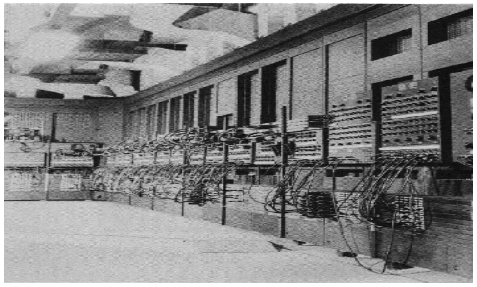
# ICME 2004 Tutorial: Audio Feature Extraction

George Tzanetakis
Assistant Professor
Computer Science Department
University of Victoria, Canada

gtzan@cs.uvic.ca
http://www.cs.uvic.ca/~gtzan

# Bits of the history of bits



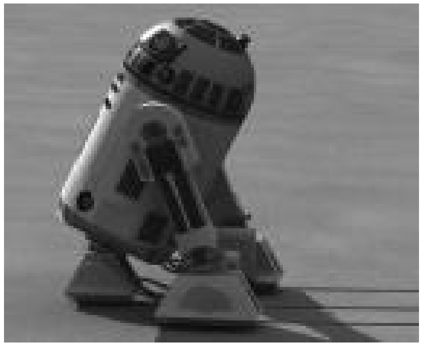
01011110101010

Hello world

Web

Multimedia

Understanding multimedia content ->



# Tutorial Goals

- Overview of state of the art
- Fundamentals
- Technical Background
  - Some math, computer science, music
- Shift emphasis from audio coding/compression to audio analysis
- There is more to audio analysis than MFCCs

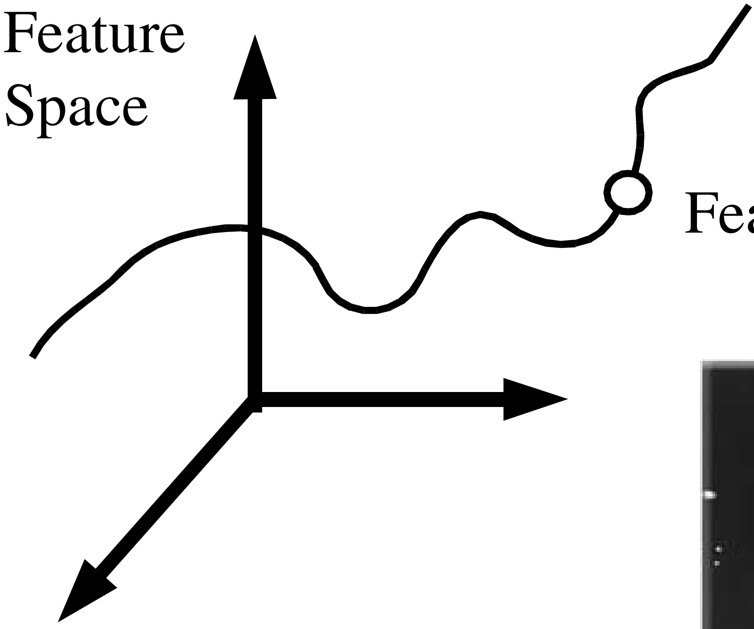
# Some simple but important observations

- Analysis/Understanding require multiple representations – no “best” one
- Coding/Compression/Processing typically search for the “best” “optimal” way to do things
- Paradigm shift is necessary to make multimedia more than just lots of numbers
- !!! MACHINE LEARNING

# My background – MIR

- Database of all recorded music
- Tasks: organize, search, retrieve, classify recommend, browse, listen, annotate
- Examples:

# Feature extraction



# Outline

| | |
|---|---|
| ➢ Introduction | 10 min |
| ➢ Signal Processing | 30 min |
| ➢ Source-based | 20 min |
| ➢ Perception-based | 30 min |
| ➢ Music-specific | 20 min |
| ➢ Fingerprinting and watermarking | 25 min |
| ➢ Sound Separation and CASA | 25 min |
| ➢ Future work and challenges | 20 min |

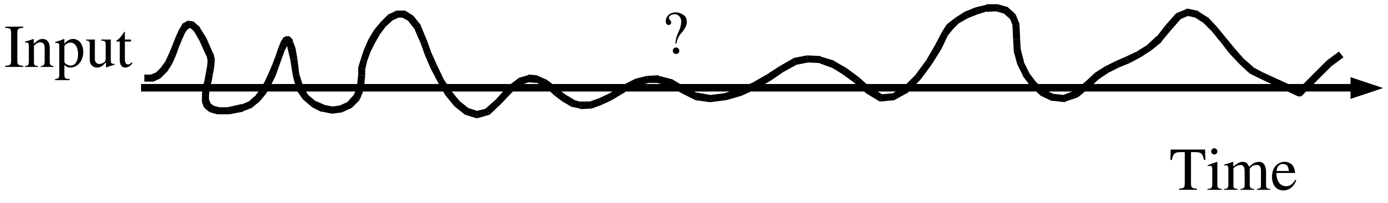
# Signal Processing

- Sound and Sine Waves
- Short Time Fourier Transform
- Discrete Wavelet Transform
- Fundamental Frequency Detection

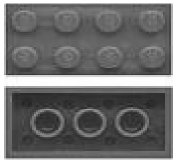
# Understanding Sound

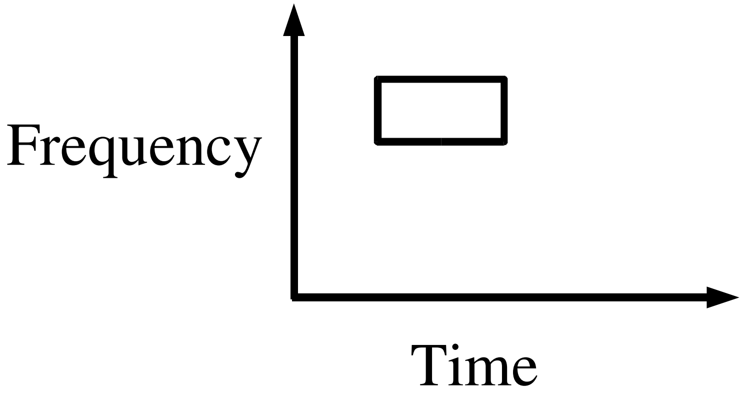
- Longitudinal wave – pulsating expanding sphere
- 344 m (1128 feet) / second (at 20 Celcius)
- Reflections
- Sound production, propagation and perception are to a certain degree linear phenomena

# Time-domain waveform

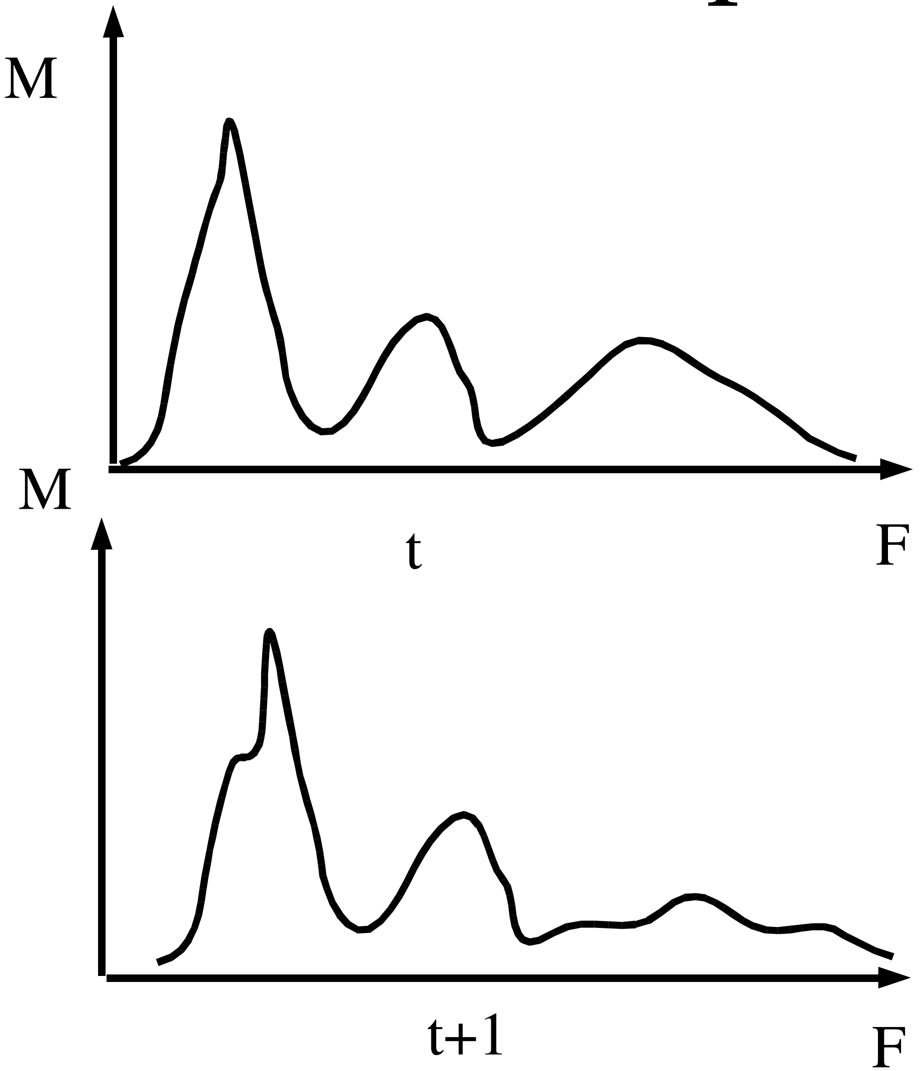


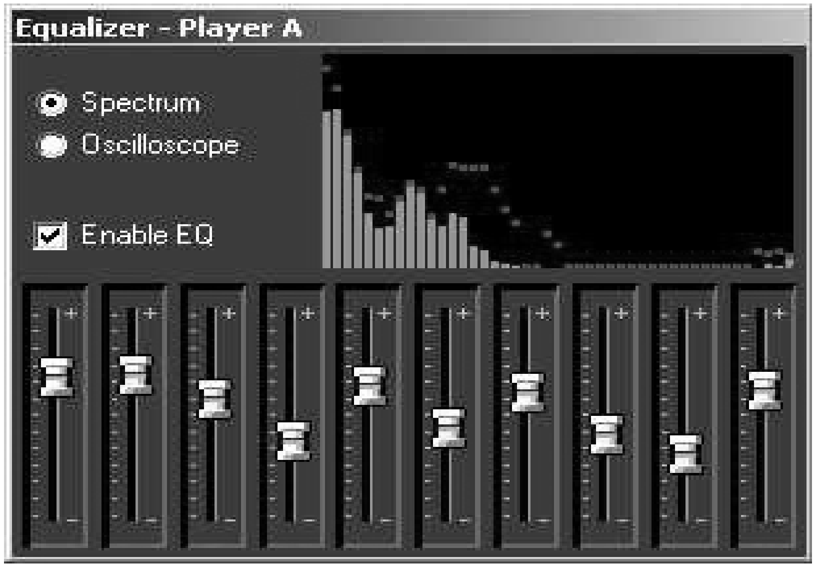
Decompose to building blocks that are created in a regular fashion





# Spectrum




Equalizer - Player A

● Spectrum

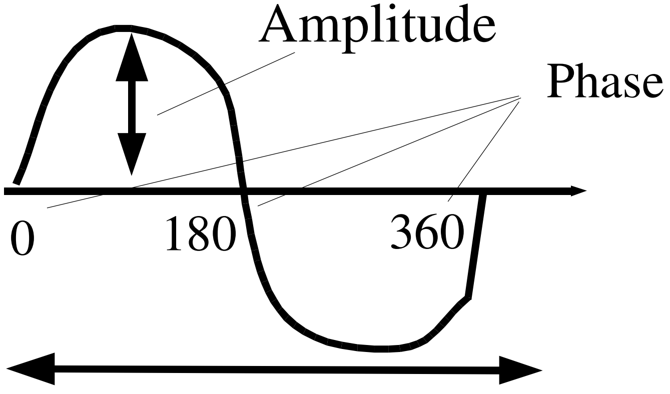
○ Oscilloscope

☑ Enable EQ

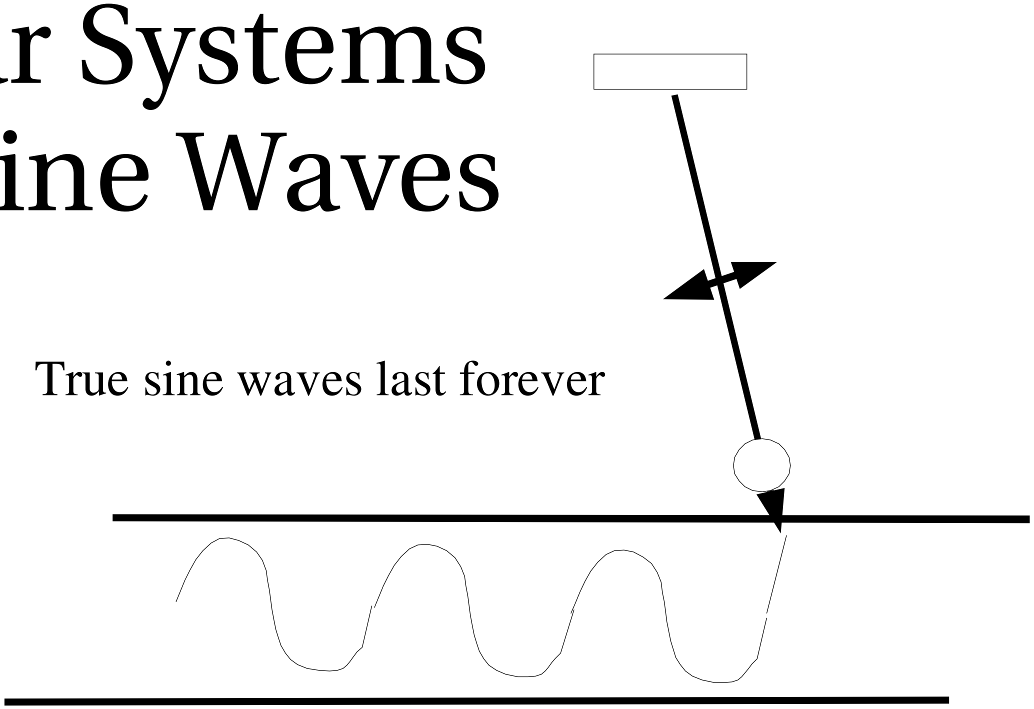
Recipe for how to combine the building blocks to form the signal

Different view of the same information

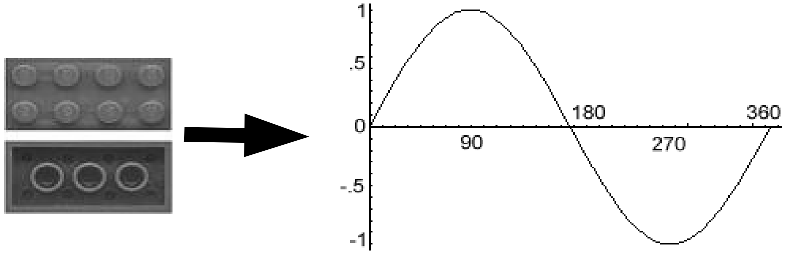
# Linear Systems and Sine Waves



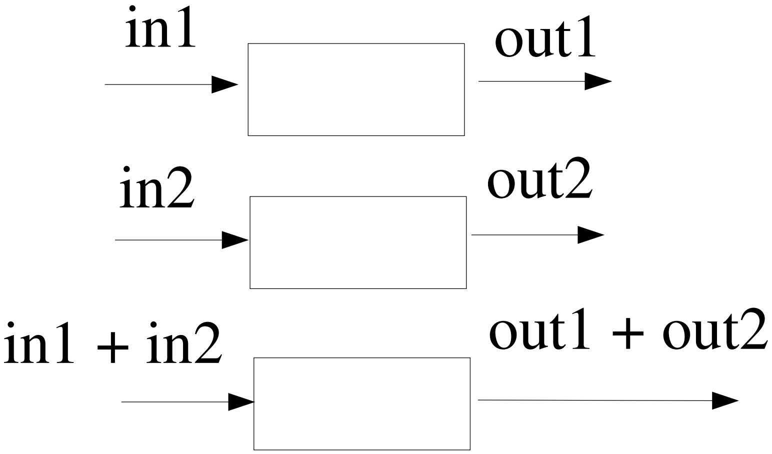
Period = 1 / Frequency



True sine waves last forever



sine wave -> LTI -> new sine wave



in1

out1

in2

out2

in1 + in2

out1 + out2

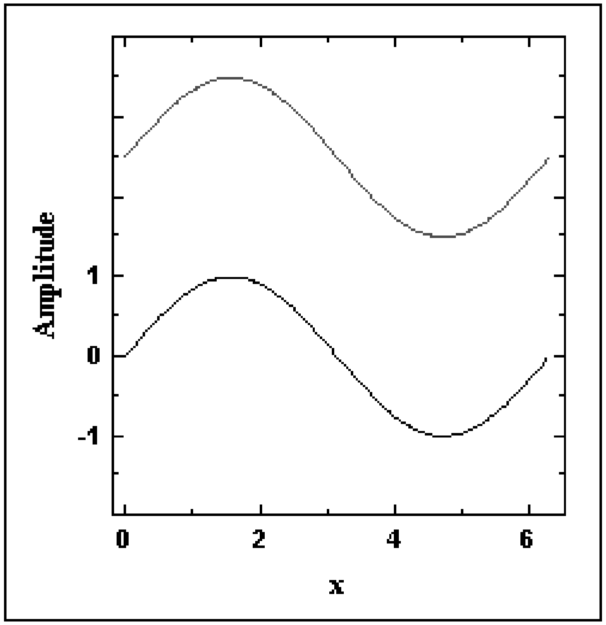
# Time-Frequency Analysis Fourier Transform

$$f \quad x \quad = \sum_{n=0}^{\infty} a_n \cos \quad n * x \quad + \sum_{n=0}^{\infty} b_n \sin \quad n * x$$

$$f \quad t \quad = \frac{1}{2\pi} \int_{-\infty}^{\infty} f \quad \omega \quad \mathrm{e}^{-i\omega t} dt$$

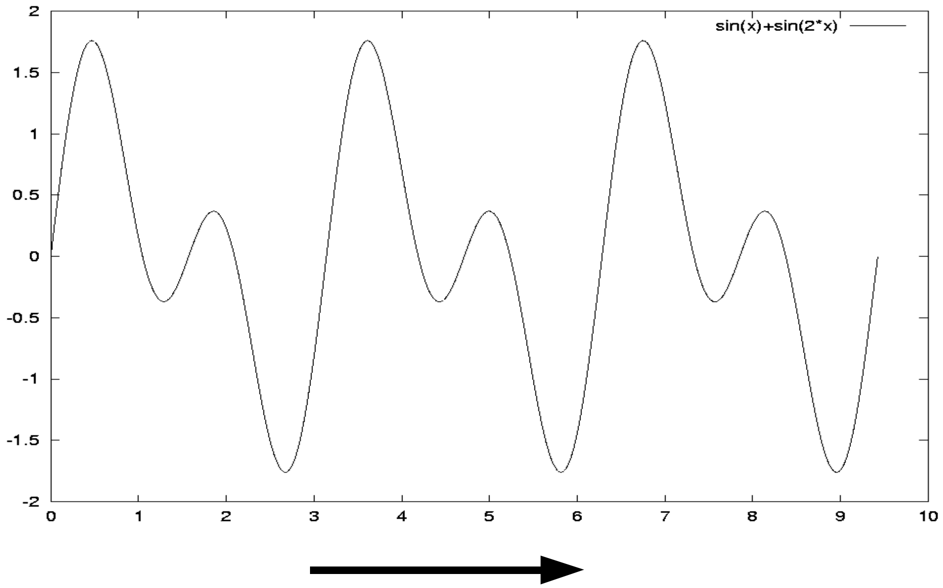
$$f \quad \omega \quad = \int_{-\infty}^{\infty} f \quad t \quad \mathrm{e}^{i\omega t} dt$$

$$\mathrm{e}^{i\theta} = \cos \quad \theta \quad + i * \sin \quad \theta$$



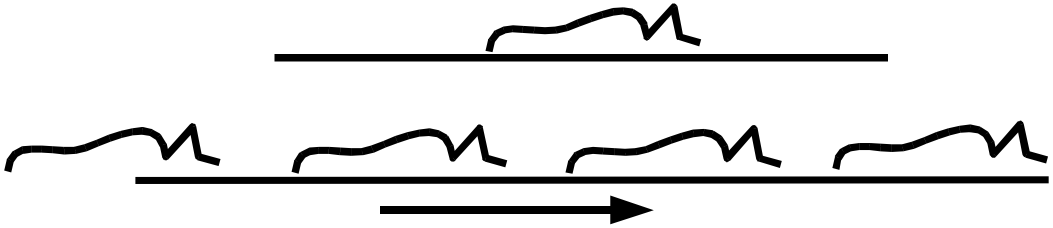
Any periodic waveform can be approximated by a number of sinusoidal components that are harmonics (integer multiples) of a fundamental frequency

# Non-periodic signals

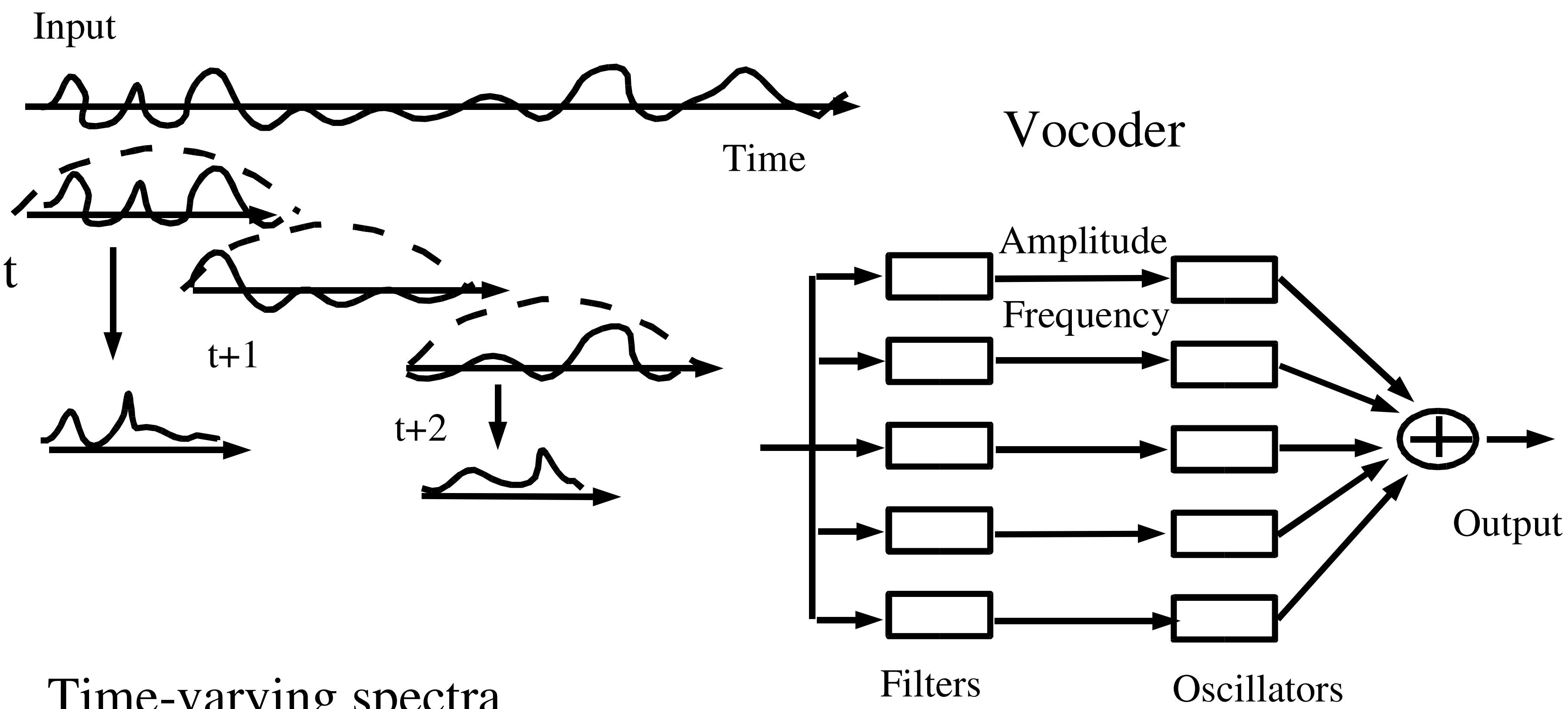


P=1/f=2π/ω

We force them to be periodic by repeating them to infinity



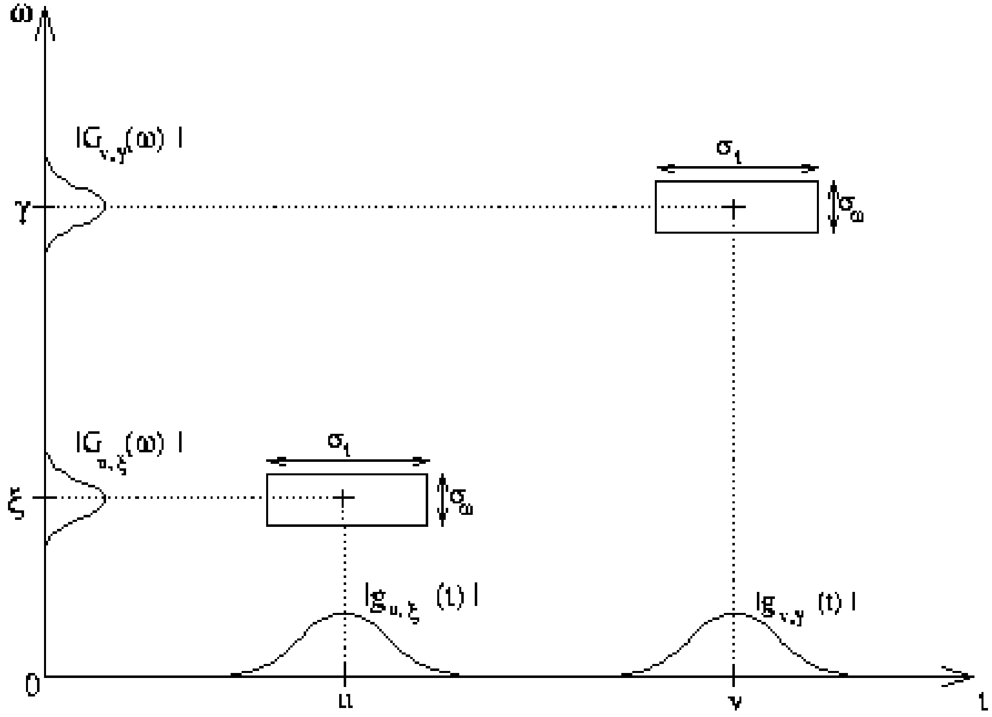
# Short Time Fourier Transform



Time-varying spectra
Fast Fourier Transform FFT

# Short Time
# Fourier Transform II

FT = global representation of frequency content



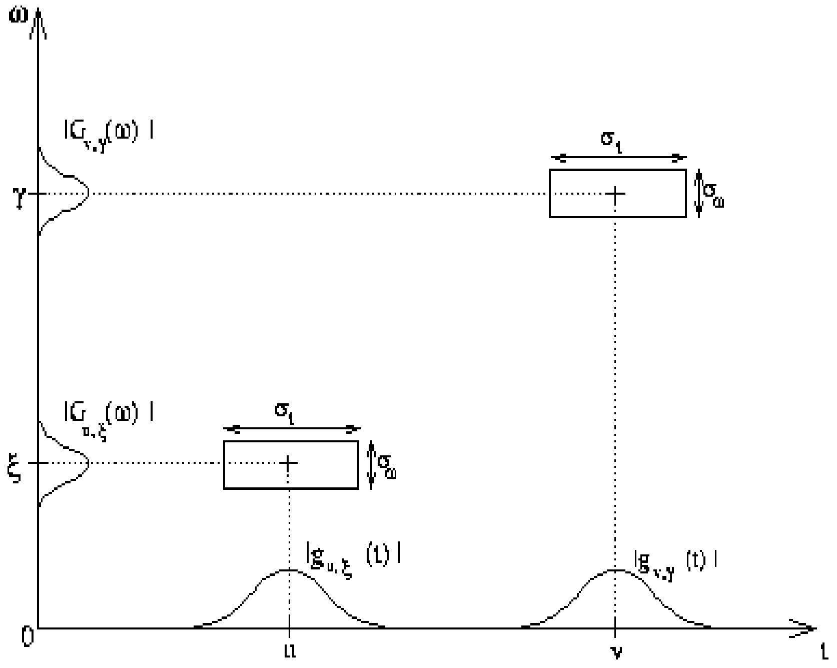
$$\boxed{Sf(u,\omega) = \int_{-\infty}^{\infty} f(t)\, g(t-u)\, e^{-i\omega t}\, dt}$$

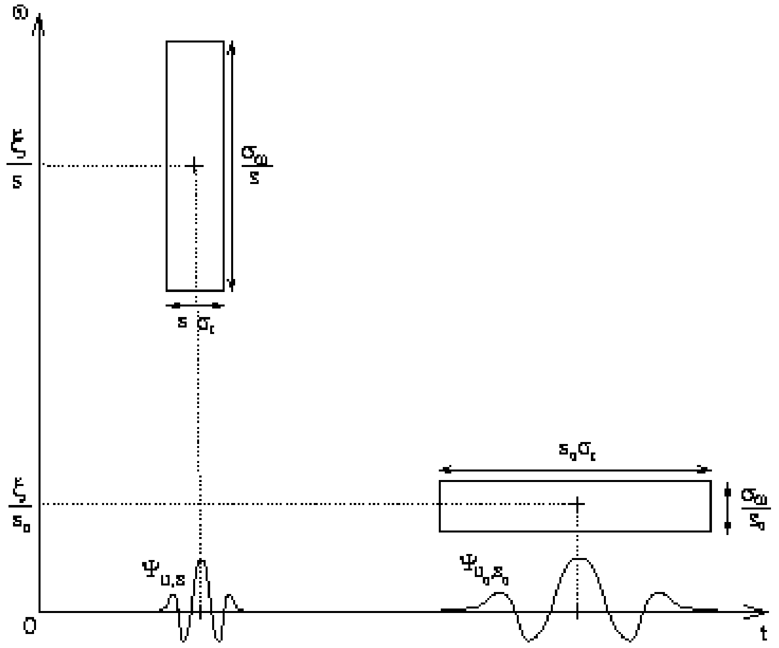
Time – Frequency

L2 Heisenberg uncertainty

$$\boxed{\sigma_t \sigma_\omega \geq 1/4}$$

# STFT- Wavelets

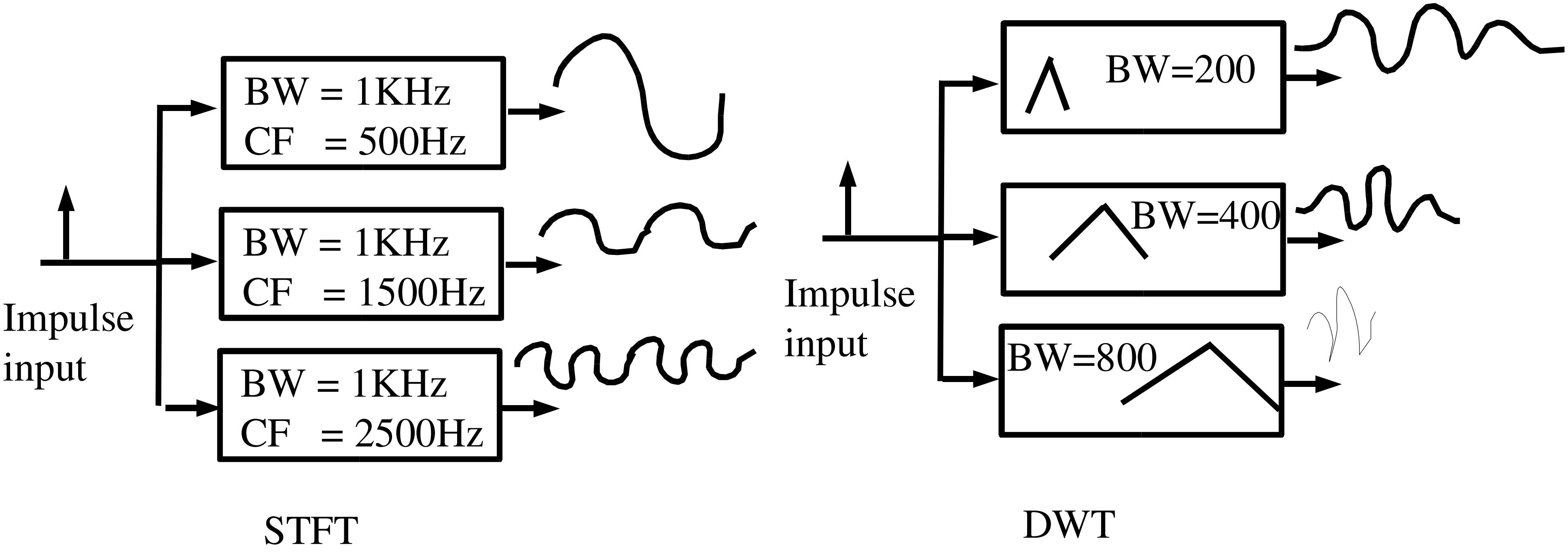




Time – Frequency Heisenberg uncertainty

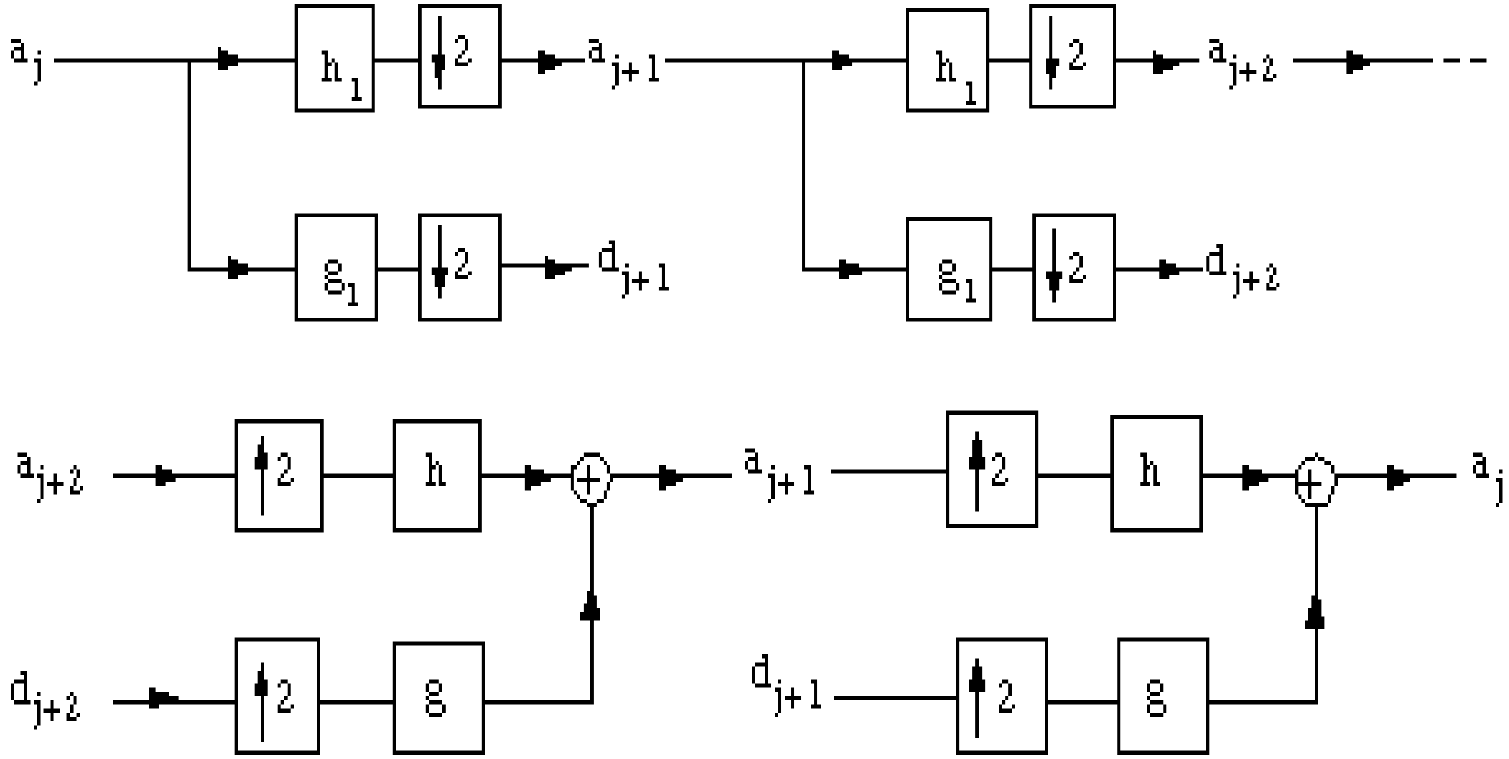
$$\boxed{\sigma_t \sigma_\omega \geq 1/4}$$

# A filterbank view of STFT and DWT



# The Discrete Wavelet Transform

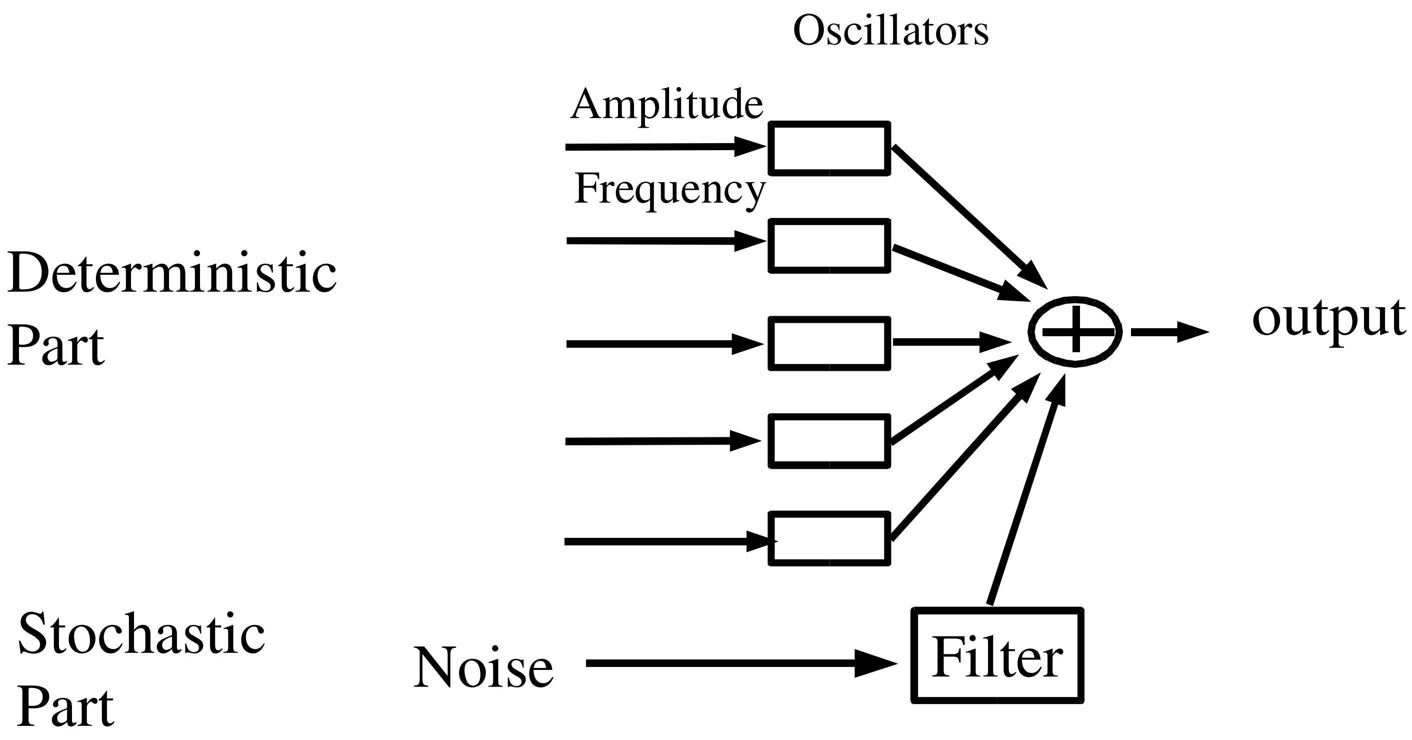
Octave filterbank



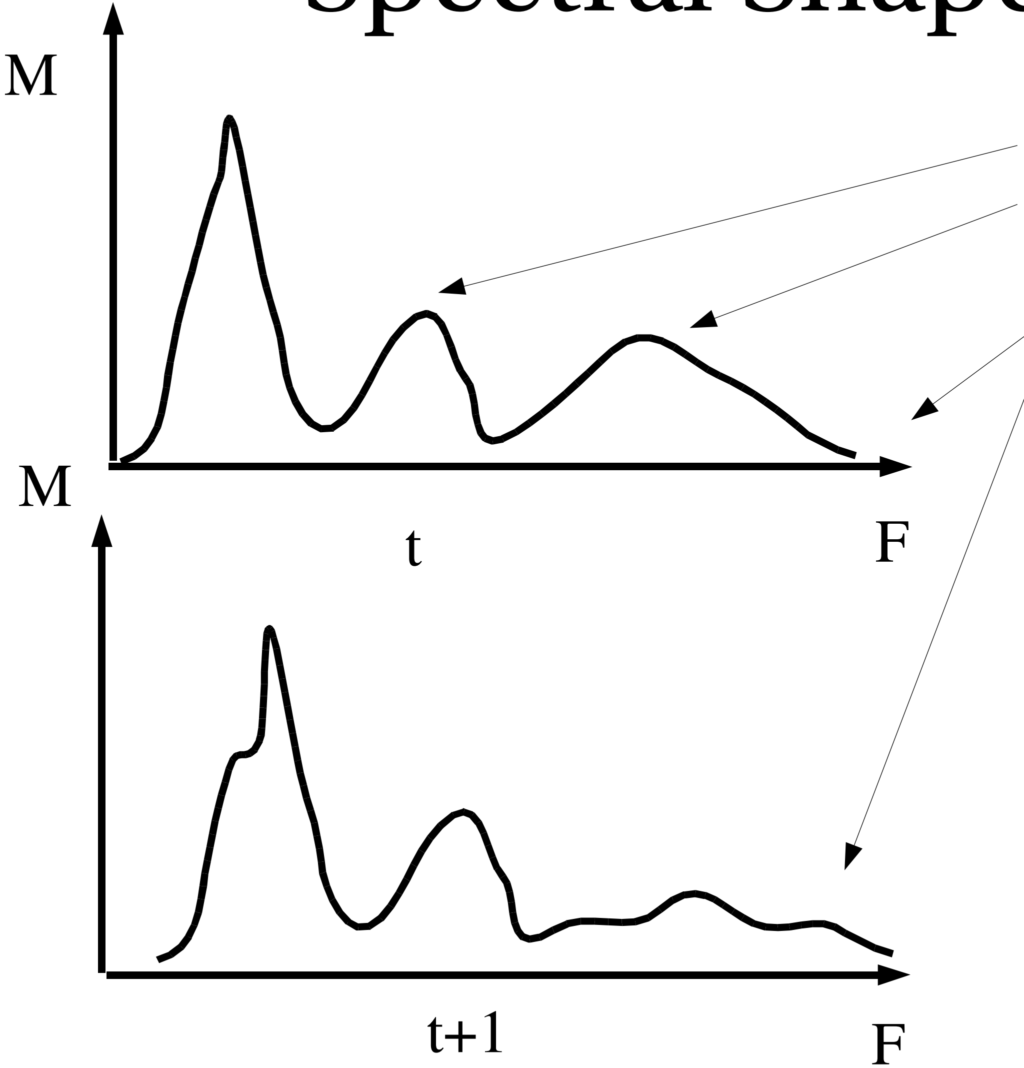
Analysis

Synthesis

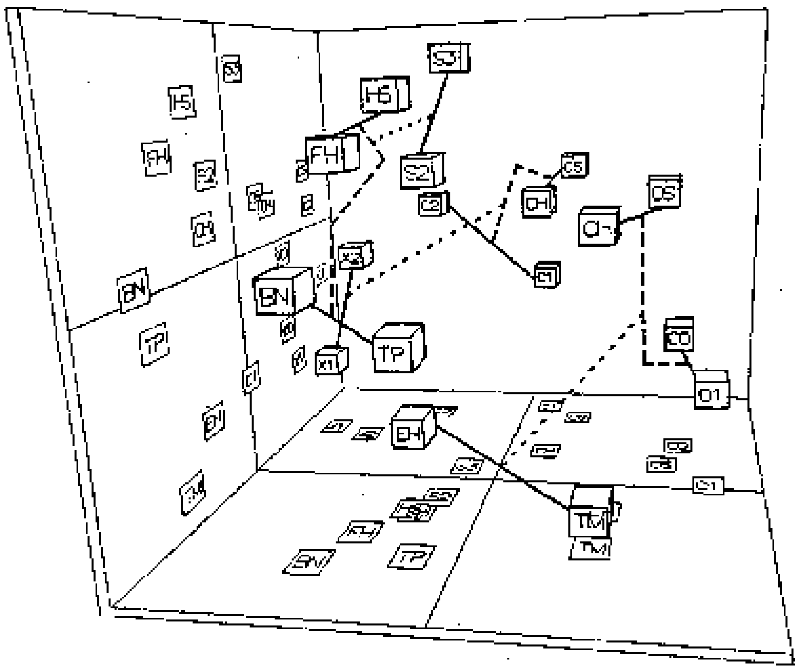
# Sinusoids + noise modeling



# Spectral Shape Features

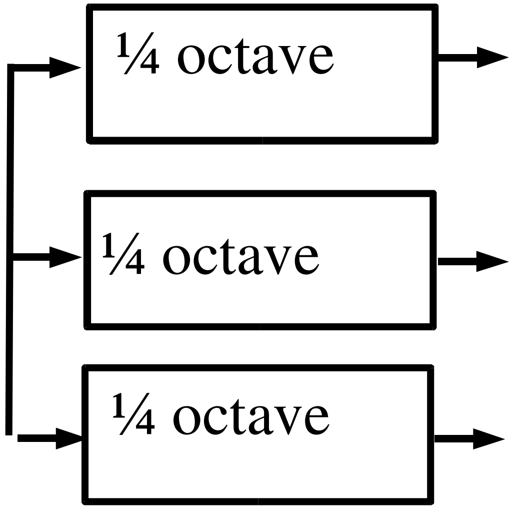


Centroid = center of gravity of spectrum
Rolloff = energy distribution low/high
Flux = short time change



Grey's timbrespace (1975)

# Spectral Flatness

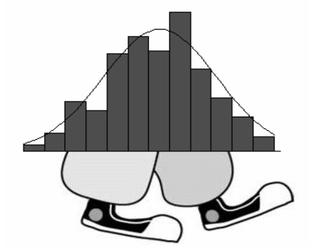


ratio of geometric mean to arithmetic mean of spectral power (presence of tonal component in band)

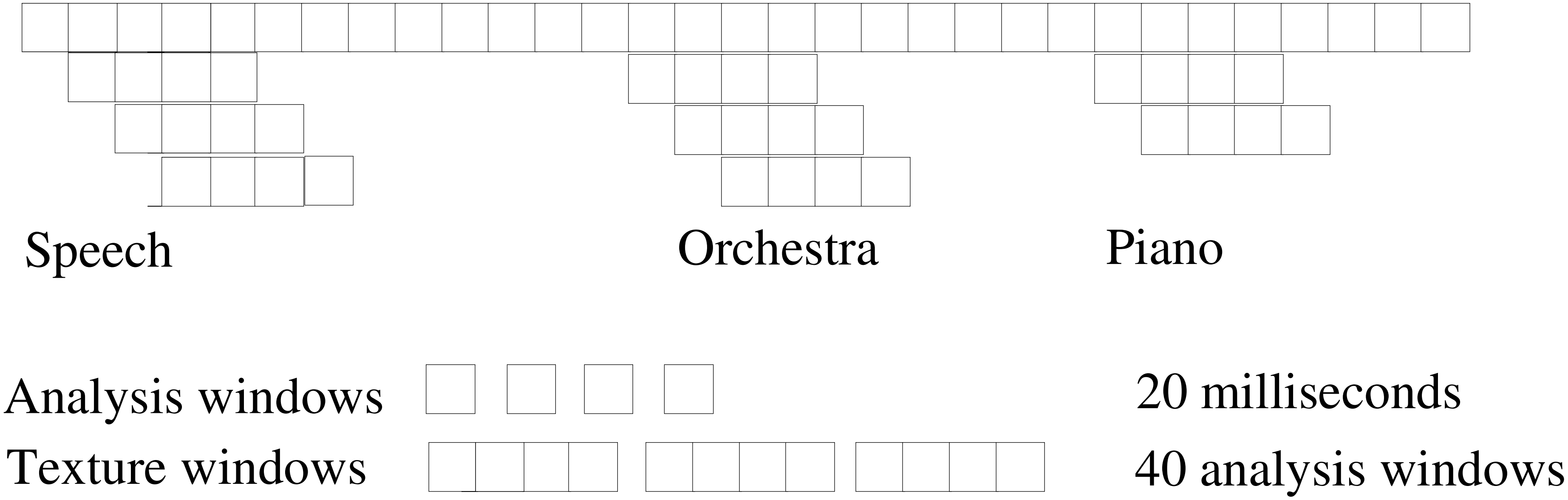
Logarithmically-spaced overlapping frequency bands

Tonality coefficients = how much tone vs noise is that particular band

# Analysis and Texture Windows



Running multidimensional Gaussian distribution (means, variances over texture window)



Speech

Orchestra

Piano

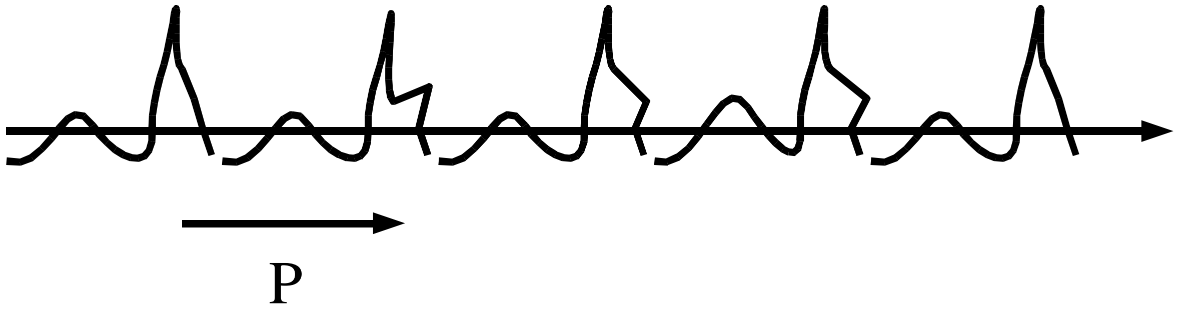
Analysis windows

20 milliseconds

Texture windows

40 analysis windows

# Fundamental Frequency Detection



Time-domain
Frequency-domain
Perceptual

Autocorrelation
Peaks at multiple of
the fundamental frequency

$$r_x = \sum_{n=0}^{N-1} x \;\; n \;\; x \;\; n+l \;\; , l = 0,1,.. \; L-1$$

ZeroCrossings

Rhythm -> ~20 Hz Pitch
(created by Roger Dannenberg)

# Demos I

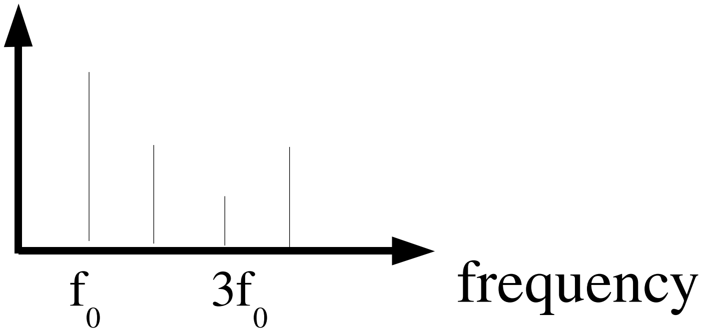
- Phase vocoder
- Spectrograms – time frequency tradeoffs
- Wavelet decomposition

# Source-based approaches

- Linear Prediction
- CELP, GSM
- Isolated tone musical instrument recognition

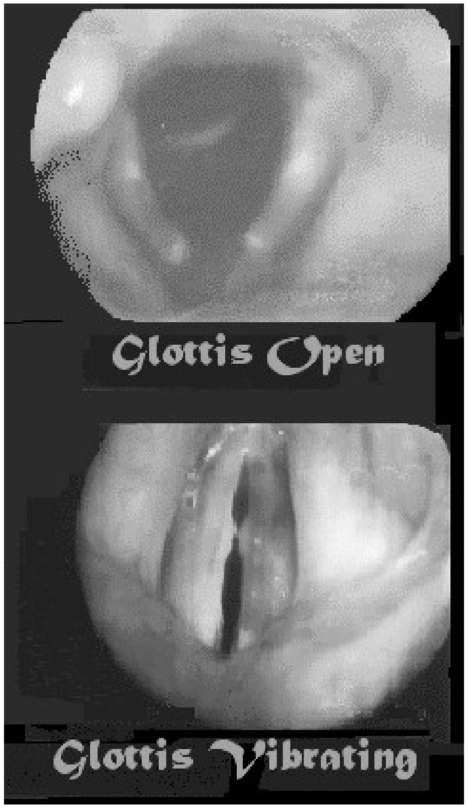
# Harmonic Partials

- Instruments and the voice are harmonic oscillators (solution to pdf)
- Partials (peaks in the spectrum)
- Harmonic Partials are integer multiples of the first partial or fundamental frequency



Helmholz – timbre is based on the relative weights of the harmonics

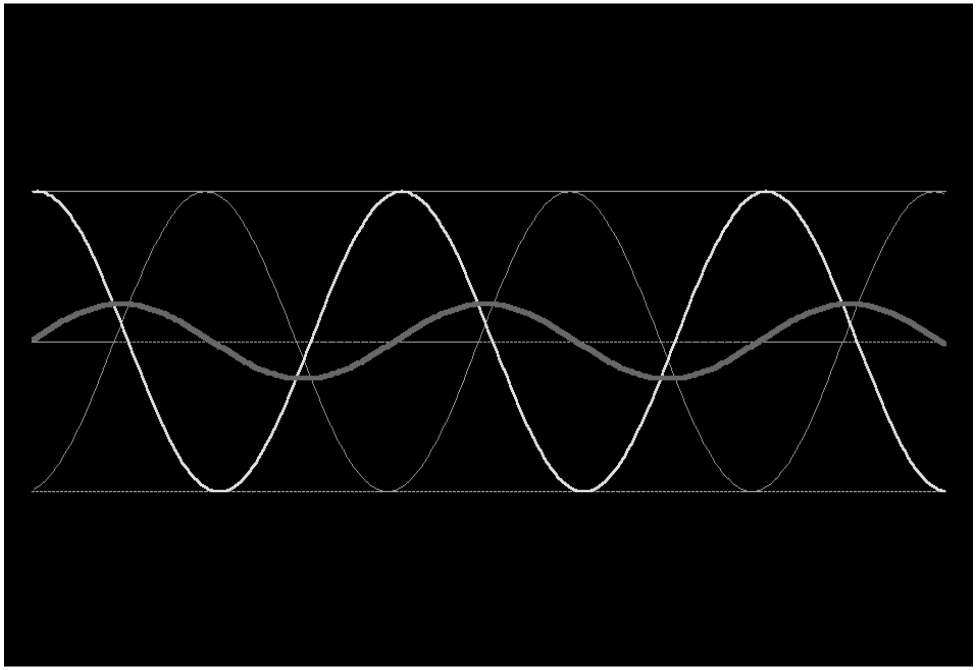
# Voice production



- Vocal Folds
  - breath pressure from lungs causes the folds to oscillate
  - oscillator driven by breath pressure acts as excitation to the vocal folds
- Vocal Tract
  - tube(s) of varying cross-section exhibiting modes of vibration (resonances)
  - resonances “shape” the excitation

# Formant Peaks - Resonances

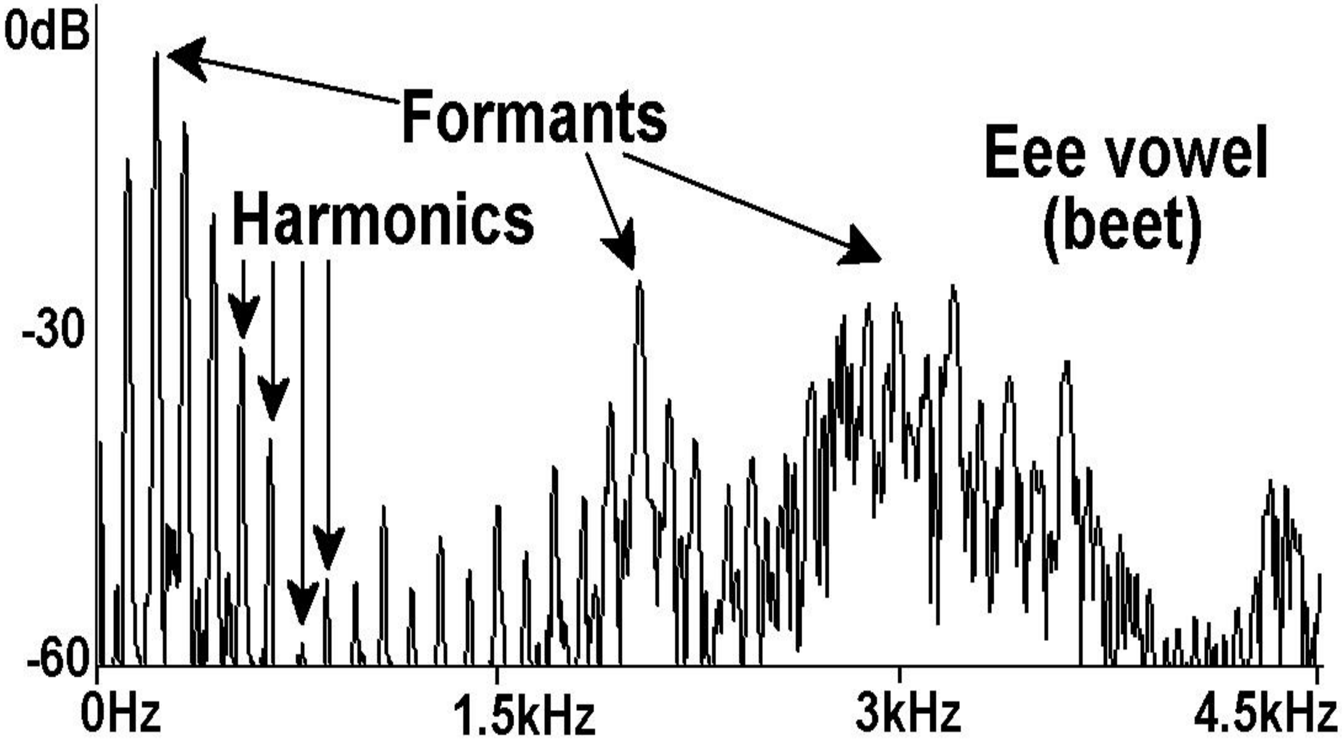
Modes/resonances are the result of standing waves constructive interference – boosted regions of frequency



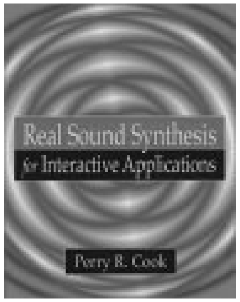
Vocal tract is essentially a tube that is closed at the vocal fold and open at the lips

modes = odd-multiples of ¼ cycle of a sine wave  (F1 = c/l/4) l=9in
375 Hz, 1125 Hz, 2075 Hz

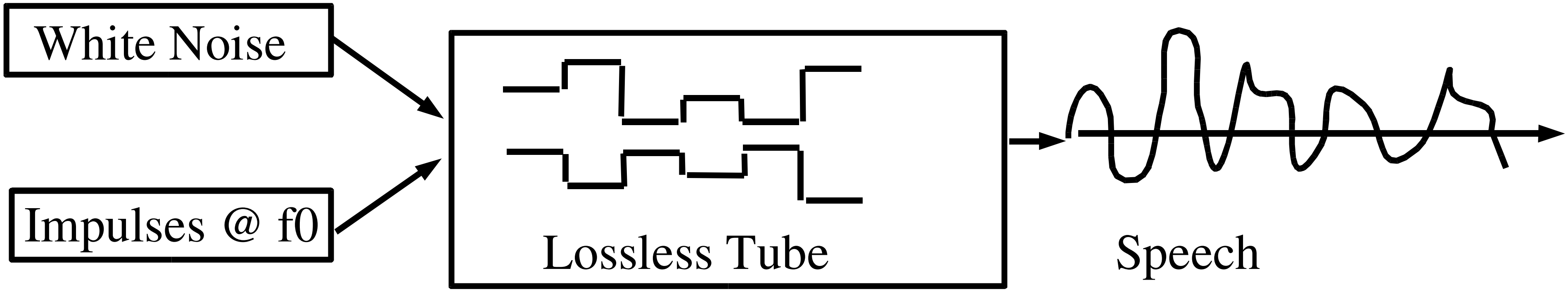
# Formants



From “Real Sound Synthesis for Interactive Applications”
P. Cook, A.K Peters Press, used by permission



# Linear Prediction Coefficients



Source

$$s'_n = \sum_{i=1}^{p} a_i s_{n-1}$$

Filter

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

Original

Resynthesized with impulses/noise

# Variations

- Perceptual Linear Prediction
- RASTA – Relative Spectral Transform -Perceptual Linear Prediction
  - Take advantage of HAS characteristics
- CELP (Code Excited Linear Prediction)
  - better modeling of excitation
- GSM

# CELP

- Problems with LPC
  - tube is not one tube but two (nose)
  - buzz is not buzz
  - everything goes into residue
- Codebook Excitation
  - table of typical residue signals
  - one fixed
  - one adaptive

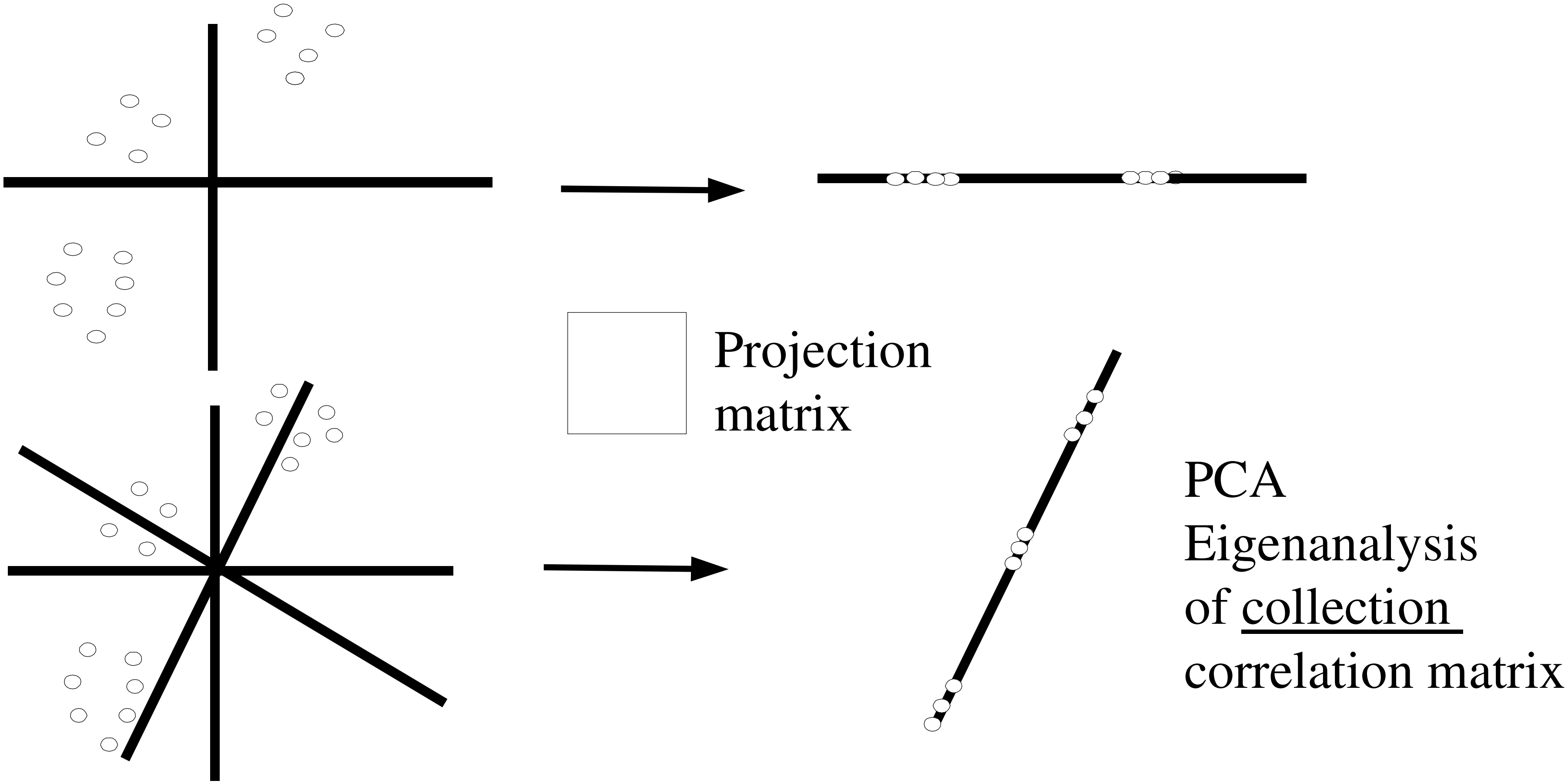
# Isolated Tone Instrument Classification

- Important step for music transcription
- Hierarchical classification
  - Family: bowed, wind etc
  - Instrument: violin, flute, piano etc
- Spectral
- Temporal
  - temporal centroid, onset time

# MPEG-7
# Audio Descriptors

- Low-Level Audio Descriptors
  - Waveform, Spectral
  - Spectral Timbral (centroid, spread)
  - Temporal Timbral (temporal cntrd, log-attack)
- High-Level Description Tools
  - Sound recognition and indexing
  - Spoken content
  - Musical instrument timbre
  - Melody description

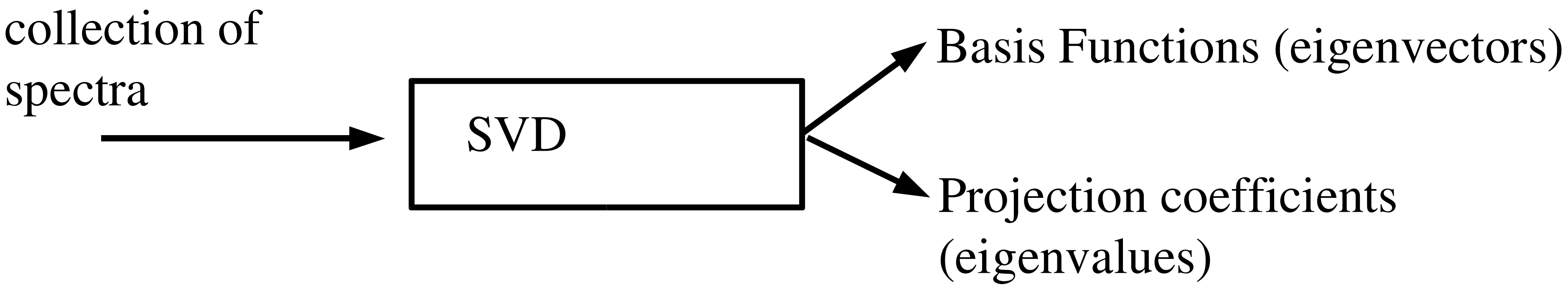
# Principal Component Analysis



Projection matrix

PCA Eigenanalysis of <u>collection</u> correlation matrix

# MPEG-7 Spectral Basis Functions



collection of spectra

SVD

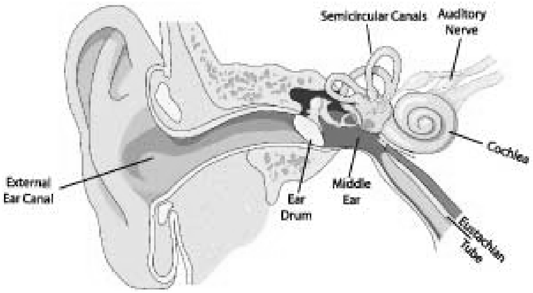
Basis Functions (eigenvectors)

Projection coefficients (eigenvalues)

typical: 70% of original 32-dimensional data is captured by 4 sets of basis functions and projection coefficients

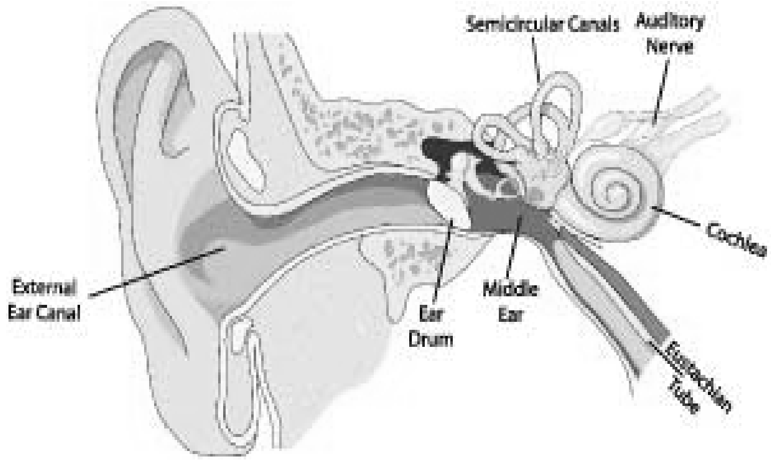
Each spectrum can be expressed as a linear combination of the basis

# Perception-based approaches

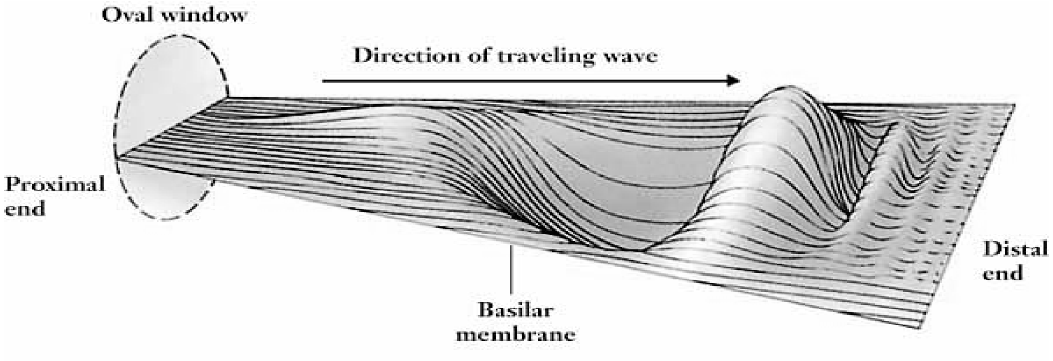


- Pitch perception
- Loudness percetion
- Critical Bands
- Mel-Frequency Cepstral Coefficients
- Masking
- Perceptual Audio Compression (MPEG)

# The Human Ear



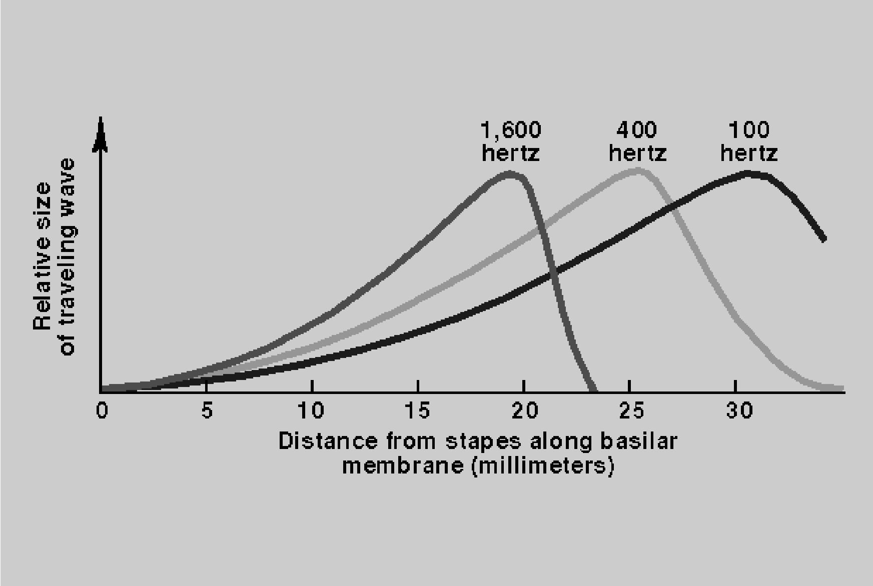
Pinna
Auditory canal
Ear Drum
Stapes-Malleus-Incus (gain control)
Cochlea (freq. analysis)
Auditory Nerve ?



Wave travels to cutoff slowing down increasing in amplitude power is absorbed

Each frequency has a position of maximum displacement

# Masking



Two frequencies -> beats
-> harsh
-> seperate

Inner Hair Cell excitation

Frequency Masking
Temporal Masking

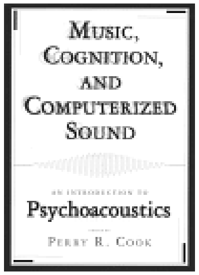
Pairs of sine waves (one softer) – how much weaker in order to be masked ? (masking curves) wave of high frequency can not mask a wave of lower frequency

# Masking Demo

High sine waves mask low: 500 Hz tone at 0dB with lower tones at -40dB, 300, 320, 340, 360, 380, 400, 420, 440, 460, 480 Hz

Low sine waves mask hih: 500Hz tone at 0dB with higher tones at -40dB, 1700, 1580, 1460, 1340, 1220, 1100, 980, 860, 740, 620 Hz

From “Music, Cognition and Computerized Sound”
P. Cook (Editor) MIT Press



Music, Cognition, and Computerized Sound

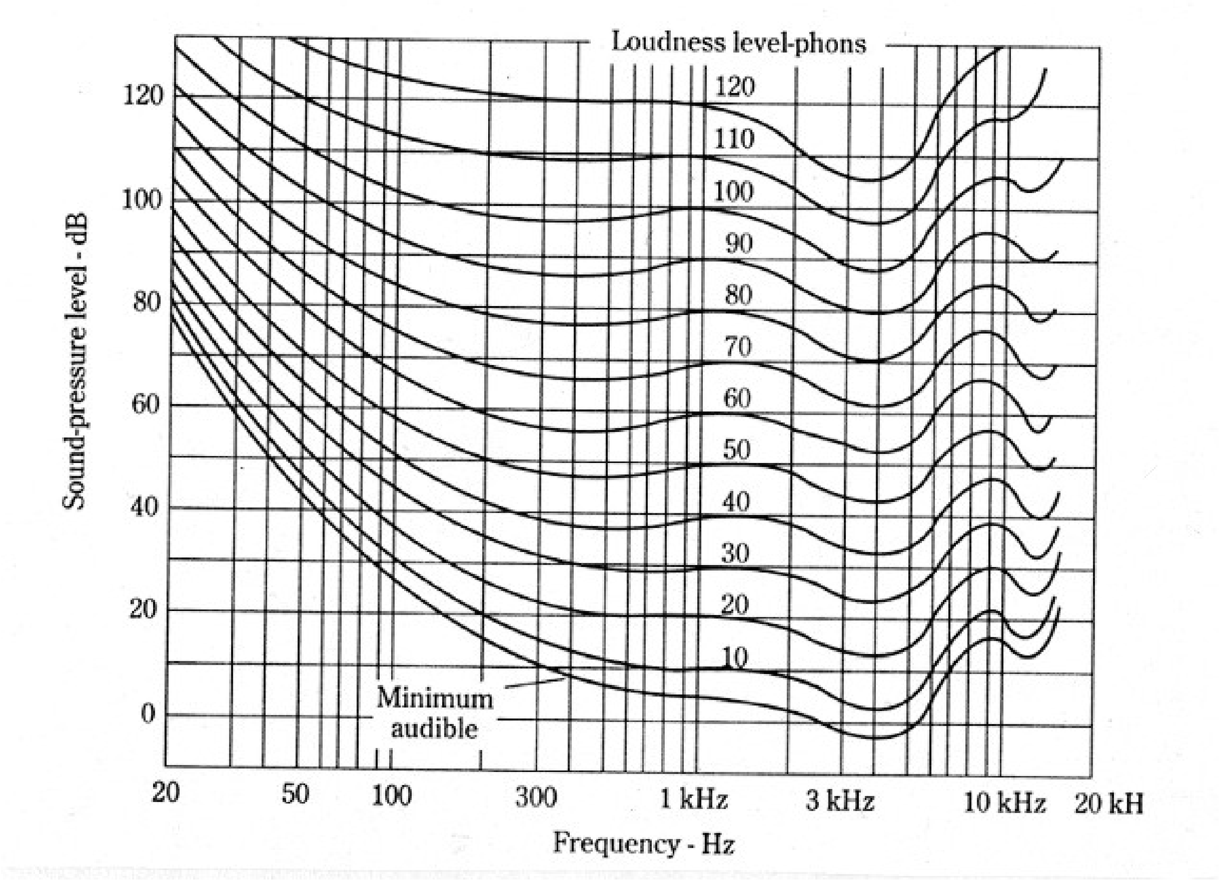
An Introduction to Psychoacoustics

Perry R. Cook

# Critical Bands

- Critical bandwidth = two sinusoidal signals interact or mask one another
- Bark scale  (24 critical bands)
  - [0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700]
  - samplings of a continuous variation in the frequency response of the ear to a sinusoid or narrow band process
  - there is no discrete filterbank in the ear

# Fletcher-Munson Curves



Loudness is a perceptual (not physical) quantity i.e two sound with same SPL different frequencies are perceived to have different loudness

(used in PLP)

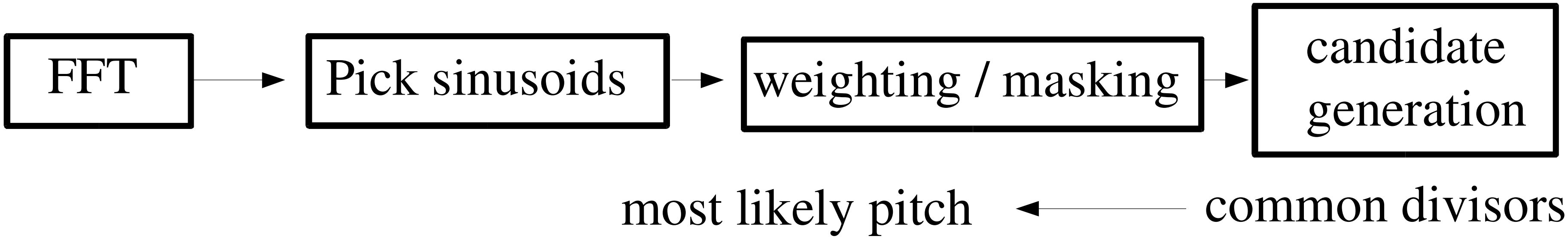
for a soft sound at 50Hz
to sound as loud as one
at 2000 Hz
50dB more intense
(100,000 times more power)

# Pitch Perception I

- Pitch is not just fundamental frequency
- Periodicity or harmonicity or both ?
- Human judgements (adjust sine method)
- 1924 Fletcher – harmonic partials missing fundamental (pitch is still heard)
  - Examples: phone, small radio
- Terhardt (1972), Licklider (1959)
  - duplex theory of pitch (virtual & spectral pitch)

# Pitch Perception II

- One perception – two overlapping mechanisms
  - Counting cycles of period < 800Hz
  - Place of excitation along basilar membrane > 1600 Hz



FFT

Pick sinusoids

weighting / masking

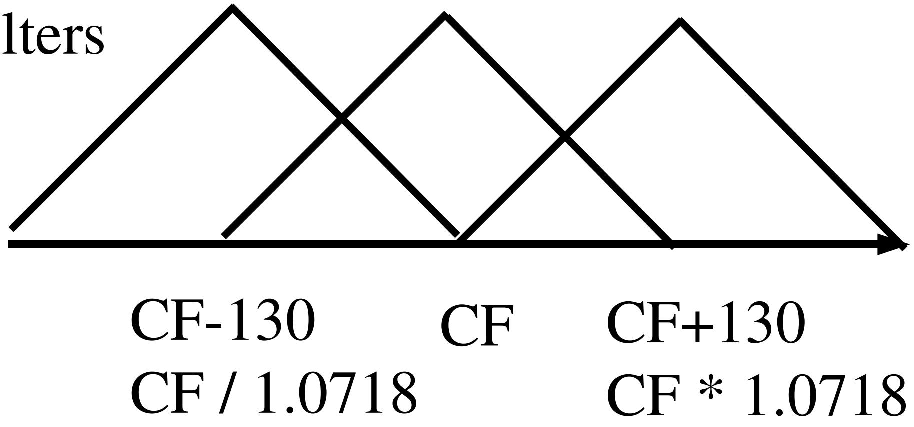
candidate generation

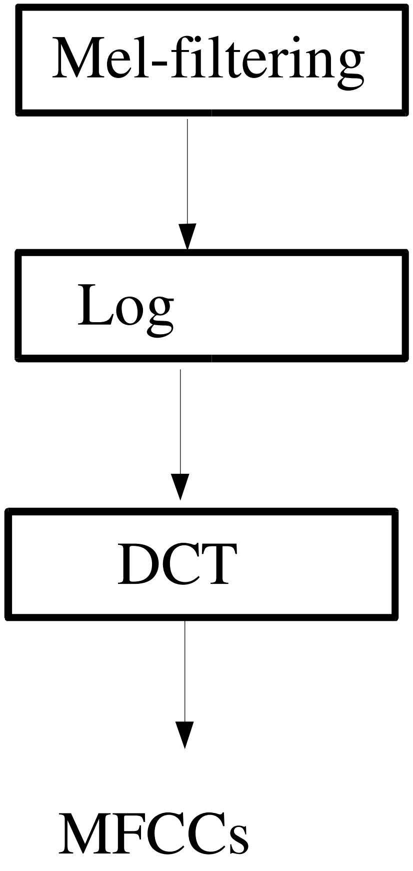
common divisors

most likely pitch

# Mel Frequency Cepstral Coefficients

Mel-scale
13 linearly-spaced filters
27 log-spaced filters





# Cepstrum

Measure of periodicity of frequency response plot

$S(e^{j\theta}) = H(e^{j\theta})\ E\ (e^{\,j\theta})$ H is linear filter, E is excitation

$\log(|S(e^{j\theta})|) = \log\ (|H(e^{j\theta})|) + \log(|E\ (e^{\,j\theta})|)$

(homomorphic transformation – the convolution of two signals becomes equivalent to the sum of their cepstra)

Aims to deconvolve the signal (low order coefficients filter shape – high order coefficients excitation with possible F0)
Cepstral coefficients can also be derived from LPC analysis

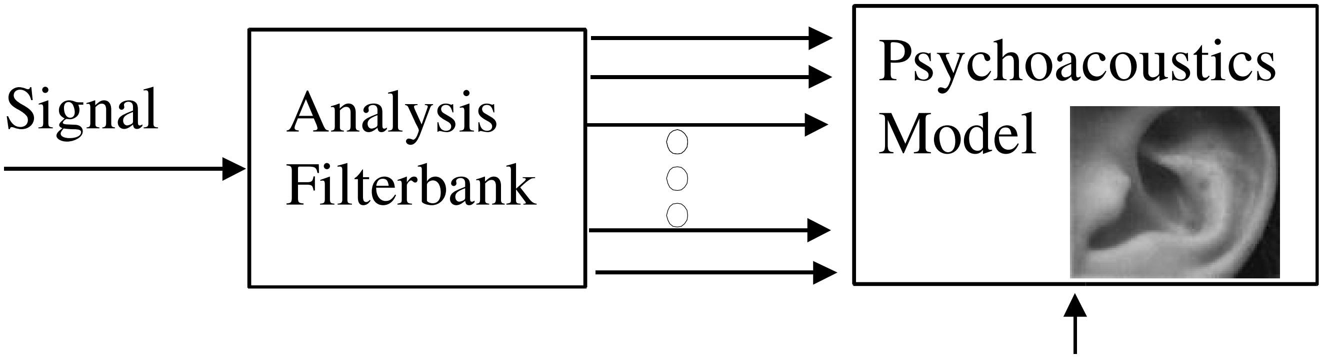
# Discrete Cosine Transform

- Strong energy compaction
- For certain types of signals approximates KL transform (optimal)
- Low coefficients represent most of the signal
- Can throw high coefficients
- MFCCs keep first 13-20
- MDCT (overlap-based) used in MP3, AAC, Vorbis audio compression

# Short MPEG Audio Coding Overview (mp3)

MPEG Perceptual Audio Coding

32-linearly spaced bands



Signal

Analysis Filterbank

Psychoacoustics Model

□ □ □
□ □
□ □ □ □ □

□ □ □ □ □
□ □ □ □ □ □ □ □

□ □ □ □ □ □ □ □
□ □ □ □ □ □ □ □
□ □ □ □ □ □ □ □

available bits

Encoder: slow, complicated
Decoder: fast, simple

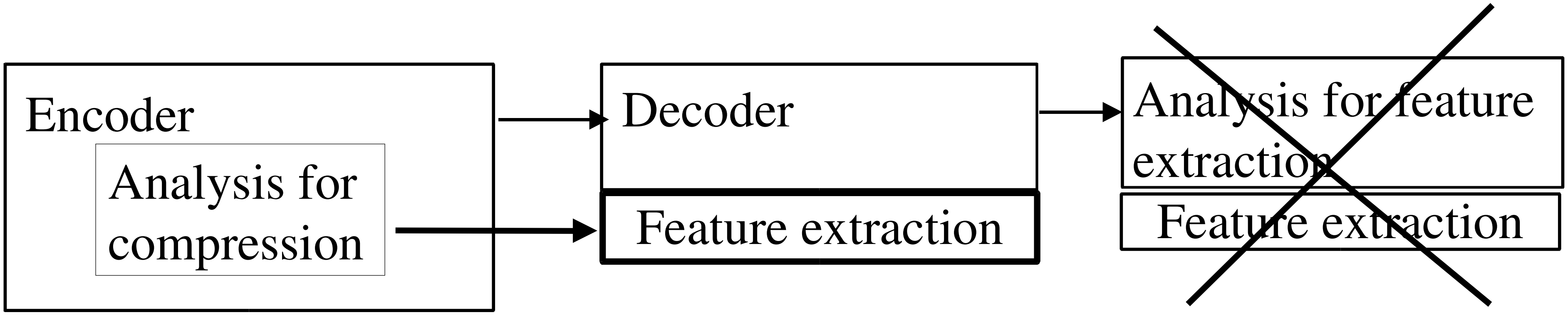
# Psychoacoustic Model

- Each band is quantized
- Quantization introduces noise
- Adapt the quantization so that it is inaudible
- Take advantage of masking
  - Hide quantization noise where it is masked
- MPEG standarizes how the quantized bits are transmitted not the psychoacoustic model – (only recommended)

# MP3 Feature Extraction

Pye ICASSP 00
Tzanetakis & Cook ICASSP 00

- Feature extraction while decoding MPEG audio compressed data (mp3 files)
- Free analysis for encoding
- Space and time savings



Encoder

Analysis for compression

Decoder

Feature extraction

~~Analysis for feature extraction~~

~~Feature extraction~~

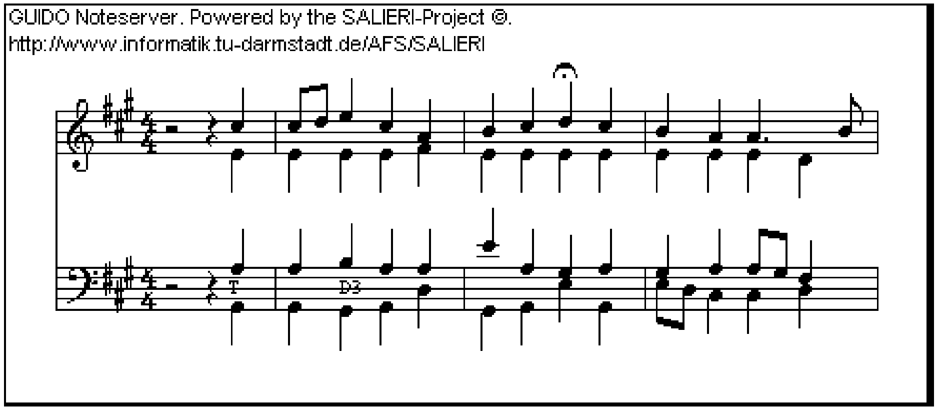
# Music-specific Audio Features

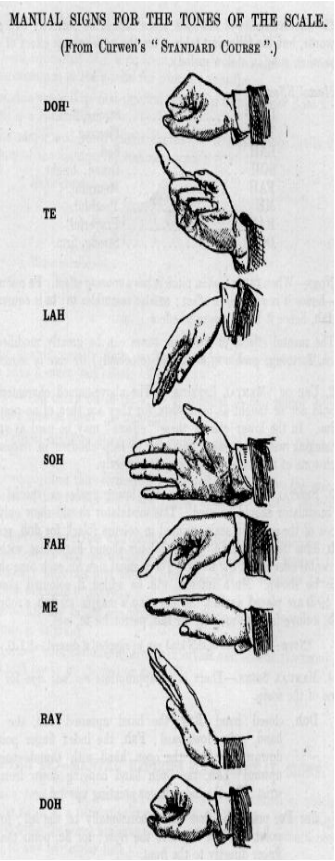
- Beat extraction and rhythm representation
- Multi-pitch analysis and transcription
- Chroma
- MPEG-4 Structured Audio
- Similarity Retrieval
- Genre Classification
- Score following

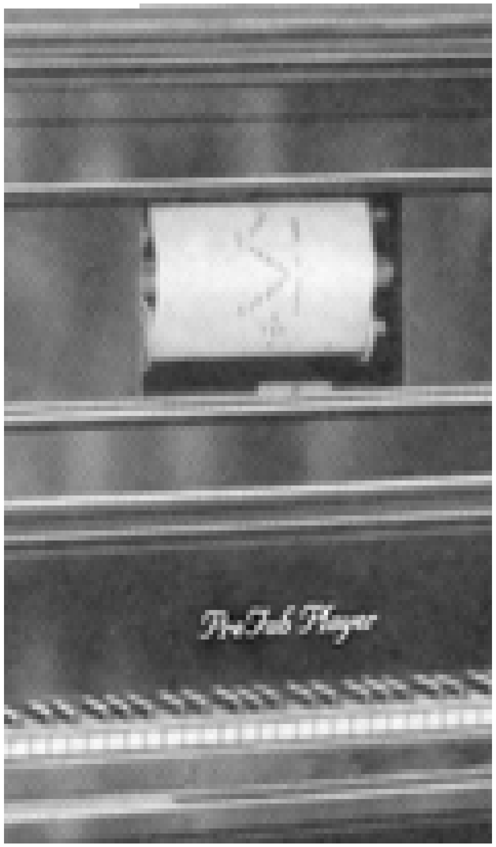
# Importance of Music

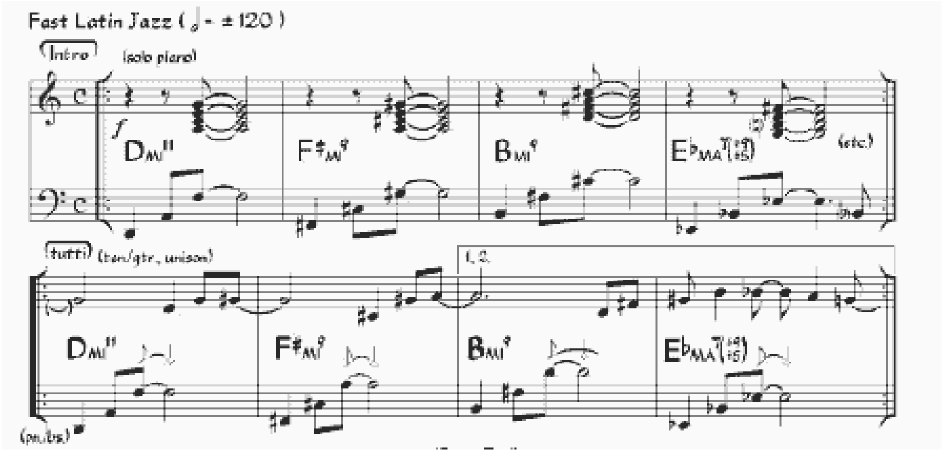
- 4 m CD tracks
- 4000 CDs / month
- 60-80% ISP bandwidth
- Napster- 1.57m sim.users (00)
- 61.3m downloaded music (01)
- Kazaa – 230 m downloads (03)
- Global, Pervasive, Complex

# Traditional Music Representations


GUIDO Noteserver. Powered by the SALIERI-Project ©.
http://www.informatik.tu-darmstadt.de/AFS/SALIERI
T
D3


MANUAL SIGNS FOR THE TONES OF THE SCALE.
(From Curwen's "STANDARD COURSE".)
DOH'
TE
LAH
SOH
FAH
ME
RAY
DOH



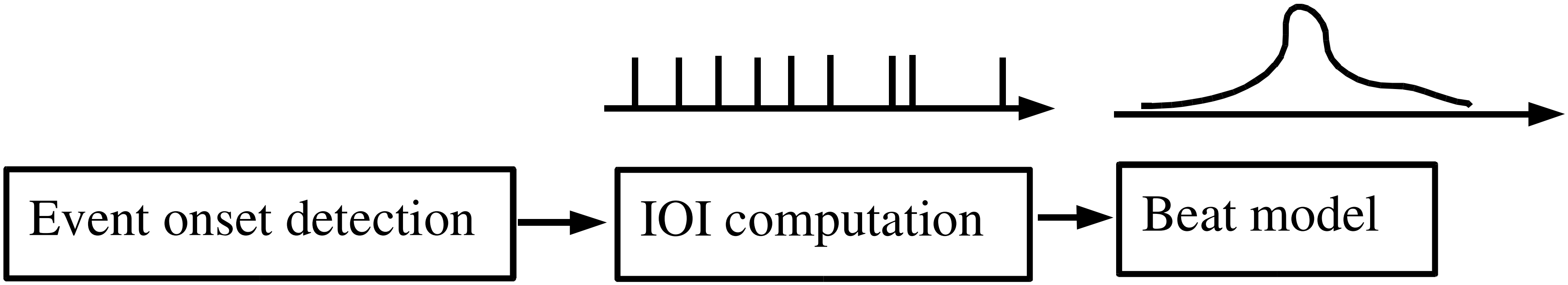


# Rhythm

- Rhythm = movement in time
- Origins in poetry (iamb, trochaic…)
- Foot tapping definition
- Hierarchical semi-periodic structure at multiple levels of detail
- Links to motion, other sounds
- Running vs global

# Event-based

| | |
|---|---|
| Alghoniemy, Tewfik | WMSP99 |
| Dixon | ICMC02 |
| Goto, Muraoka | IJCA97 |
| Gouyon et al | DAFX 00 |
| Laroche | WASPAA 01 |
| Seppanen | WASPAA 01 |



Event onset detection

IOI computation

Beat model

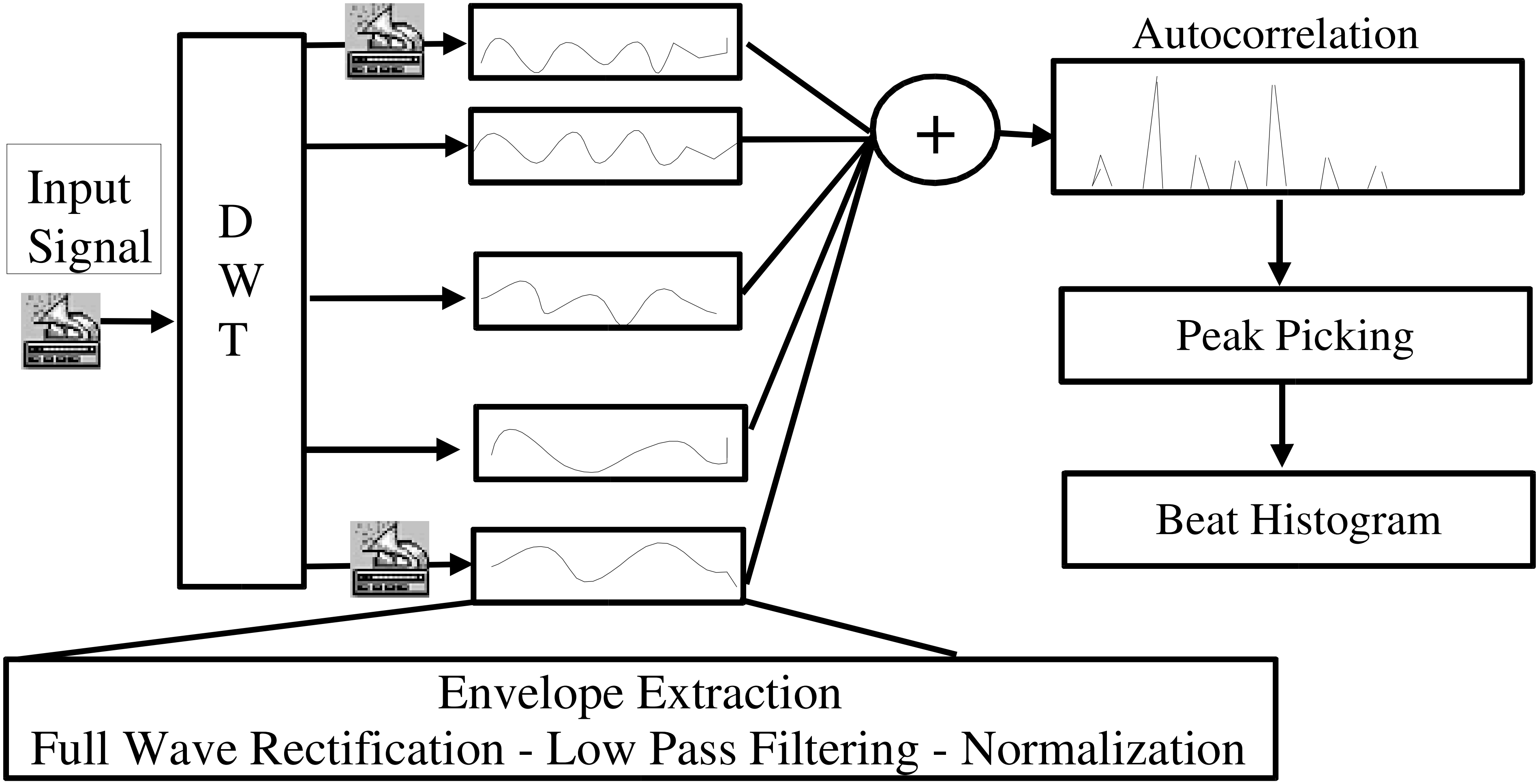
Single band
Multiple band
Thresholding
Chord changes

Successive onset pairs
All onset pairs
Quantization

Constant-tempo
Measure changes
Computational cost
Multiple hyphotheses

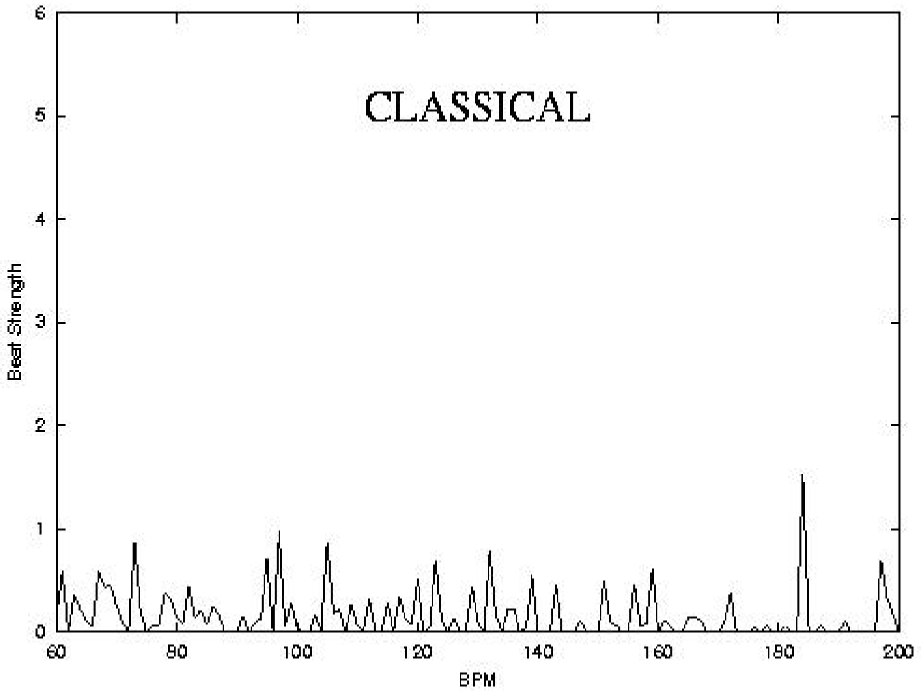
# Self-similarity

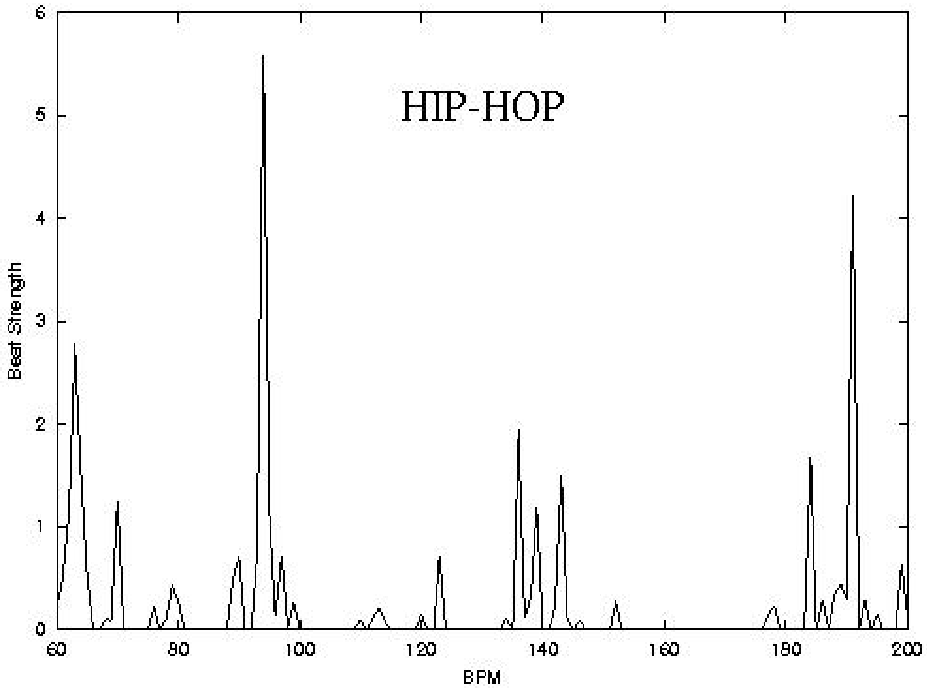
| | |
|---|---|
| Goto, Muraoka | CASA98 |
| Foote, Uchihashi | ICME01 |
| Scheirer | JASA98 |
| Tzanetakis et al | AMTA01 |



# Beat Histograms

Tzanetakis et al AMTA01





max(h(i)), argmax(h(i)) → ○

# Beat Spectrum

Foote, Uchihashi 01

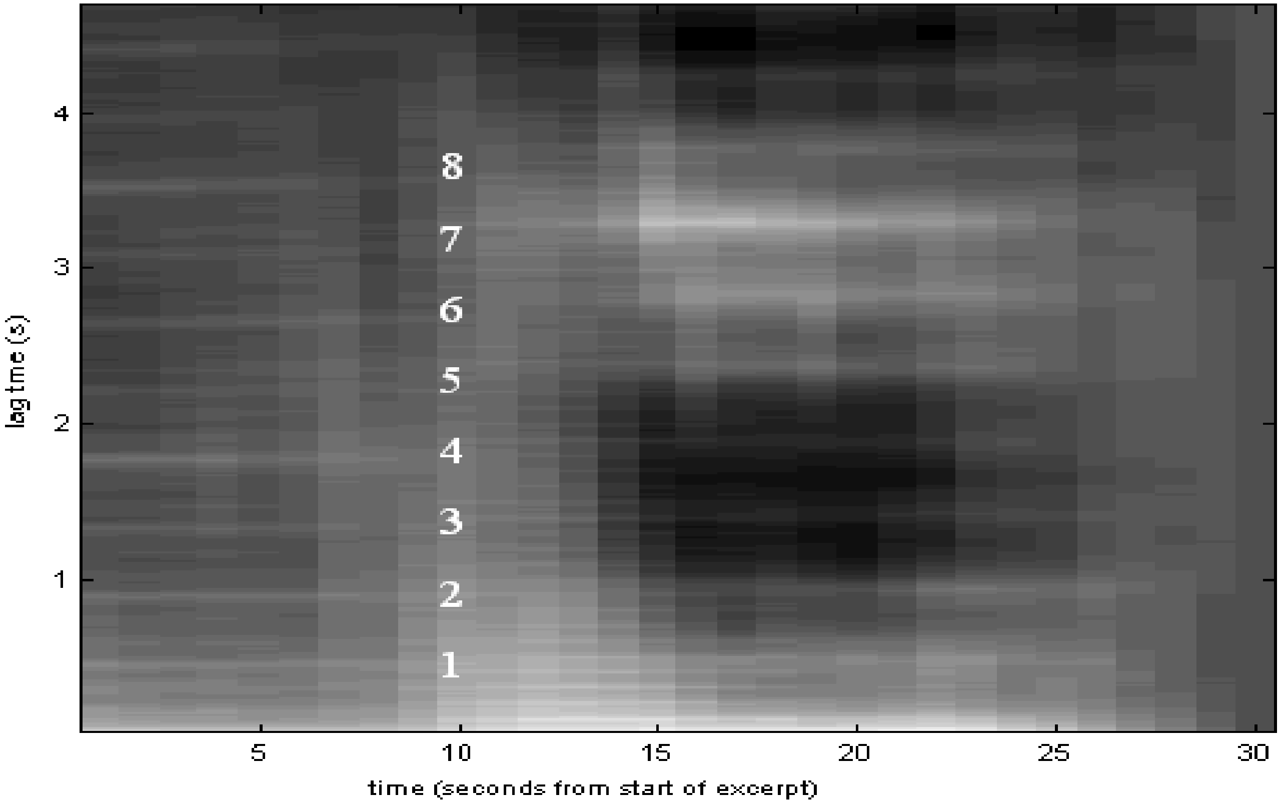


**Figure 4.** Beat spectrogram of Pink Floyd's *Money* (excerpt), showing transition from 4/4 to 7/4 time

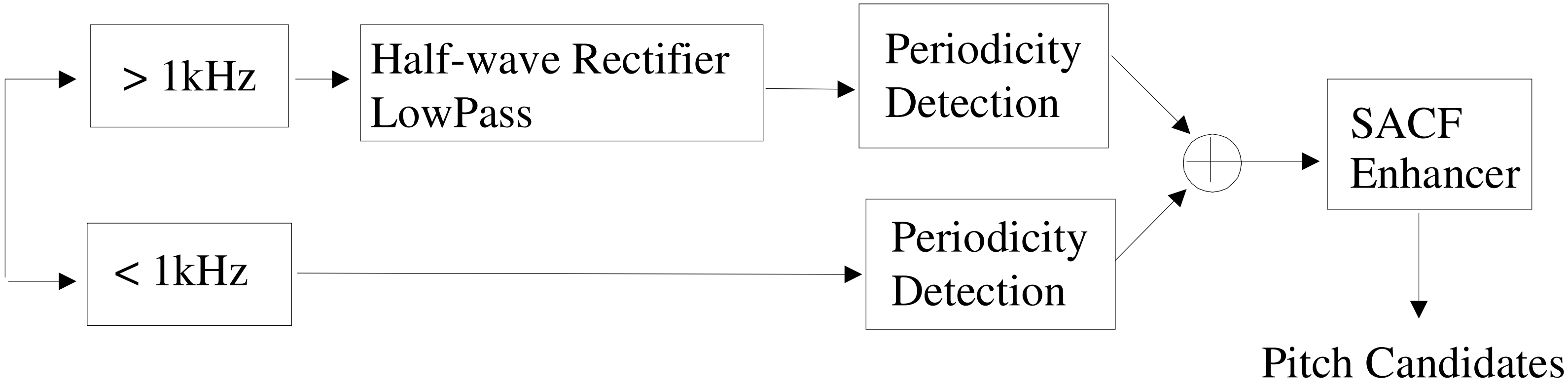
# Rhythmic content features

- Main tempo
- Secondary tempo
- Time signature
- Beat strength
- Regularity

# Multiple Pitch Detection

Tolonen and Karjalainen, TSAP 00

Tzanetakis et al, ISMIR 01



> 1kHz

Half-wave Rectifier LowPass

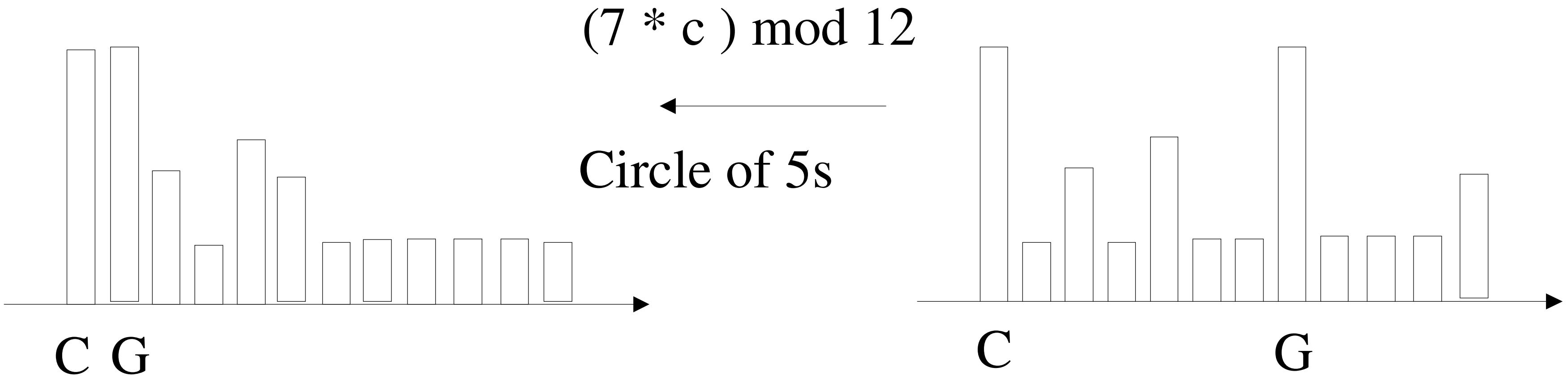
Periodicity Detection

< 1kHz

Periodicity Detection

SACF Enhancer

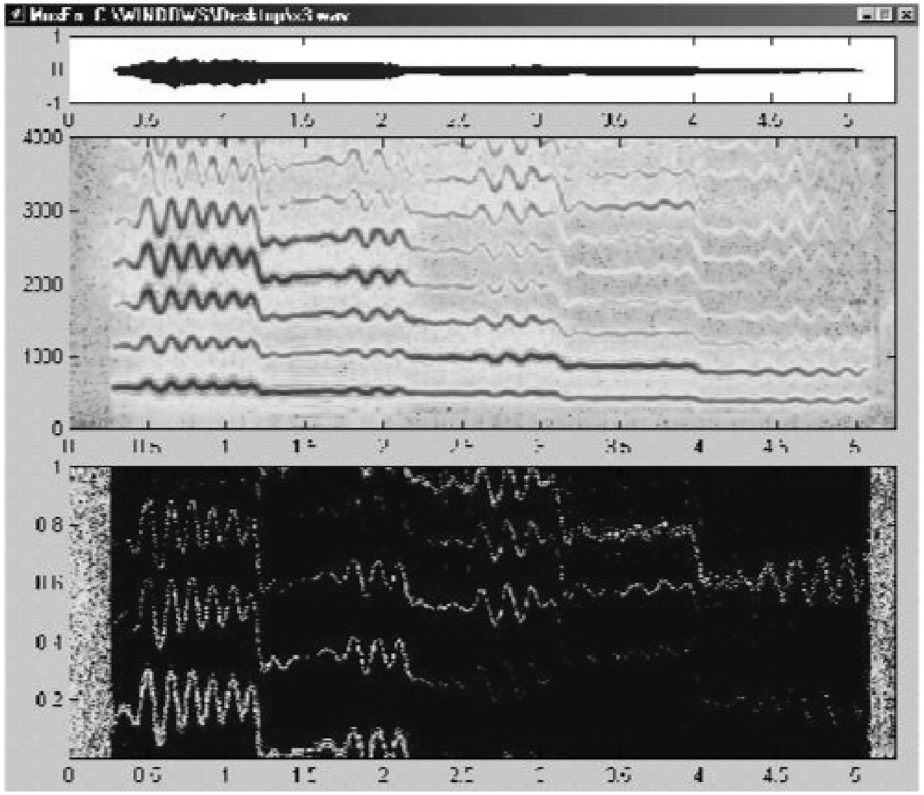
Pitch Candidates

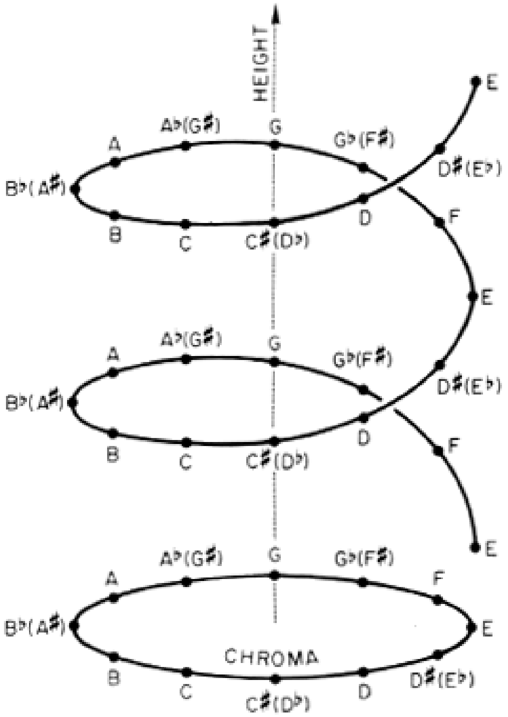


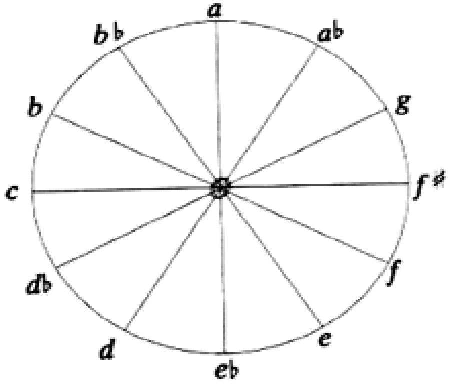
(7 * c ) mod 12

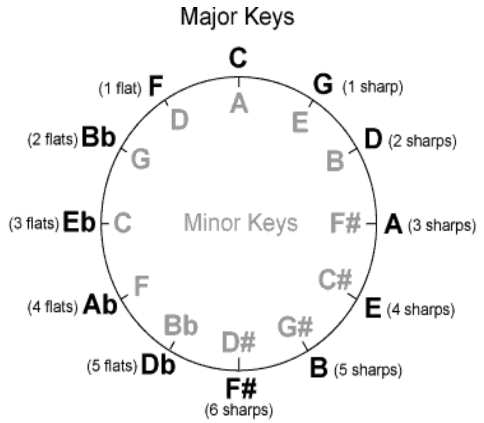
Circle of 5s

# Chroma – Pitch perception









# MIDI

- Musical Instrument Digital Interfaces
  - Hardware interface
  - File Format
- Note events
  - Duration, discrete pitch, "instrument"
- Extensions
  - General MIDI
  - Notation, OMR, continuous pitch

# Structured Audio

MPEG-4 SA
Eric Scheirer

Instead of samples store sound as a computer program that generates audio samples

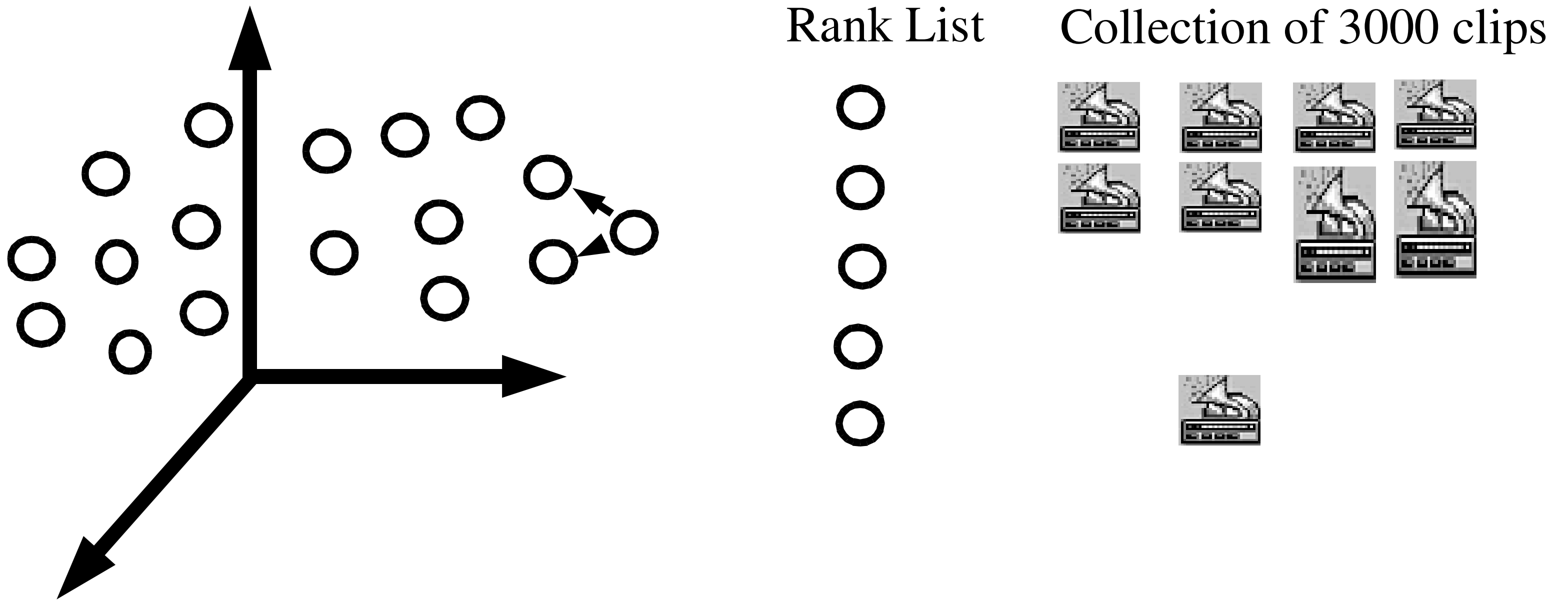
SAOL

```
 instr tone ()
{
   asig x, y, init;
   if (init = 0)
   {  init=1;
      x=0;}
x=x - 0.196307* y;
y=y + 0.196307* x;
output(y);
}
```

SASL

```
0.25 tone 4.0
4.50 end
```

# Query-by-Example Content-based Retrieval



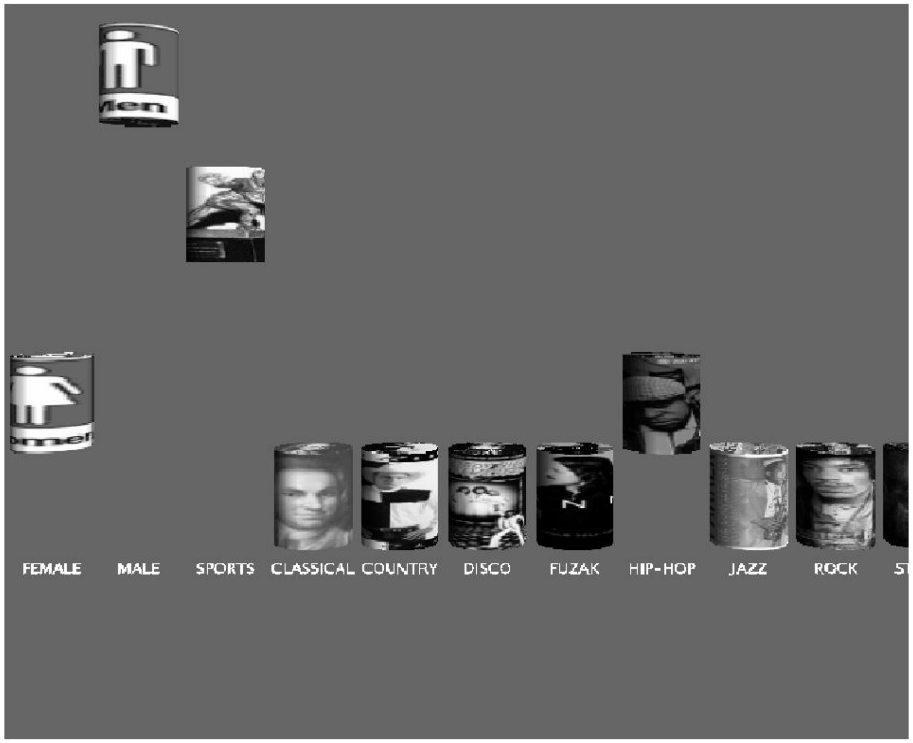
Rank List

Collection of 3000 clips

# Automatic Musical Genre Classification

- Categorical music descriptions created by humans
  - Fuzzy boundaries
- Statistical properties
  - Timbral texture, rhythmic structure, harmonic content
- Automatic Musical Genre Classification
  - Evaluate musical content features
  - Structure audio collections

# GenreGram DEMO

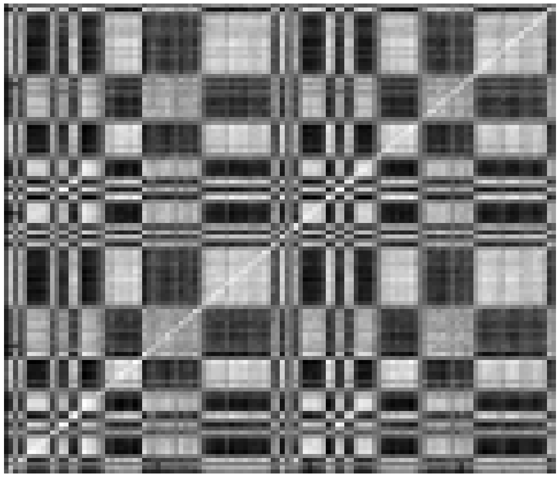


Dynamic real time 3D display
for classification of radio signals

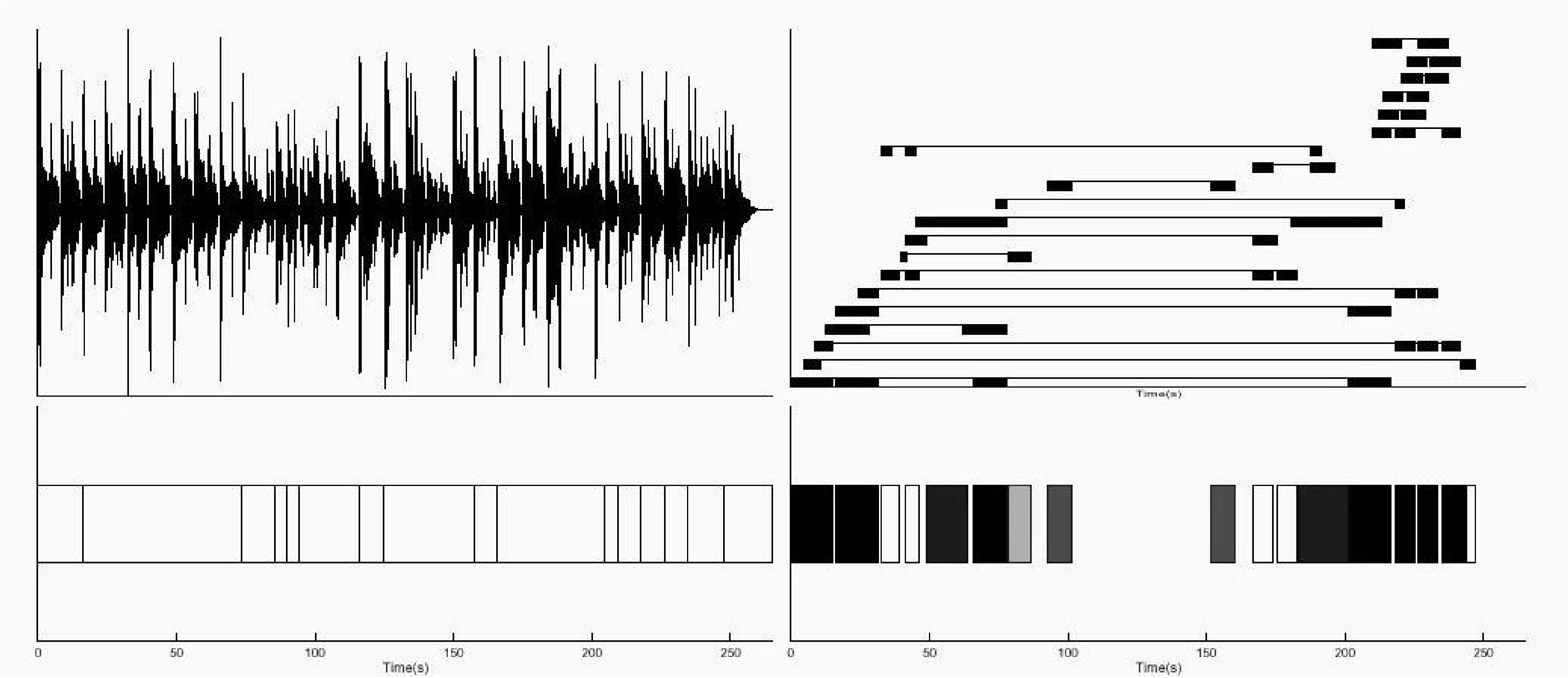
# Structural Analysis

Dannenberg & Hu, ISMIR 2002
Tzanetakis, Dannenberg & Hu, WIAMIS 03

- Similarity matrix
- Representations
  - Notes
  - Chords
  - Chroma
- Greedy hill-climbing algorithm
  - Recognize repeated patterns
- Result = AABA (explanation)



# An example – Naima (demo ?)

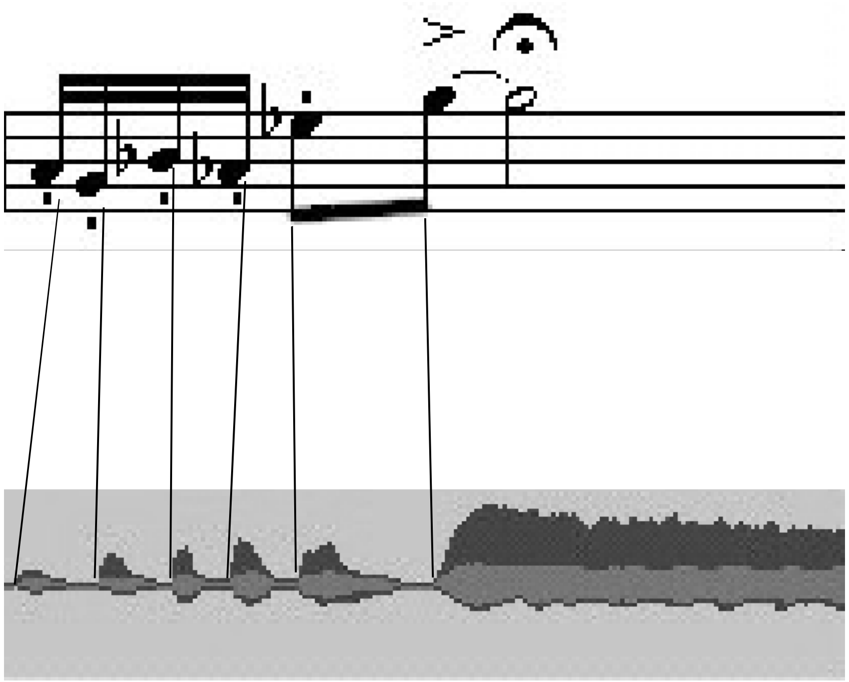


# Music Representations

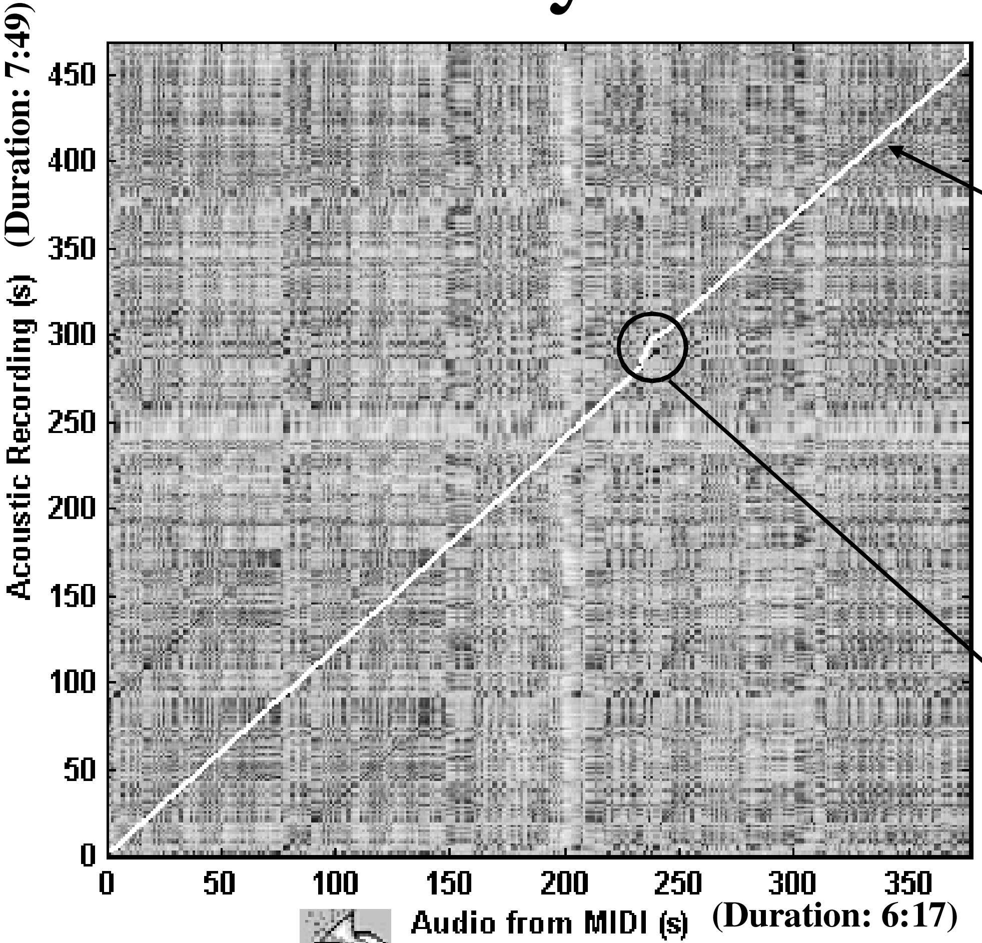
- Symbolic Representation
  - easy to manipulate
  - “flat” performance

Align

- Audio Representation
  - expressive performance
  - opaque & unstructured



# Similarity Matrix



Optimal Alignment Path

Oboe solo:
- Acoustic Recording
- Audio from MIDI

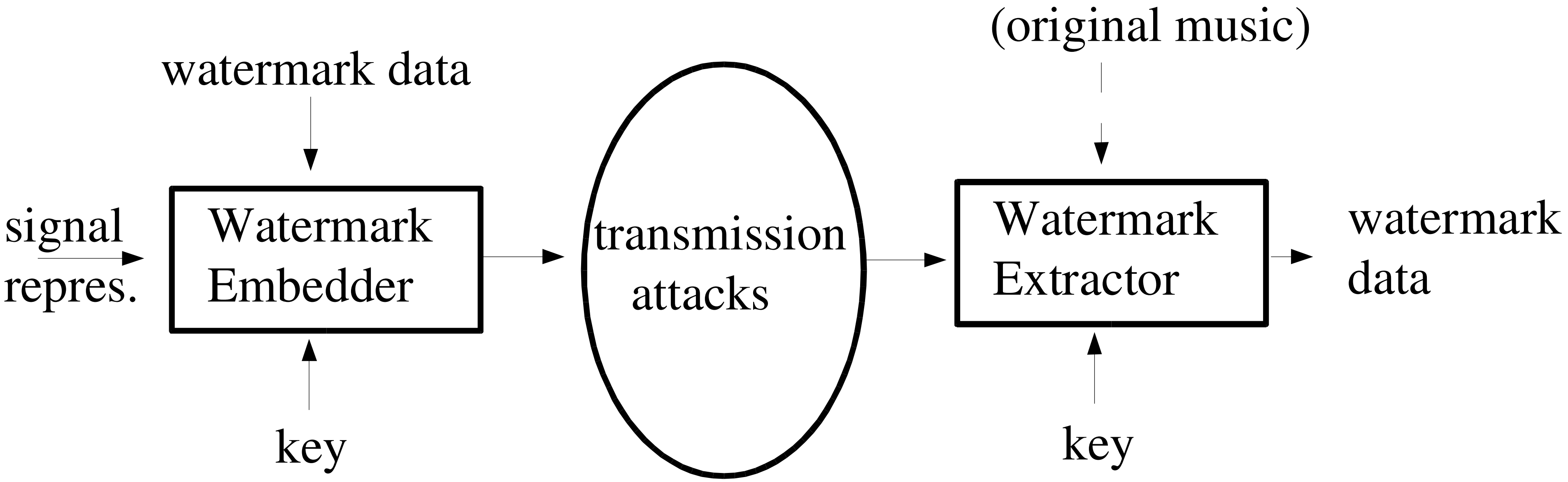
Similarity Matrix for Beethoven’s 5th Symphony, first movement

# Audio Fingerprinting and Watermarking

- Watermarking
  - Copyright protection
    - Proof of ownership
    - Usage policies
  - Metadata hiding
- Fingerprinting
  - Tracking
  - Copyright protection
  - Metadata linking

# Watermarking

- Steganography (hiding information in messages – invisible ink )



# Desired Properties

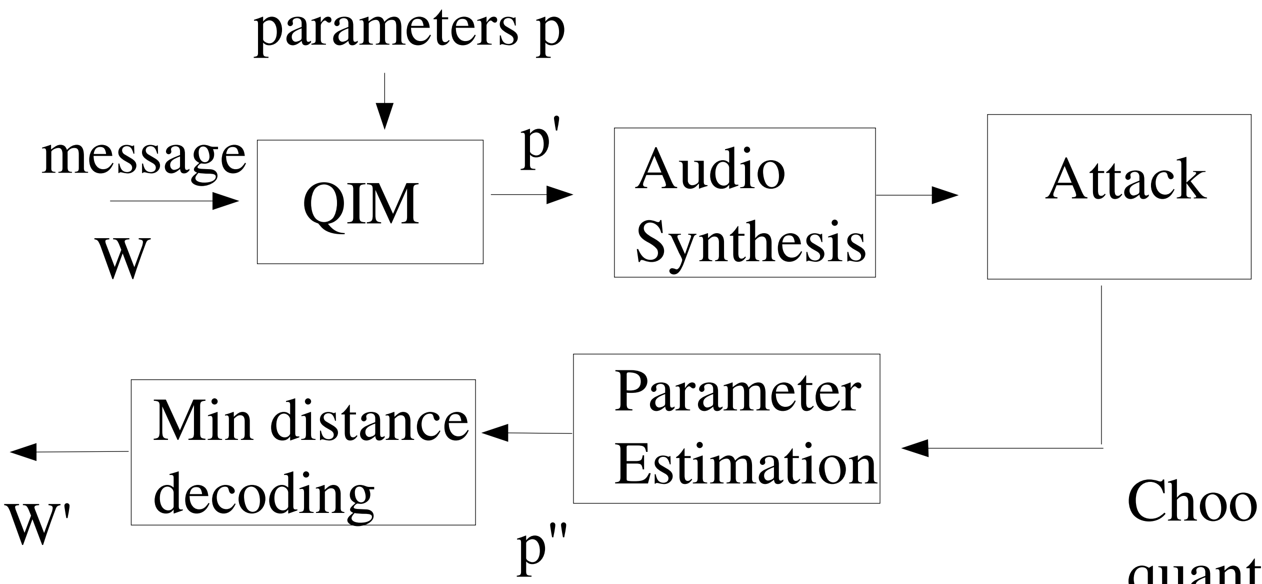
- Perceptually hidden (inaudible)
- Statistically invisible
- Robust against signal processing
- Tamper resistant
- Spread in the music, not in header
- key dependent

# Representations for Watermarking

- Basic Principles
  - Psychoacoustics
  - Spread Spectrum
    - redundant spread of information in TF plane
- Representations
  - Linear PCM
  - Compressed bitstreams
  - Phase, stereo
  - Parametric representations

# Watermarking on parametric representations

Yi-Wen Liu
J. Smith 2004



Choose attack tolerance quantize so perceptual distortion < t and lattice finding possible

# Problems with watermarking

- The security of the entire system depends on devices available to attackers
  - Breaks Kerckhoff's Criterion: A security system must work even if reverse-engineered
- Mismatch attacks
  - Time stretch audio – stretch it back (invertible)
- Oracle attacks
  - Poll watermark detector

# Audio Fingerprinting

- Each song is represent as a fingerprint (small robust representation)
- Search database based on fingerprint
- Main challenges
  - highly robust fingerprint extraction
  - efficient fingerprint search strategy
- Information is summarized from the whole song – attacks degrade unlike watermarking

# Hash functions

- H(X) -> maps large X to small hash value
- compare by comparing hash value
- Perceptual hash function ?
    - impossible to get exact matching
- Perceptually similar objects result in similar fingerprints
- Detection/false alarm tradeoff

# Properties

- Robustness
- Reliability
- Fingerprint size
- Granularity
- Search speed and scalability

# Fraunhofer

Allamanche Ismir
2001

- LLD Mpeg-7 framework (SFM)
- Vector quantization (k-means)
  - Codebook of representative vectors
- Database target signature is the codebook
- Query -> sequence of feature vectors
- Matching by finding “best” codebook
- Robust not very scalable (O(n) search))

# Philips Research

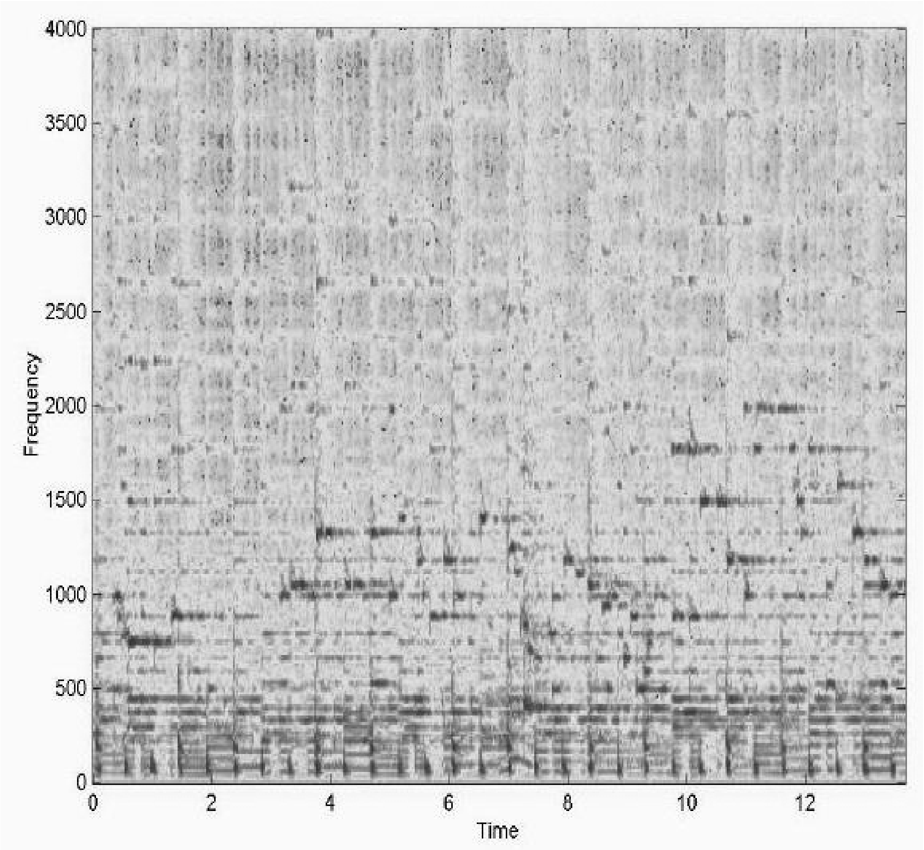
Haitsa & Kalker
Ismir 2002

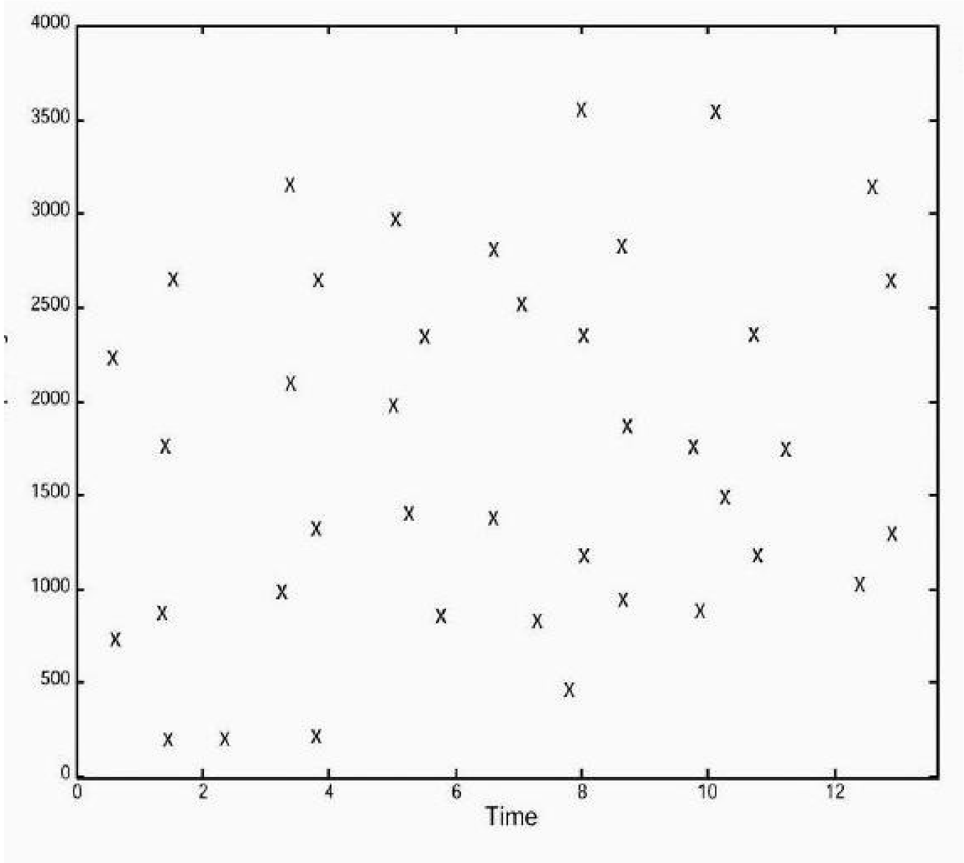
- 32-bit subfingerprints for every 11.6 msec
- overlapping frames of 0.37 seconds (31/32 overlap)
- PSD -> logarithmic band spacing (bark)
- bits 0-1 sign of energy
- looks like a fingerprint
- assume one fingerprint perfect – hierarchical database layout (works ok)

# Shazam Entertainment

- Pick landmarks on audio – calculate fingerprint
- histogram of relative time differences for filtering
- Spectrogram peaks (time, frequency)

# Spectrogram Peaks

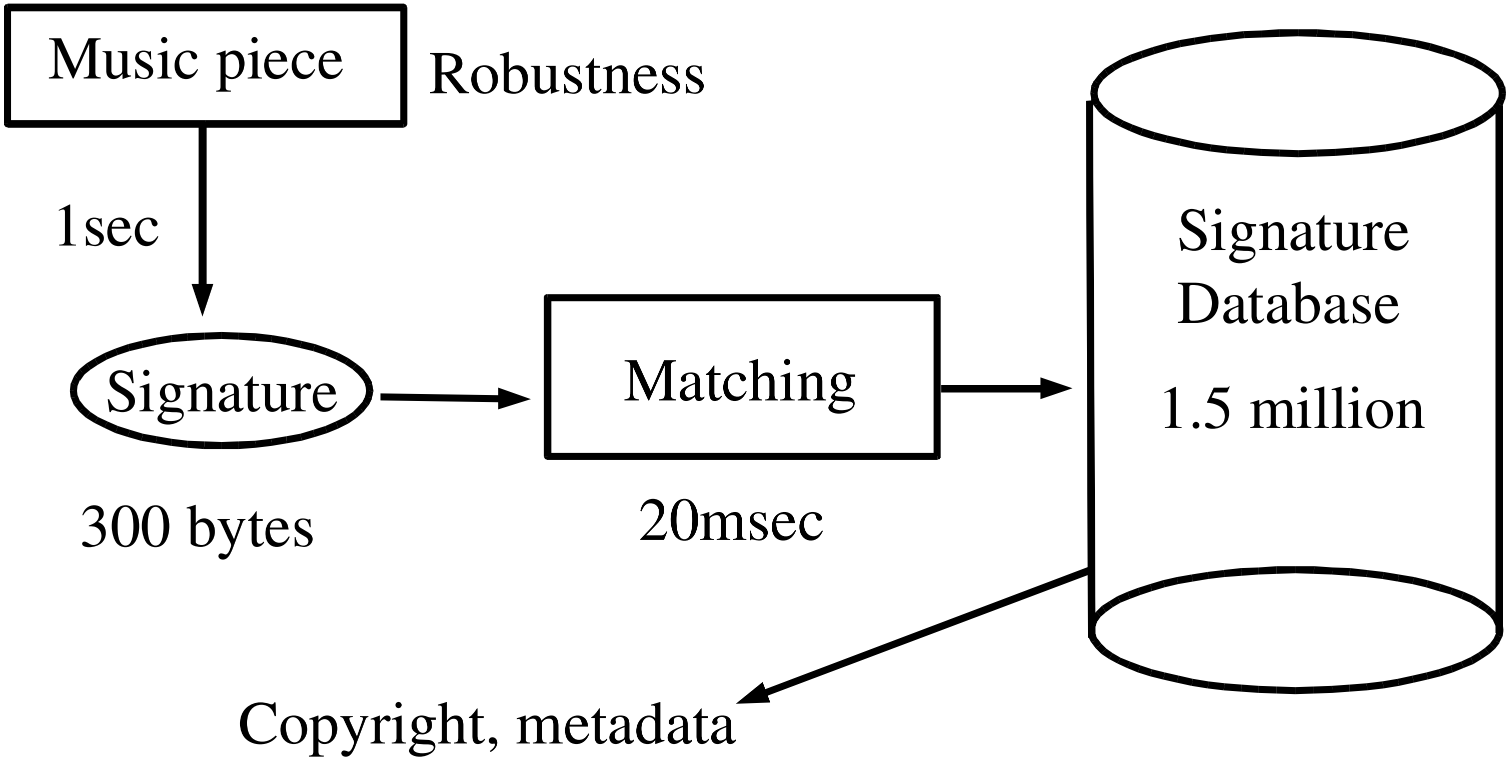




Very robust – even over noisy cell phones

# Audio Fingerprinting

moodlogic.net



Music piece

Robustness

1sec

Signature

300 bytes

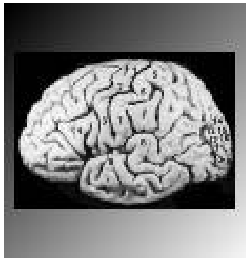
Matching

20msec

Signature Database

1.5 million

Copyright, metadata



# Auditory Scene Analysis

- Music and Sound Cognition
- Onset detection
- Toward Transcription

# Auditory Scene Analysis

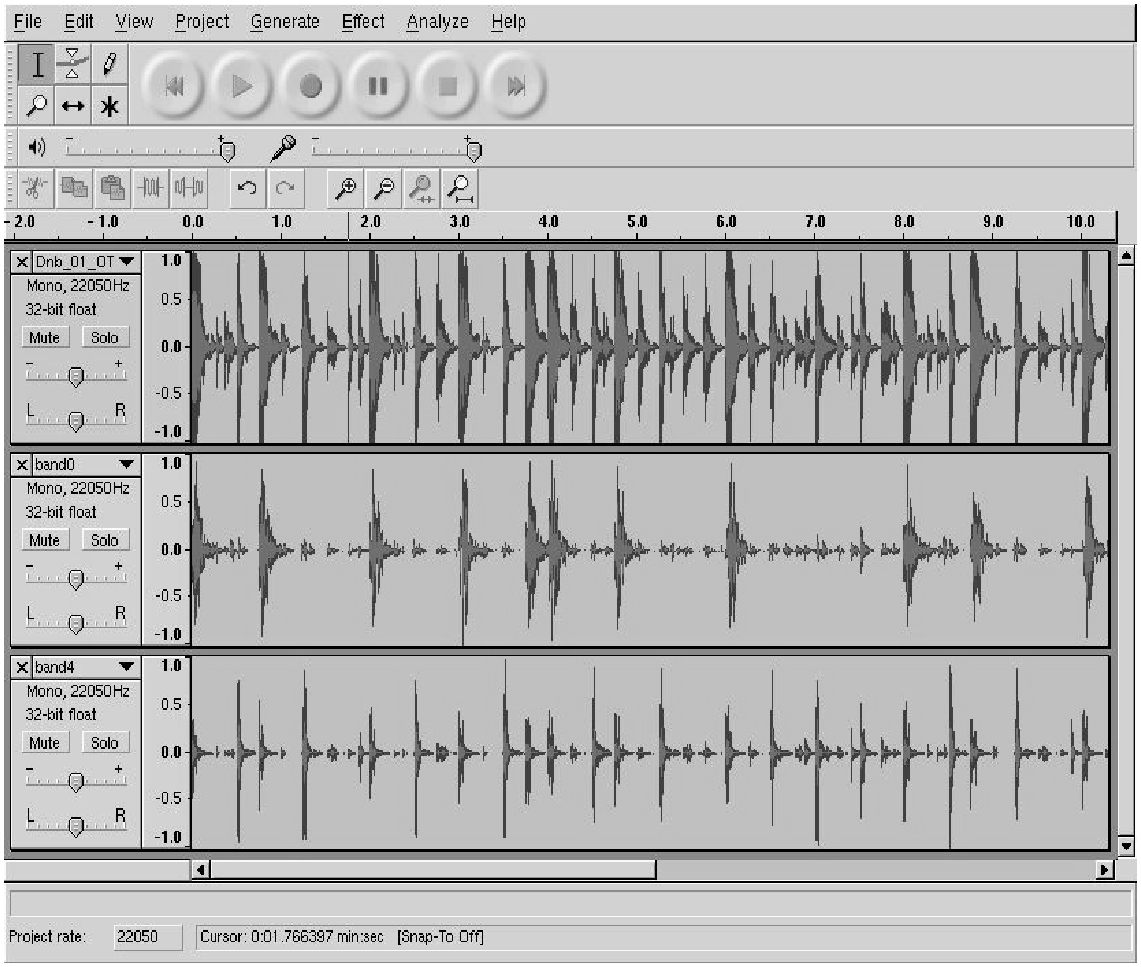
Bregman

- Auditory stream
  - perceptual grouping of parts of the neural spectrogram that go together
- Sound is a mixture and is transparent
  - Primitive process of streaming
  - Schemas for particular classes of sounds
- Grouping
  - across time (sequential)
  - across freq (simultaneous)

# Onset detection

Naive: peaks in power
Multiband
(wavelet, filterbanks)

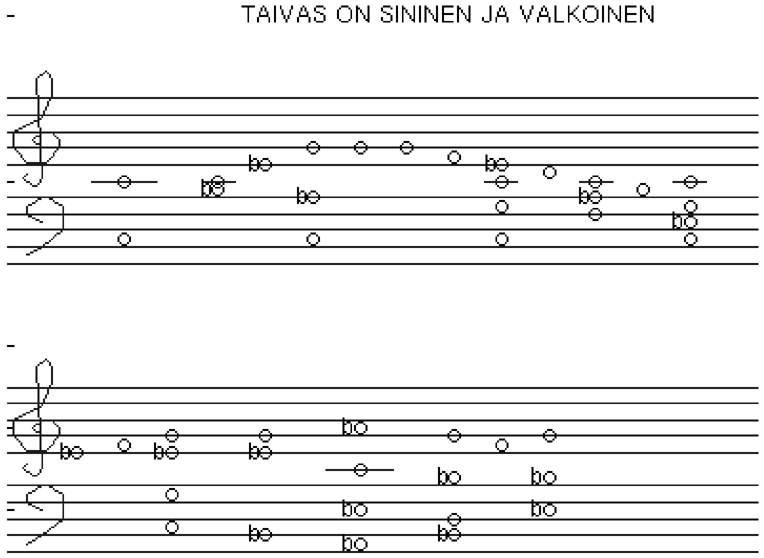
Synchronicity
Temporal Continuity
Common Fate
Proximity



Project rate: 22050 Cursor: 0:01.766397 min:sec [Snap-To Off]

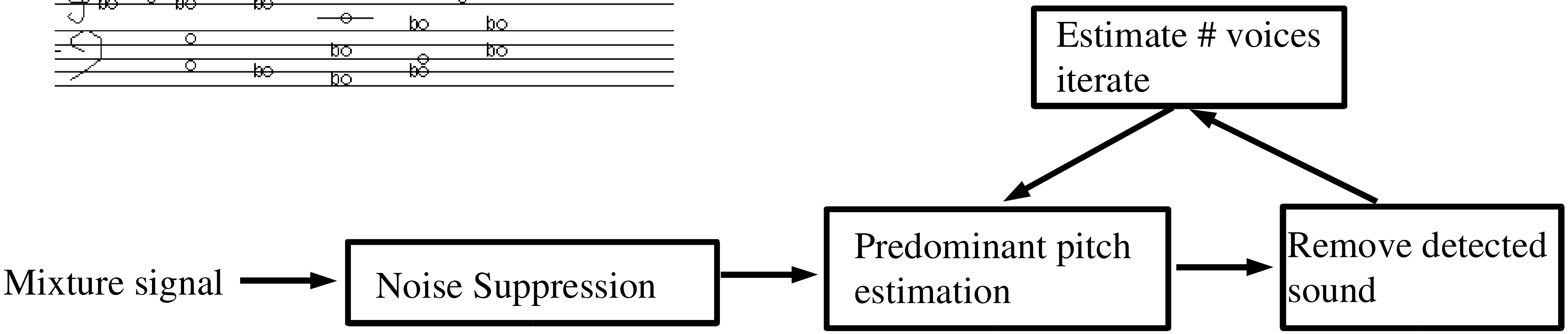
# Polyphonic Transcription

Klapuri et al, DAFX 00



Original

Transcribed



Mixture signal

Noise Suppression

Predominant pitch estimation

Remove detected sound

Estimate # voices iterate

# Summary

- Applications and especially analysis have different requirements -> different features
- wide variety of proposed audio features
- still many to be found .... hopefully by you :-)

# Future Challenges

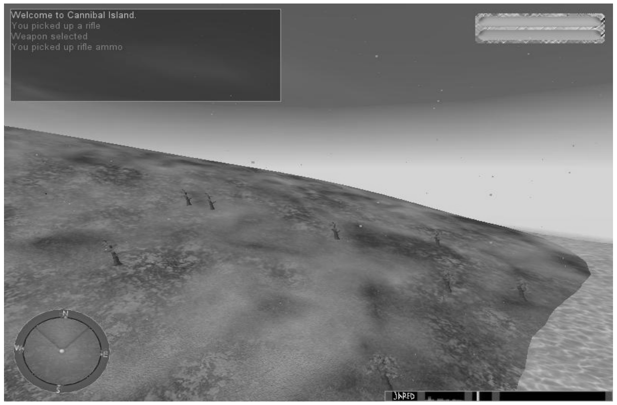
- Main challenges
  - escape HMM and MFCC
  - tackle the general problem of auditory scene analysis
  - “real learning”
  - active audition - search for evidence rather than try to find

# Implementation

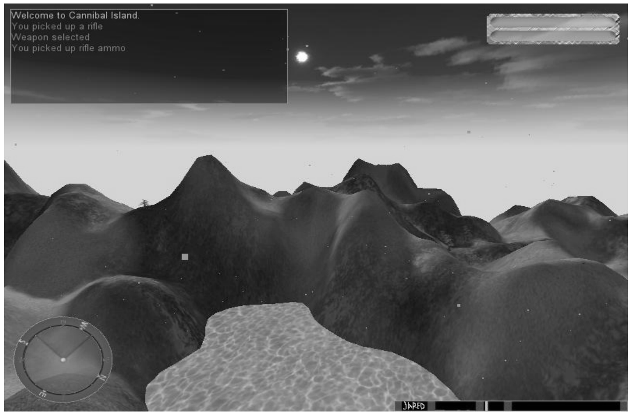
Tzanetakis & Cook Organized Sound 4(3) 00

- *MARSYAS* : free software framework for computer audition research
  - marsyas.sourceforge.net
  - Server in C++ (numerical signal processing and machine learning)
  - Client in JAVA (GUI)
  - Linux, Solaris, Irix and Wintel (VS , Cygwin)

# Marsyas users



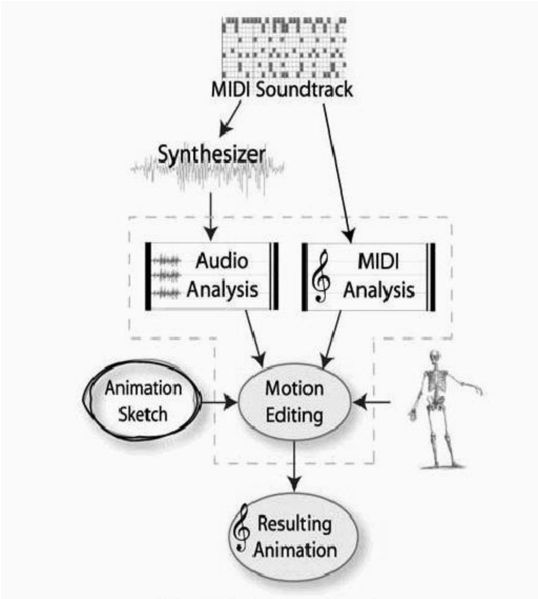
Welcome to Cannibal Island.
You picked up a rifle
Weapon selected
You picked up rifle ammo



Welcome to Cannibal Island.
You picked up a rifle
Weapon selected
You picked up rifle ammo

Desert Island

Jared Hoberock
Dan Kelly
Ben Tietgen



Music-driven
motion editing
Marc Cardle

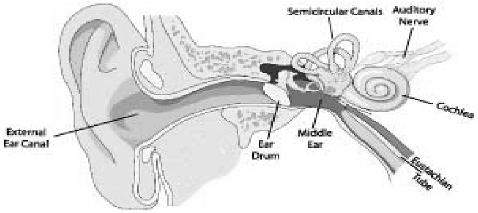
moodlogic
The Mix Maker for your MP3s

Real time
music-speech
discrimination

ALL MUSIC
PUBLISHING

# Auditory Scene Analysis

Albert Bregman



# THE END

- Perry Cook, Robert Gjerdingen, Ken Steiglitz
- Malcolm Slaney, Julius Smith, Richard Duda
- Georg Essl, John Forsyth
- Andreye Ermolinskiy, Doug Turnbull, George Tourtellot, Corrie Elder
- ISMIR, WASPAA, ICMC, DAFX, ICASSP , ICME