ISMIR 2002 Tutorial: Music Information Retrieval for Audio Signals

George Tzanetakis
PostDoctoral Fellow
Computer Science Department
Carnegie Mellon University

gtzan@cs.princeton.edu
http://www.cs.cmu.edu/~gtzan

MIR Music History

9000 BC

1000

1700

1877

1960

2000

Music

- 4 million recorded CDs
- 4000 CDs / month
- MP3 Bandwidth %
- Global
- Pervasive
- Why ?

The future of MIR

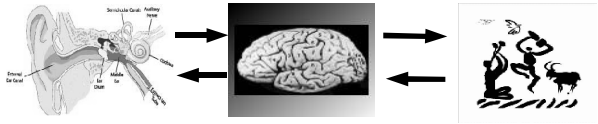
- Database of all recorded music
- Tasks: organize, search, retrieve, classify recommend, browse, listen, annotate
- Examples:

Audio MIR Pipeline

Hearing
Representation

Understanding
Analysis

Acting
Interaction



Signal Processing

Machine Learning

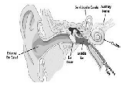
Human Computer
Interaction

Tutorial Goals

- Overview of state of the art
- Guide to bibliography
- Fundamentals
- Technical Background
  - Some math, computer science, music
- Link audio MIR to symbolic MIR
- Understand all ISMIR papers

Outline

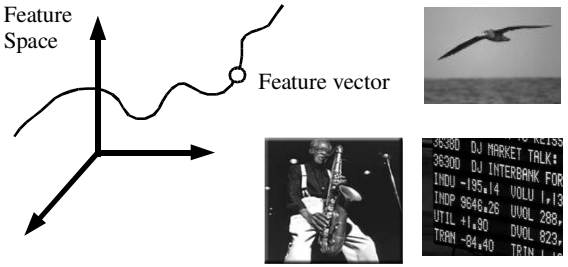
- Representation 50 min
- Analysis 50 min
- Interaction 50 min
- Discussion 30 min
- ISMIR 5 days
- MIR Research years



Representation Outline

- Overview
- Timbral Features
  - STFT, DWT, LPC, MFCC, MP3
- Pitch Analysis
  - Autocorrelation, sinusoids, transcription
- Beat Analysis
  - Event-based, similarity-based, running, global

Feature extraction

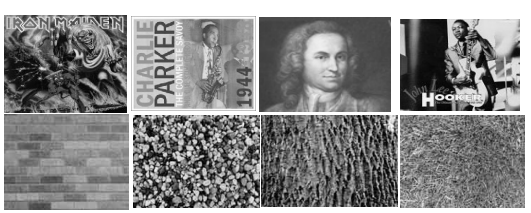


Timbral Texture

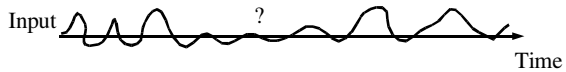
Timbre = differentiate sounds of same pitch and loudness

Timbral Texture = differentiate mixtures of sounds (possibly with the same or similar rhythmic and pitch content)

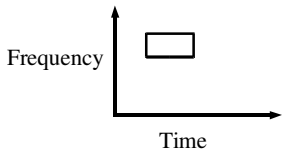
Global, statistical and fuzzy properties



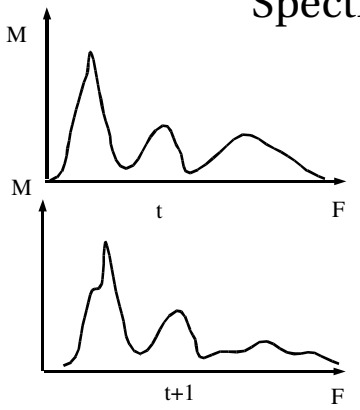
Time-domain waveform



Decompose to building blocks



Spectrum



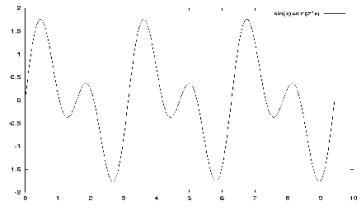
Fourier Transform

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

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\omega)\, e^{-i\omega t}\, dt$$

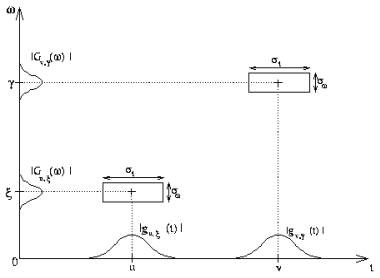
$$f(\omega) = \int_{-\infty}^{\infty} f(t)\, e^{i\omega t}\, dt$$

$$e^{i\theta} = \cos(\theta) + i * \sin(\theta)$$



Short Time Fourier Transform I

FT = global representation of frequency content



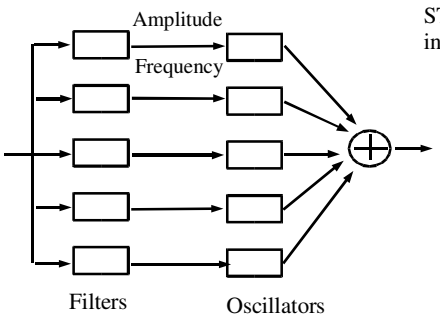
$$Sf(u,\omega) = \int_{-\infty}^{\infty} f(t)\, g(t-u)\, e^{-i\omega t}\, dt$$

Time – Frequency

L2 Heisenberg uncertainty

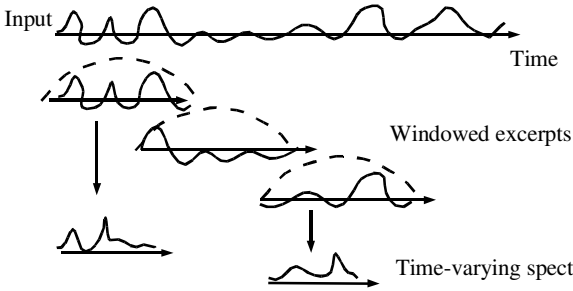
$$\sigma_t \sigma_\omega \geq 1/4$$

Short Time Fourier Transform II



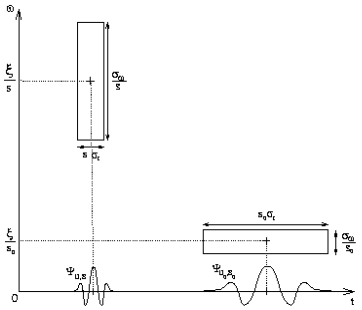
STFT filter-bank interpretation

Short Time Fourier Transform III



Fast Fourier Transform FFT

Wavelets



$$Wf(u,s) = \int_{-\infty}^{\infty} f(t)\,\frac{1}{s}\,\psi^{*}\left(\frac{t-u}{s}\right)$$

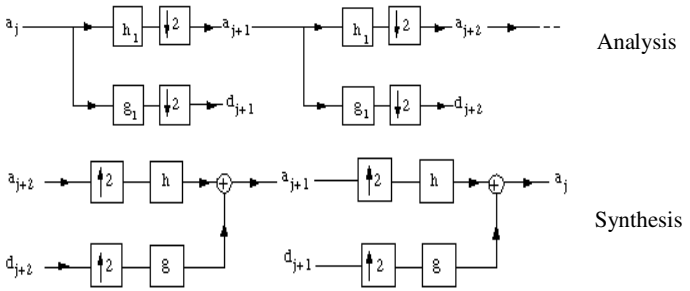
Time – Scale

L2 Heisenberg uncertainty

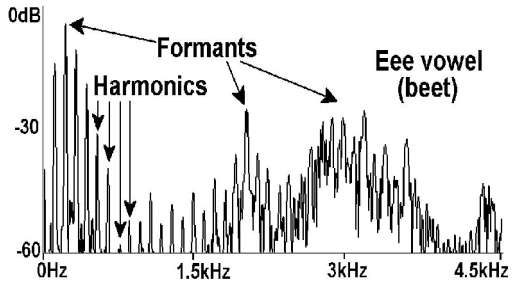
$$\sigma_t \sigma_\omega \geq 1/4$$

The Discrete Wavelet Transform

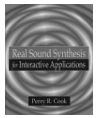
Octave filterbank



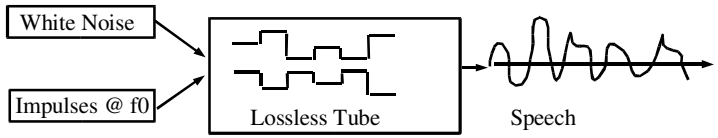
Formants



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



Linear Prediction Coefficients



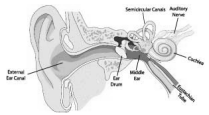
Source

Filter

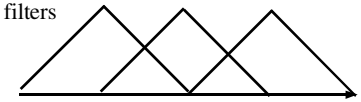
$$s'_n = \sum_{i=1}^{p} a_i s_{n-1}$$

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

Mel Frequency Cepstral Coefficients



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

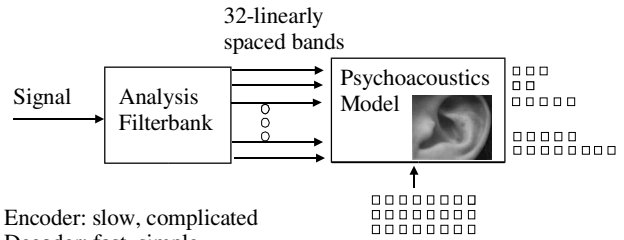


CF-130 CF CF+130
CF / 1.0718 CF * 1.0718

Log base 10
Discrete Cosine Transform

Short MPEG Audio Coding Overview (mp3)

MPEG Perceptual Audio Coding

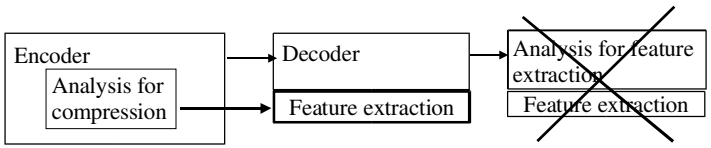


Encoder: slow, complicated
Decoder: fast, simple

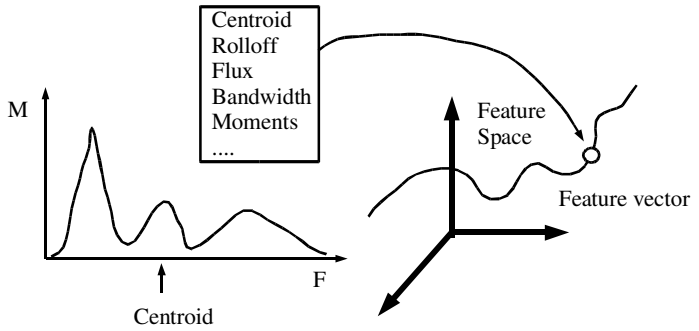
MP3 Feature Extraction

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

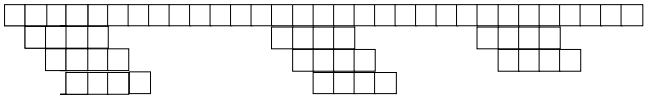


Spectral Shape



Analysis and Texture Windows

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



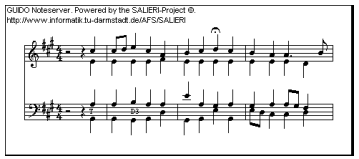
Analysis windows ☐ ☐ ☐ ☐ 20 milliseconds

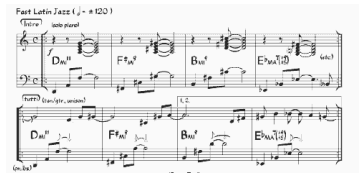
Texture windows ☐☐☐ ☐☐☐ ☐☐☐ 40 analysis windows

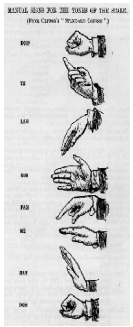
Summary of Timbral Texture Features

- Time-Frequency analysis
- Signal processing (STFT, DWT)
- Source-filter (LPC)
- Perceptual (MFCC, MPEG)
- Statistics over "texture" window
- Feature vector(s)

Traditional Music Representations





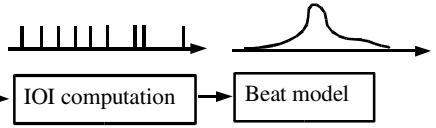


Rhythm

- 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



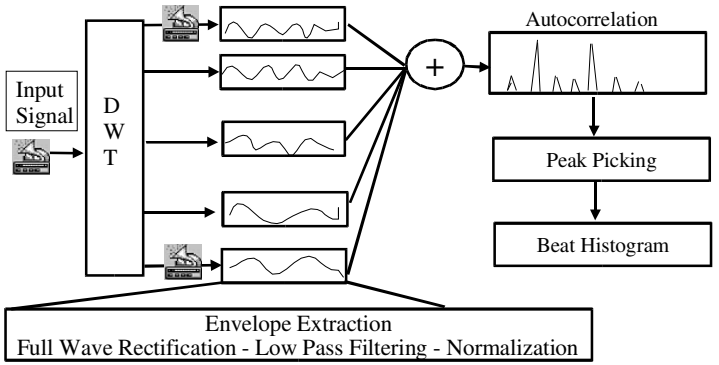
Single band
Multiple band
Thresholding
Chord changes

Successive onset pairs
All onset pairs
Quantization

Constant-tempo
Measure changes
Computational cost
Multiple hyphotheses

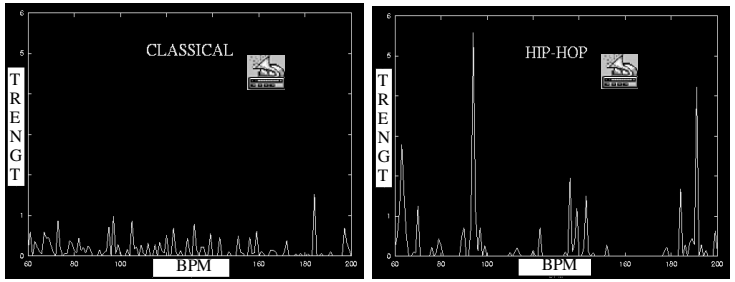
Self-similarity

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



Beat Histograms

Tzanetakis et al AMTA01



Beat Spectrum

Foote, Uchihashi 01

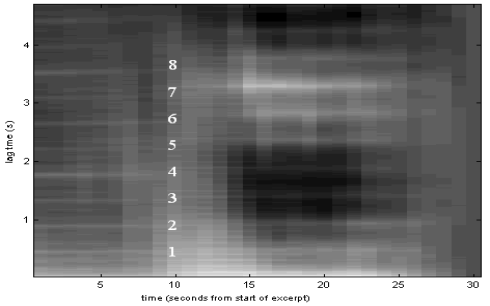


Figure 4. Beat spectrogram of Pink Floyd's *Money* (excerpt), showing transition from 4/4 to 7/4 time

Rhythmic content features

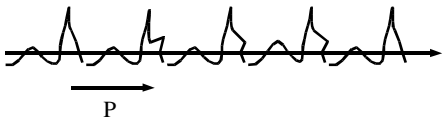
- Main tempo
- Secondary tempo
- Time signature
- Beat strength
- Regularity

Pitch content

- Harmony, melody = pitch concepts
- Music Theory Score = Music
- Bridge to symbolic MIR
- Automatic music transcription
- Non-transcriptive arguments

Split the octave to discrete logarithmically spaced intervals

Pitch Detection



Time-domain
Frequency-domain
Perceptual

Autocorrelation
Peaks at multiple of
the fundamental frequency

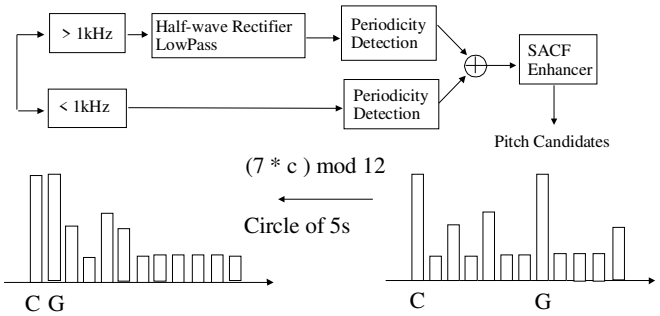
$$r_x = \sum_{n=0}^{N-1} x\ n\ x\ n+l\ ,l = 0,1,..\ L-1$$

ZeroCrossings

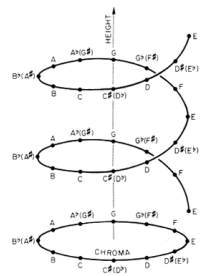
Multiple Pitch Detection

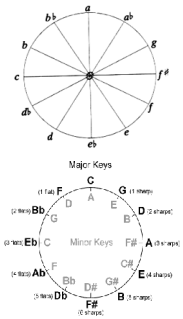
Tolonen and Karjalainen, TSAP 00

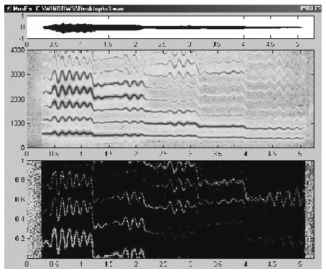
Tzanetakis et al, ISMIR 01



Chroma – Pitch perception

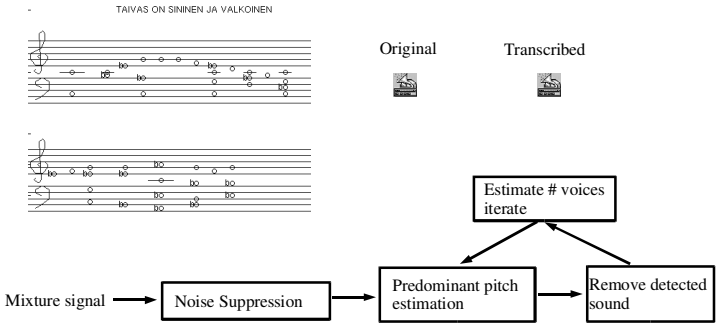






Polyphonic Transcription

Klapuri et al, DAFX 00



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

SASL

```
0.25 tone 4.0
4.50 end
```

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);
}
```

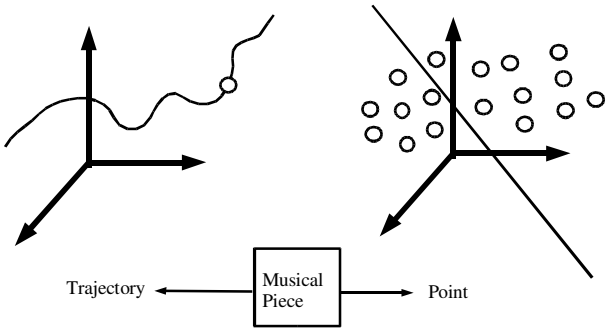
Analysis Outline

- Overview
- Similarity retrieval
- Classification
- Clustering
- Segmentation
- Thumbnailing
- Fingerprinting

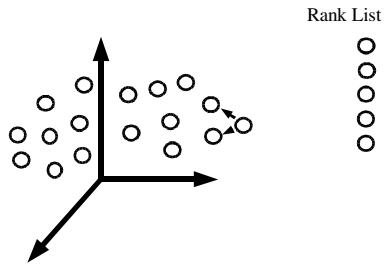
Musical Content Features

- Timbral Texture
  - Spectral Shape
  - MFCC (perceptually motivated features, ASR)
- Rhythmic structure
  - Beat Histogram Features
- Harmonic content
  - Pitch Histogram Features

Analysis Overview



Query-by-Example Content-based Retrieval

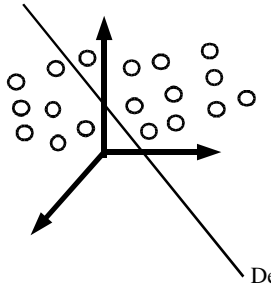


QBE Examples

- Collection of 3000 clips (30s)

	Query	Results
Rock: Beatles		
Jazz : Bobby Hutcherson		
Funk : Mano Negra		
Ethnic: Tibetan singer		
Computer Music ? : P.Lansky		

Statistical Supervised Learning



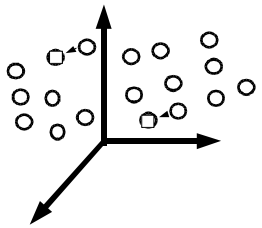
Partitioning of feature space

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

Decision boundary

□ Music
□ Speech

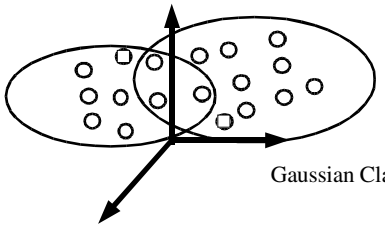
Non-parametric classifiers



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

Nearest-neighbor classifiers

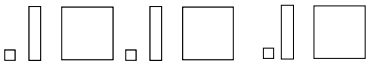
Parametric classifiers



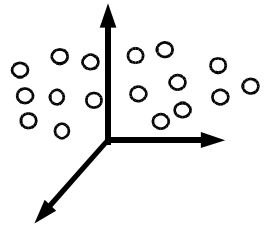
$$P(\square|\bigcirc) = \frac{p(\bigcirc|\square) * P(\square)}{p(\bigcirc)}$$

Gaussian Classifier

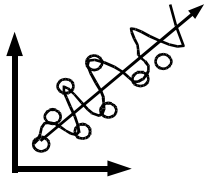
Gaussian Mixture Models



Cross-validation Overfitting



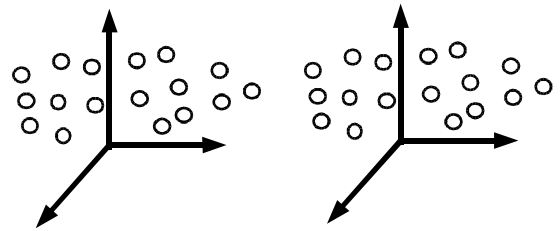
Overfitting – generalization



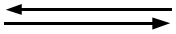
Supervised Learning

- Labeled data
  - Training set, testing set
  - Cross validation
- Classifiers
  - Gaussian
  - Gaussian Mixture Model
  - K Nearest Neighbors
  - Backpropagation Artificial Neural Network

Unsupervised Learning Clustering



Classify



Re-estimate

k-means

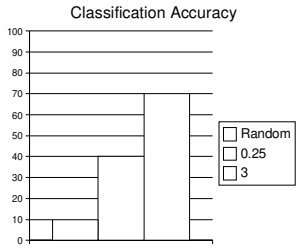
Automatic Musical Genre Classification

- 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

Classification Evaluation – 10 genres

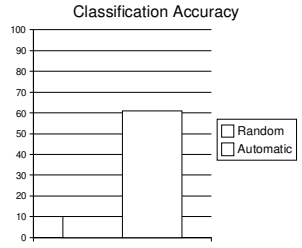
Manual (52 subjects)
Perrot & Gjerdingen, M.Cognition 99

0.25 seconds 40%
3 seconds 70%

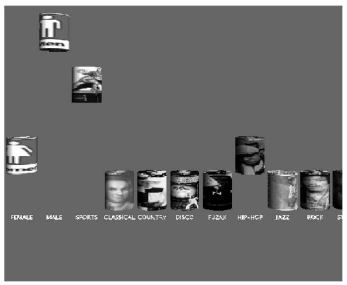


Automatic (different collection)
Tzanetakis & Cook, ISMIR 01

Gaussian Mixture Model (GMM)
10-fold cross-validation 61%



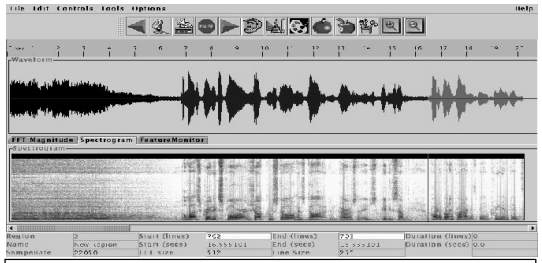
GenreGram DEMO



Dynamic real time 3D display
for classification of radio signals

Audio Segmentation

- Segmentation = changes of sound "texture"



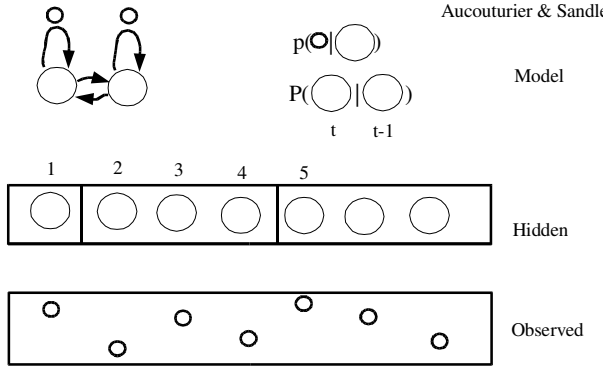
News:

Segmentation

- Model-based — Aucouturier & Sandler, AES 01
  - HMM
  - Fixed # of "textures", no RMS
- Metric-based — Zang & Kuo, TSAP 01; Tzanetakis & Cook, WASPAA 99
  - Detect abrupt changes
  - Arbitrary # of "textures", RMS
  - Sensitive to transients
- Hybrid

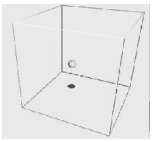
HMM segmentation

Aucouturier & Sandler, AES 01

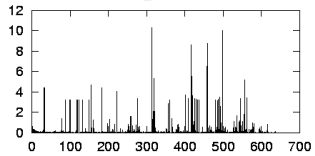


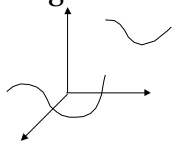
Multifeature Segmentation Methodology

Tzanetakis & Cook, WASPAA 99

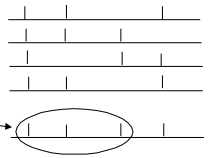


- Time series of feature vectors V(t)
- f(t) = d(V(t), V(t-1))
  - $d(x,y) = (x-y)C^{-1}(x-y)^t$ (Mahalanobis)
- df/dt peaks correspond to texture changes





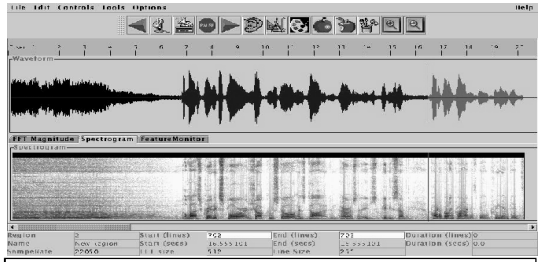
Segmentation Evaluation User Study



- 20 subjects, fixed # segments
- Manual is consistent
  - 75% segments >50% subjects
- Automatic approximates manual
  - 70% segments automatically detected
- Editable automatic segmentation doesn't bias results
- Errors: no semantic information

Audio Segmentation

- Segmentation = changes of sound "texture"



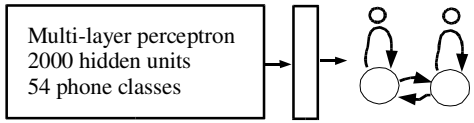
News:

Segment Annotation

- Short description (2-8 words) of segment
  - average ~4
  - 2200 meaningful
  - 620 unique
  - 100 = occur more than 5 times 64% of total word count
- Word types
  - source of sound
  - structural terms
  - basic acoustic parameters
- Possible to automate

Locating singing voice segments

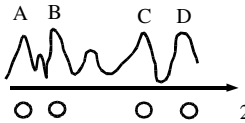
Berenzweig & Ellis, WASPAA 99



80% accuracy

Performance matching

Yang, WASPAA 99



Power plot

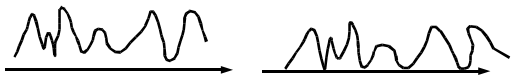
24d "pitch" vectors from FFT

Nearest neighbor with Locality-Sensitive Hashing

Identical, different copy, different vocals, different performance (80%)

Characteristic sequence

Foote, ISMIR 00



Dynamic programming

Symphonies

Audio Thumbnailing

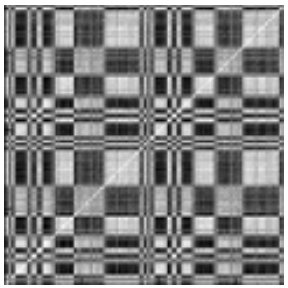
- Representative short summary of piece
  - Segmentation-based
  - Repetition-based
- Hard to evaluate

Segmentation-based Thumbnailing

- Begin and end times of a 2 second thumbnail that best represents the segment
  - 62% first two seconds of the segment
  - 92% two seconds within the first five seconds of the segment
- Automatic thumbnailing
  - first 5 sec + best effort about 80% "correct"

Repetition-based thumbnailing

Logan, B., ICASSP 00
Bartch and Wakefield,WASPAA99

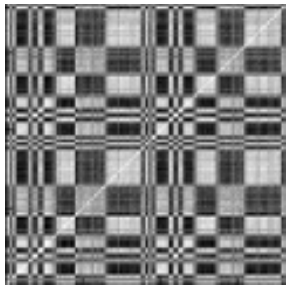


Thumbnail = maximum repeated segment

Alternatives: Clustering, HMM

Structure from similarity

Foote et al, ISMIR 02
Dannenberg et al, ICMC 02

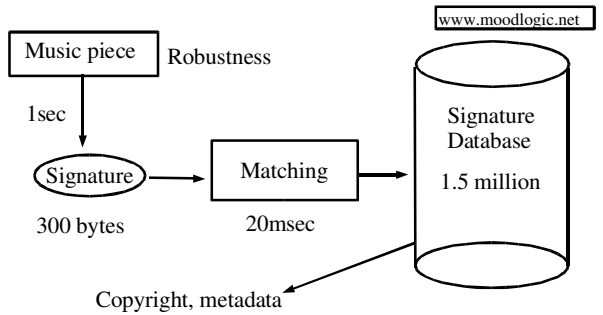


Feature vector trajectory
Correlation at various time lags

ABAA'

Audio Fingerprinting

Allamanche, ISMIR 01



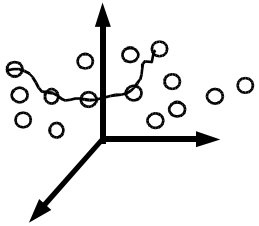
Playlist generation

Tewfik, ICASSP 99
Pachet, IEEE Multimedia00

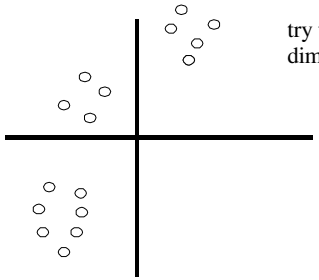
(s1,s2,s3, ... , sn) 20% slow songs, 80% fast, female jazz singers

Constraint-satisfaction problem
Smooth transitions

Technical attributes (artist, album, name)
Content attributes (jazz singer, brass)



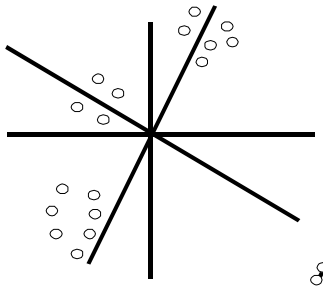
Dimensionality Reduction PCA



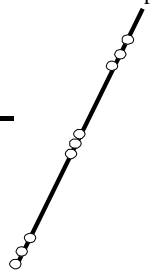
try to represent the feature space with fewer dimensions

Naïve approach keep only one feature axis

PCA

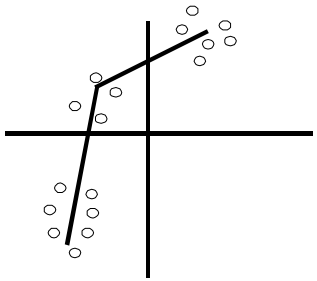


Principal Component Analysis



Pick the axis
that passes through
the centroid of the data
such that the variance of
the projected points
is maximum

Principal Curves



Curves such that the local
means of the points fall
on the curve (piecewise
linear in our implementation)

Interaction Outline

- Motivation
- Content & Context Aware UIs
  - Editors
  - Displays
- Query UIs
  - Audio-based
  - Midi-based

Content & Context Aware User Interfaces

- Automatic results not perfect
- Music listening subjective
- Browsing vs retrieval
- "Overview, Zoom and Filter, Details"
- Adapt UI to audio "Content & Context"
  - Computer audition
  - Visualization

Content and Context

- Content ~ file
  - Genre, male voice, high frequency
- Context ~ file and collection
  - Similarity
  - Slow – fast
- Multiple visualizations
  - Same content, different context

Billie Holiday

Betty Carter

Traditional Audio UI

CoolEdit

Music production and recording

Waveform and Spectrogram Displays

Cut, paste, effects, etc

Limited content no context

Frequency to color

www.comparisonics.com

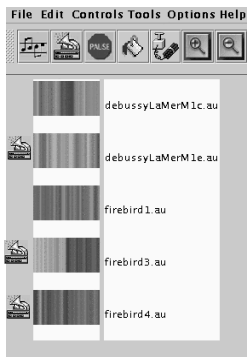
Female-Male Bass-Snare

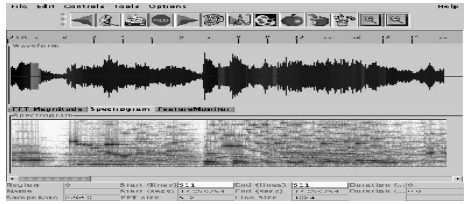
Low frequencies darker
High frequencies lighter

Only content-sensitive

Timbregrams

Tzanetakis & Cook DAFX00, ICAD01



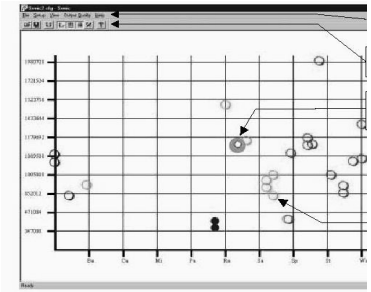


Content and context similarity and periodic structure using color

Principal Component Analysis

Enhanced Audio Editor

Sonic Browser



Direct Sonification

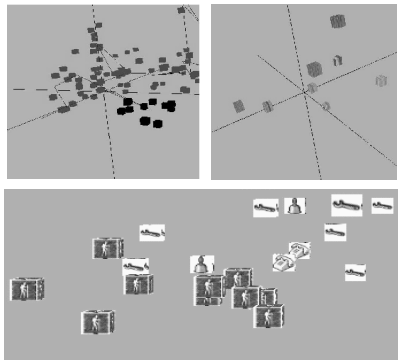
Spatial-audio aura

Manual placement
drag-drop or simple attributes

Sonic Browser (Univ. Limerick)
Fernstrom & Brazil ICAD 01

TimbreSpace Browser 2D,3D DEMO

Tzanetakis & Cook DAFX00, ICAD01



Automatic coloring

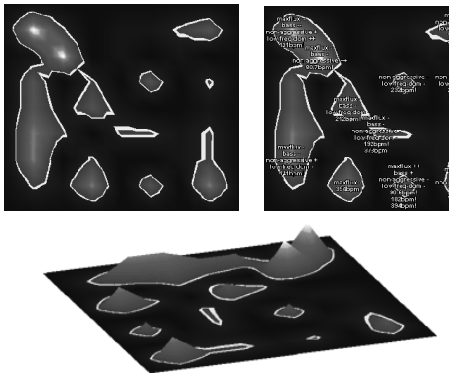
Hierarchical zooming

Automatic positioning

Principal Component Analysis for dimensionality reduction

Islands of Music

Pampalk, ISMIR 02



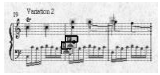
Automatic analysis

Feature vectors

Self-Organizing Map (SOM)

Beyond the QBE paradigm

- Activate the user
- Browsing – filter part
- Direct audio feedback
- Alternatives to QBE
  - Audio-based
  - Midi-based



Audio-based QUIs DEMO

Tzanetakis et al, ICMC02

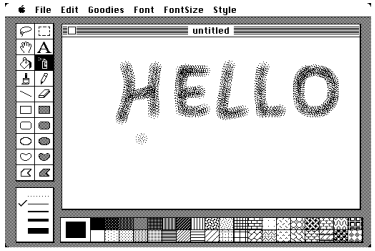
- Use collection to provide continuous audio feedback
- Kill the search button
- Volume control example
- Sounding objects

SoundSliders and SoundLists

Midi-based QUIs

Tzanetakis et al, ICMC02

- Sketchpad for music
- Style modeling – generate query

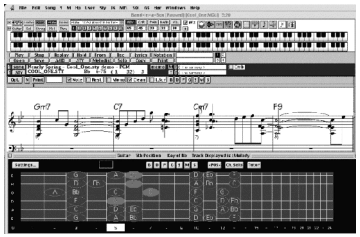


Examples

Generated using Band-in-a-Box
Converted to audio
Retrieval only by Beat Histogram features

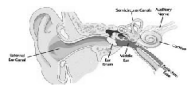
Query

Best Match (4000)

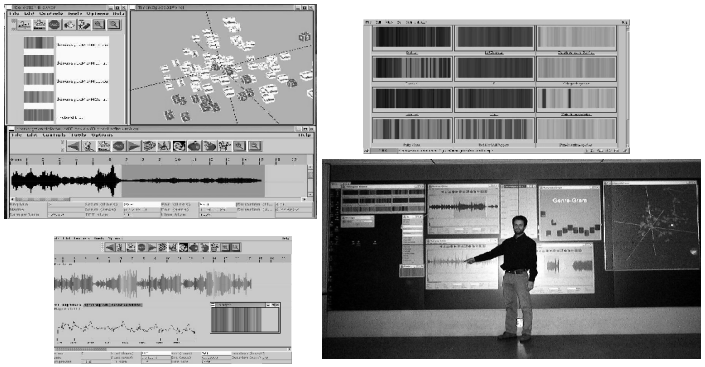


Auditory Scene Analysis

Albert Bregman



Integration



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

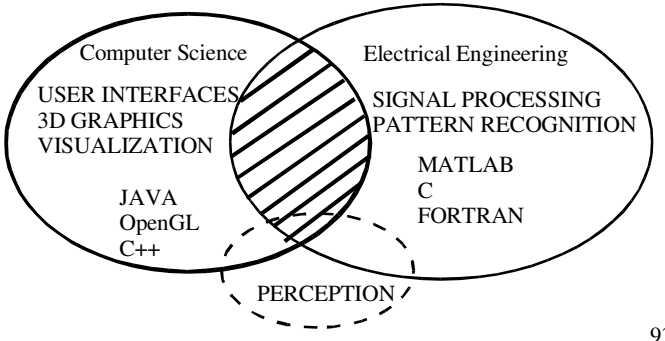
MARSYAS

- Manipulation Analysis and Retrieval Systems for Audio Signals
- Musical Analysis and Retrieval Systems for Audio Signals
- MusicAl Research System for Analysis and Synthesis
- www.cs.princeton.edu/~gtzan/marsyas.html
- C++ server (signal processing, machine learning)
- Java client (GUIs)
- 1600 different host downloads

Collections

- Typical size ~30 seconds, mono, 22050
- MusicSpeech (~200)
- Radio (~100)
- Jazz (~100)
- Instruments (~1000) (McGill samples)
- Sfx (~200)
- Rock (~1000)
- Genres (~1200)
- Music Library (~5000)

Lack of software tools for audio collections



Genre Classification Confusion Matrix

	Classical	Country	Disco	Hiphop	Jazz	Rock	Blues	Reggae	Pop	Metal
Classical	73	0	0	0	6	2	0	0	0	0
Country	0	43	1	0	1	6	2	4	3	1
Disco	0	6	43	10	0	4	9	3	3	2
Hiphop	0	4	9	49	0	3	2	17	10	2
Jazz	21	5	1	0	71	6	5	0	2	3
Rock	4	16	6	1	8	41	11	5	11	17
Blues	2	18	2	1	7	7	61	5	1	2
Reggae	0	2	12	30	2	13	6	63	6	0
Pop	0	3	25	8	3	7	0	3	64	2
Metal	0	3	1	1	1	11	4	0	0	71

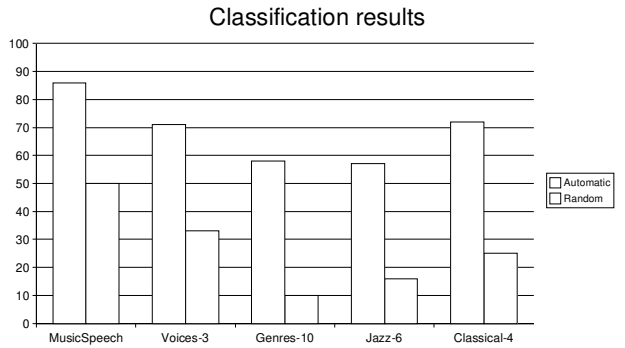
Figure 3. Automatic genre classification confusion matrix

Perrot & Gjedringen
Music Cognition 1999

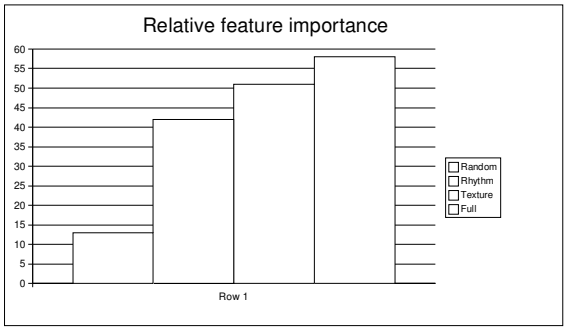
Humans – Genre Classification

- 10 Genres (2 different (R & B, Latin)
- 70% at 3 seconds
- 40% at 250 milliseconds
- 10% chance
- 52 College students
- Fuzzy nature of genre
- Demonstration available for the sceptical

Classification Evaluation (10-fold cross validation, classifier)



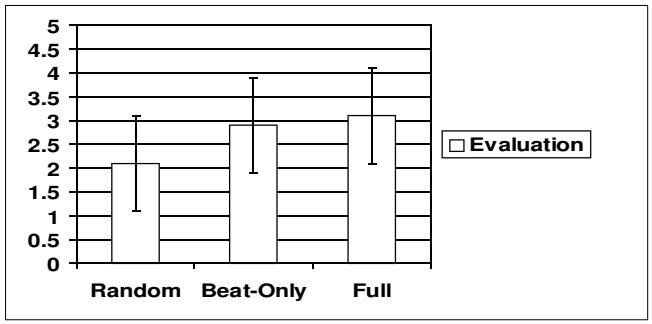
Relative importance of feature sets



Content-based similarity retrieval

- Small user study
  - 12 queries * 5 matches * 3 algorithms * 7 subjects
  - Random, BPM only, BMP + Spectral
- Single Vector approach
  - Spectral Features
  - Mean BPM (Scheirer 1999)
- 1000 Rock songs

Rock Retrieval Study Results



The Princeton Scalable Display Wall

- 6 x 3 meters rear projection screen
- 4096 x 1536 resolution
- Custom-built 16-speaker sound system and sound server PC (8 x 2 channels)
- 6 x 3 array of projectors (4 x 2)
- Each projector is driven by a commodity PC with an off-the-shelf graphics accelator
- Variety of input methods (multiple users interactive collaboration)