Carnegie Mellon

Music Analysis and Retrieval for Audio Signals

George Tzanetakis
PostDoctoral Fellow
Computer Science Department
Carnegie Mellon University

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

Bits of the history of bits


01011110101010


Hello world


Web


Multimedia

Understanding multimedia content =

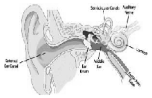
Music

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

The not so far future of MIR

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

Talk Outline



Hearing
Signal Processing

MP3 feature extraction
DWT Beat Histograms

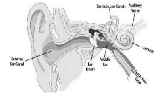
Understanding
Machine Learning

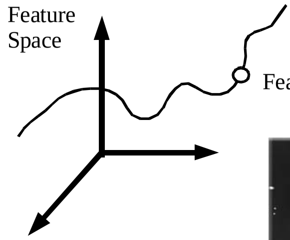
QBE similarity retrieval
Genre Classification

Dancing
Human Computer
Interaction

Content & Context Aware
Timbregrams
Timbrespaces

Hearing - Feature extraction





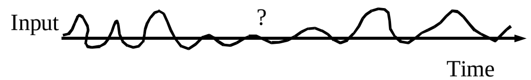
Timbral Texture

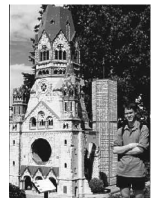
Timbre = differentiate sounds of same pitch and loudness

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

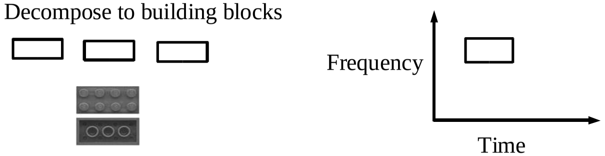
Global, statistical and fuzzy properties

Time-domain waveform

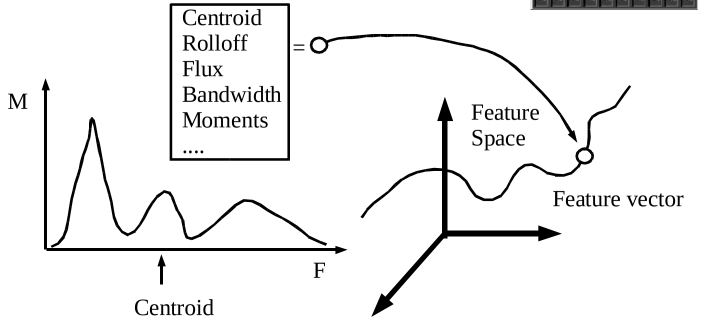




Decompose to building blocks



Spectrum and Shape Descriptors



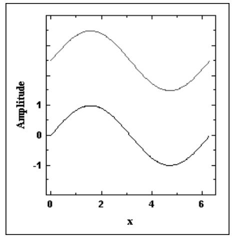
Time-Frequency Analysis Fourier Transform

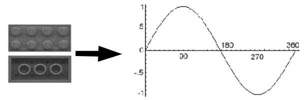
$$f(x) = \sum_{n=0}^{\infty} a_n \cos(n*x) + \sum_{n=0}^{\infty} b_n \sin(n*x)$$

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

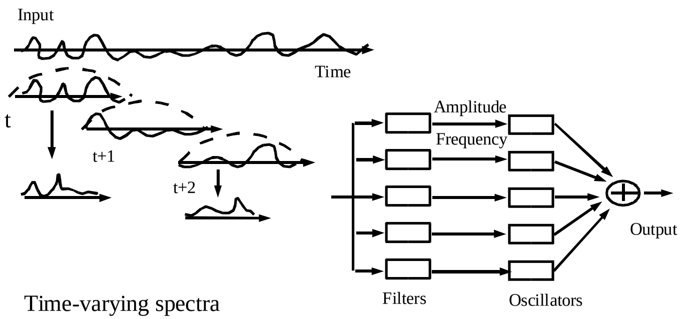
$$f(\omega) = \int_{-\infty}^{\infty} f(t)\, e^{i\omega t}\, dt$$

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



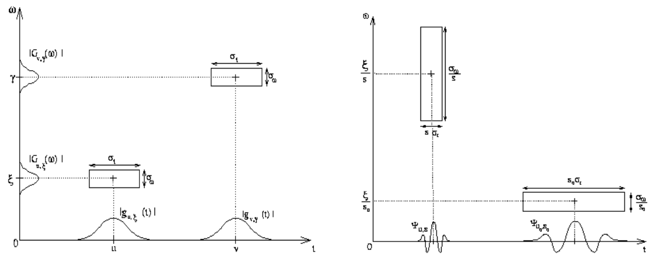


Short Time Fourier Transform



Time-varying spectra
Fast Fourier Transform FFT

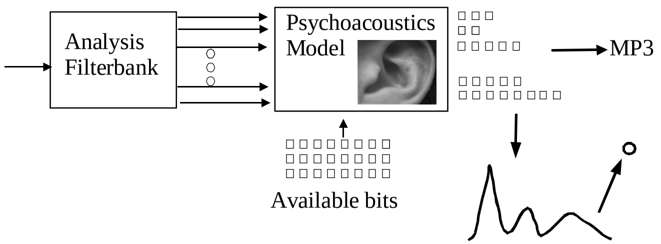
STFT- Wavelets



Time – Frequency   Heisenberg uncertainty $\sigma_t \sigma_\omega \geq 1/4$

MPEG Audio Feature Extraction

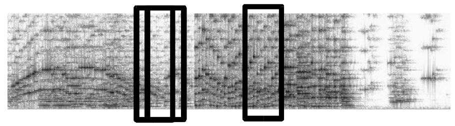
Pye ICASSP 00
Tzanetakis & Cook ICASSP 00

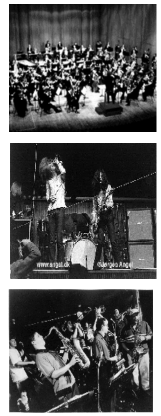


Perceptual Audio Coding (slow encoding, fast decoding)

Summary of Timbral Texture Features

- Time-Frequency analysis
- Signal processing (STFT, DWT)
- Perceptual (MFCC, MPEG)
- Statistics over "texture" window



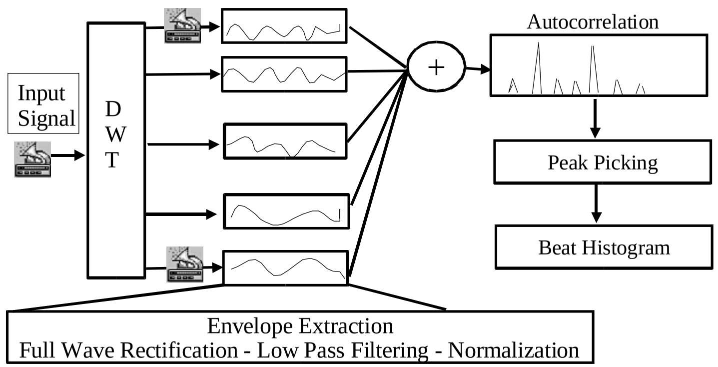


Rhythm

- Rhythm = movement in time
- Origins in poetry (iamb, trochaic...)
- Foot tapping
- Hierarchical semi-periodic structure
- Linked to motion
- Running vs global

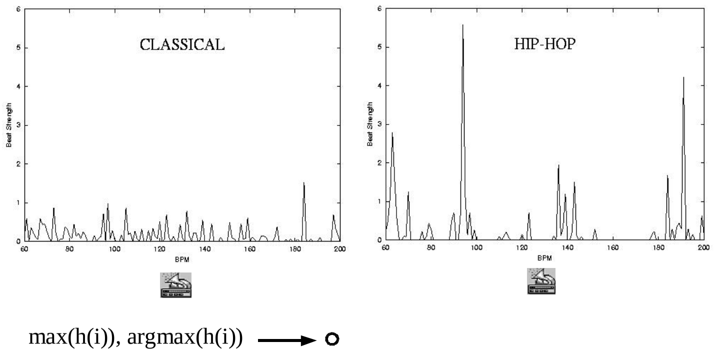
Wavelet-based Rhythm Analysis

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



Beat Histograms

Tzanetakis et al   AMTA01

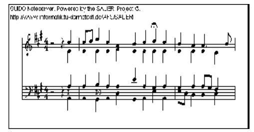


max(h(i)), argmax(h(i))

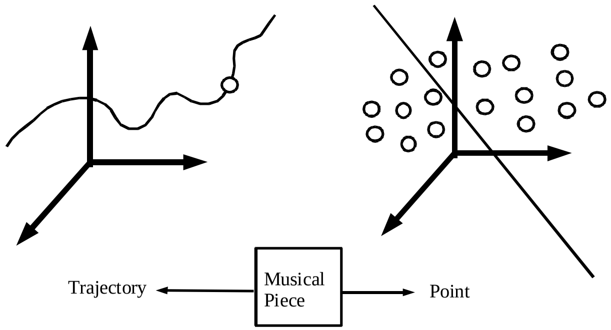
Musical Content Features

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

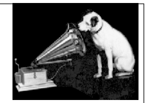
- Rhythmic structure (6)
  - Beat Histogram Features
- Harmonic content (5)
  - Pitch Histogram Features

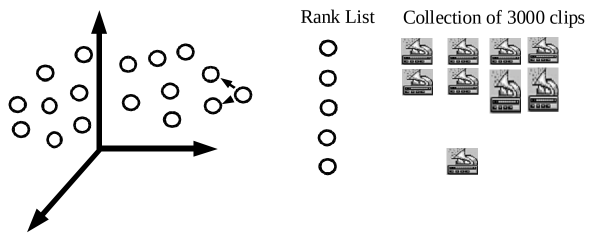


Understanding



Query-by-Example Content-based Retrieval

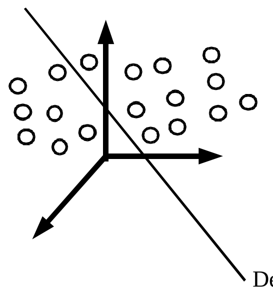




Automatic Musical Genre Classification

- 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

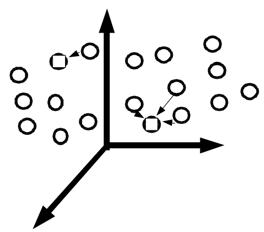


Partitioning of feature space

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

- □ Music
- □ Speech

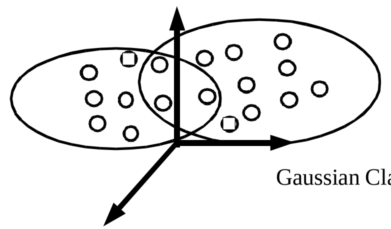
Non-parametric classifiers



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

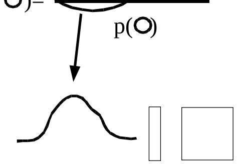
Nearest-neighbor classifiers (K-NN)

Parametric classifiers

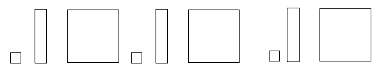


$$P(\square|O) = \frac{p(O|\square) * P(\square)}{p(O)}$$

Gaussian Classifier



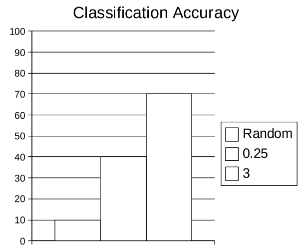
Gaussian Mixture Models



Classification
Evaluation – 10 genres

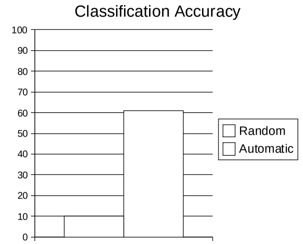
Manual (52 subjects)
Perrot & Gjerdingen, M.Cognition 99

0.25 seconds 40%
3 seconds 70%

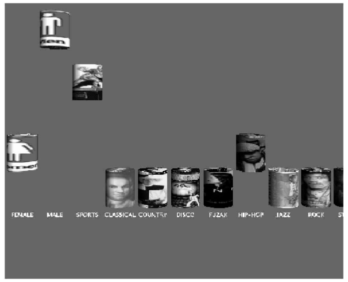


Automatic (different collection)
Tzanetakis & Cook, TSAP 10(5) 2002

Gaussian Mixture Model (GMM)
10-fold cross-validation 61% (70%)



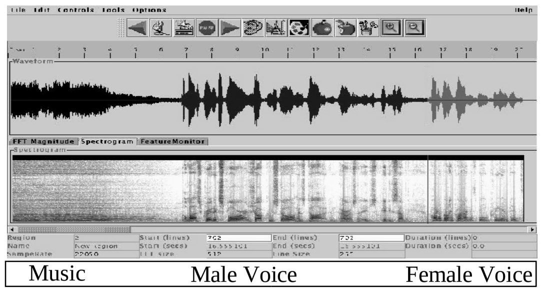
GenreGram DEMO



Dynamic real time 3D display
for classification of radio signals

Audio Segmentation

- Segmentation = changes of sound "texture"

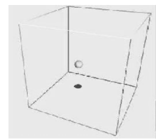


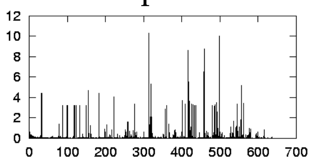
News:

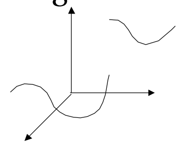
Multifeature Segmentation Methodology

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

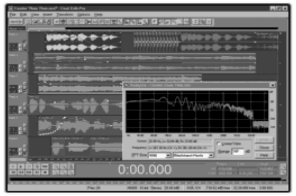






Interaction

- Automatic results not perfect
- Music listening subjective
- Browsing vs retrieval
- Adapt UI to audio "Content & Context"
  - Computer Audition
  - Visualization



Cooledit

Content and Context

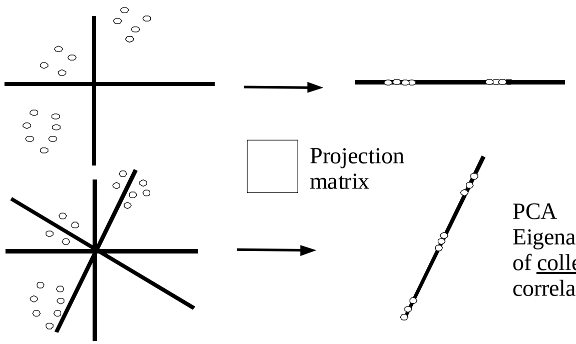
- Content ~ file
  - Genre, male voice, high frequency
- Context ~ file and collection
  - Similarity
  - Slow – fast
- Multiple visualizations
  - Same content, different context

Christina Aguilera

Billie Holiday

Ella Fitzerald

Principal Component Analysis

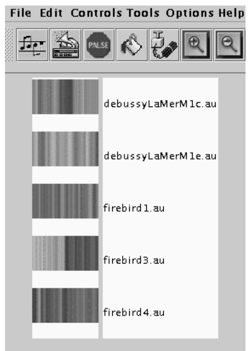


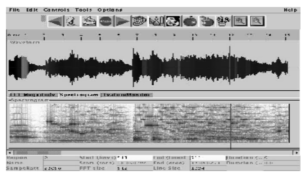
PCA
Eigenanalysis
of <u>collection</u>
correlation matrix

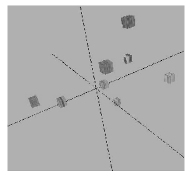
Timbregrams and Timbrespaces

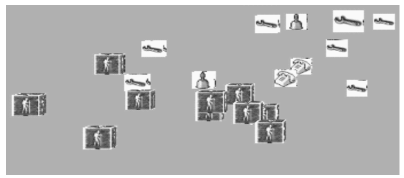
Tzanetakis & Cook DAFX00, ICAD01

PCA = content & context

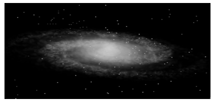


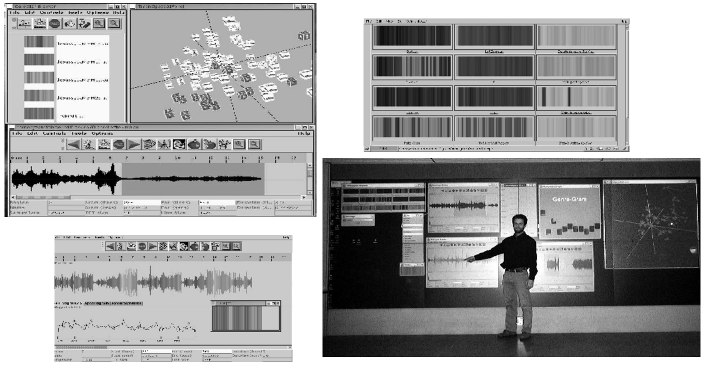






Integration



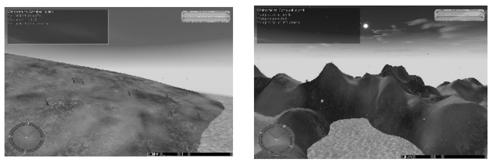


Implementation

Tzanetakis & Cook Organized Sound 4(3) 00

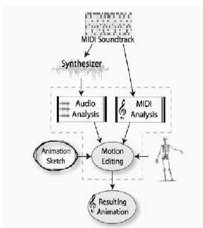
- *MARSYAS* : free software framework for computer audition research
  - Server in C++ (numerical signal processing and machine learning)
  - Client in JAVA (GUI)
  - Linux, Solaris, Irix and Wintel (VS , Cygwin)
- Apr. 5500 downloads, 2300 different hosts, 30 countries since March 2001

Marsyas users



Desert Island

Jared Hoberock
Dan Kelly
Ben Tietgen



Music-driven motion editing
Marc Cardle

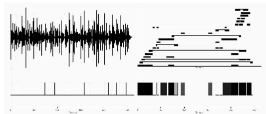
moodlogic
The Mix Maker for your MP3s

Real time music-speech discrimination

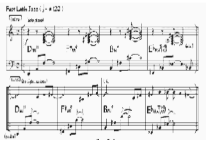
ALL MUSIC PUBLISHING

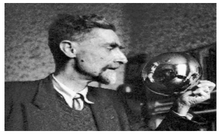
Current Work-Collaborations

Tzanetakis, Hu and Dannenberg, WIAMIS 03



- CMU
  - Structural analysis
  - Query -by-humming
  - MIR over P2P networks (Jun Gao)
  - Informedia
- Princeton: sound fx analysis-synthesis(P.Cook)
- Rochester: machine learning (Tao Li)
- Northwestern: perception of musical genre (R.Gjerdingen)

Future Work

- Music
  - Singer identification
  - Chord progression detection
  - Intermediate representations
- Motion capture signals
- Biological signals – time series in general
- Content and context aware multimedia UI



Auditory Scene Analysis

Albert Bregman