SOUND!!!!

prc

(really gtzan, with many others (Ge, Ananya, Matt))

Representing Raw Sound

- So Many Bits, So Little Time (Space)
  - **CD audio rate: 2 * 2 * 8 * 44100 = 1,411,200 bps**
  - **CD audio storage: 10,584,000 bytes / minute**
  - **A CD holds only about 70 minutes of audio**
  - **An ISDN line can only carry 128,000 bps**
- Security: Best compressor removes all recognizable about the original sound
- Graphics people eat up all the space

New Audio Formats

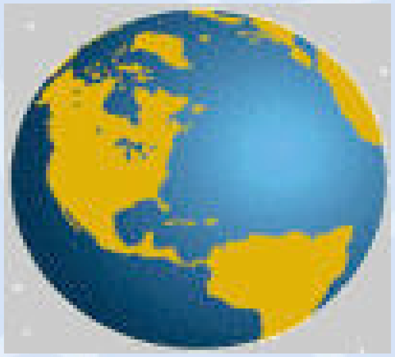
SIGGRAPH 2003 SAN DIEGO

- 24 bit, common on soundcards
- 48 KHz (standard DVD) and other
- 96 Khz
- 192 KHz
- 5.1, 7.2, 14.2
- Highest spec to date:
  - 192KHz, 24 bit, 14.2, uncompressed (SACD, DVDAudio)
  - This is 9MBytes per second!!
  - 552,950,000 bytes per minute
  - 33,177,600,000 per hour

SIGGRAPH 2003
SAN DIEGO

Music

- 4 million recorded CDs
- 4000 CDs / month
- 60-80% ISP bandwidth
- Global
- Pervasive
- Complex



Sound in life

- Capture work hours:
  - 8-10 hours per day
  - 5-6 days per week
  - 16KHz, 16 bit
  - Over average work life (40 years)
    - 10*5*60*60*16k*2*40 = 230,400,000,000 bytes
    - (compare to Steve Jobs' 1989 256MByte)

Compression/ Representation

- Classical Data Compression View:
- Take advantage of
  - **Redundancy/Correlation**
  - **Statistics (Local/Global)**
  - **Assumptions / Models**
- Problem: Much of this doesn’t work directly on sound waveform data
  - **Redundancy, nope**
  - **Correlation, not really**

One View of Sound

Sound is a waveform,
we can record it, store it,
and play it back accurately

PCM playback is all we need for
interactions, movies, games, etc.

Features and statistics of the raw data, or
waveform shape, is enough to classify.

But, take some visual analogies:

"If I take lots of polaroid images, I can flip through them real fast and make any image sequence"

"We should be able to use correlations, similar to color in images, to compress, segment, etc. sound"

We Can Compute Sound!!

SIGGRAPH 2003
SAN DIEGO

Views of Sound:

- Time Domain x(t)
  (from physics, and time's arrow)

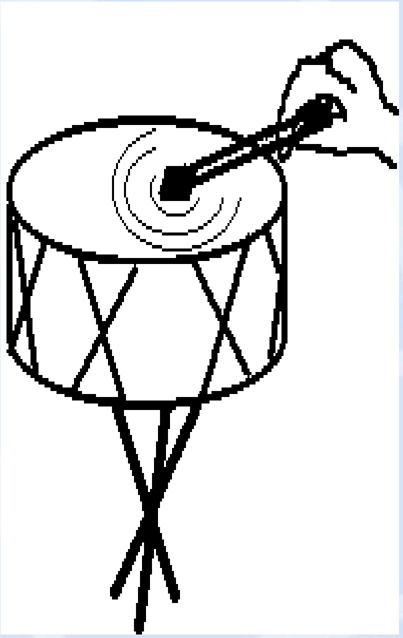
- Frequency Domain X(f)
  (from math, and perception)

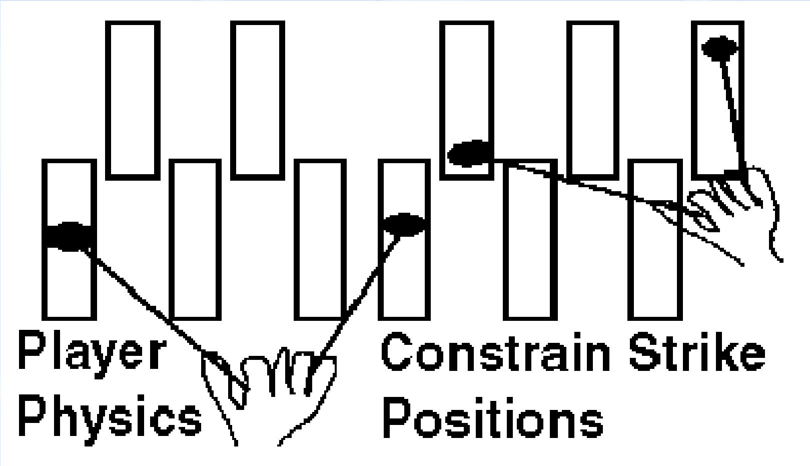
- Production what caused it

- Perception our "image" of it

Views of Sound: Production

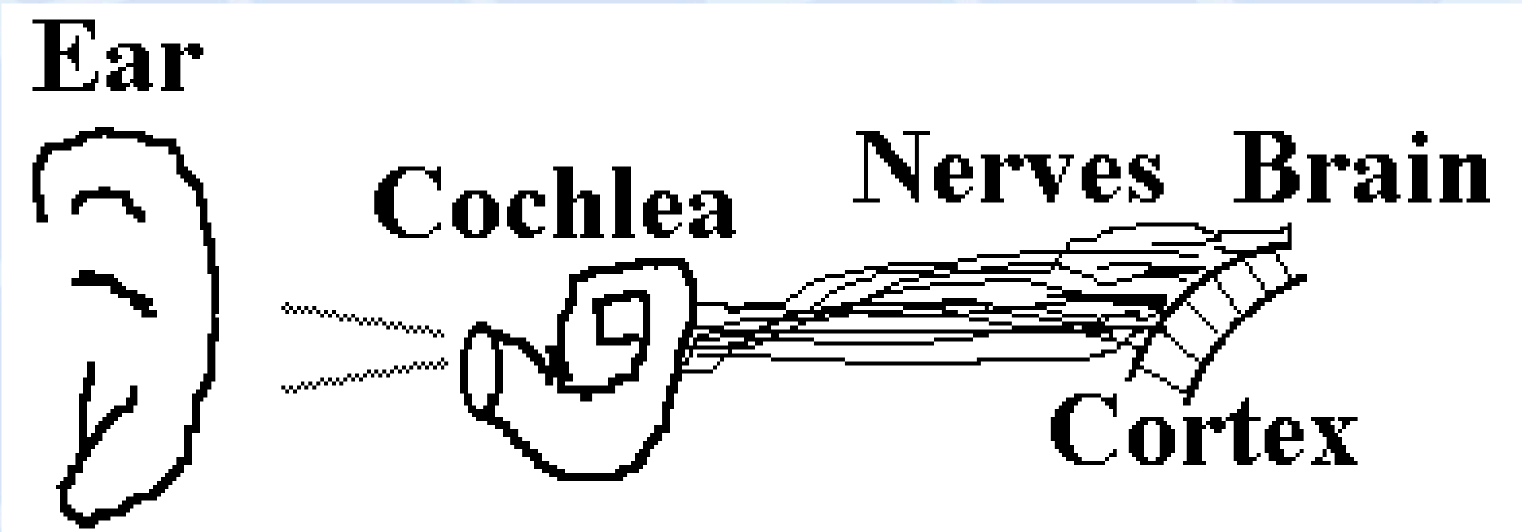
- Throughout most of history, some physical mechanism was responsible for sound production.
- From our experience, certain <u>gestures</u> produce certain audible results





Examples:
Hit harder --> louder AND brighter
Can't move instantaneously
Can't do exactly the same thing twice

Views of Sound: Perception



receive 1-D waves	convert to frequency dependent nerve firings	further refine time & frequency information	High level cognition, object formation, interpretation

Auditory system does time to frequency conversion

Views of Sound

SIGGRAPH 2003
SAN DIEGO

- The Time Domain
  is most closely related to
  Production

- The Frequency Domain
  is most closely related to
  Perception

Limits of Human Hearing

- Time and Frequency

Events longer than 0.03 seconds are resolvable *in time*

shorter events are perceived as *features in frequency*

20 Hz. < Human Hearing < 20 KHz.
(for those under 15 or so)

"Pitch" is <u>PERCEPTION</u> related to <u>FREQUENCY</u>
Human Pitch Resolution is about 40 - 4000 Hz.

Limits of Human Hearing

SIGGRAPH 2003 SAN DIEGO

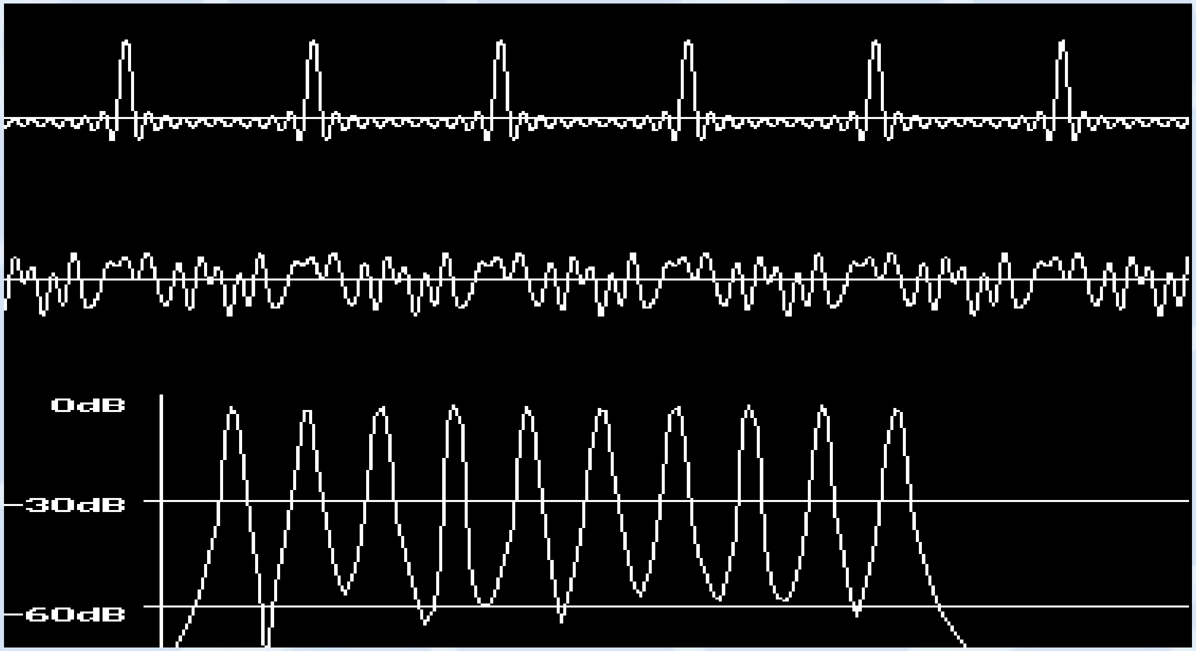
- Amplitude or Power???
  - **"Loudness" is <u>PERCEPTION</u> related to <u>POWER</u>, not <u>AMPLITUDE</u>**
  - **Power is proportional to (integrated) square of signal**
  - **Human Loudness perception range is about 120 dB,**

    where +10 db = 10 x power = 20 x amplitude
  - **Waveform shape is of little consequence. Energy at each frequency, and how that changes in time, is the most important feature of a sound.**

Limits of Human Hearing

- Waveshape or Frequency Content??
- Here are two waveforms with identical power spectra, and which are (nearly) perceptually identical:

- Wave 1
- Wave 2
- Magnitude Spectrum of Either



Limits of Human Hearing

- Masking in Amplitude, Time, and Frequency
  - **Masking in Amplitude: Loud sounds ‘mask’ soft ones. Example: Quantization Noise**
  - **Masking in Time: A soft sound just before a louder sound is more likely to be heard than if it is just after. Example (and reason): Reverb vs. “Preverb”**
  - **Masking in Frequency: Loud ‘neighbor’ frequency masks soft spectral components. Low sounds mask higher ones more than high masking low.**

Limits of Human Hearing

- Masking in Amplitude
- *Intuitively, a soft sound will not be heard if there is a competing loud sound. Reasons:*
  - **Gain controls in the ear**
    - **stapedes reflex and more**
  - **Interaction (inhibition) in the cochlea**
  - **Other mechanisms at higher levels**

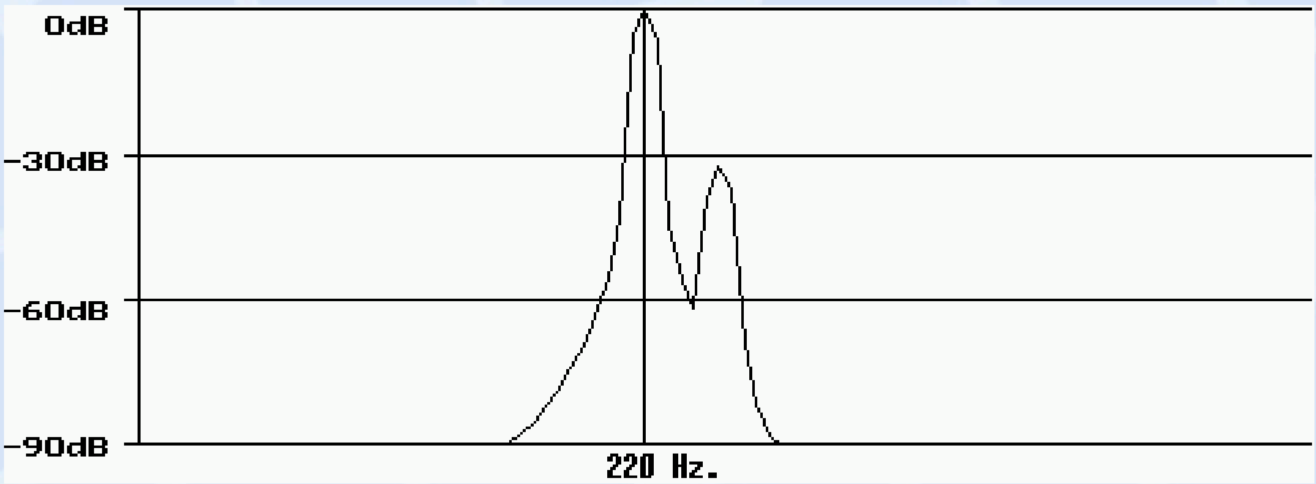
Limits of Human Hearing

- Masking in Time
  - **In the time range of a few milliseconds:**
  - **A soft event following a louder event tends to be grouped perceptually as part of that louder event**
  - **If the soft event precedes the louder event, it might be heard as a separate event (become audible)**

Limits of Human Hearing

- Masking in Frequency

Only one component in this spectrum is audible because of frequency masking



Sound Views: Frequency Domain

- Many physical systems have <u>modes</u> (damped oscillations)
- Wave equation (2nd order) or
  Bar equation (4th order)
  need 2 or 4 “boundary conditions”
  for solution
- Once boundary conditions are set, solutions are sums of exponentially damped sinusoidal modes
- One more important aspect of frequency:

The (discrete) Fourier Series

SIGGRAPH 2003 SAN DIEGO

A time waveform is a sum of sinusoids

(A_m is complex)

$$x(n) = \sum_{n=0}^{N-1} A_m \exp(\frac{j2\pi nm}{N})$$

$$= \sum_{n=0}^{N-1} B_m \sin(\frac{2\pi nm}{N}) + C_m \cos(\frac{2\pi nm}{N})$$

$$= \sum_{n=0}^{N-1} D_m \cos(\frac{2\pi nm}{N} + \theta_m)$$

The (discrete) Fourier Transform

$$A(m) = X(SRATE * m / N) = \sum_{n=0}^{N-1} x(n) \exp(\frac{-jnm2\pi}{N})$$

A "Spectrum" is a

unique and invertible

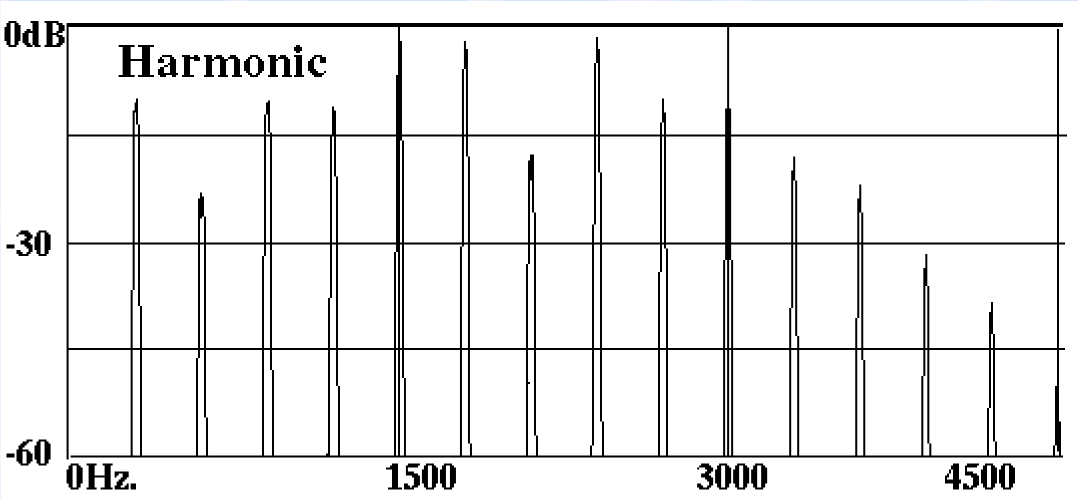
Sinusoidal decomposition of a signal

Spectra: Magnitude and Phase

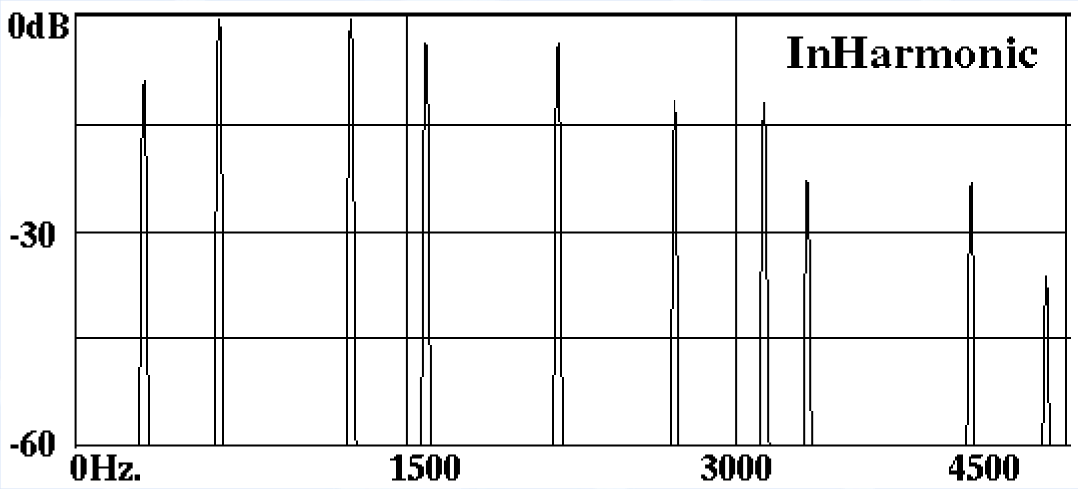
- Often only magnitude is plotted
  - Human perception is most sensitive to magnitude
    - Environment corrupts and changes phase
  - 2 (pseudo-3) dimensional plots easy to view
- Phase is important, however
  - Especially for transients (attacks, consonants, etc.)
- If we know instantaneous amplitude and frequency, we can derive phase

Common Types of Spectra

- Harmonic
  - sines at integer
  - multiple freqs.



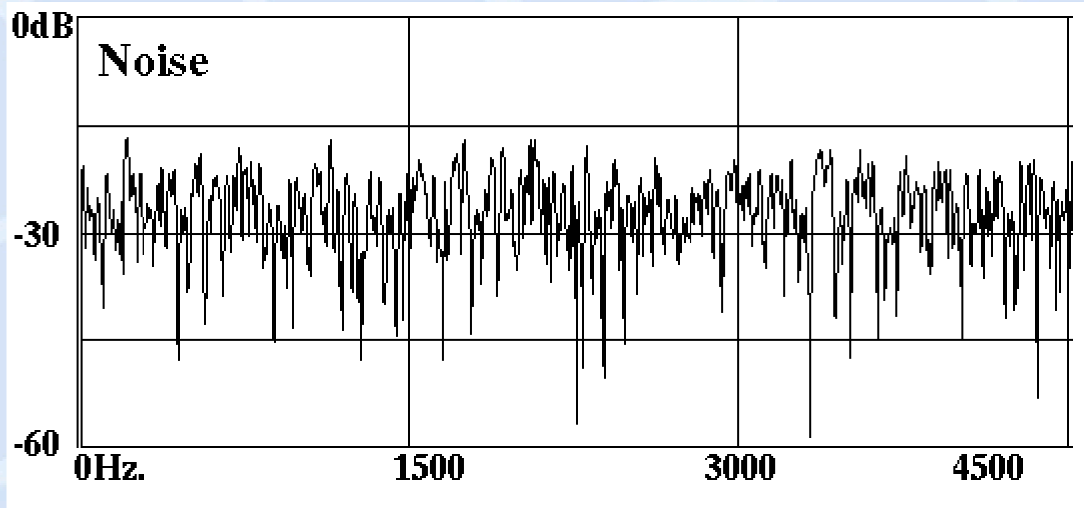
- Inharmonic
  - sines (modes),
  - but not integer
  - multiples

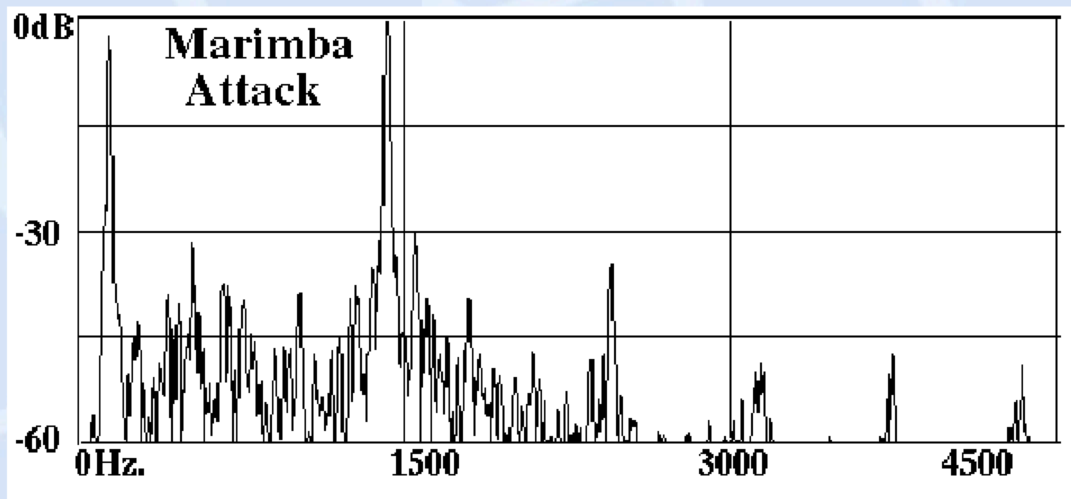


Common Types of Spectra

- Noise
  - random
  - amplitudes
  - and phases

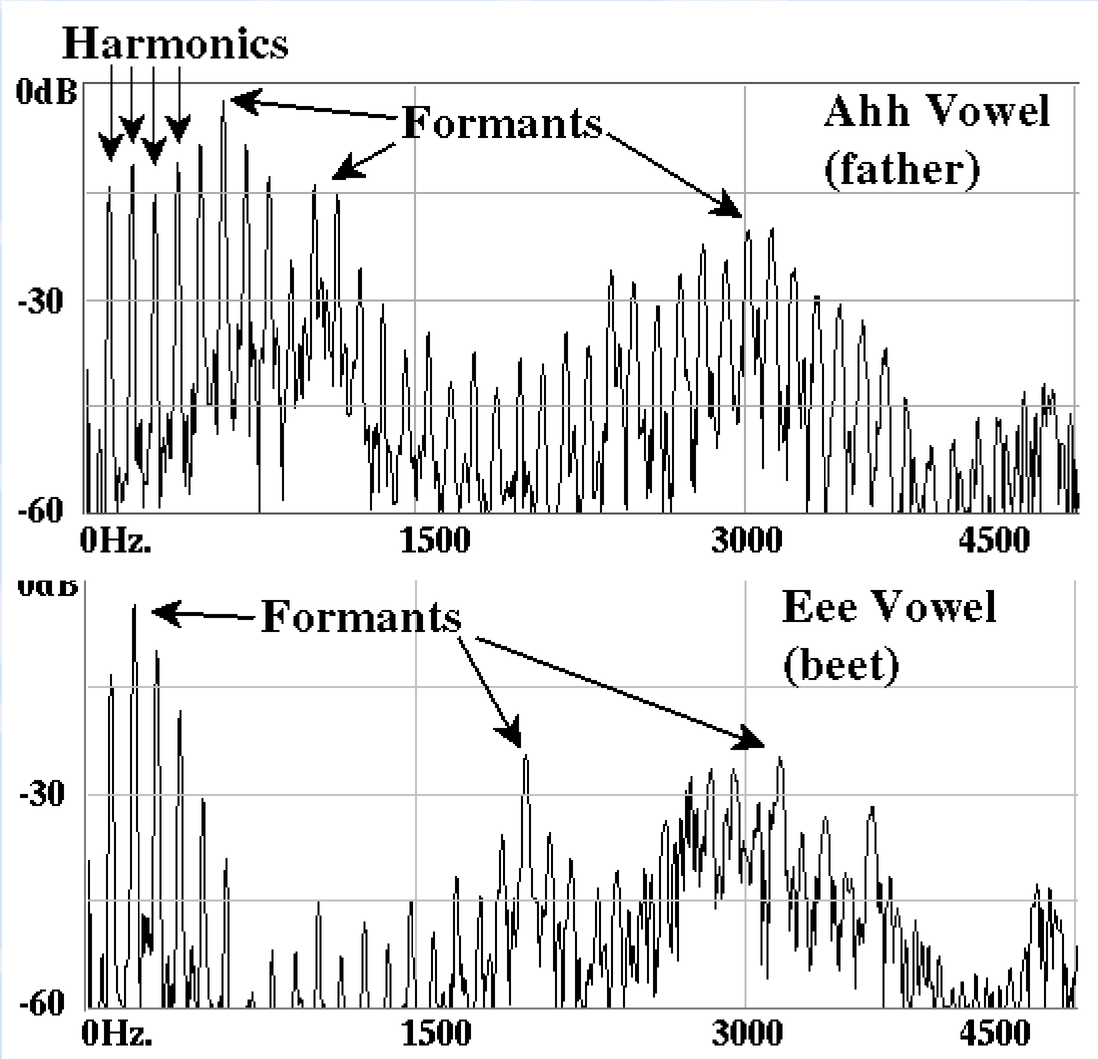
- Mixtures
  - (most real-
  - world sounds)





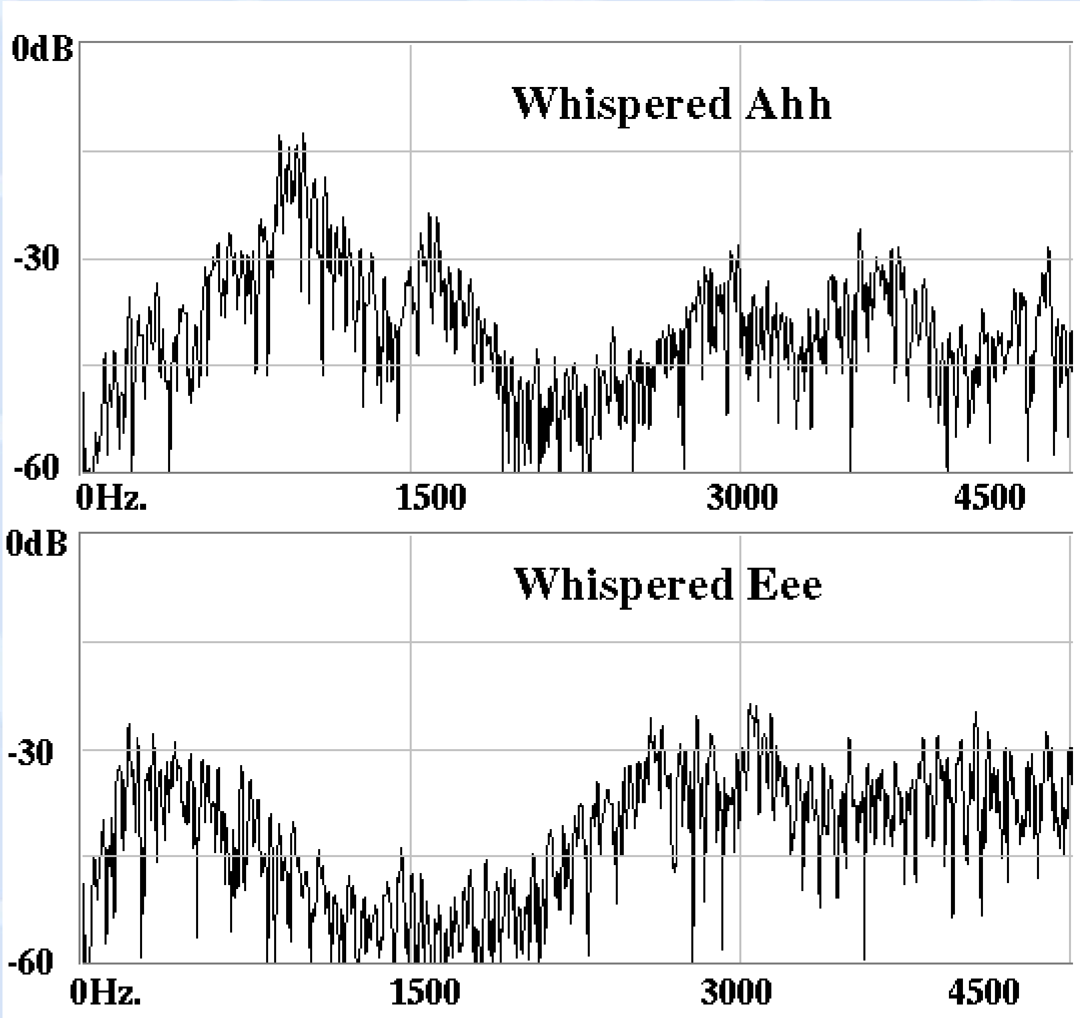
Perception: Spectral Shape

- Formants (resonances) are peaks in spectrum.
- Human ear is sensitive to these peaks.



Spectral Shape and Timbre

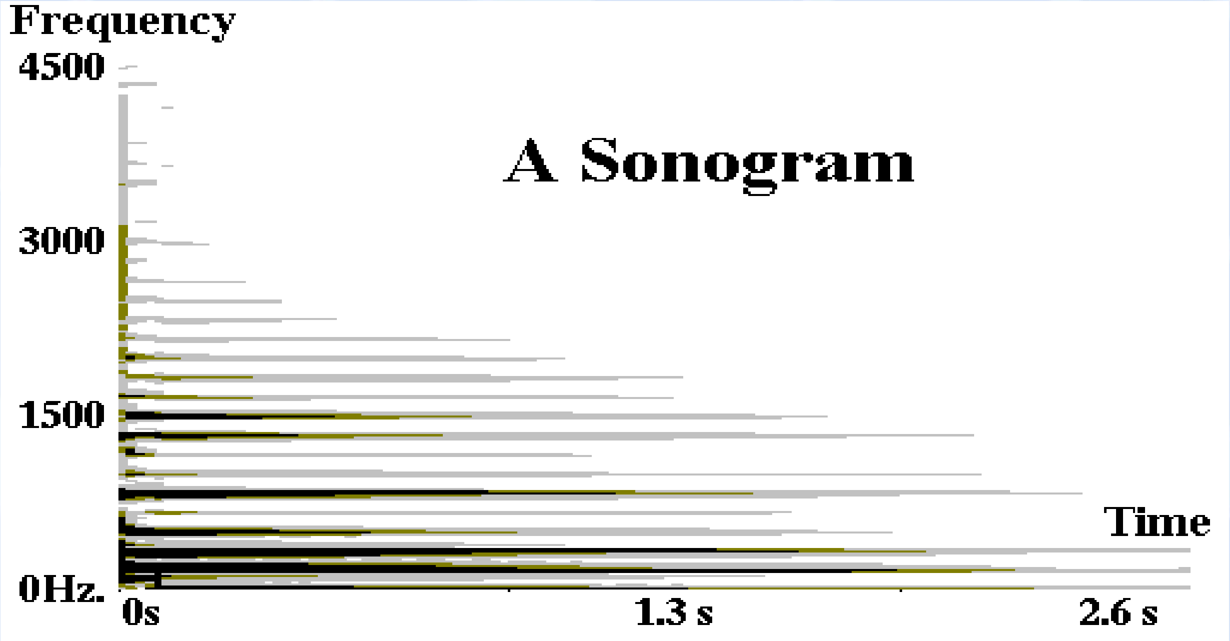
- Quality of a sound is determined by many factors
- Spectral shape is one important attribute



Spectra Vary in Time

- Spectrogram (sonogram)

  amplitude as darkness (color)
  vs. frequency and time

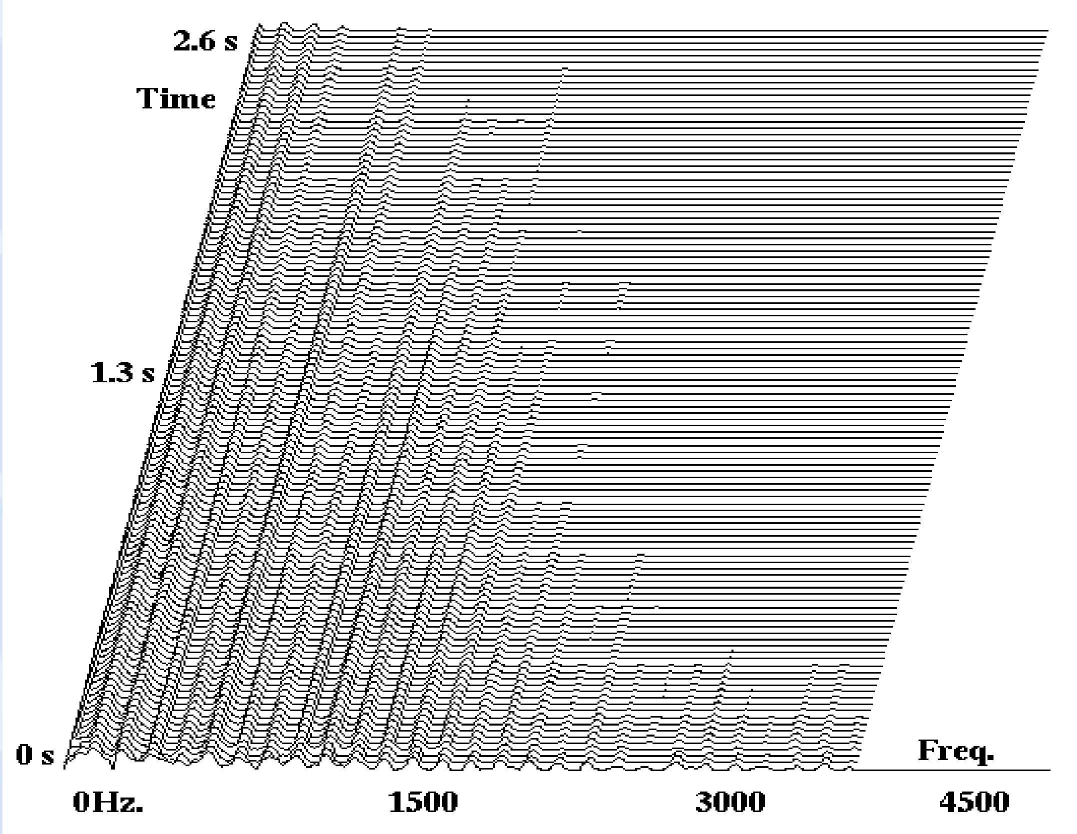


Spectra in Time (cont.)

- Waterfall Plot

pseudo 3-d amplitude as height vs. freq. and time

- Each horizontal slice is an amplitude vs. time magnitude spectrum



sndpeek demo

Sound Perception

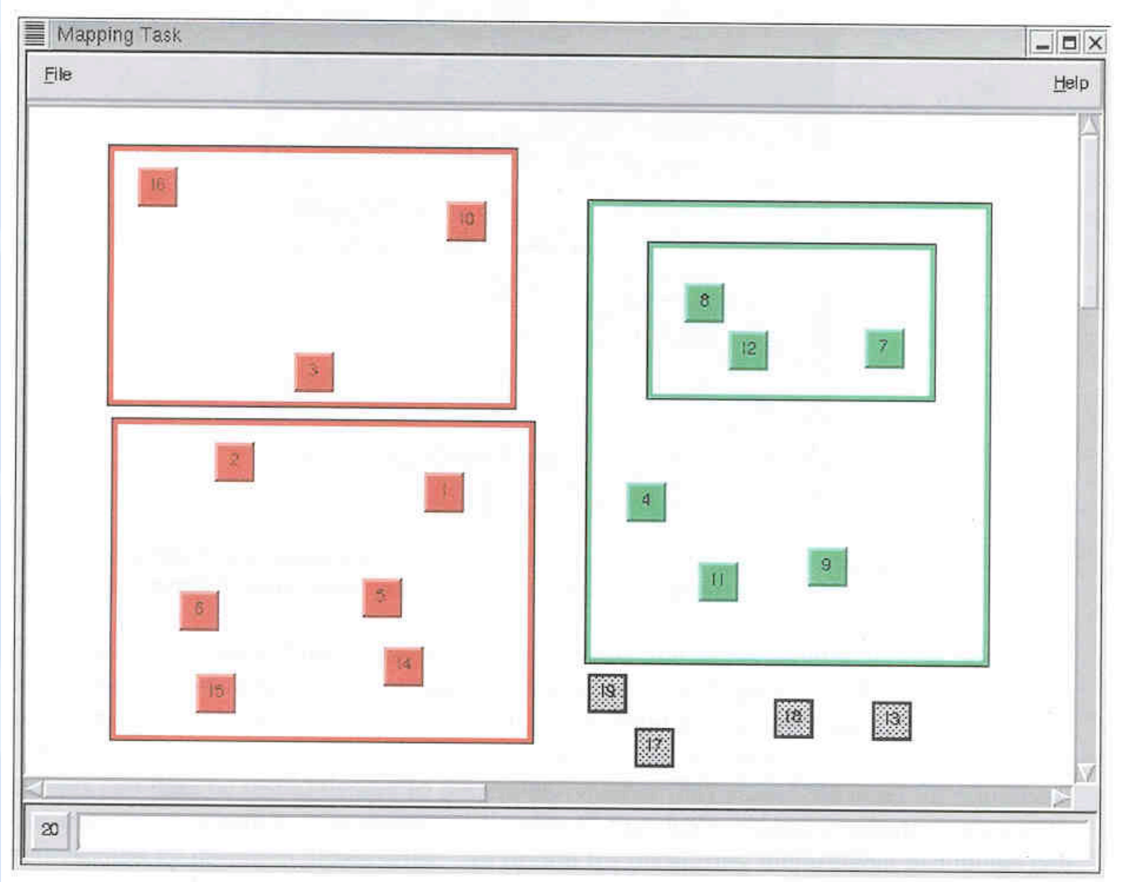
SIGGRAPH 2003
SAN DIEGO

- What are human mechanisms for identifying sounds?
- How do humans classify sounds as to similarity, difference, quality, etc.?
- If the auditory system doesn't care, we might not need to compute it.
- (How) does sound interact with other sensory modalities?
- How can we say it sounds "right,", "real," "good," "effective," etc.

Perception

Clustering and categorization of sound effects

(with Lakatos, Scavone, Harbke)

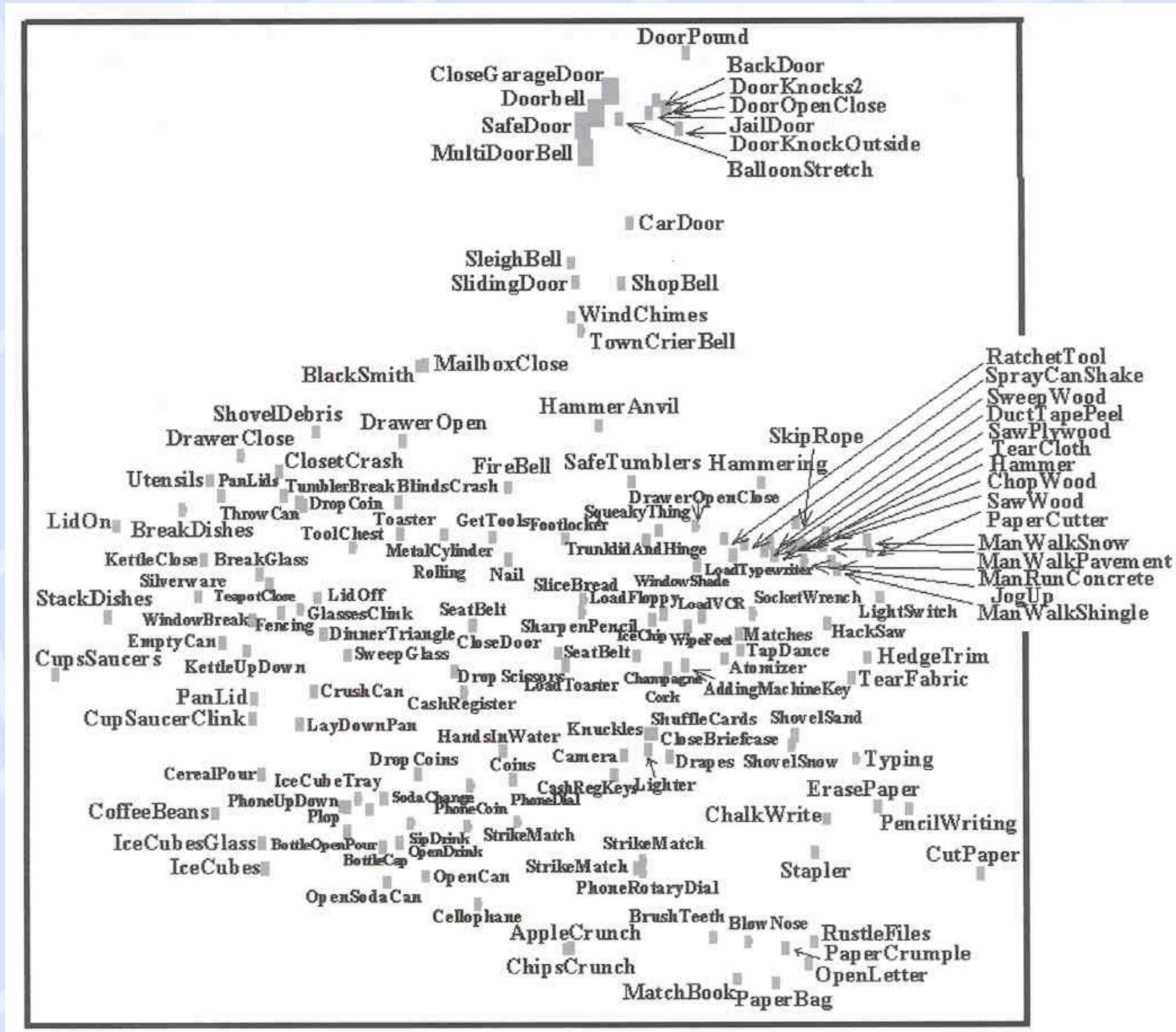


The Sonic Mapper

Clustering Results

MDS matches pair-wise

Ecological vs. abstract

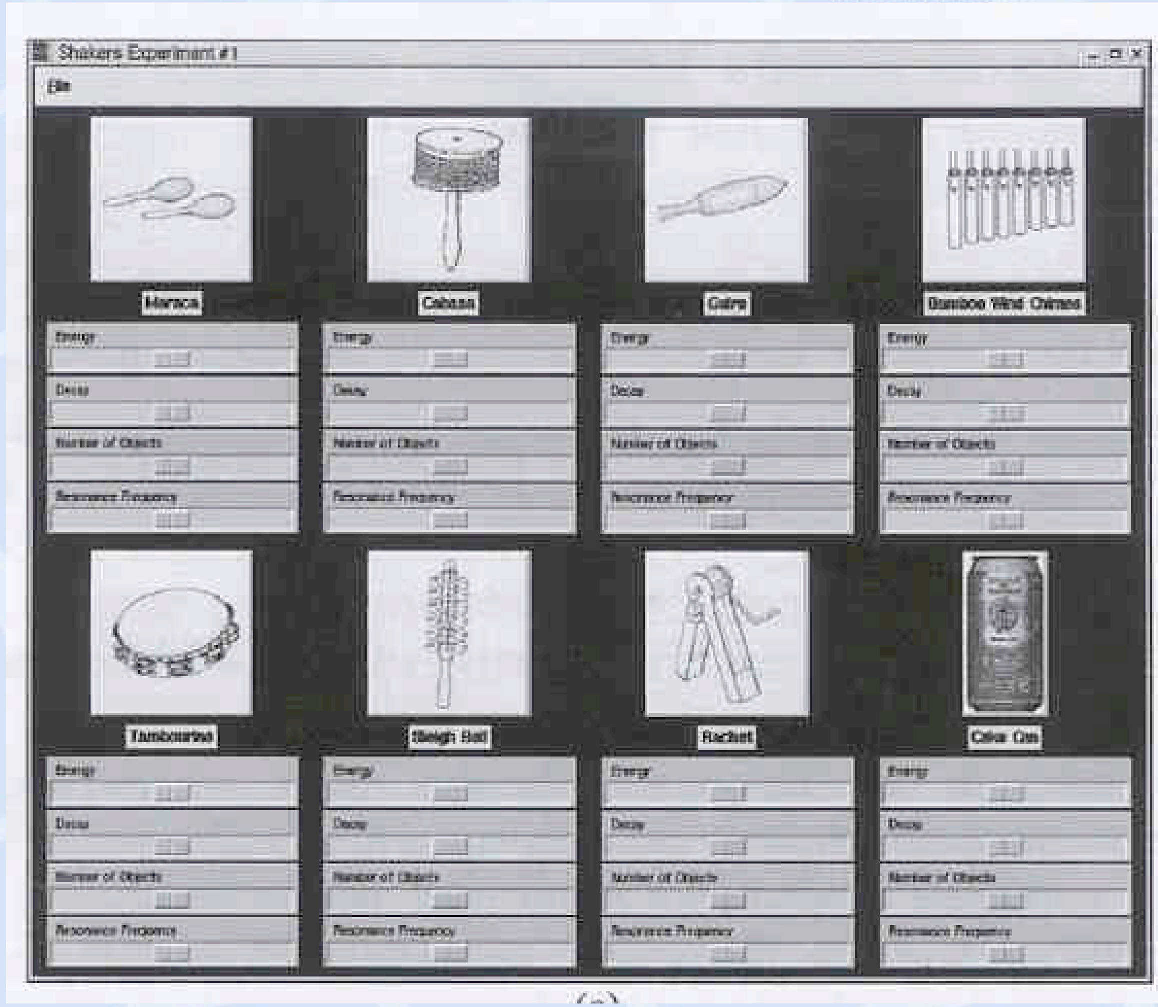


Perception

Learning by interacting with physical models

(with Lakatos, Scavone, Harbke)

Learning is proportional to structure of interface



PhISEM interface

Machine "perception"

SIGGRAPH 2003
SAN DIEGO

- Low level audio features
  - Power (loudness), sometimes/not
  - Spectral Centroid (brightness)
  - Spectral Rolloff (tilt, shape)
  - Zero Crossings (a hack, but works) DEMO
  - Spectral Flux (Δ of adjacent spectra)
  - Minimum Energy (% silence)
  - Means and standard deviations of all these

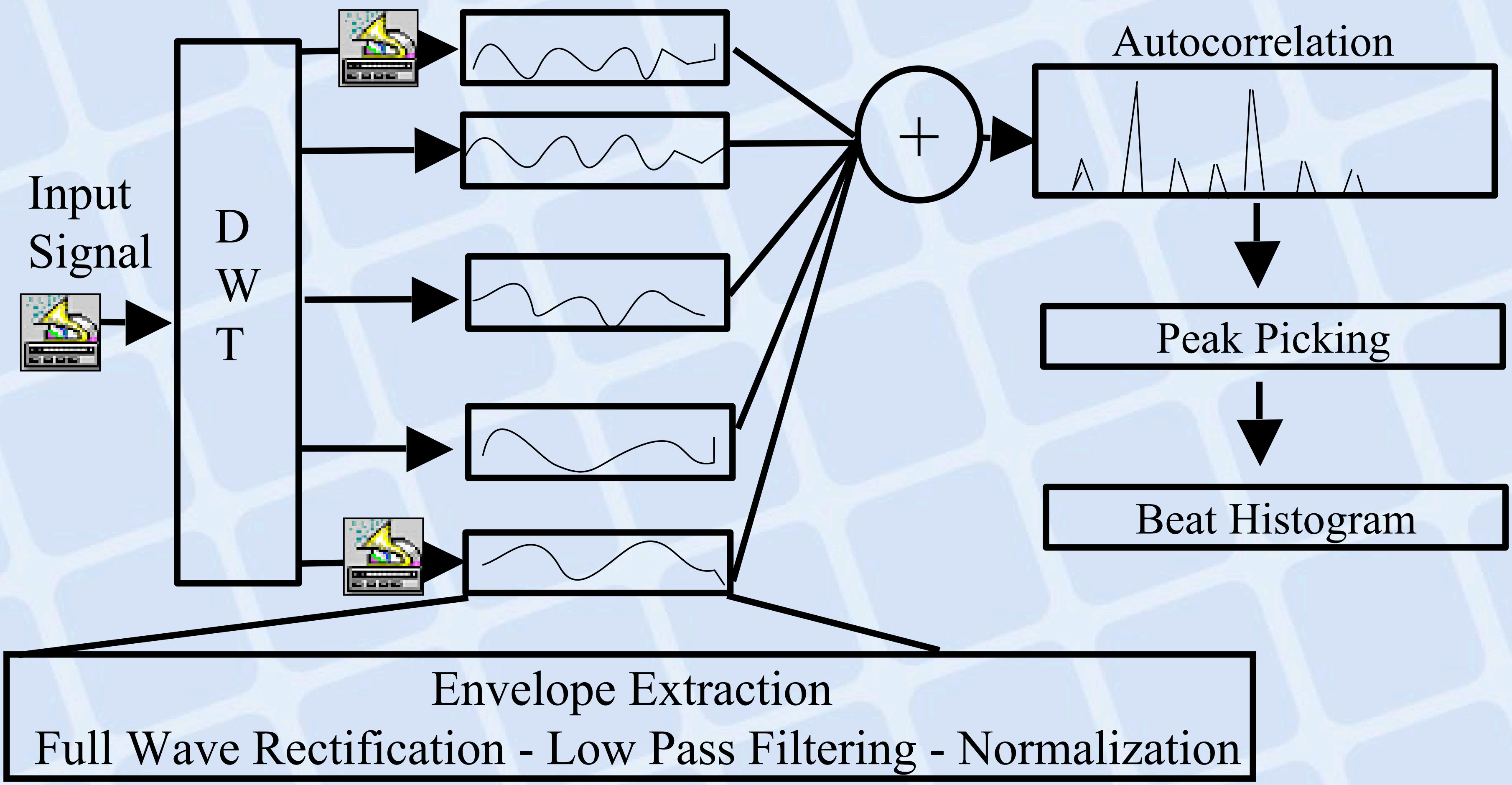
Machine perception 2

- Higher level features:
  - Mel Frequency Cepstral Components
  - Multi-band time periodicity (rhythm) the “beat histogram”
  - Pitch histogram
- “Cognitive” level features:
  - Style, Genre, Scene, Situation, …
- Multi-Resolution
  - Short “event” windows
  - Longer “texture” windows

Wavelet-based Rhythm Analysis

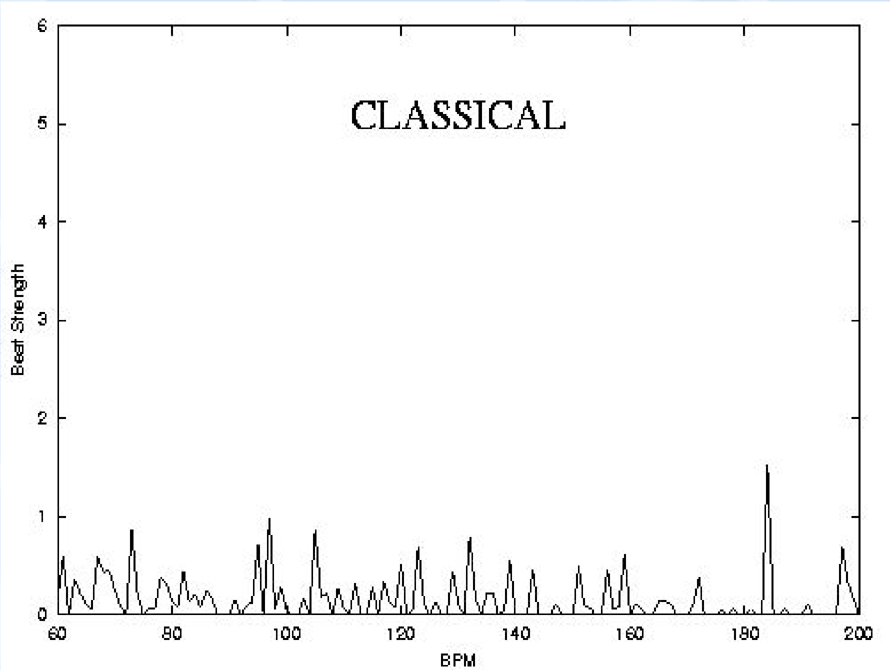
Tzanetakis et al AMTA01
Goto, Muraoka CASA98
Foote, Uchihashi ICME01
Scheirer JASA98

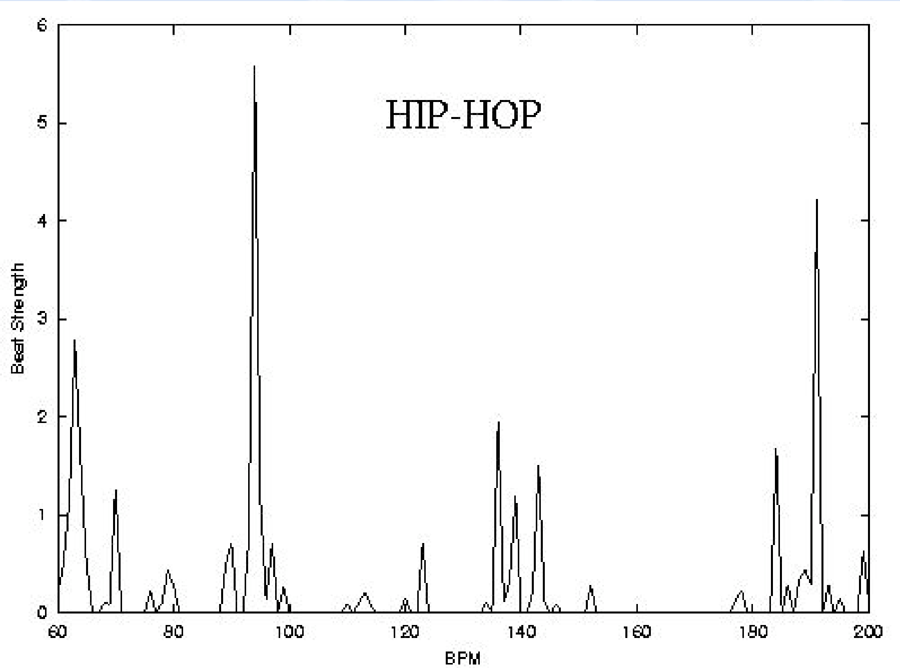
SIGGRAPH 2003 SAN DIEGO



Beat Histograms

Tzanetakis et al AMTA01

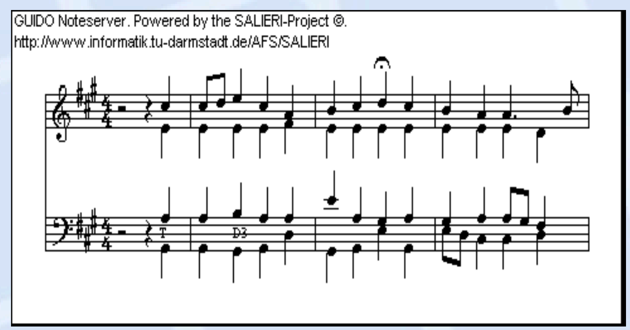




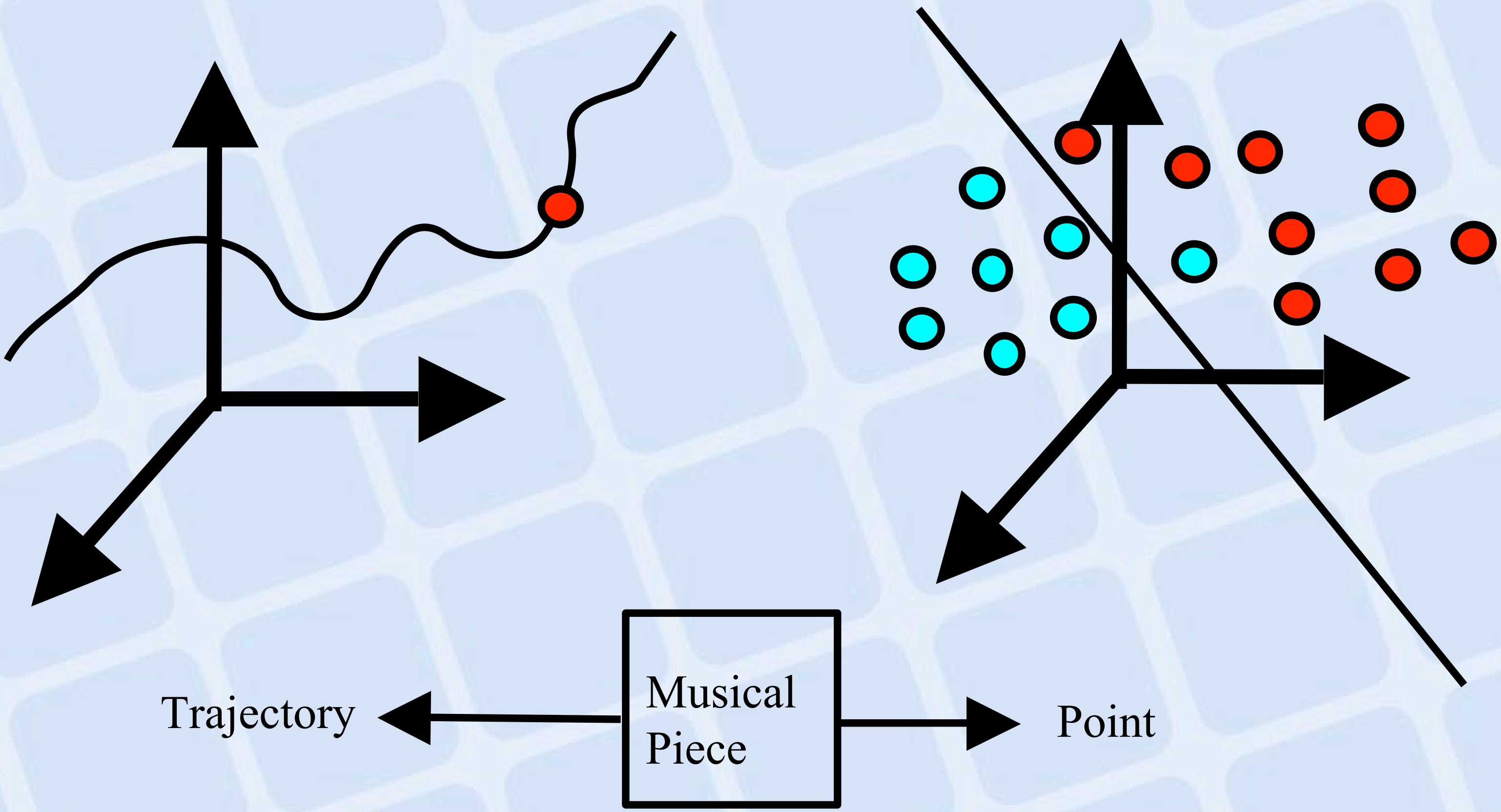
Musical Content Features

- Timbral Texture (19)
  - Spectral Shape
  - MFCC (perceptually motivated features, ASR)

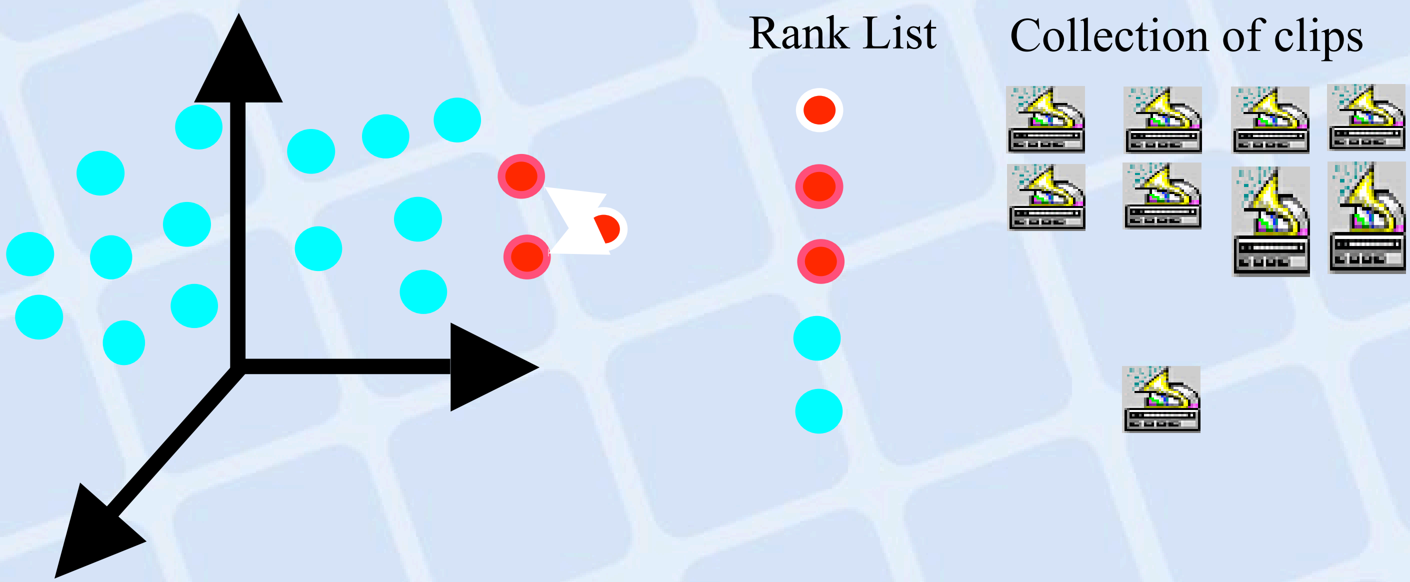
- Rhythmic structure (6)
  - Beat Histogram Features
- Harmonic content (5)
  - Pitch Histogram Features



Understanding



Query-by-Example Content-based Retrieval



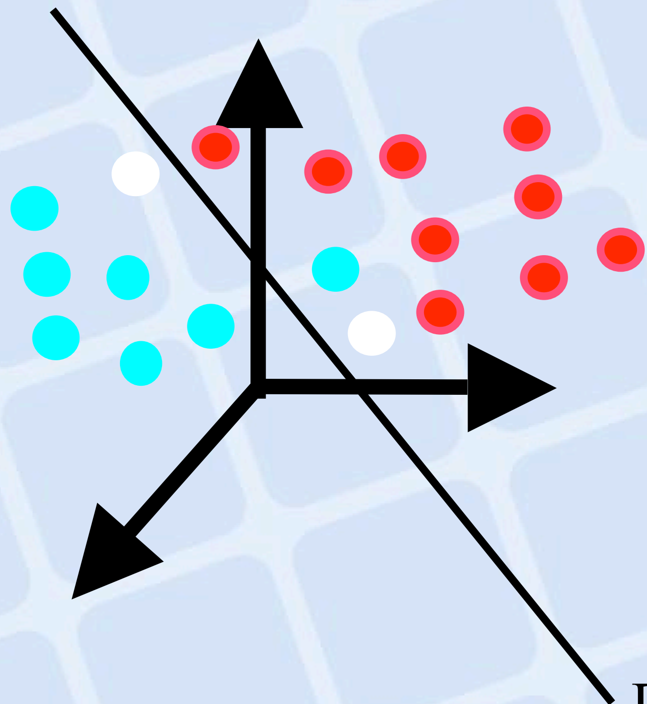
Demo

Automatic Musical Genre Classification

SIGGRAPH 2003 SAN DIEGO

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

Statistical Supervised Learning

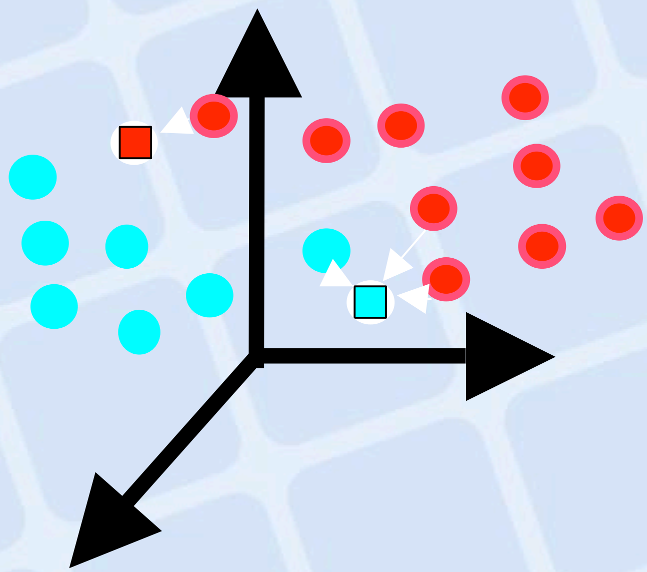


Partitioning of feature space

$$P(\blacksquare \mid \bullet) = \frac{p(\bullet \mid \blacksquare) * P(\blacksquare)}{p(\bullet)}$$

- ■ Music
- ■ Speech

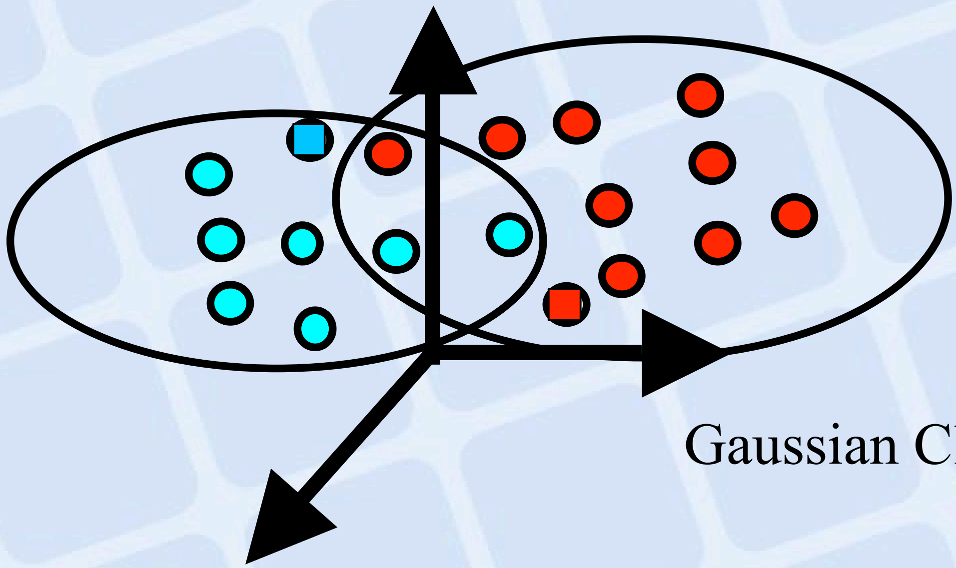
Non-parametric classifiers

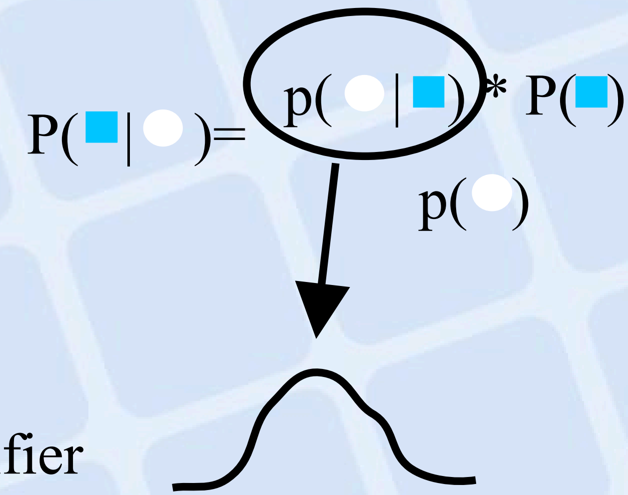


$$P(\square|\circ) = \frac{p(\circ|\square) * P(\square)}{p(\circ)}$$

Nearest-neighbor classifiers (K-NN)

Parametric classifiers





Gaussian Classifier

Gaussian Mixture Models

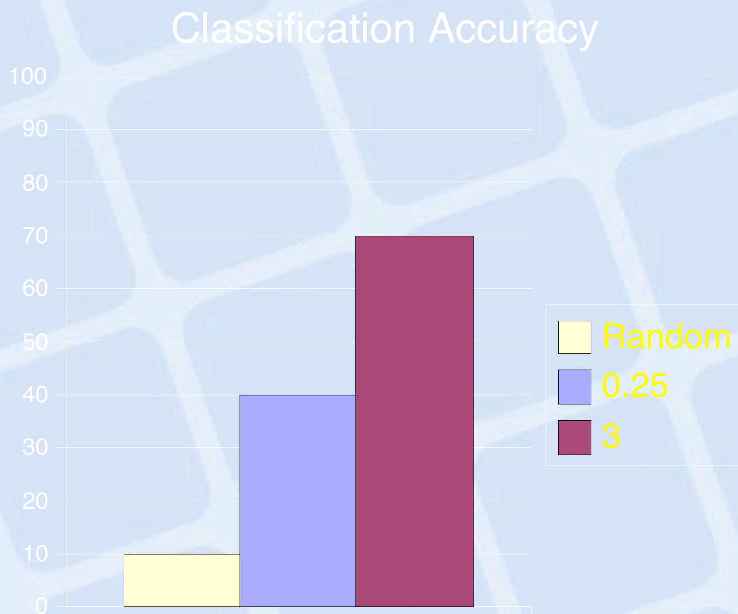
Classification Evaluation – 10 genres

Manual (52 subjects)

Perrot & Gjerdingen, M.Cognition 99

0.25 seconds 40%

3 seconds 70%

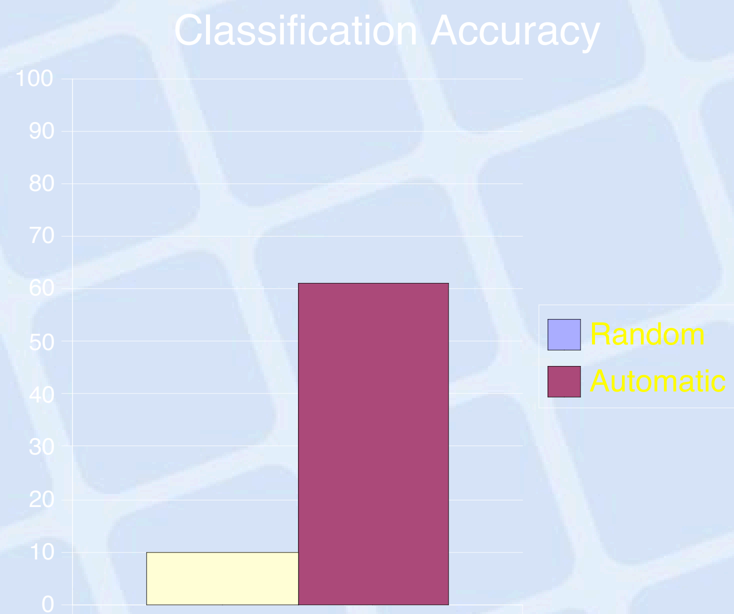


Automatic (different collection)

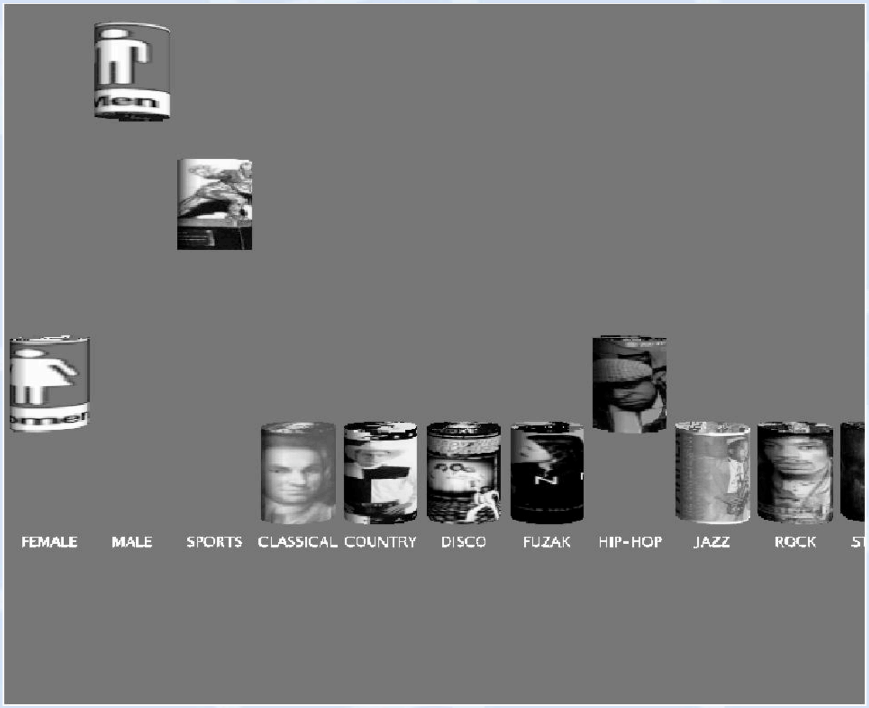
Tzanetakis & Cook, TSAP 10(5) 2002

Gaussian Mixture Model (GMM)

10-fold cross-validation 61% (70%)



GenreGram DEMO

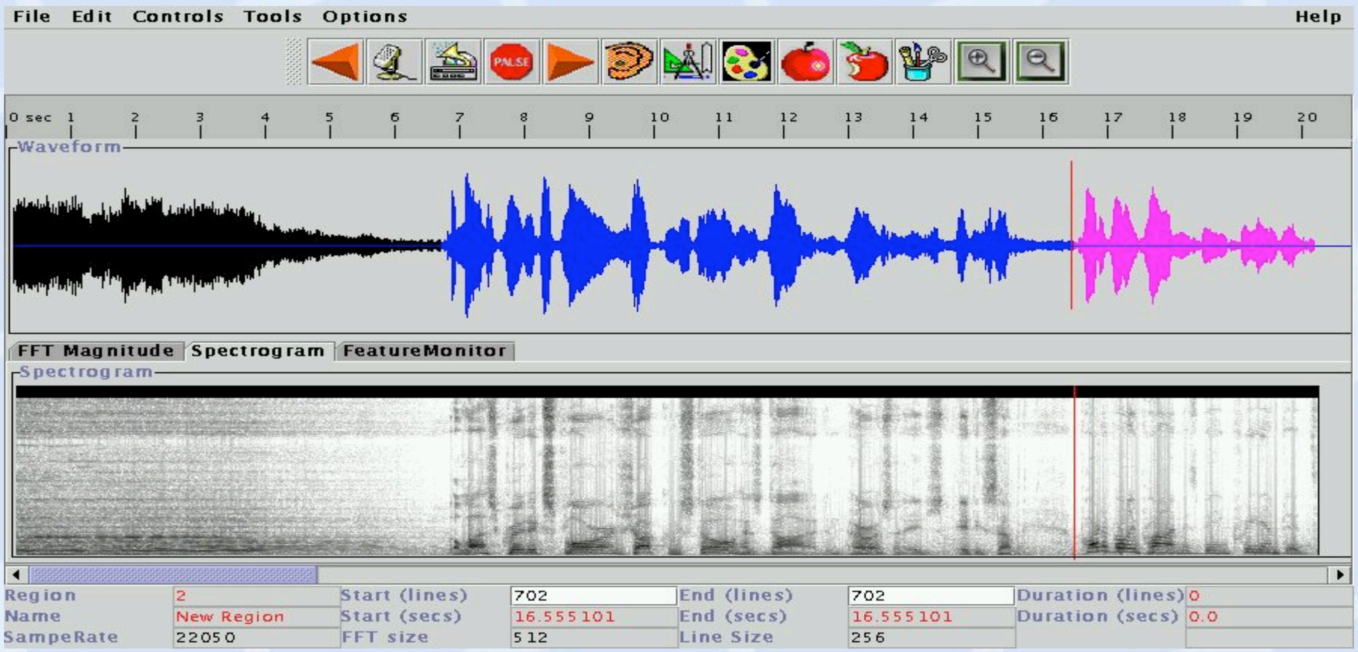


Dynamic real time 3D display
for classification of radio signals

Audio Segmentation

SIGGRAPH 2003 SAN DIEGO

- Segmentation = changes of sound "texture"



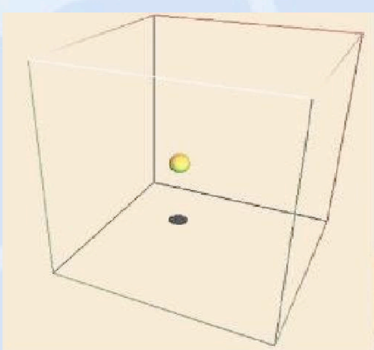
Music Male Voice Female Voice

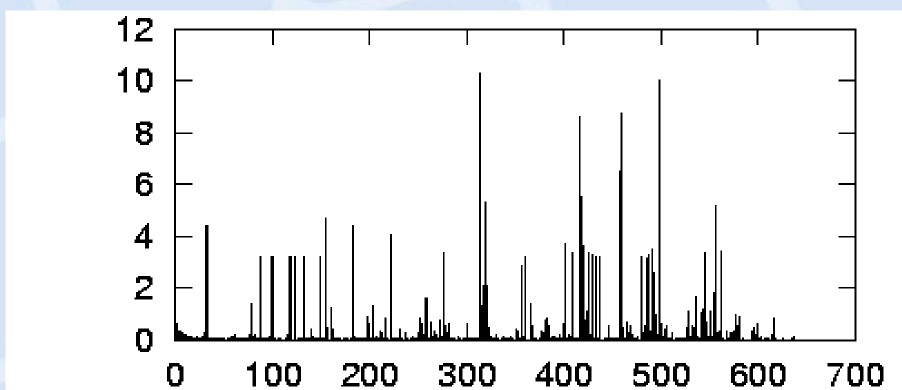
News:

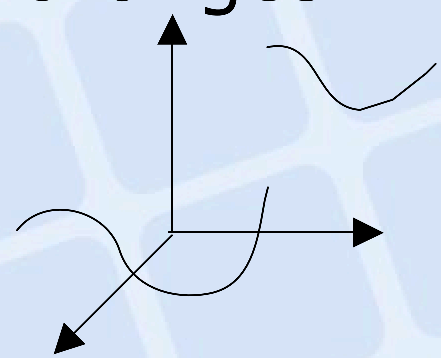
Multifeature Segmentation Methodology

Tzanetakis & Cook, WASPAA 99

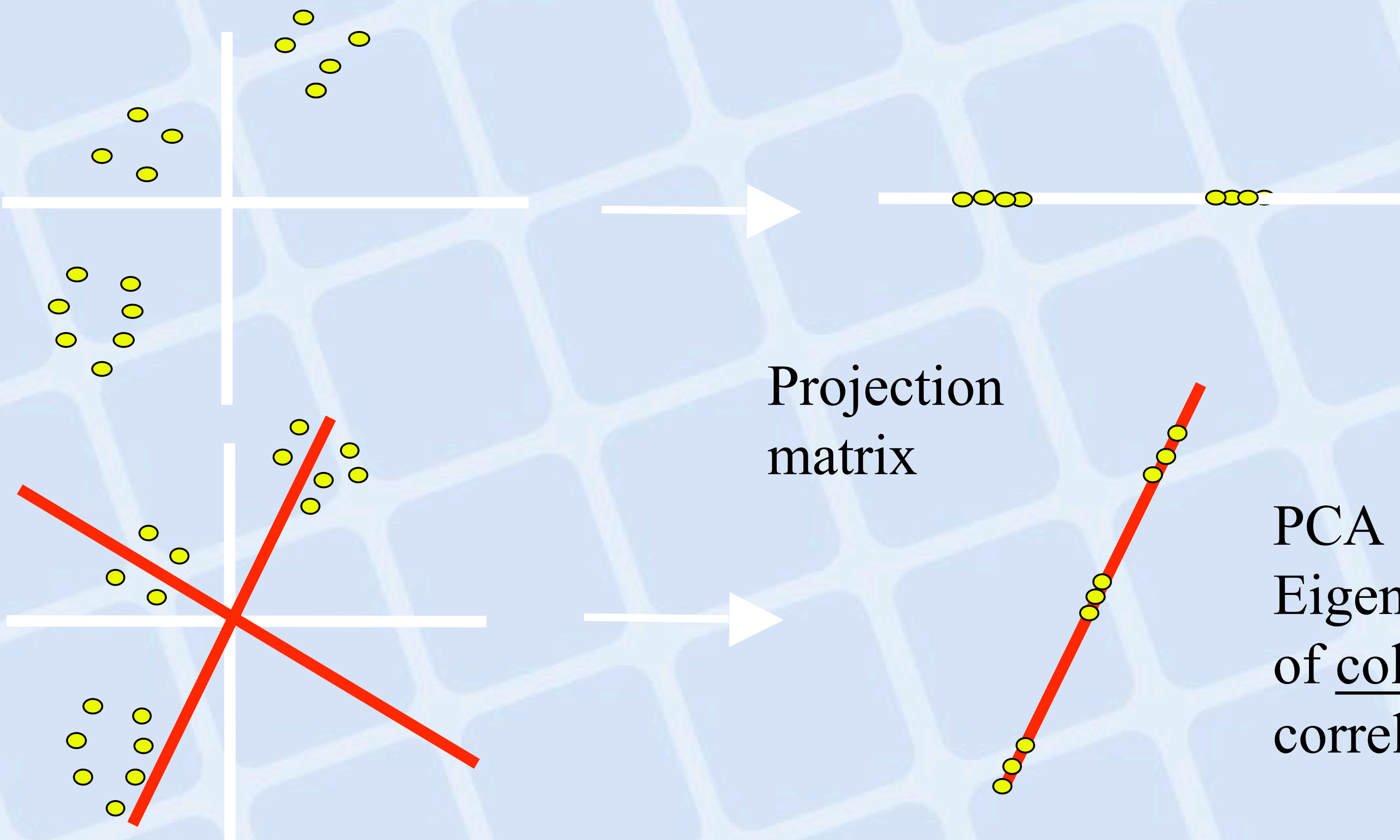
- Time series of feature vectors V(t)
- f(t) = d(V(t), V(t-1))
  - $D(x,y) = (x-y)C^{-1}(x-y)^t$ **(Mahalanobis)**
- df/dt peaks correspond to texture changes







Principal Components Analysis



Projection matrix

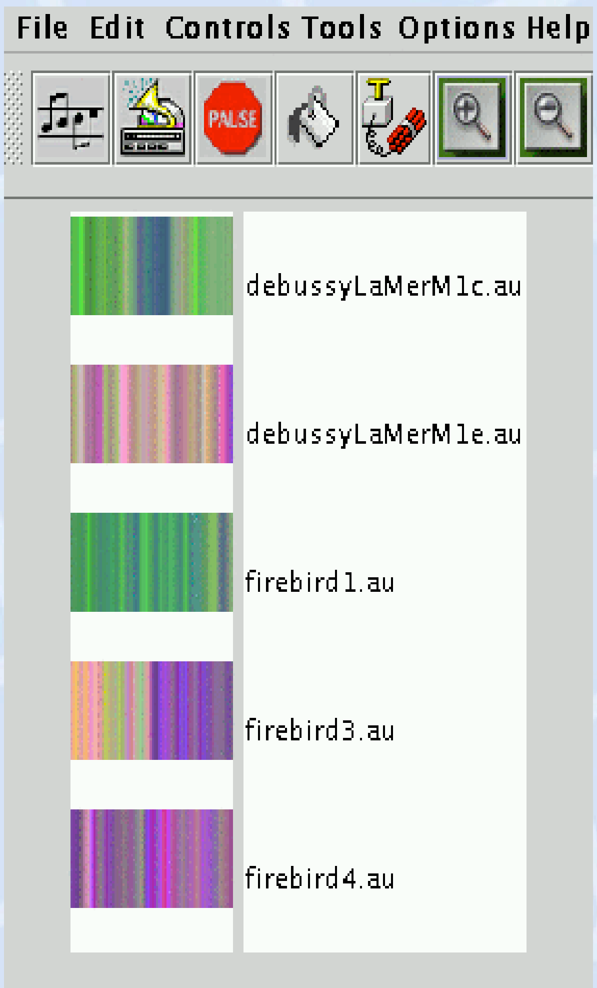
PCA Eigenanalysis of collection correlation matrix

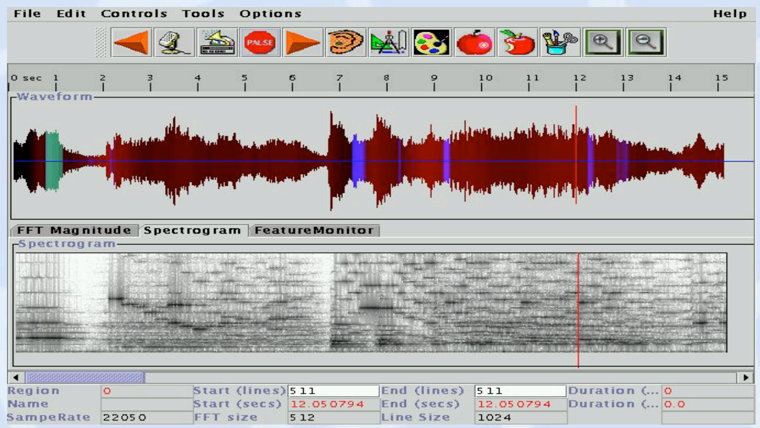
Timbregrams and Timbrespaces

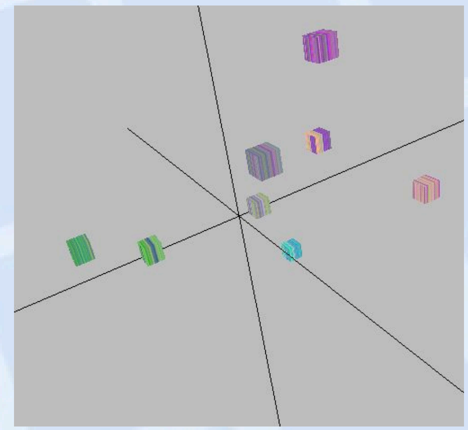
SIGGRAPH 2003
SAN DIEGO

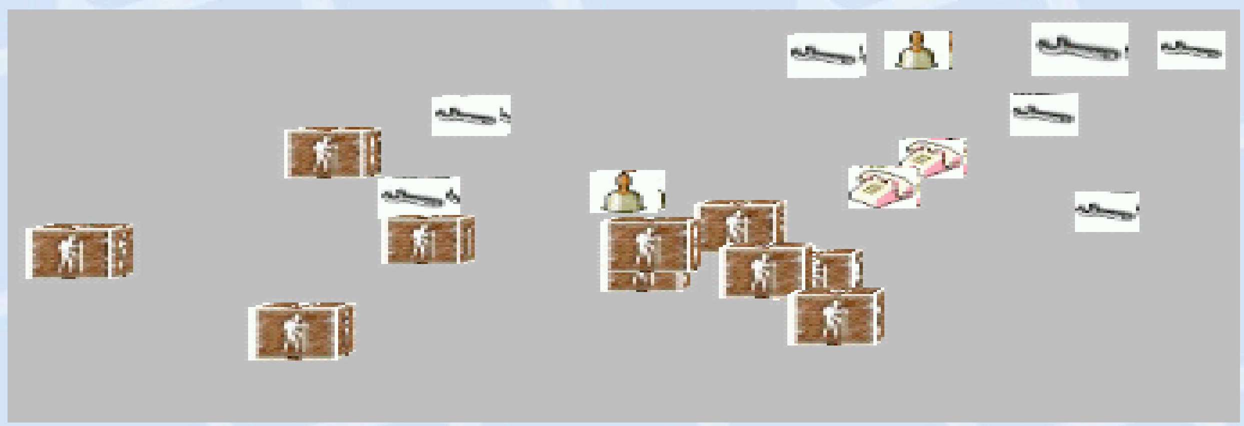
Tzanetakis & Cook DAFX00, ICAD01

PCA = content & context

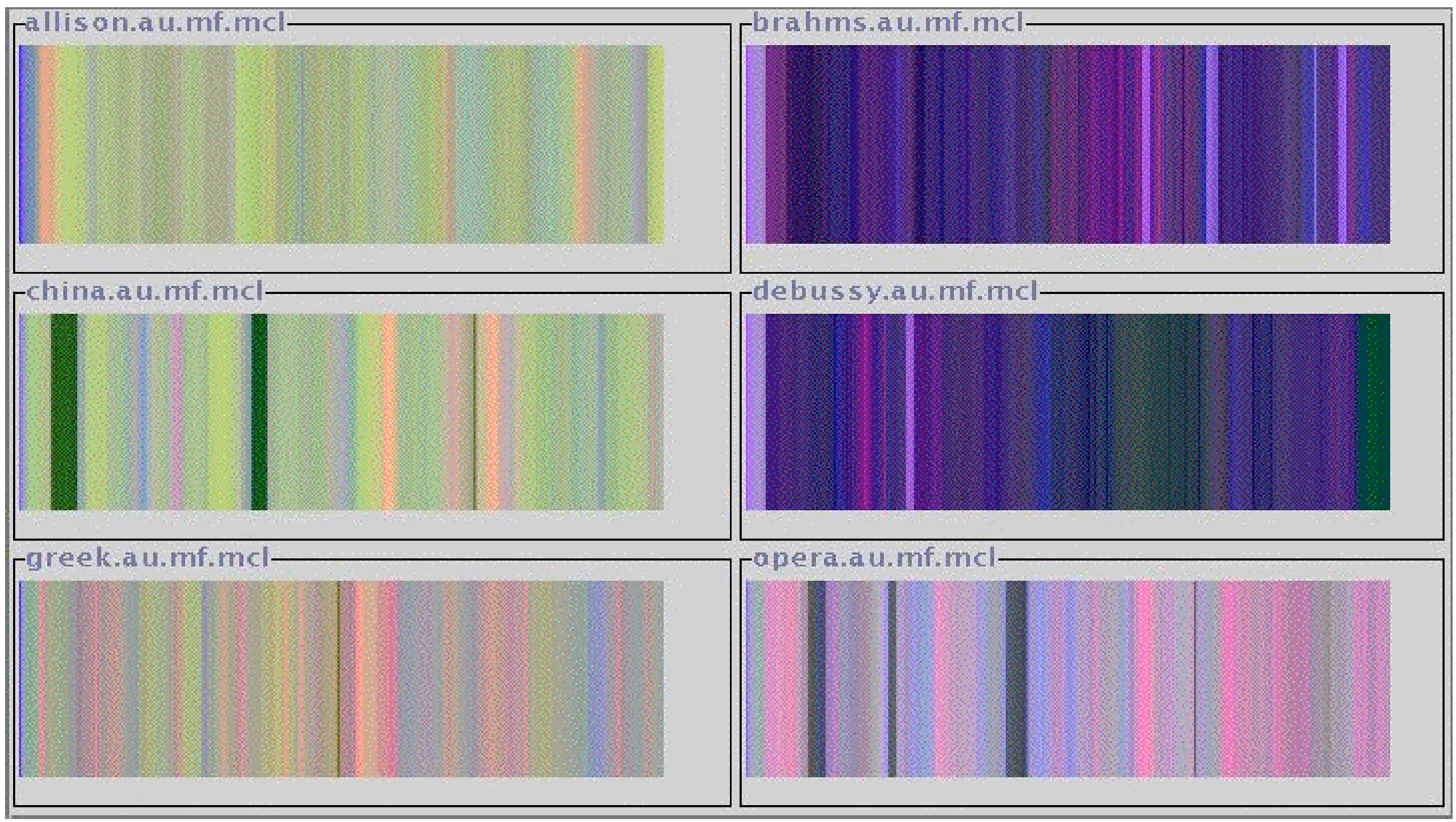






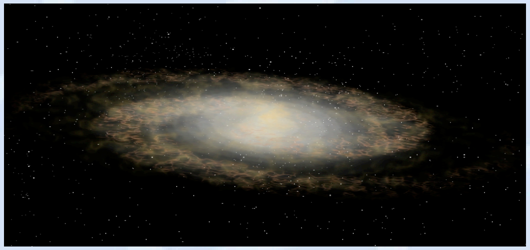


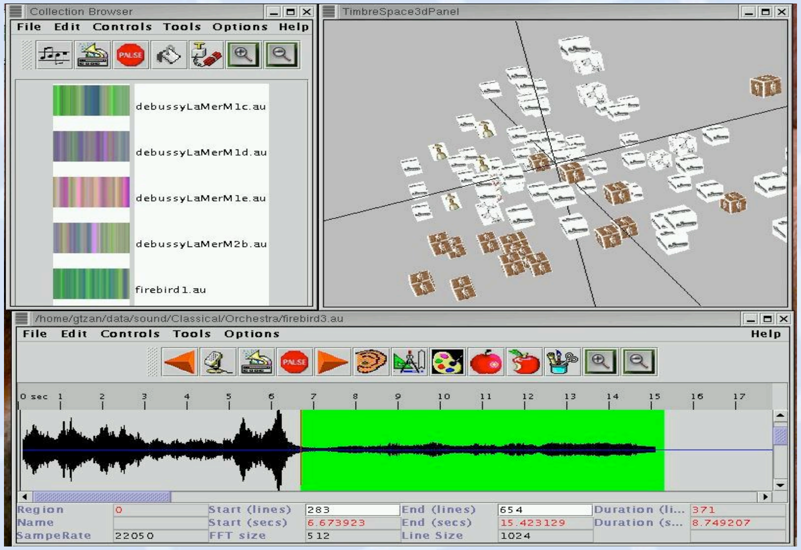
Timbregram Classes

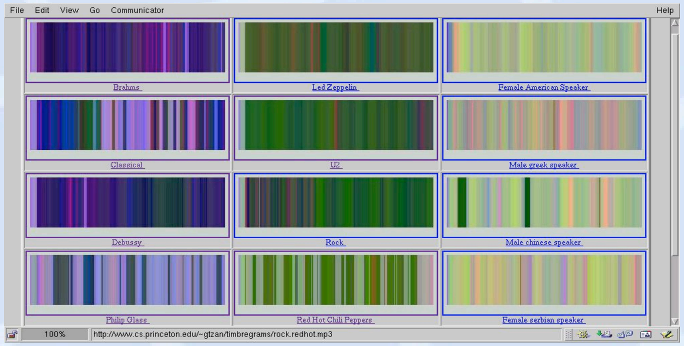


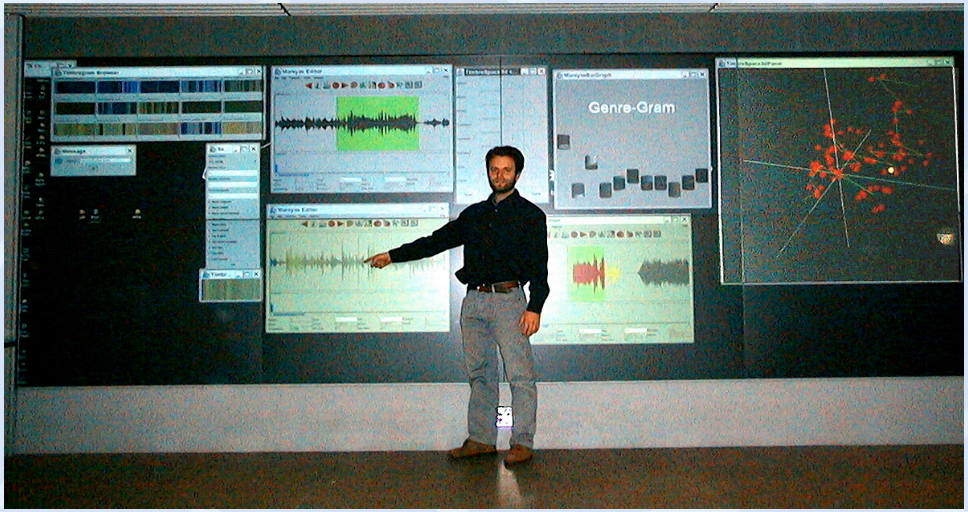
Speech (different languages) **Music (orch, or opera (lower))**

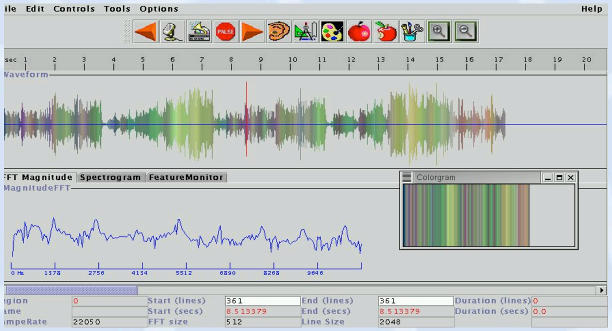
Integration











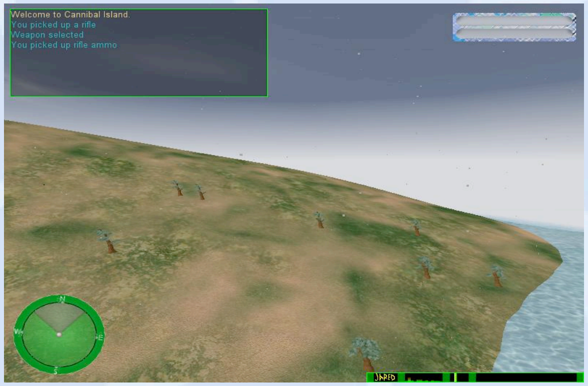
Implementation

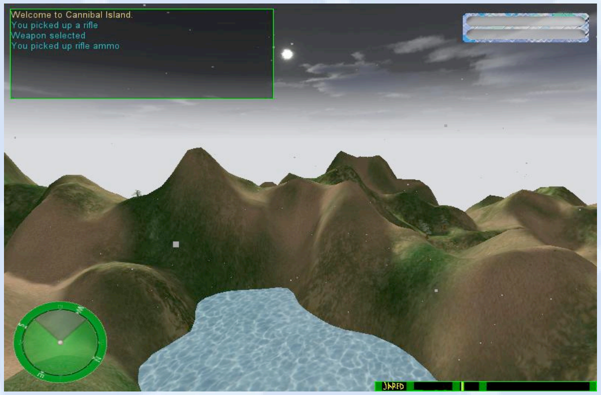
Tzanetakis & Cook Organized Sound 4(3) 00

- *MARSYAS* : free software framework for computer audition research
  - Server in C++ (numerical signal processing and machine learning)
  - Client in JAVA (GUI)
  - Linux, Solaris, Irix and Wintel (VS , Cygwin)
- Apr. 2004, 5500 downloads, 2300 different hosts, 30 countries since March 2001
- Recent ISMIR conference, 80% citations, and 65% users

Marsyas users

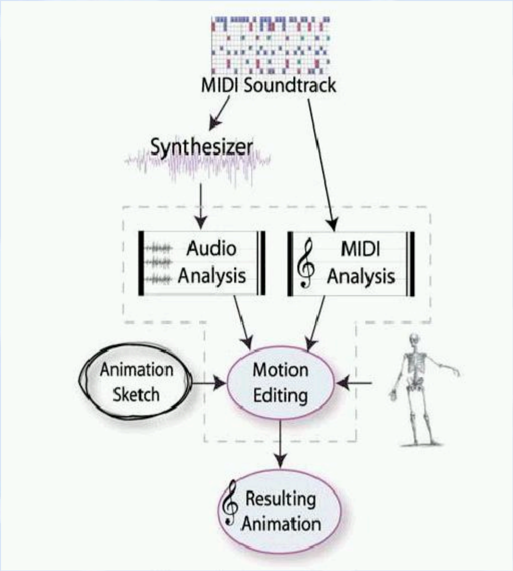
SIGGRAPH 2003
SAN DIEGO





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

What we can(‘t) do

What we can(‘t) do

SIGGRAPH 2003
SAN DIEGO

- Identify Genres
- Identify scenes, situations
- Speaker/singer identification
- Query

- Separate sounds (polyphony)
- Model high level human ranking
- “understand”