

# Perceptual Audio Coding

Henrique Malvar  
Communication, Collaboration,  
and Signal Processing Group

Microsoft  
**Research**

UW Lecture - March '04

## Contents

- Motivation
- Auditory Masking
- Block & Lapped Transforms
- Audio compression
- Examples

# Motivation

3

## Many applications need digital audio

- Business
  - Internet call centers
  - Multimedia presentations
- Communication
  - Telephony & teleconferencing
  - Voice mail, voice annotations on e-mail
- Entertainment
  - solid-state music players
  - 150 songs on standard CD
  - 1,500 songs on portable Jukebox
  - Internet radio
  - Games

4

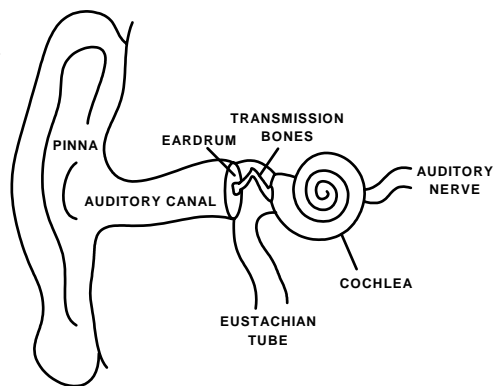
# Auditory Masking

Model the sink, not the source

5

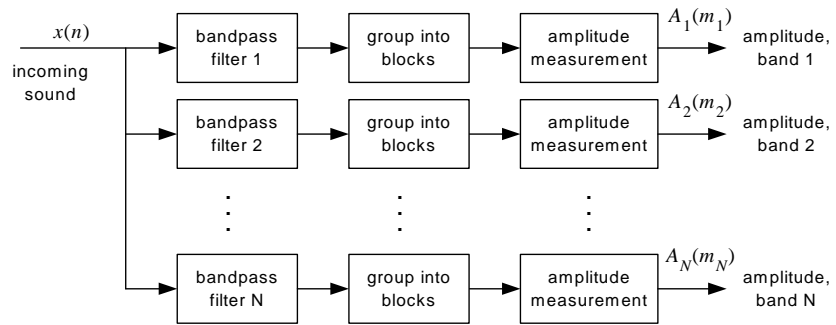
## Physiology of the ear

- Thousands of "microphones"
  - hair cells in cochlea
- Automatic gain control
  - muscles around transmission bones
- Directivity
  - pinna
- Boost of middle frequencies
  - auditory canal
- Nonlinear processing
  - auditory nerve



6

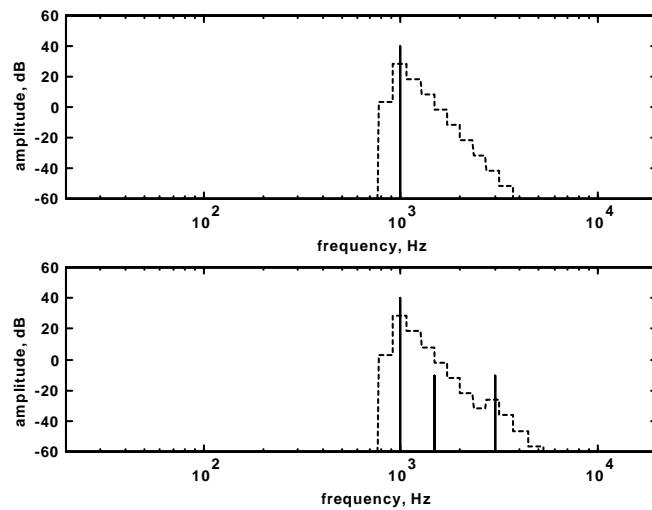
## Filter bank model



- Explains frequency-domain masking

7

## Frequency-domain masking



8

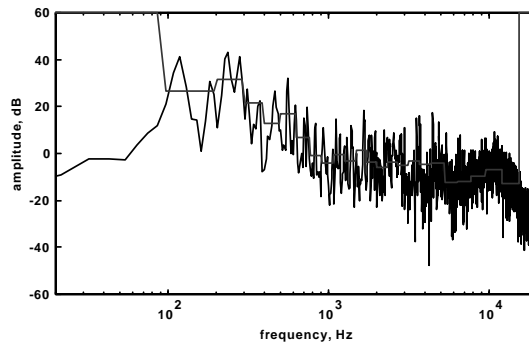
## Example of masking

- Typical spectrum & masking threshold

- Original sound:



- Sound after removing components below the threshold (1/2 to 1/3 of the data):

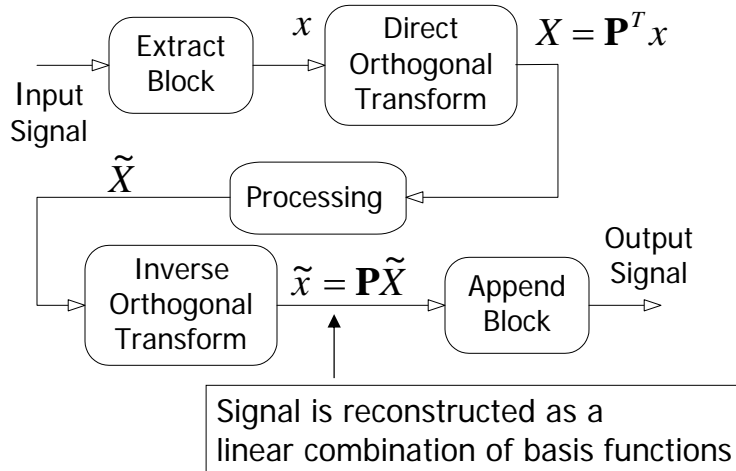


9

## Block & Lapped Transforms

10

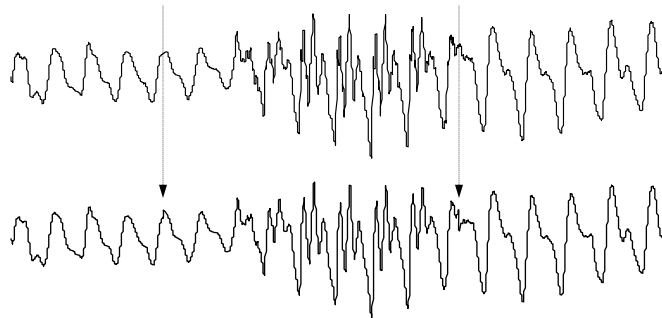
## Block signal processing



11

## Block processing: good and bad

- Pro: allows adaptability



- Con: blocking artifacts

12

## Why transforms?

- More efficient signal representation
  - Frequency domain
  - Basis functions ~ “typical” signal components
- Faster processing
  - Filtering, compression
- Orthogonality
  - Energy preservation
  - Robustness to quantization

13

## Compactness of representation

- Maximum energy concentration in as few coefficients as possible
- For stationary random signals, the optimal basis is the Karhunen-Loève transform:

$$\lambda_i p_i = R_{xx} p_i, \quad \mathbf{P}^T \mathbf{P} = \mathbf{I}$$

- Basis functions are the columns of  $\mathbf{P}$
- Minimum geometric mean of transform coefficient variances

14

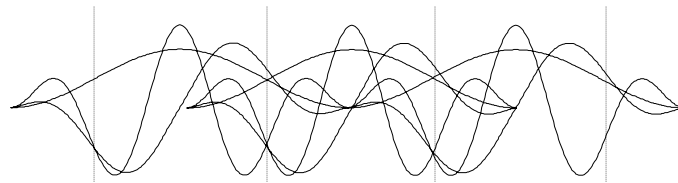
## Sub-optimal transforms

- KLT problems:
  - Signal dependency
  - **P** not factorable into sparse components
- Sinusoidal transforms:
  - Asymptotically optimal for large blocks
  - Frequency component interpretation
  - Sparse factors - e.g. FFT

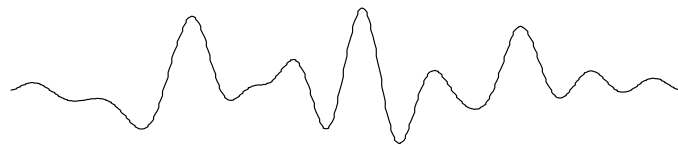
15

## Lapped transforms

- Basis functions have tails beyond block boundaries
  - Linear combinations of overlapping functions such as



- generate smooth signals, without blocking artifacts



16



## Modulated lapped transforms

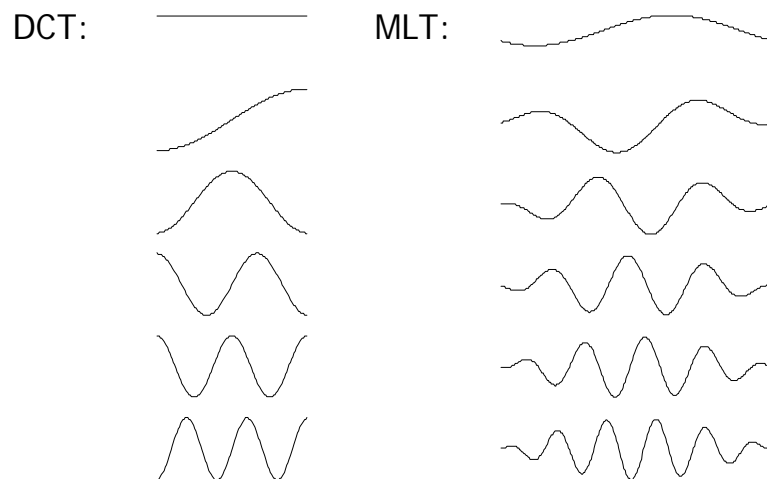
- Basis functions = cosines modulating the same low-pass (window) prototype  $h(n)$ :

$$p_k(n) = h(n) \sqrt{\frac{2}{M}} \cos \left[ \left( n + \frac{M+1}{2} \right) \left( k + \frac{1}{2} \right) \frac{\pi}{M} \right]$$

- Can be computed from the DCT or FFT
- Projection  $X = \mathbf{P}^T x$  can be computed in  $O(\log_2 M)$  operations per input point

17

## Basis functions

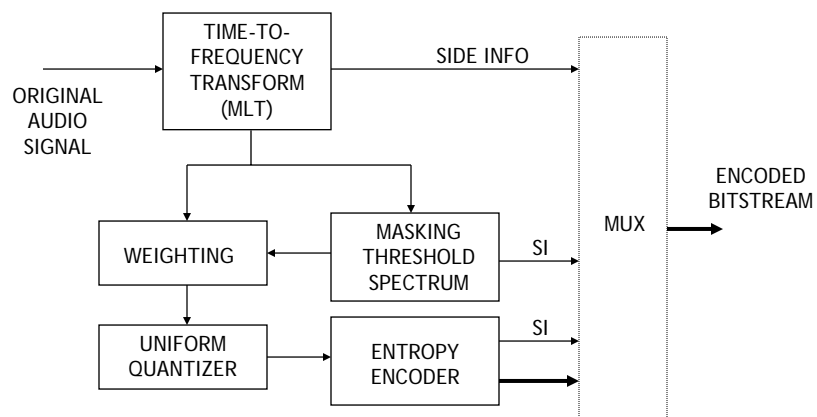


18

# Audio compression

19

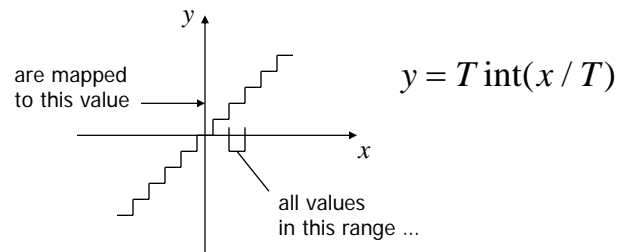
## Basic architecture



20

## Quantization of transform coefficients

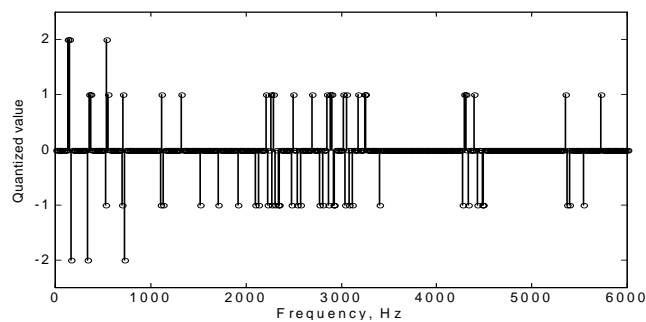
- Quantization = rounding to nearest integer.
- Small range of integer values = fewer bits needed to represent data
- Step size  $T$  controls range of integer values



21

## Encoding of quantized coefficients

- Typical plot of quantized transform coefficients



- Run-length + entropy coding

22

## Basic entropy coding

- Huffman coding: less frequent values have longer codewords



- More efficient if groups of values are assembled in a vector before coding

Value	Codeword
-7	'1010101010001'
-6	'10101010101'
-5	'101010100'
-4	'10101011'
-3	'101011'
-2	'1011'
-1	'01'
0	'11'
+1	'00'
+2	'100'
+3	'10100'
+4	'1010100'
+5	'1010101011'
+6	'101010101001'
+7	'1010101010000'

23

## Side information & more about EC

- SI: model of frequency spectrum
  - e.g. averages over subbands
- Quantized spectral model determines weighting
  - masking level used to scale coefficients
- Run-length + Vector Huffman works
  - Context-based AC can be better
  - Room for better context models via machine learning?
- Backward adaptation removes need for SI






24

## Other aspects

- Stereo coding
  - $(L+R)/2$  & L-R coding, expandable to multichannel
  - Intensity + balance coding
  - Mode switching – extra work for encoder only
- Encoding of mostly speech signals
  - May need to introduce source models, as in LPC
- Lossless coding
  - Easily achievable via lifting-based MLT

25

## WMA examples:

- Original clip (~1,400 kbps)      64 kbps (MP3)      64 kbps (WMA)  
            
- Original clip      WMA @ 32 kbps (Internet radio)  
      
- More examples at [windowsmedia.com](http://windowsmedia.com)

26