CODING OF AUDIO SIGNALS WITH OVERLAPPING BLOCK TRANSFORMS AND ADAPTIVE WINDOW FUNCTIONS

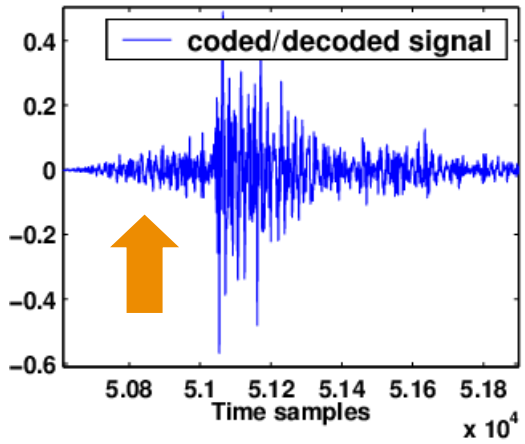
Reducing pre-echoes

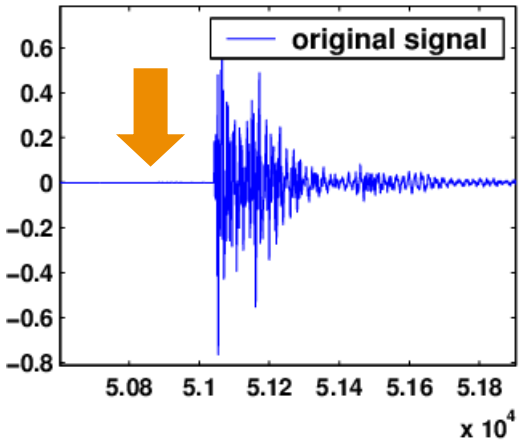
Stefan Ringer

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

What are pre-echoes?





What will be covered?

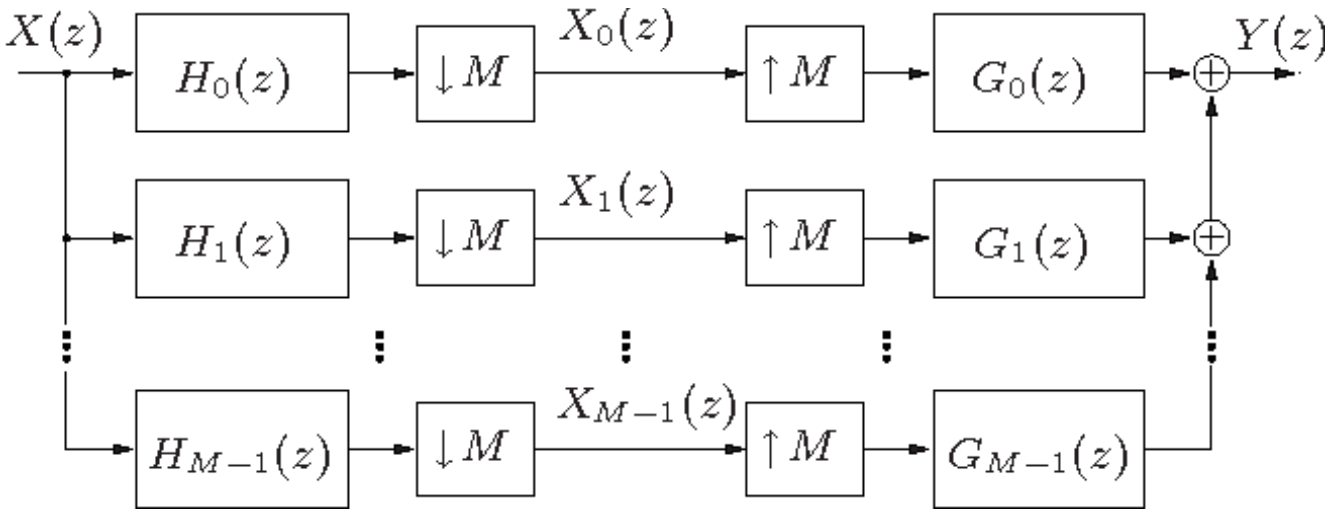
1. Introduction
   - Filterbanks & overlapping blocktransforms
   - The problem of pre-echoes

2. Adaptive Windowing
   - Novelty of choosing distinct window functions
   - A polyphase matrix based derivation
   - Alternatives in the literature

3. Discussion

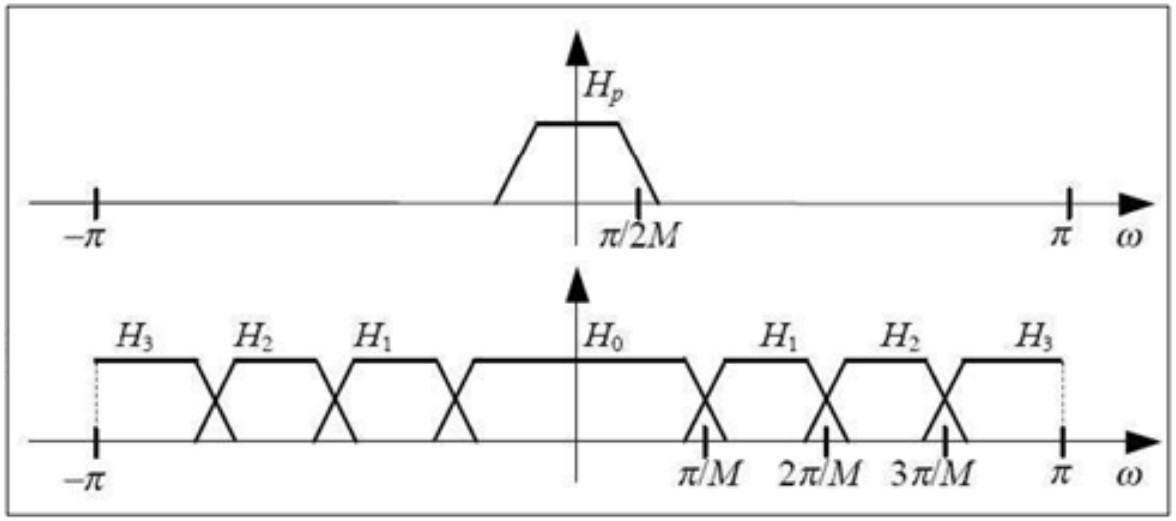
“Tiling” the frequency domain – Filterbanks I

- “Tiling” in the frequency domain is desirable
  - Easier to pair with a perceptual model (signal power over frequencies)
  - Allow for exploitation of frequency masking when quantizing
  - Reduction of statistical redundancies (harmonics...)
- Filterbanks are an often-used tool for that matter



“Tiling” the frequency domain – Filterbanks II

- Creating Sub-bands in the frequency domain



- Constraint: Perfect reconstruction
  - Cancel the frequency domain aliasing
  - Choice of G and H constrained
  - G and H not necessarily equal

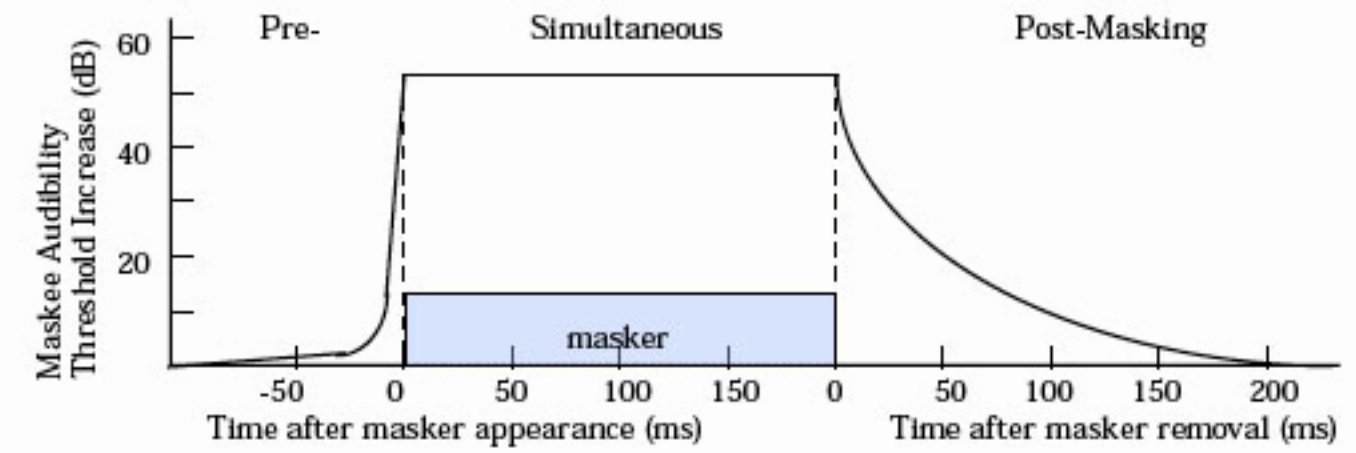
$$G_0(z)H_0(-z) + G_1(z)H_1(-z) = 0$$
$$G_0(z)H_0(z) + G_1(z)H_1(z) = 2$$

Where do pre-echoes come from? I

- Transient is broad in frequency domain
- Psycho Acoustic Model determines quantization levels
  - Suitable if signal statistics are similar within processed (time) signal
    e.g. sum of sinusoidals
  - Quantization noise is spread across whole time axis of block
  - In quiet areas before transient masking is not as strong and thus noise is audible

Where do pre-echoes come from? II

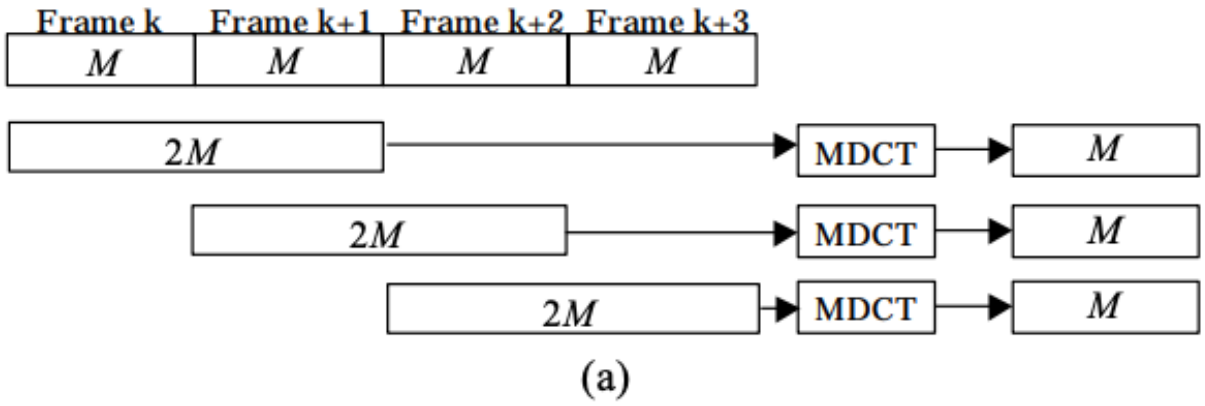
- Why is the effect called **pre**-echo?



- Forward (time direction) temporal masking is stronger than backward temporal masking
- Audible in percussions, German male speech etc.

“Tiling” the time domain – Overlapped Transforms I

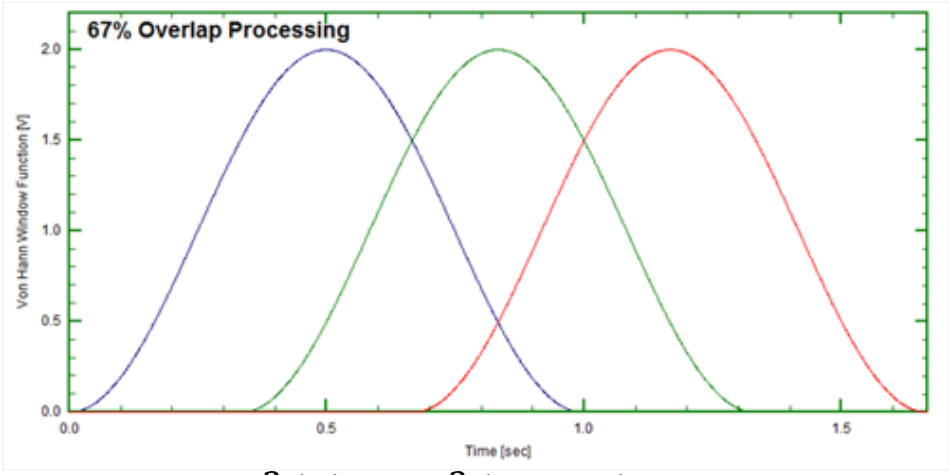
- Temporal ”tiling” is needed
  - Audio material is highly non-stationary
  - Only limited processing delay permitted
- Overlapping Transform often chosen for for that matter: MDCT



- Just like before: $Rect$ functions lead to blocking artefacts

“Tiling” the time domain – Overlapped Transforms II

- Windowing the blocks in the time domain



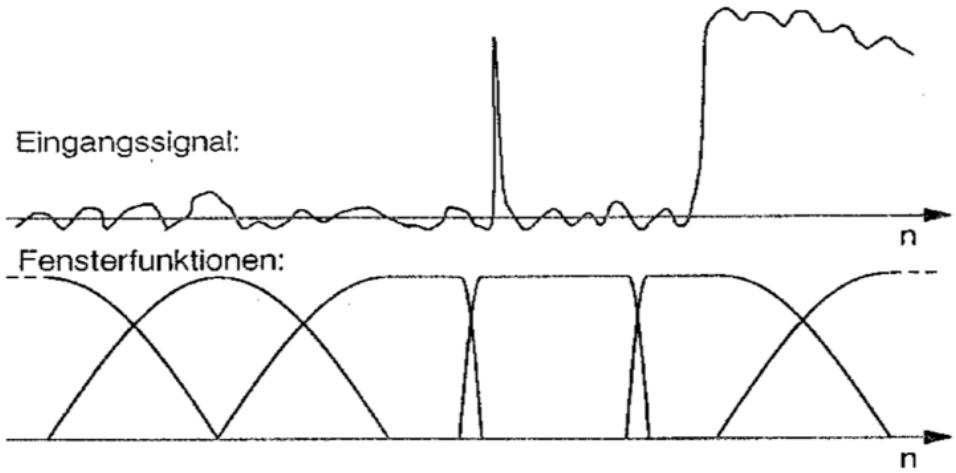
$$w^2(n) + w^2(M+n) = 1$$

$$\mathrm{w}(n)\, w(M-1-n) = w(M+n)\, w(2M-1-n)$$

- Constraint: Perfect Reconstruction
  - TDAC: **Time** domain **aliasing** cancelling ($\boldsymbol{2M \longrightarrow M}$)
  - Princen-Bradley Condition (1987)
  - M free to choose but forced constant because $w(t)$ is supposed to be the same

“Tiling” the time domain – Overlapped Transforms III

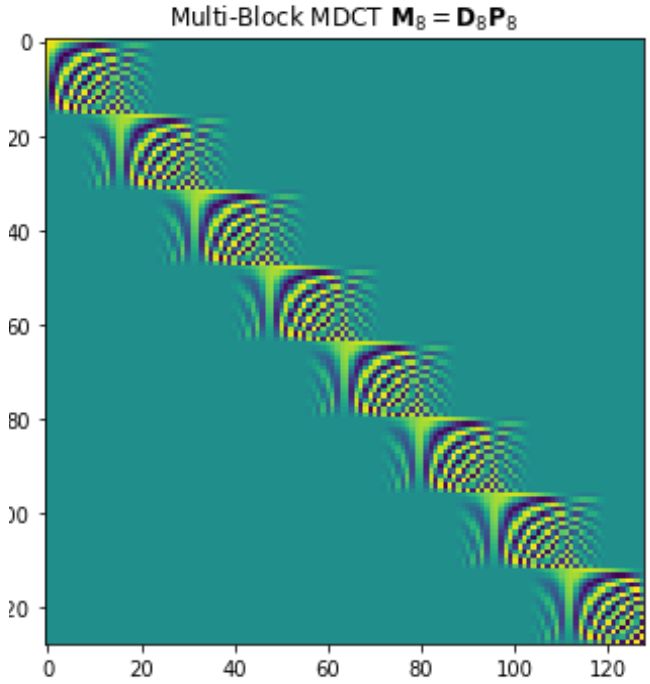
- If we use shorter and longer windows, we can adapt to transients and stationary signal parts:
- Still perfectly reconstructable if we use two distinct window functions?

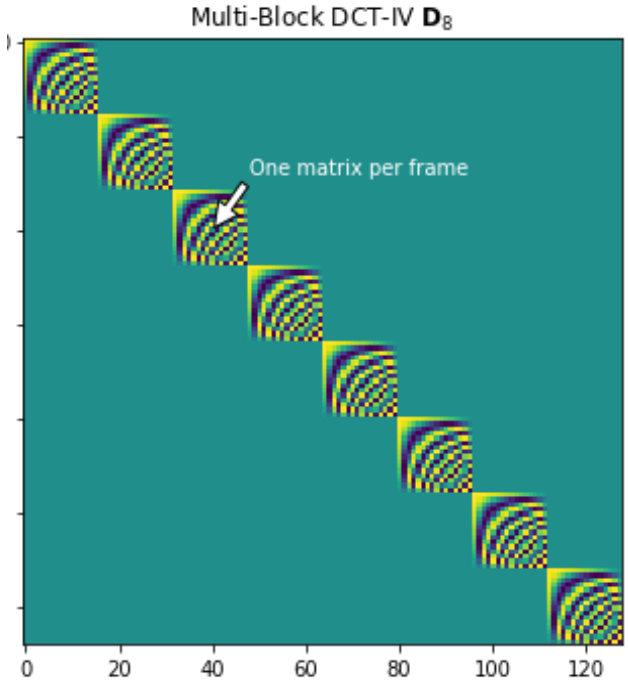


- **YES!**
  This is the novelty of Prof. Edler’s paper:

$$f(N+n)\,f(2N-1-n) = g(n)\,g(N-1-n)$$
$$f^2(N+n) + g^2(n) = 1$$

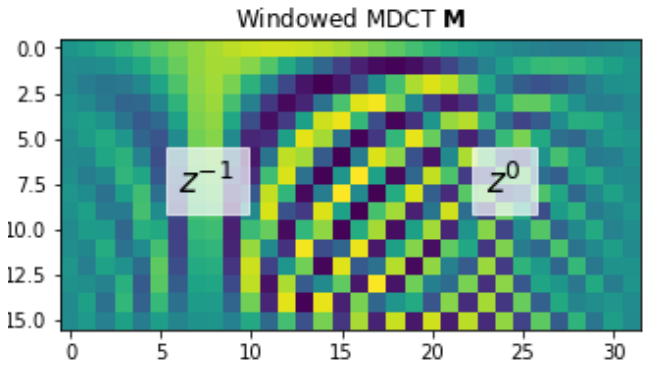
Deriving the window constraints I

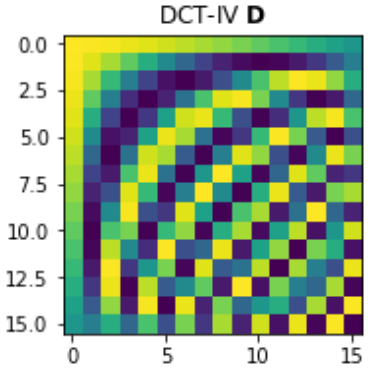


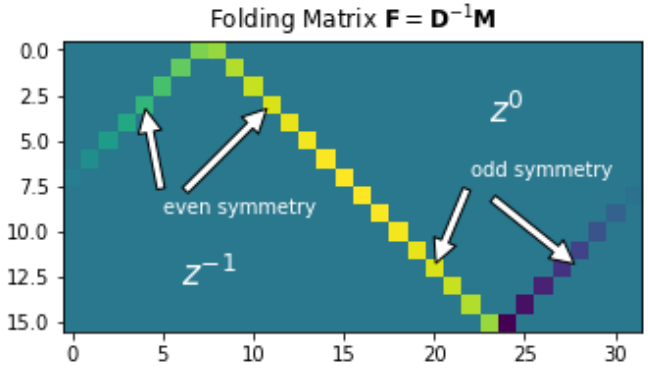


- $\vec{X} = \boldsymbol{M}\,\vec{x}$
- $\vec{X} = \boldsymbol{D}\,\boldsymbol{F}\,\vec{x}$
- $\boldsymbol{M}$: **M**DCT-Matrix
- $\boldsymbol{D}$: **D**CT-IV- Matrix

Deriving the window constraints II





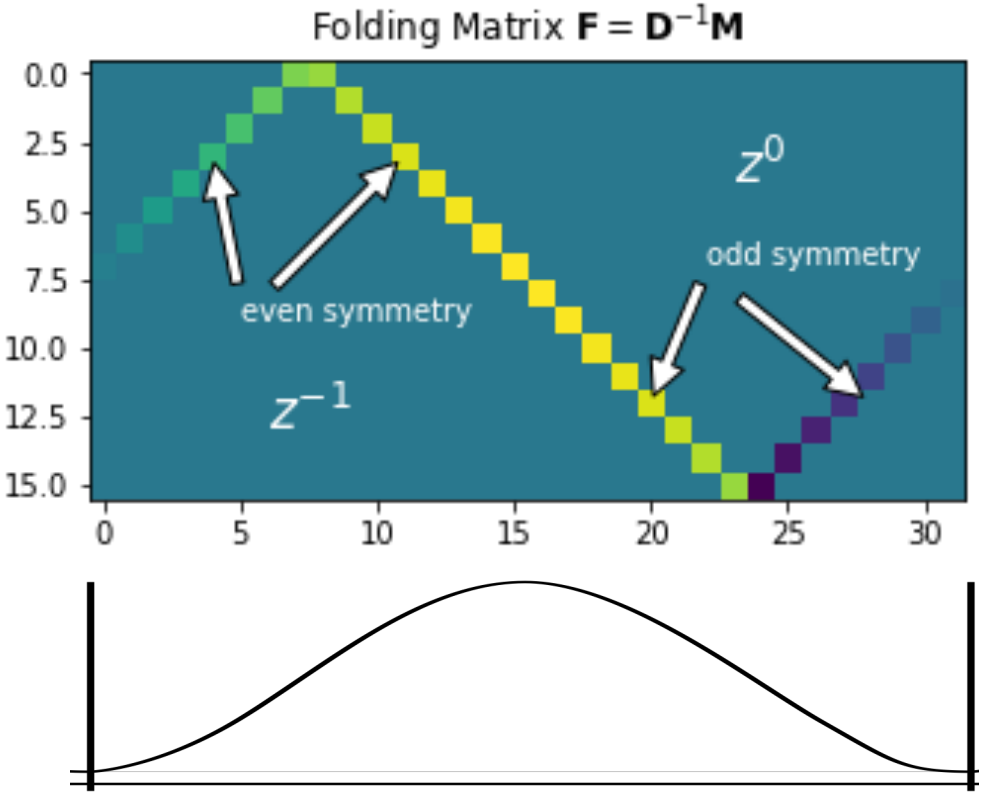


- $\boldsymbol{M} = \boldsymbol{D}\,\boldsymbol{F}(z)$
- z^{-1}= previous frame z^{0} = current frame
- $\boldsymbol{F}(z)$: window functions & sequence of block lengths

$$\vec{X}(z) = \boldsymbol{D}\,\boldsymbol{F}(z)\,\vec{x}(z)$$

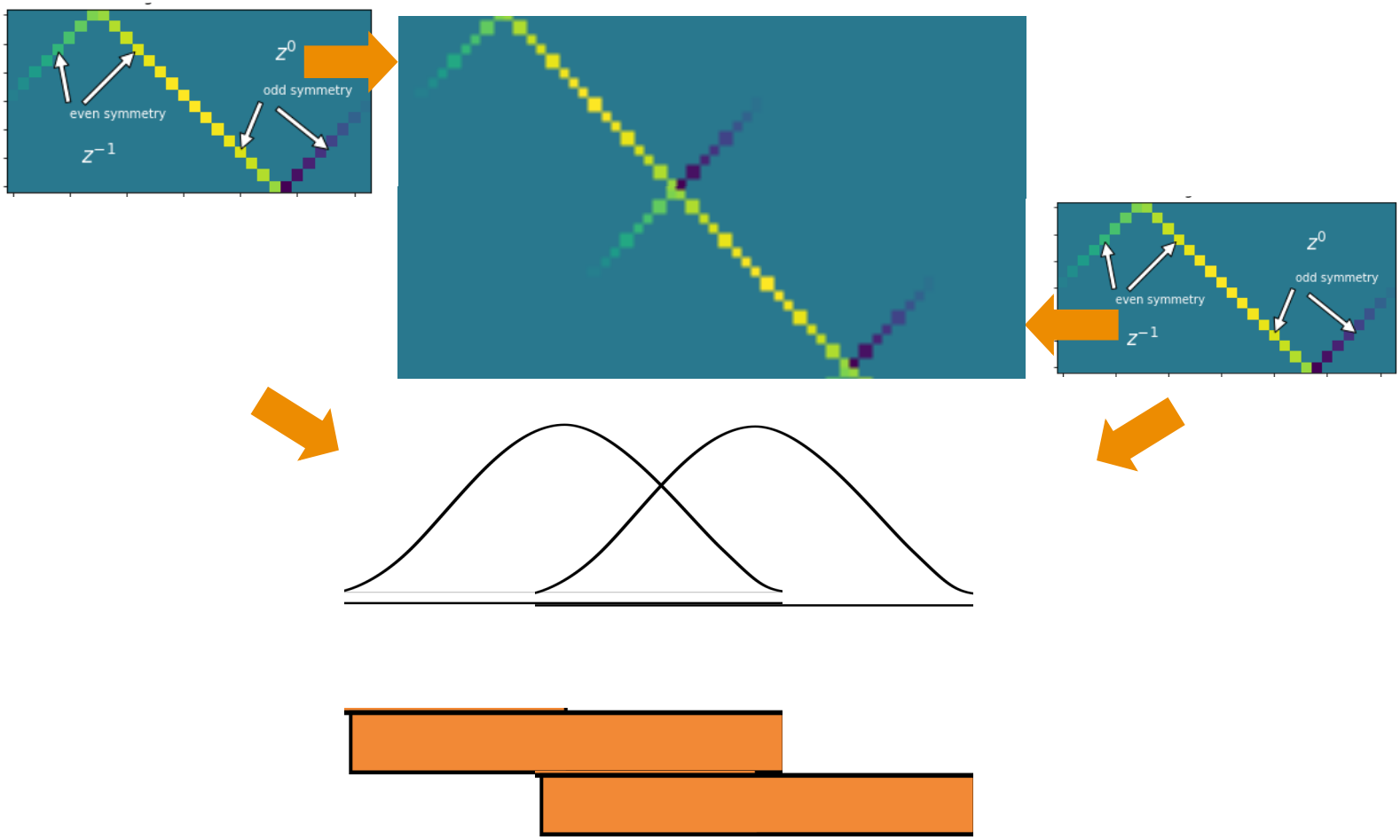
- Perfect Reconstruction if $\boldsymbol{D}\,\boldsymbol{F}(z)$ invertible (in total)
- $\boldsymbol{D}$ is orthogonal, so $\boldsymbol{D}^{-1} = \boldsymbol{D}^{T}$
- When is $\boldsymbol{F}(z)$ invertible? (It is rectangular?!)

Deriving the window constraints III

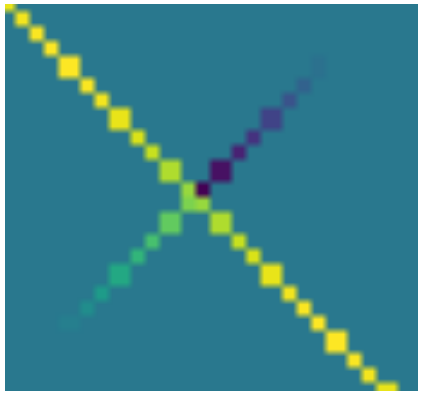


- The Matrix may be seen as a continuation of the DCT basis functions (the window extends the range of the transformation)
- A single Block/Window is not reconstructable, so let's look at a sequence of overlapping ones

Deriving the window constraints IV

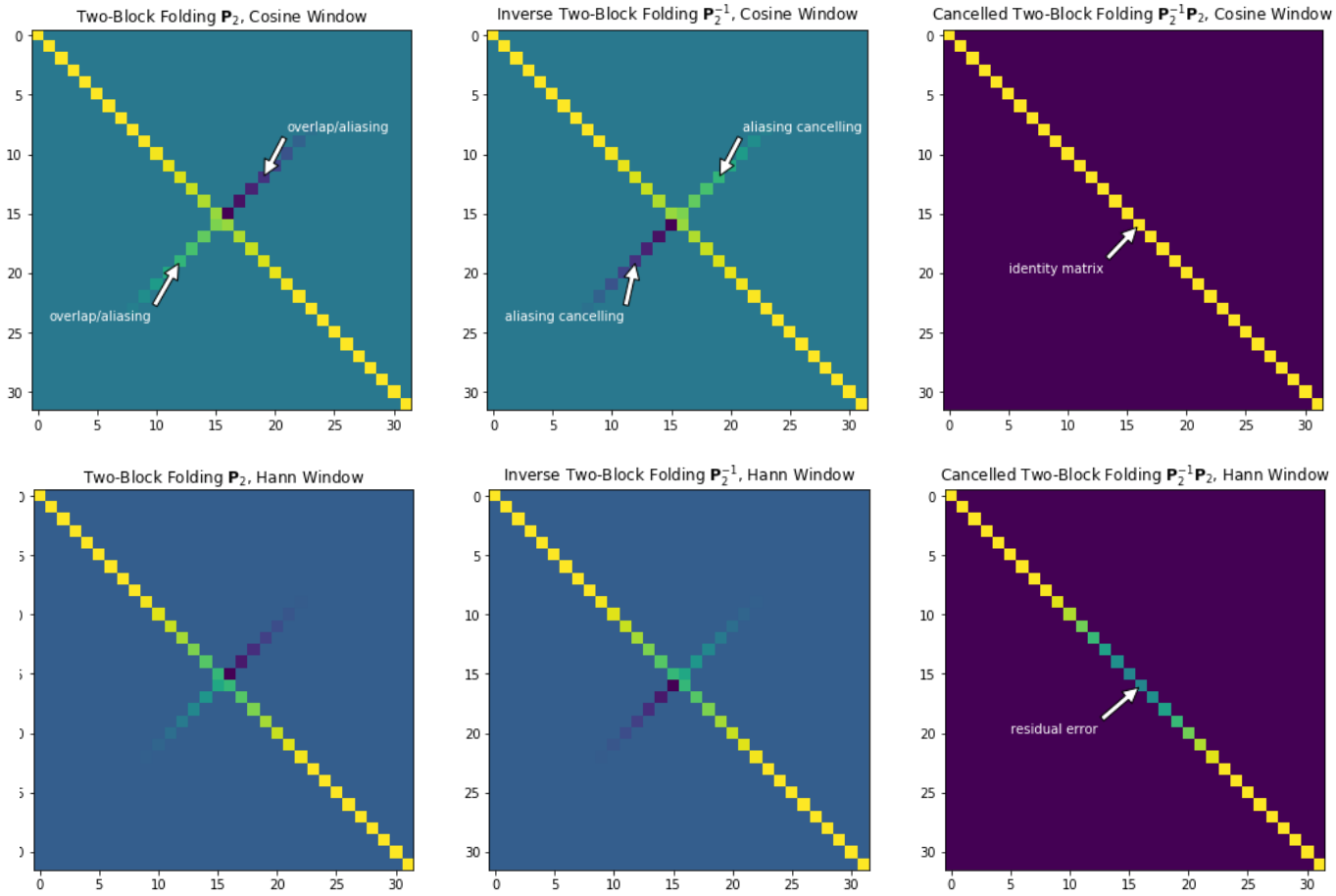


Deriving the window constraints V



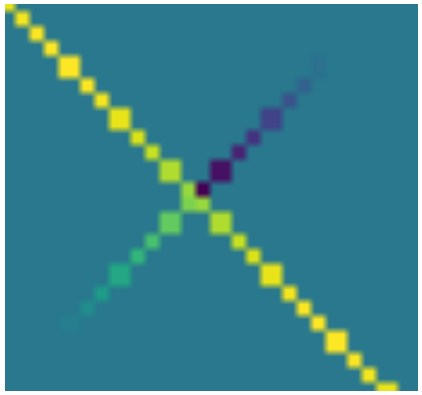
- This matrix is generally **square**
- Represents: Overlap of blocks & window functions
- If this Matrix is orthogonal: Perfect reconstruction: $\boldsymbol{F}^T = \boldsymbol{F}^{-1}$
  - Resulting transformed signal resilient to quantization
  - Eg white noise orthogonally transformed still white noise (not true for “only” invertible transform): Only rotation not scaling of base vectors
- Sufficient to only check this matrix since rest are identity matrices

Deriving the window constraints VI



- (Numerical) check for perfect reconstruction: $\boldsymbol{F}\,\boldsymbol{F}^T = \boldsymbol{I}$

Deriving the window constraints VII



- Possible to arrange the matrix into 2x2 matrices

$$\begin{bmatrix} f(N+n) & -f(2N-1-n) \\ g(n) & g(N-1-n) \end{bmatrix}$$

Deriving the window constraints VIII

- $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$

$$\begin{bmatrix} f(N+n) & g(n) \\ -f(2N-1-n) & g(N-1-n) \end{bmatrix} = \begin{bmatrix} g(N-1-n) & f(2N-1-n) \\ -g(n) & f(N+n) \end{bmatrix}$$

- $Det(\boldsymbol{A}) = 1$
- $f(N+n)g(N-1-n) + g(n)f(2N-1-n) = 1$

- Equations:

$$f(N+n) = g(N-1-n)$$
$$f(2N-1-n) = g(n)$$
$$f^2(N+n) + g^2(n) = 1$$

Deriving the window constraints IX

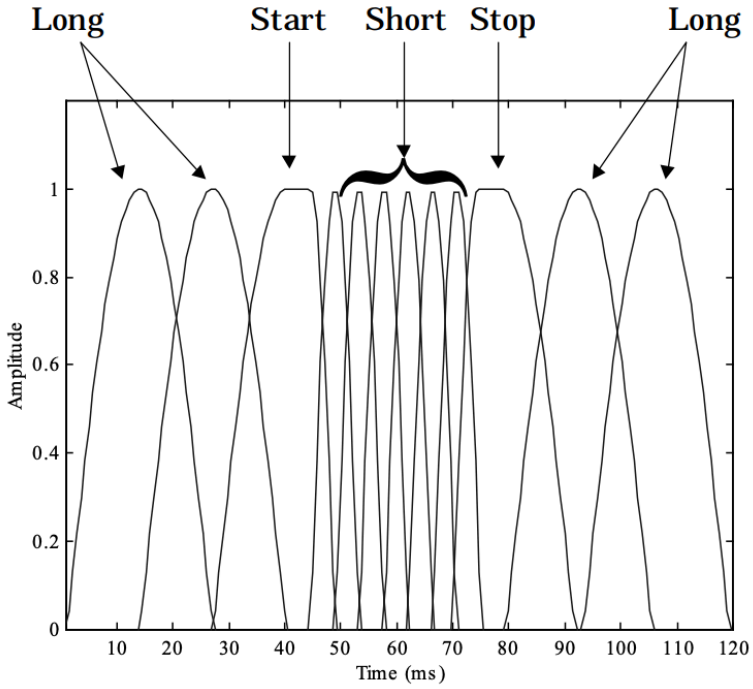
$$f(N+n)\, f(2N-1-n) = g(n)\, g(N-1-n)$$
$$f^2(N+n) + g^2(n) = 1$$

- Notice how the previously known Princen-Bradely condition from $f(n) = g(n) = w(n)$, $N = M$ follows from this more general result:

$$w^2(n) + w^2(M+n) = 1$$
$$\mathrm{w}(n)\, w(M-1-n) = w(M+n)\, w(2M-1-n)$$

- Transform lengths of successive blocks can now be chosen independently

Application of adaptive windowing I



- Signal part without peaks: Long transform length
  - Optimizing coding gain
- Signal peaks: Switch to shorter transform length
  - Reducing pre-echoes
- Peak detection needed

Advantages and Drawbacks of Adaptive Windowing

Advantages

- Improved perceptual quality while still achieving good coding gain
- Compatible with MDCT for which fast FFT based implementations are available

Drawbacks

- Perceptual model and lossless coding algorithms need to support multiple time resolutions
- High coder delay
- Overusage of short windows for “pitched” signals (degraded coding gain)

Alternatives in the literature I

- **Bit reservoir:**
  - Transients need more bits to be properly encoded
  - Idea: Store bits from resource intensive parts in less intensive parts
  - Problem: In signals with lots of transients unfeasibly large bit reservoirs are needed
- **Hybrid and Switched Filter Banks:**
  - Window Switching: Only transform and window length are adapted.
  - Switched Filter Banks: Filter banks themselves are exchanged depending on the current signal statistics
  - Hybrid Filter Banks: Cascade filters for optimal time-frequency tiling

Questions?

Peak detection I

- Proposed algorithm by Edler detects the existence but not the position of a peak by two relative metrics.
- $c_1(v) = \frac{\sum_{n=0}^{N-1} |d(vN+n)|}{\sum_{n=0}^{N-1} |d(vN-N+n)|}$
  - $d(n) = x(n) - x(n-1)$. v is the block number.
  - Quotient of energies (L_2 replaced by L_1 norm) of two successive blocks
  - Big ratio $c_1(v)$ is an indicator of a peak in the current block v
  - $d(n)$ is a discretized derivative and with $\frac{df(t)}{dt} \multimap\!\bullet\ i\omega \mathrm{G}(\mathrm{j}\omega)$ recognizable as highpass filtering that improves detection performance

Peak detection II

- $c_2(\nu) = \frac{\sum_{n=0}^{N-1} \max\{|x(\nu N+n)|\} \cdot N}{\sum_{n=0}^{N-1} |x(\nu N+n)|}$
  - $c_1(\nu)$ not suitable for detecting small impulses that contain few energy
  - maximum abs value in relation to the mean of the abs values within a block

- Complement one another and can be used for switching of the lengths as one of them passes a certain threshold.
- The transform length is then reduced from $N_1 = 512$ to four transforms of the length $N_2 = 128$ and window overlap $L = 32$
- Only 1 control bit for the adaption has to be given for each block of 512 samples.

Alternatives in the literature I

- **Gain Modification [9]:**
  - Before coding, transients are smoothed and then processed like stationary blocks
  - Since the block is smoothed the temporal spread of quantization noise is not as significant and thus results in a reduced pre-echo.
  - This approach can lead to issues of broadening the filter bank responses at low frequencies beyond critical bandwidth while measures to tackle this problem struggle with coder complexity

Derivation of constraints following original paper

- See original paper