio impulse responses
   (b) Coding
   (c) Transforms
3. **Proposed codec**
   (a) Overview
   (b) Windowing
   (c) Spectral coding
4. **Results**
   (a) Plots and observations
   (b) Listening test
5. **Conclusions & suggestions**
   (a) Conclusions
   (b) Suggestions
   (c) Questions

Theory - Audio impulse responses

- 1 impulse response $=>$ reaction of a system on a pulse (δ)
- Can be measured with noise-like or sweep signal (and deconvolving)
- Multiple impulse responses define an ’acoustic environment’ for an enclosure
- Software packages can be used to approximate impulse responses using ray-tracing and mirror image source models

Theory - Audio impulse responses



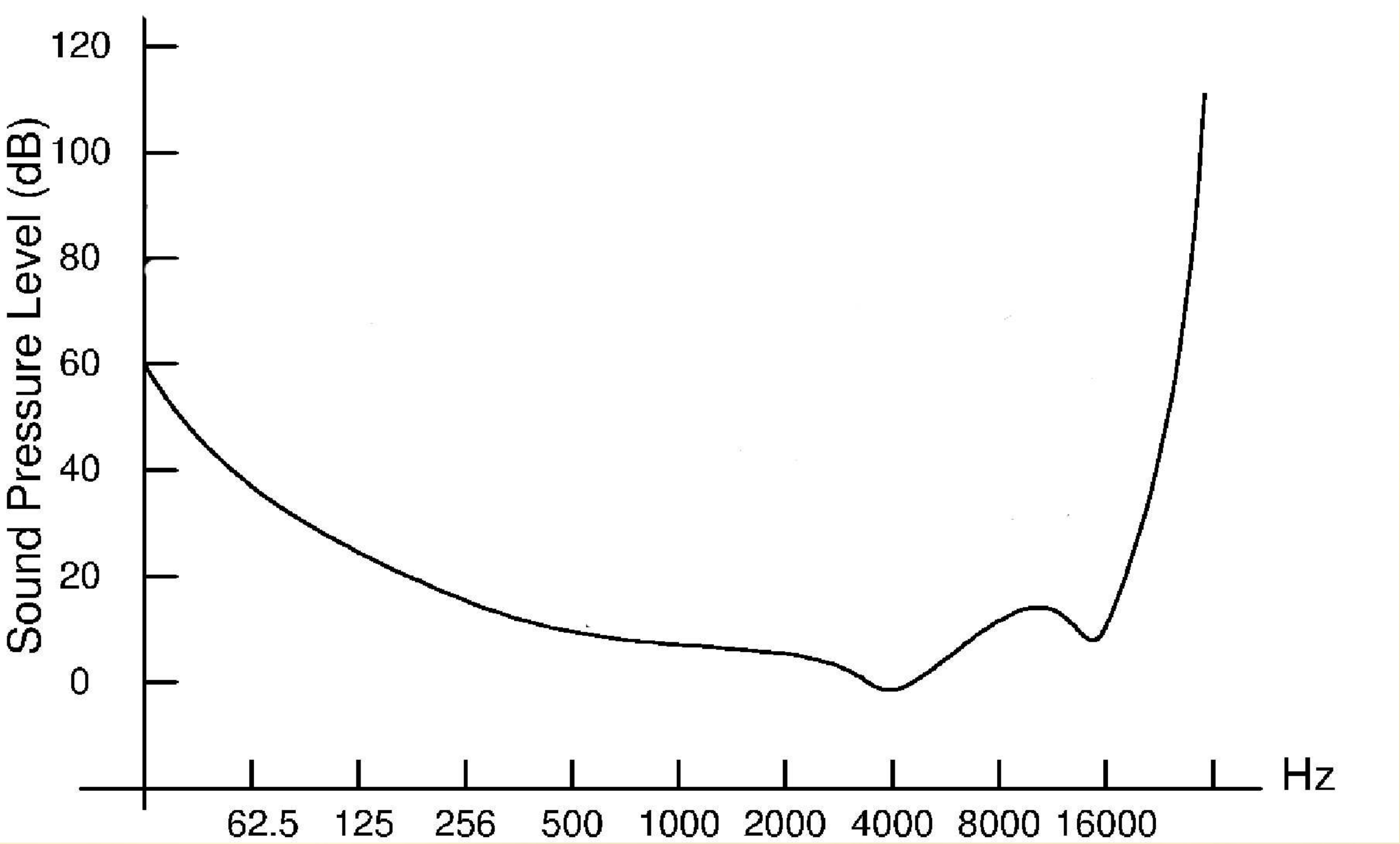
Theoretical and measured impulse response

Theory - Coding

Psycho-acoustic analysis

- Threshold of hearing and high frequency limit
- Temporal masking
  - Forward masking (length: 50-200 ms).
  - Backward masking (length: 5 ms).
- Spectral masking $=>$ existence of 'critical bands'

Theory - Coding



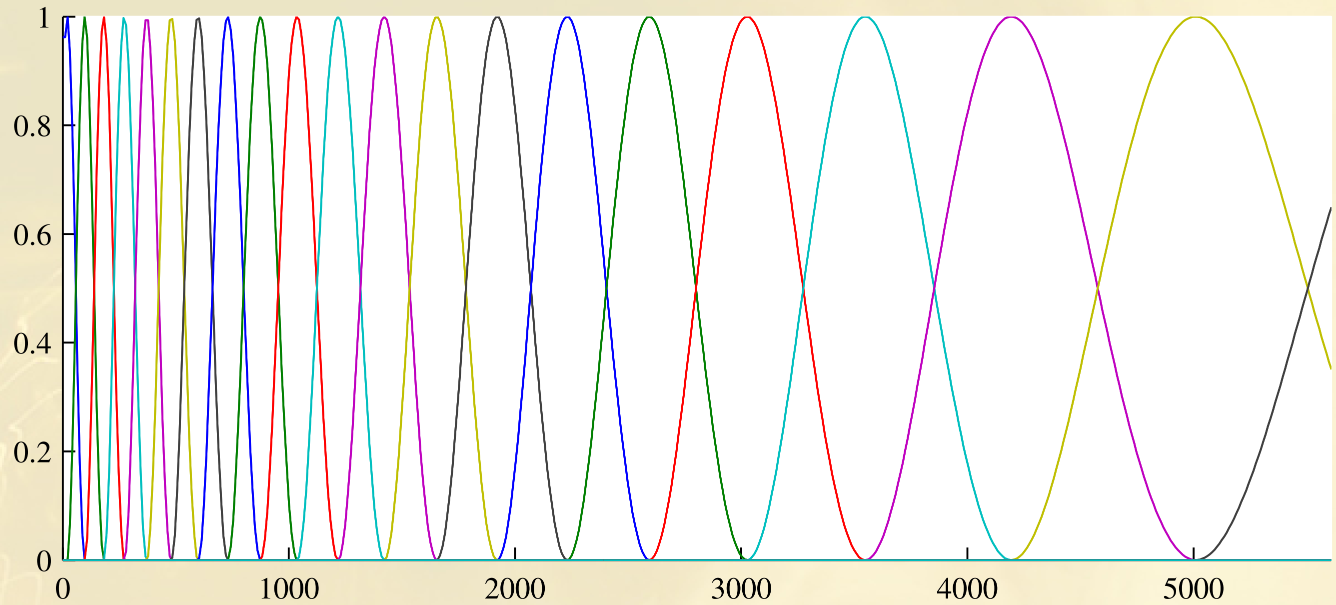
Threshold of hearing & high frequency limit

Theory - Coding

Psycho-acoustic analysis

- Threshold of hearing and high frequency limit
- Temporal masking
  - Forward masking (length: 50-200 ms).
  - Backward masking (length: 5 ms).
- Spectral masking $=>$ existence of 'critical bands'

Theory - Coding



Bark scale converts frequency to critical band number. Zwicker:

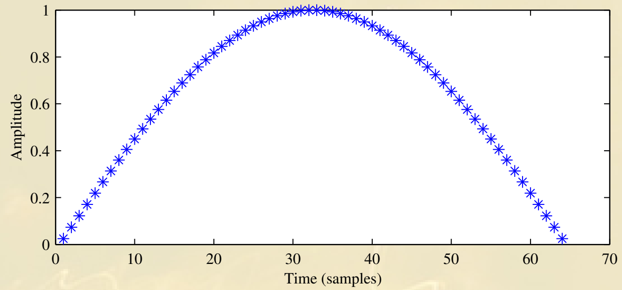
$$z(f) = 13\arctan(0.00076f) + 3.5\arctan[(\frac{f}{7500})^2]$$

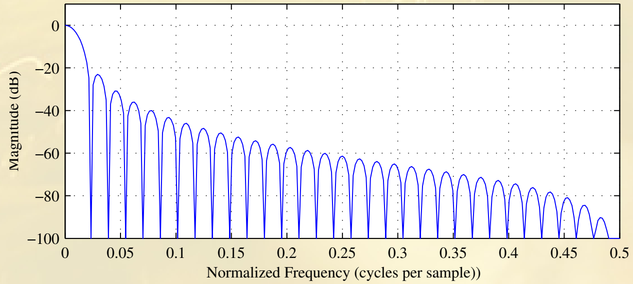
Theory - Transforms

- Transform will be applied in blocks
- Discrete Fourier Transform needs overlap-add or overlap-save
- Windowing gives a better frequency response

$$X(k) = \frac{1}{M}\sum_{n=0}^{M-1} x(n)e^{-j\frac{\pi jkn}{M}}$$

Theory - Transforms





$$h_{hf}(n) = \sin\left[\frac{\pi}{N}(n+\frac{1}{2})\right]$$

Half-sine window (as used in MPEG-2)

Theory - Transforms

Problems with the DFT

- Block-band edge effects
- No perfect reconstruction in conjunction with a filterbank
- Time domain aliasing for different window sizes
- DFT coefficients are not uncorrelated (energy compaction is not optimal)

Theory - Transforms

Solution: Use Modulated Lapped Transform

- Basis of double length: critical sampling (2N samples provide N coefficients)
- Perfect reconstruction with filterbank and 2N samples
- Time domain aliasing cancellation
- Can be calculated using the DFT $=>$ fast
- Energy compaction is also not optimal

Theory - Transforms

Modified Discrete Cosine Transform (MDCT)

Transformation:

$$X(m) = \sum_{k=0}^{N-1} x(k)h(k)\cos\left(\frac{\pi}{2N}(2k+1+\frac{N}{2})(2m+1)\right) \quad m = 0, \cdots, \frac{N}{2}-1$$

Inverse transformation:

$$y(p) = \frac{4}{N}h(p)\sum_{m=0}^{\frac{N}{2}-1} X(m)\cos\left(\frac{\pi}{2N}(2k+1+\frac{N}{2})(2m+1)\right) \quad m = 0, \cdots, N-1$$

Theory - Transforms

Perfect reconstruction conditions:

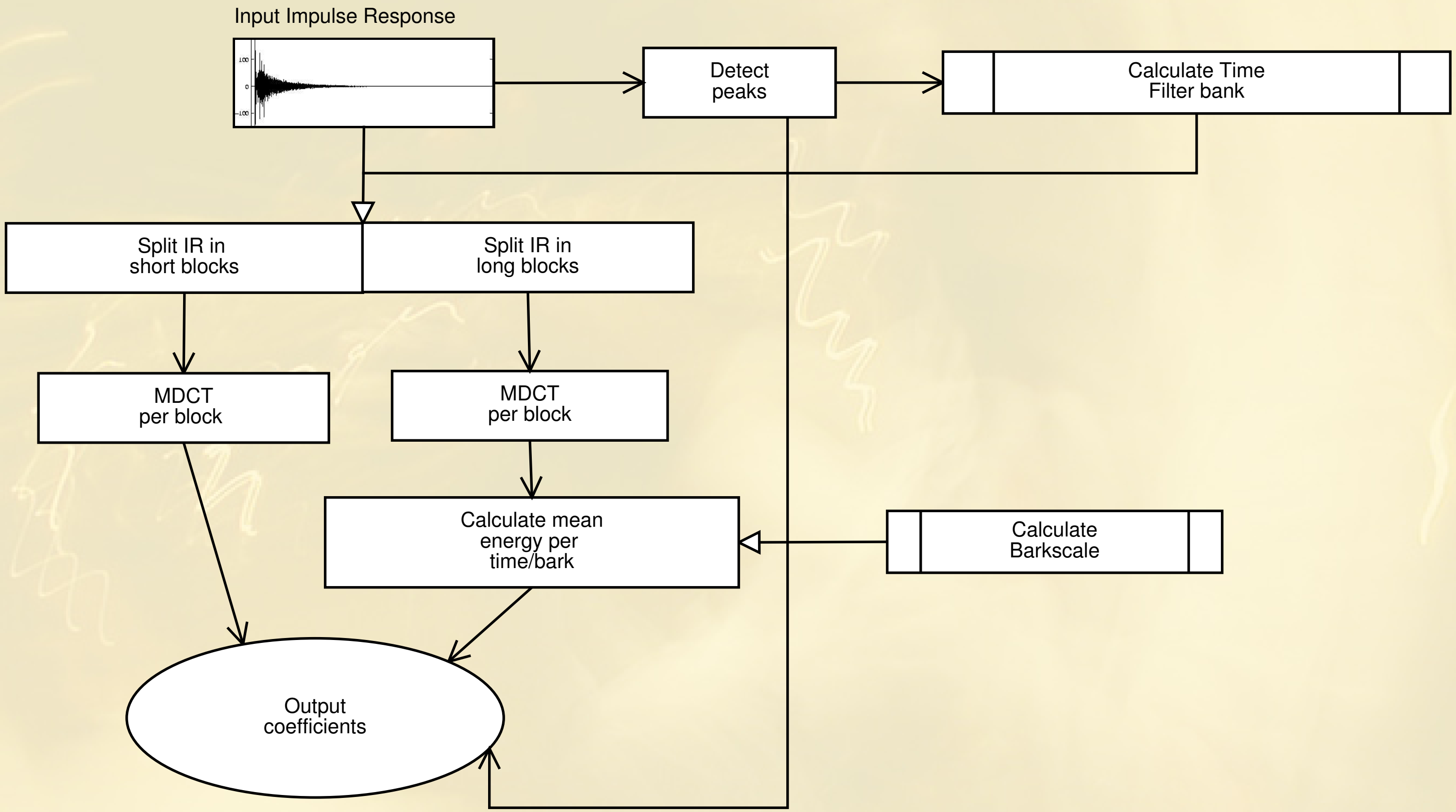
$$h^2(n) + h^2(n+M) = 1$$

$$h(2M-1-n) = h(n)$$

For example the half sine window:

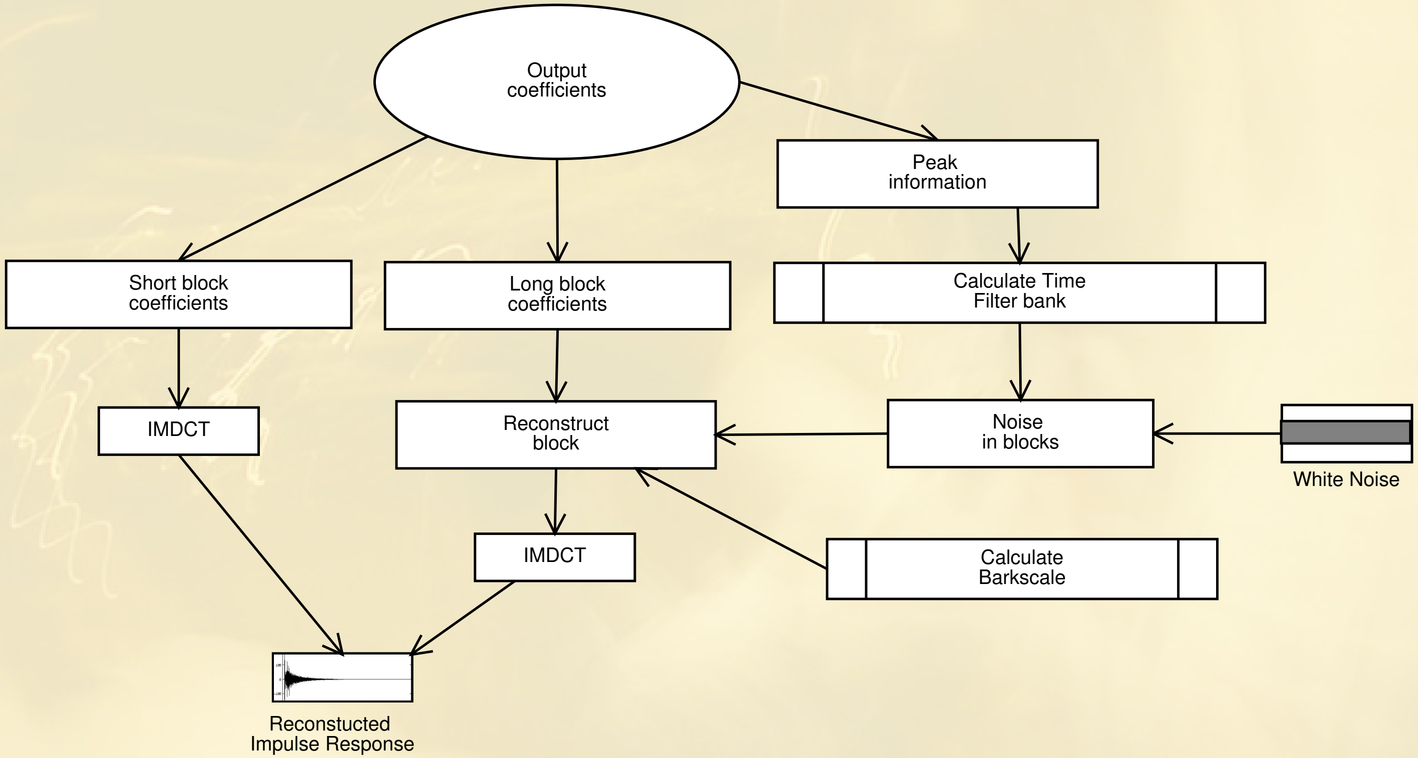
$$h(n) = \sin[\frac{\pi n}{N}] \quad \Big\} \quad n = 0 \cdots N-1$$

Proposed codec - Overview



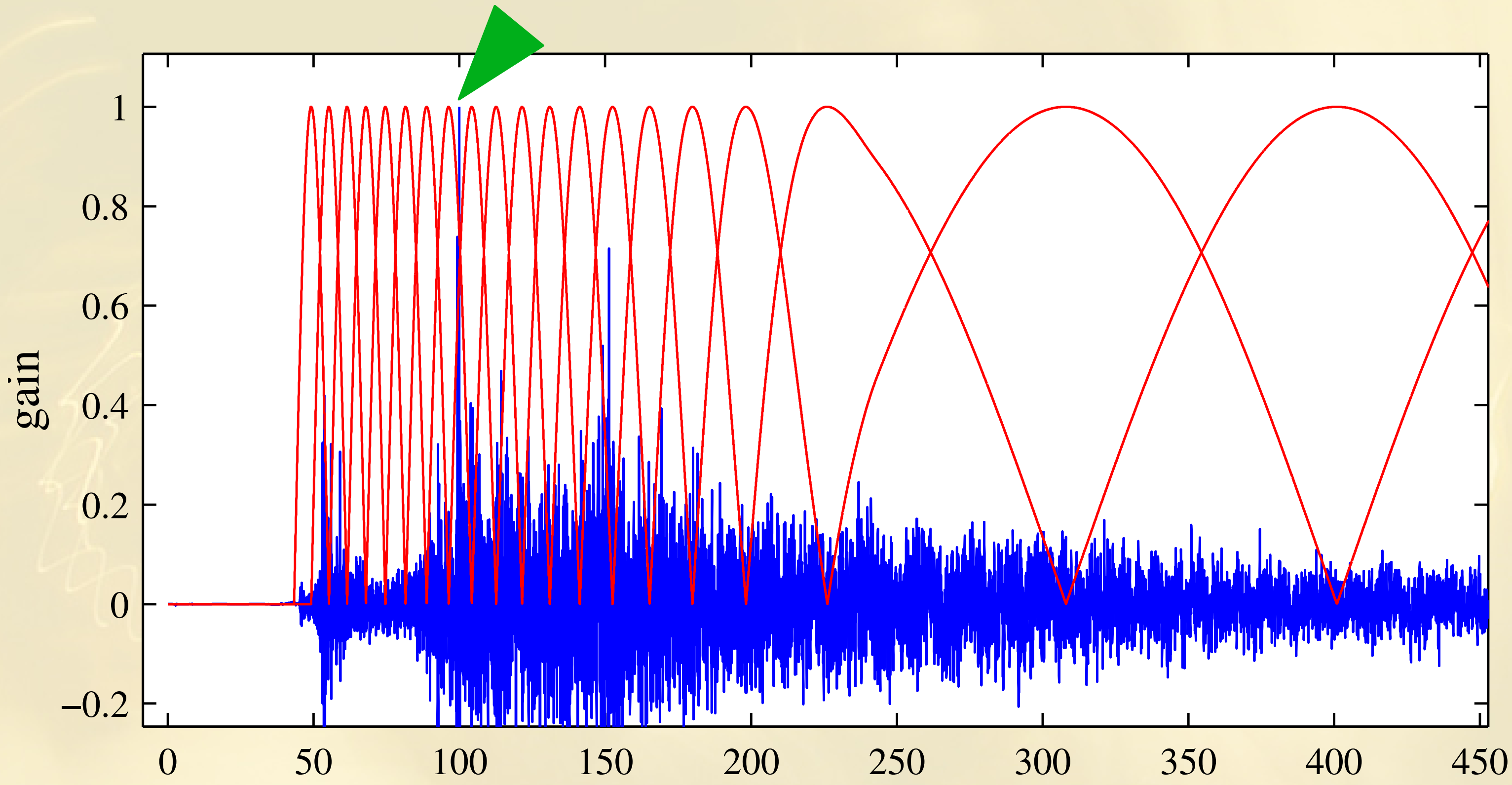
Overview of the transform encoder.

Proposed codec - Overview



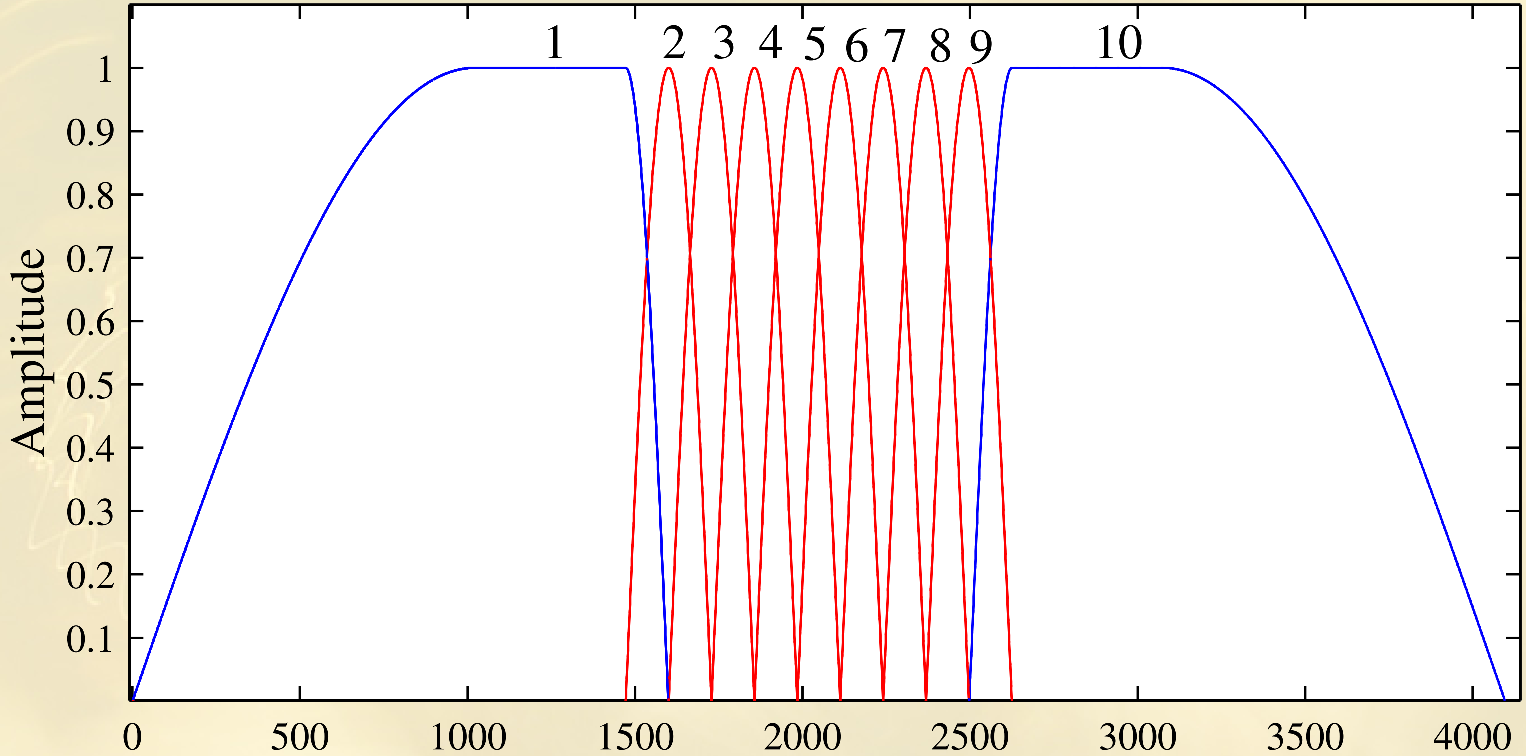
Overview of the transform decoder.

Proposed codec - Windowing



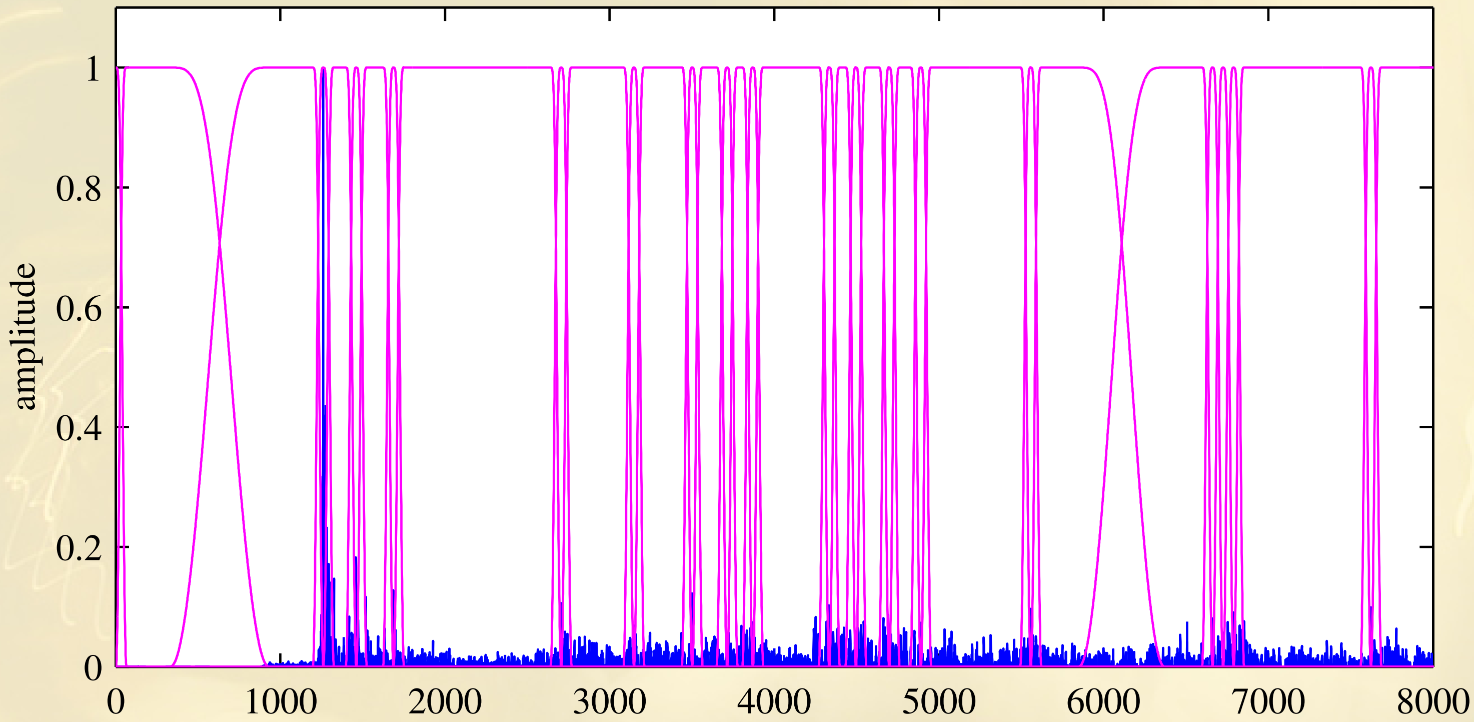
Using gradually longer windows

Proposed codec - Windowing



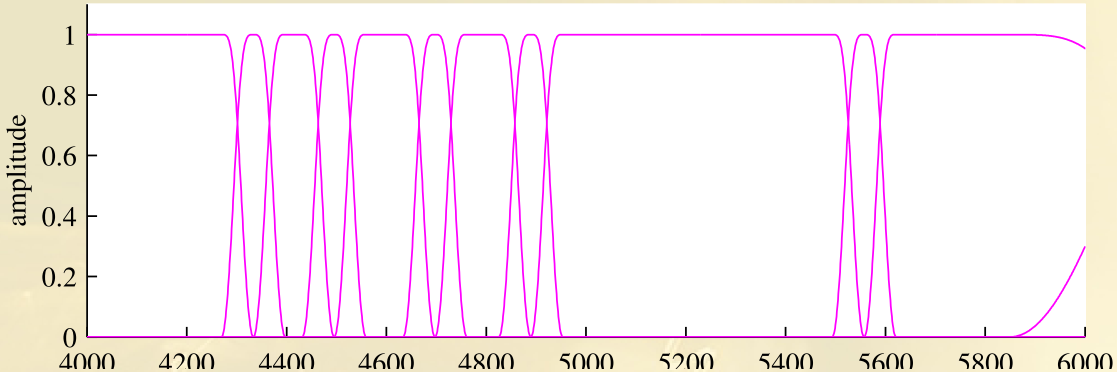
Window switching scheme in MPEG-2

Proposed codec - Windowing



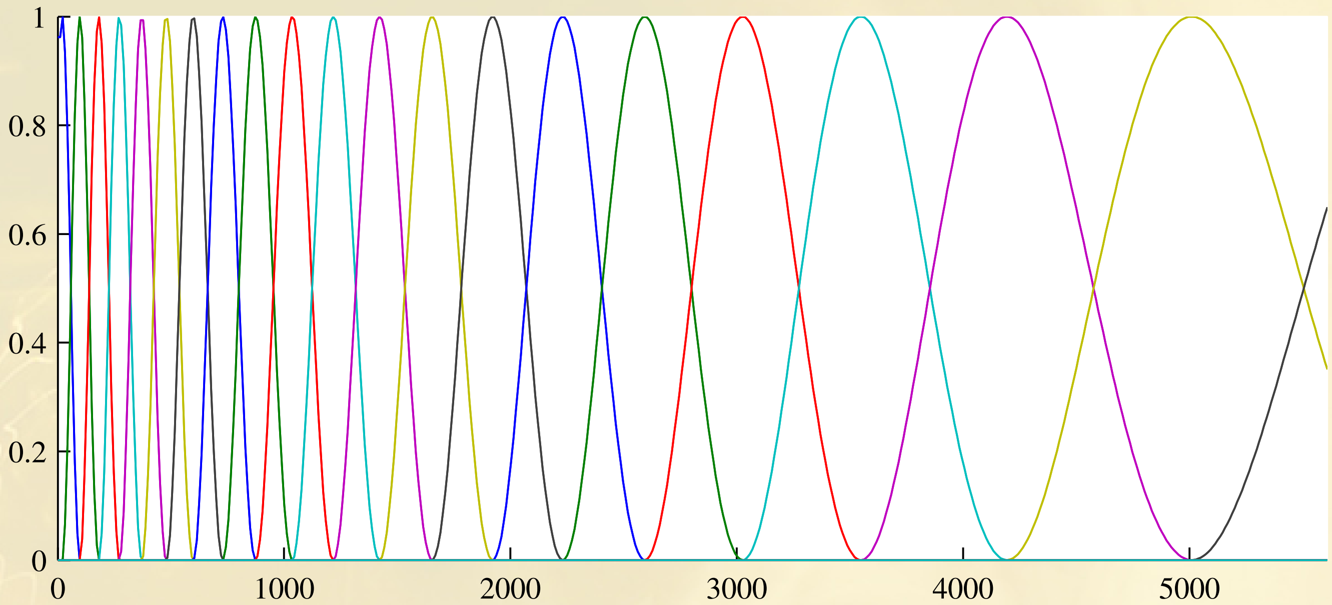
Using a window switching scheme

Proposed codec - Windowing



$$h_{start} = \begin{cases} h_{long}(n), & 0 \leq n \leq M-1 \\ 1, & M \leq n \leq M + \frac{M}{3} - 1 \\ h_{short}(n-M), & M + \frac{M}{3} \leq n \leq M + \frac{2M}{3} - 1 \\ 0, & M + \frac{2M}{3} \leq n \leq 2M - 1 \end{cases}$$

Proposed codec - Spectral coding



Bark scale converts frequency to critical band number. Zwicker:

$$z(f) = 13\arctan(0.00076f) + 3.5\arctan[(\frac{f}{7500})^2]$$

Proposed codec - Spectral Coding

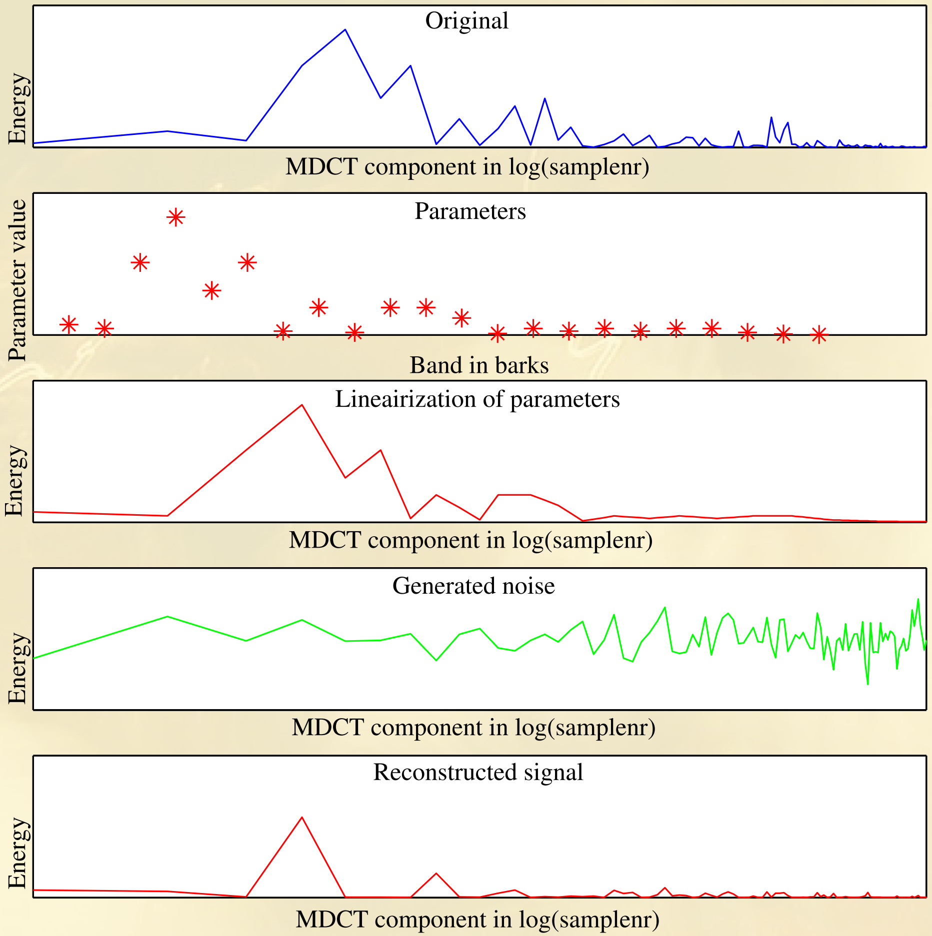
Encoding

1. Sum the energies of one Bark band in the spectrum
2. Divide by number of samples in that band

Decoding

1. Parameters are placed at the Bark center frequency
2. These points provide a spectrum line with linear interpolation
3. Spectrum line is multiplied with white noise

Proposed codec - Spectral Coding



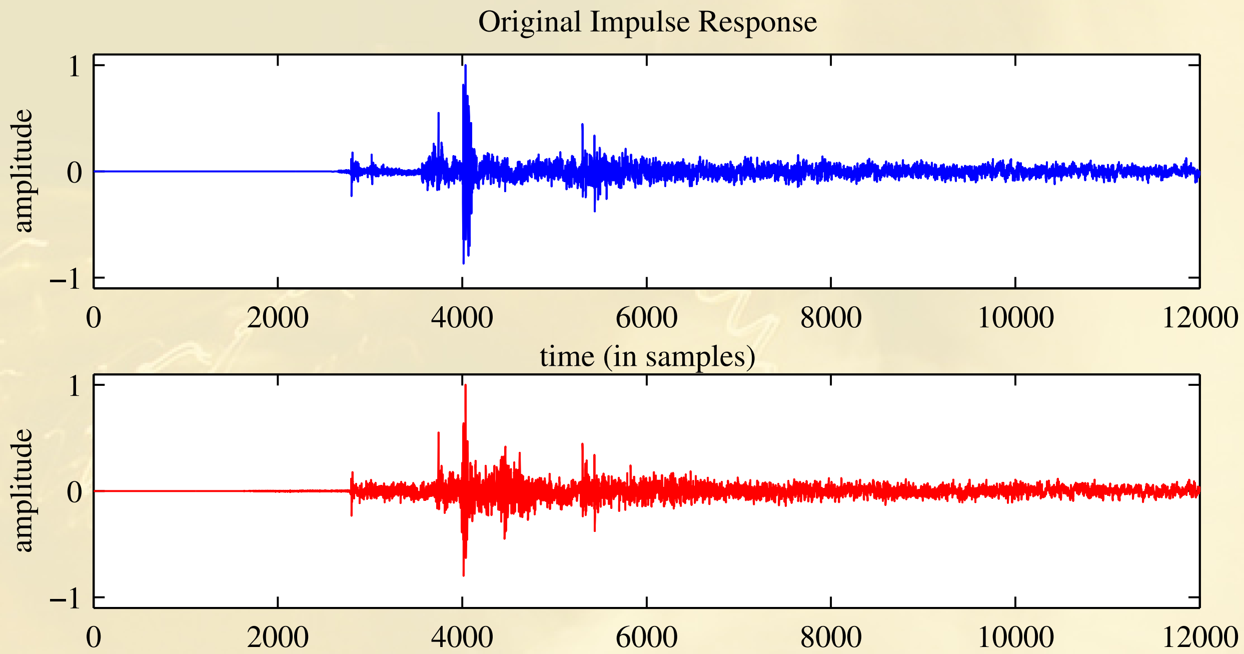
1. **Introduction**
   (a) Context
   (b) Research goals
2. **Theory**
   (a) Audio impulse responses
   (b) Coding
   (c) Transforms
3. **Proposed codec**
   (a) Overview
   (b) Windowing
   (c) Spectral coding
4. **Results**
   (a) Plots and observations
   (b) Listening test
5. **Conclusions & suggestions**
   (a) Conclusions
   (b) Suggestions
   (c) Questions

Results - Plots and observations

Parameter	Transform coder
Number of frequency bands	26
Smallest time window	128
Longest time window	2048
Percentage short windows	12.5 %
Total number of parameters	3488

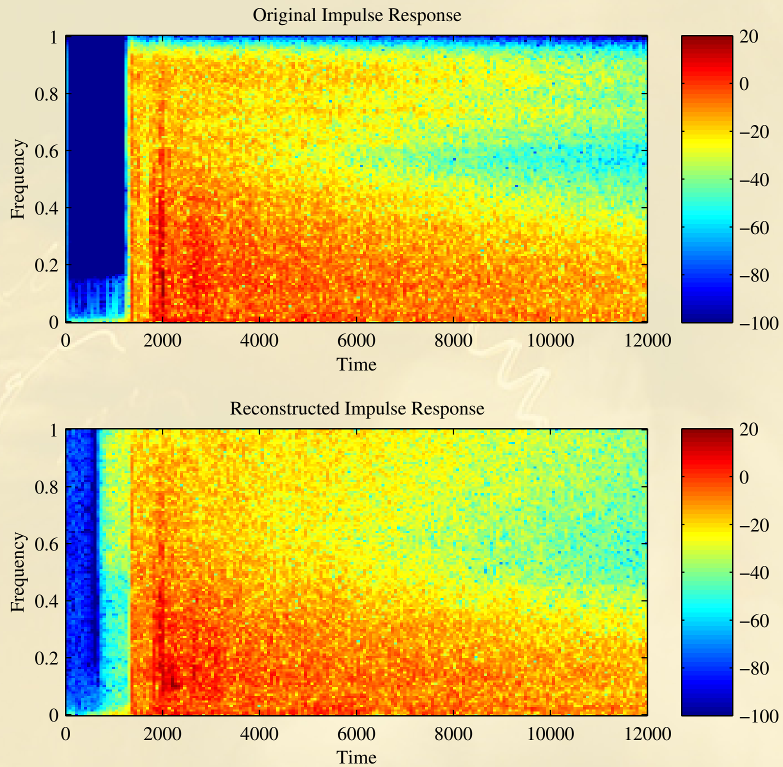
Some typical values of coder parameters, leading to a compression of 150x for an impulse response of 44.1 KHz, 16 bit

Results - Plots and observations



Time domain representation of the original and reconstructed impulse response (much reverb)

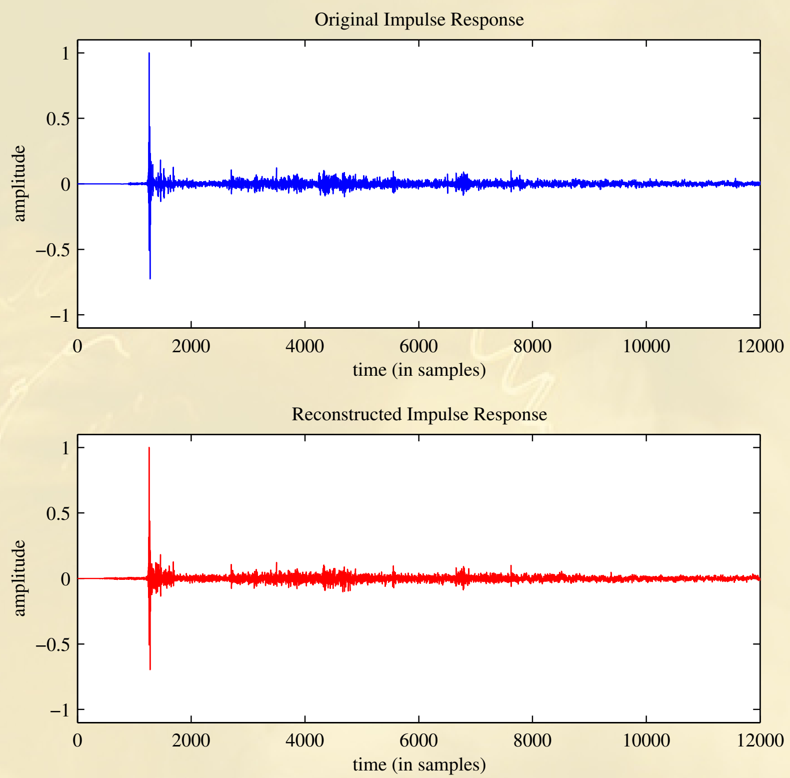
Results - Plots and observations



Spectrum of the original and reconstructed impulse response

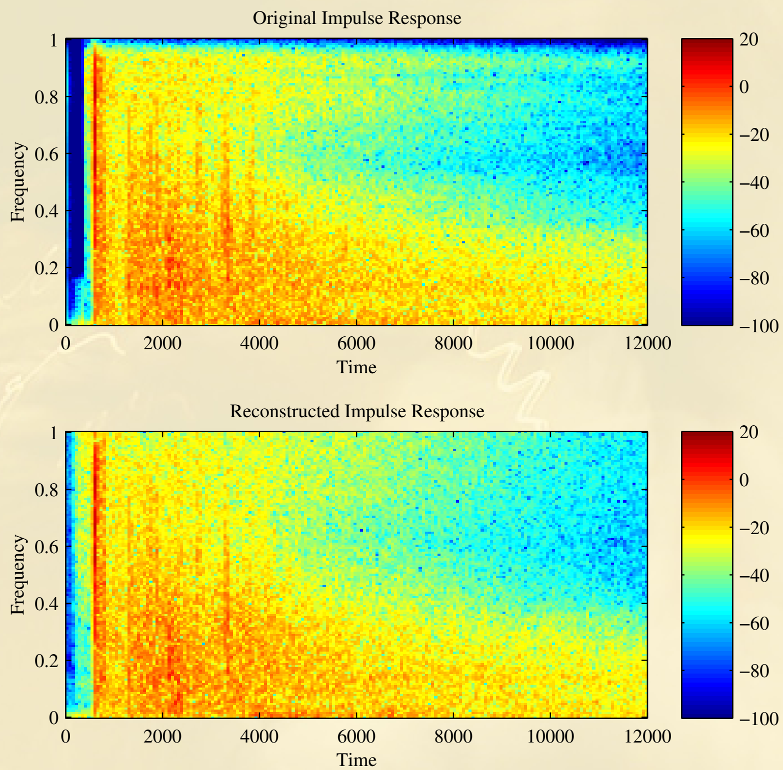
(much reverb)

Results - Plots and observations



Time domain representation of the original and reconstructed impulse response (less reverb)

Results - Plots and observations



Spectrum of the original and reconstructed impulse response

(less reverb)

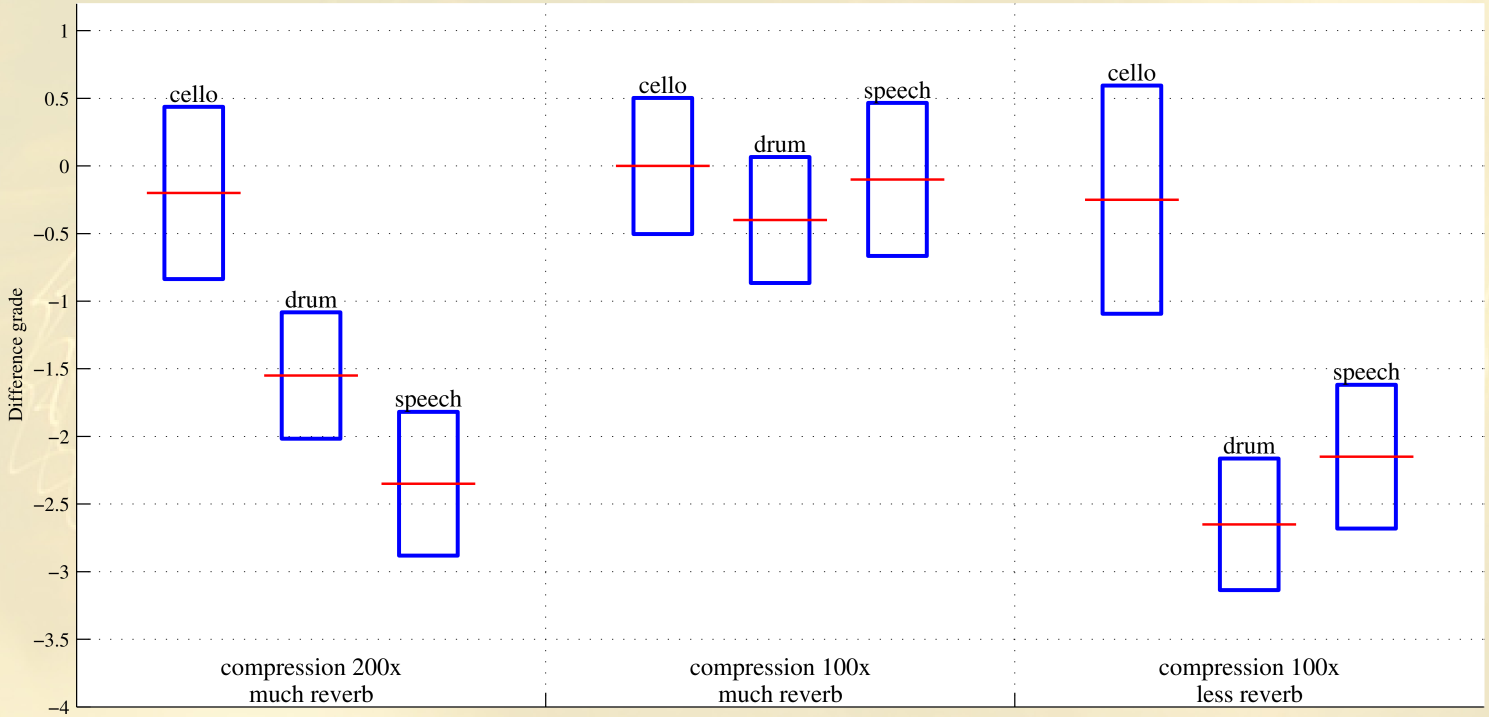
Results - Listening test

- Idea derived from ITU-R BS.1116-1, *Methods for the subjective assessments of small impairments in audio systems including multichannel sound systems*
- 21 listeners took part in the test
- Nine different sessions, done with 'double-blind triple stimulus with hidden reference'
- Expertise of listeners measured with t-test
- Statistical analysis with ANOVA model

Results - Listening test

Session	Reflections	Small h(n)	Large h(n)	Environment	Dry signal
1	8	64	4096	Much reverb	Cello
2	8	64	4096	Much reverb	Drums
3	8	64	4096	Much reverb	Speech
4	16	128	2048	Much reverb	Cello
5	16	128	2048	Much reverb	Drums
6	16	128	2048	Much reverb	Speech
7	16	128	2048	Less reverb	Cello
8	16	128	2048	Less reverb	Drums
9	16	128	2048	Less reverb	Speech

Results - Listening test



1. **Introduction**
   (a) Context
   (b) Research goals
2. **Theory**
   (a) Audio impulse responses
   (b) Coding
   (c) Transforms
3. **Proposed codec**
   (a) Overview
   (b) Windowing
   (c) Spectral coding
4. **Results**
   (a) Plots and observations
   (b) Listening test
5. **Conclusions & suggestions**
   (a) Conclusions
   (b) Suggestions
   (c) Questions

Conclusions & Suggestions - Conclusions

- The modulated lapped transform is a proper transform for coding of audio impulse responses (IR's)
- Window switching $=>$ short windows should overlap with the reflections in the IR
- Reconstructed IR approximates original if reverb is above certain level
- The compression factor can be 150x - 100x
- Below this level of reverb reconstruction the IR can be distinguished from the original IR

Conclusions & Suggestions - Suggestions

Enhance current coder:

- More research of proper parameters like, number of reflections, window size and quantization of the parameters $=>$ large scale listening test
- Research better algorithm for encoding and decoding of the spectrum (instead of linear interpolation)
- Use of vector quantization and codebooks for more (lossless) compression (useful, but falls outside physics field)

Conclusions & Suggestions - Suggestions

Different approaches to try:

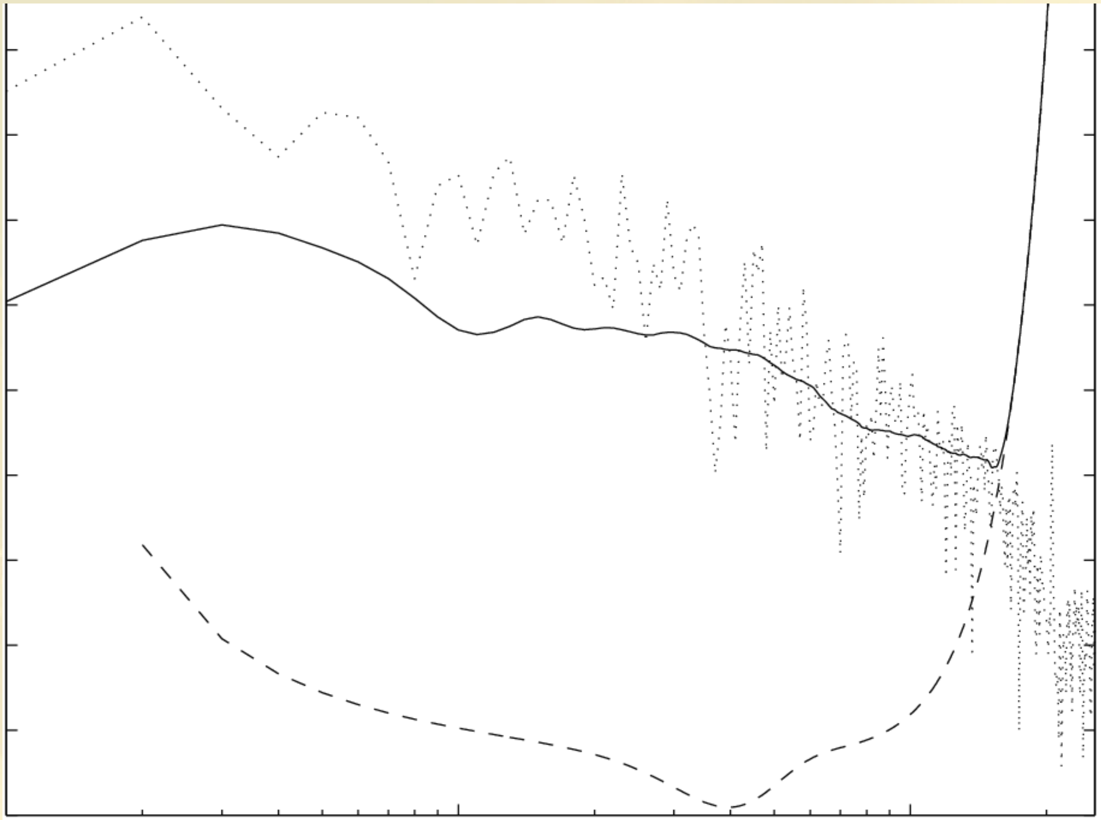
- Use a parametric coder in time domain (for example MPEG-2 CELP).
- Optimize the coder for fast convolution, instead of bandwidth, by combining the coder with partitioned convolution

Questions?

- Ask questions now!
- Read my thesis report http://vorm.net/pdf/verslag.pdf
- Contact me: jochem@njbg.nl

Thanks for your attention!

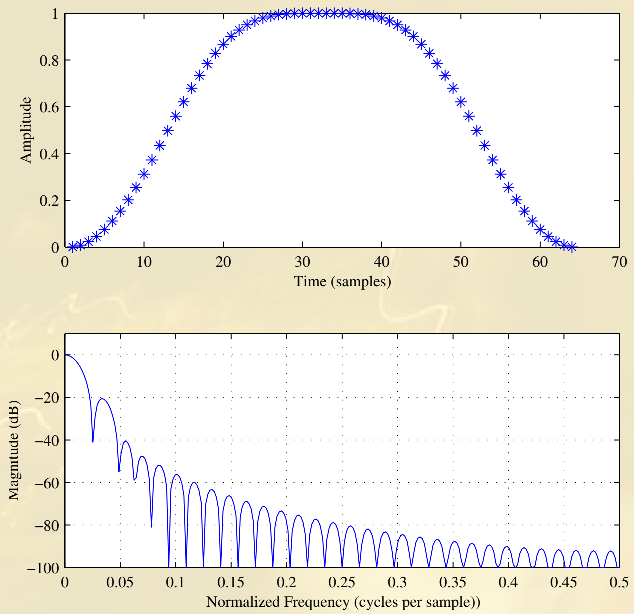
Extra sheet I



Music encoding algorithms save the calculated floor and the residue.

Calculations are done in the frequency domain.

Extra sheet II



$$h(n) = \sin\left[\frac{1}{2}\pi \sin\left(n+\frac{1}{2}\right)^2\right]$$

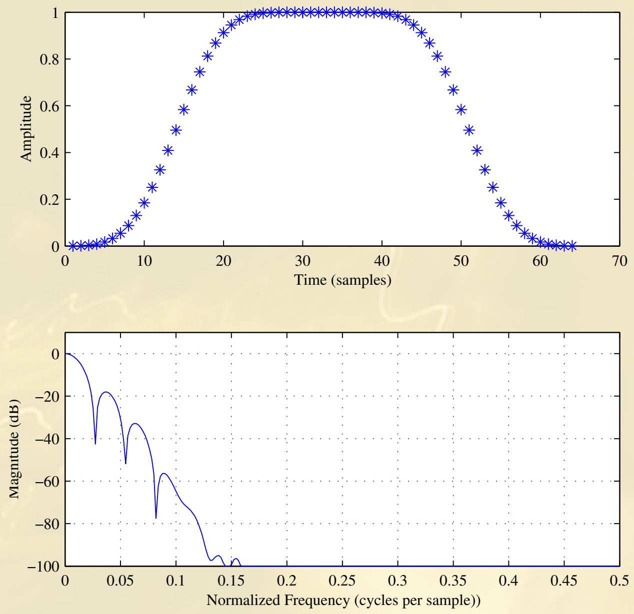
Designed window and its frequency selectivity

Extra sheet III

Using gradually longer windows

- Scales naturally with different impulse responses
- Does not exactly match peaks
- Does not fulfill perfect reconstruction conditions but can be used in conjunction with FFT

Extra sheet IV



$$h_{kbd}(n) = \sqrt{\frac{\sum_{i=0}^{n} \mathcal{W}(i)}{\sum_{i=0}^{N-1} \mathcal{W}(i)}}$$

Kaiser Bessel $\mathcal{W}(i)$ *Derived (KBD) window (*$\nu = 6$*)*

Extra sheet V

Zwicker's formula for frequency to Bark scale is:

$$z(f) = 13\,\mathrm{arctan}(0.00076f) + 3.5\,\mathrm{arctan}[(\frac{f}{7500})^2]$$

Traunmüller proposes:

$$z(f) = \frac{26.81f}{1960 + f} - 0.53$$

(and additional equations for low and high frequencies)