On the use of statistical tools for speech and musical audio processing

Mathieu Lagrange
Analyse / Synthèse Team, IRCAM
Mathieu.lagrange@ircam.fr

ATIAM

Outline

1. Introduction
    1. Context and challenges
2. Past and Present
    1. Speech
        1. Model
        2. Applications (coding, speaker recognition, speech recognition)
    2. Audio (Music)
        1. Sound models
        2. Retrieving information within songs
        3. Retrieving information across songs
3. Future
    1. Polyphony handling
    2. Building and using priors
    3. Joint estimation of several musical parameters

Outline

1. Introduction
    1. Context and challenges

Technological Context

« We are drowning in information and starving for knowledge »

R. Roger

- Needs:
  - Measurement
  - Transmission
  - Access
- Aim of a numerical representation:
  - Precision
  - Efficiency
  - Relevance
- Means
  - Mechanical biology
  - Psycho-acoustic
  - Cognition

Challenges

« Forty-two! yelled Loonquawl. Is that all you've got to show for seven and a half million years' work? »

D. Adams

Music is a great material to study as it is both:

- An object : arrangement de sons et de silences au cours du temps
- A function: more or less codified form of expression of :
  - Individual feelings (mood)
  - Collective feelings (party, singing, dance)

Audio Processing: Past and Present

Vocabulary ?

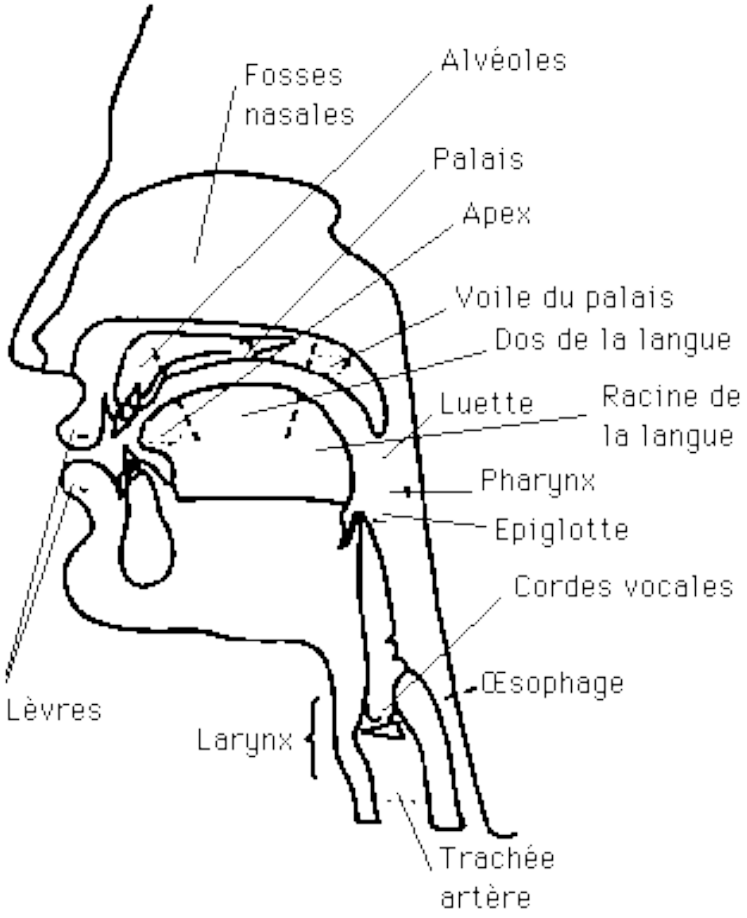
- STFT
- MFCCs
- Chromas
- K-means
- GMMs
- HMMs

Outline

1. Introduction
   1. Context and challenges
2. Past and Present
   1. Speech
      1. Model
      2. Applications (coding, speaker recognition, speech recognition)

Speech signal

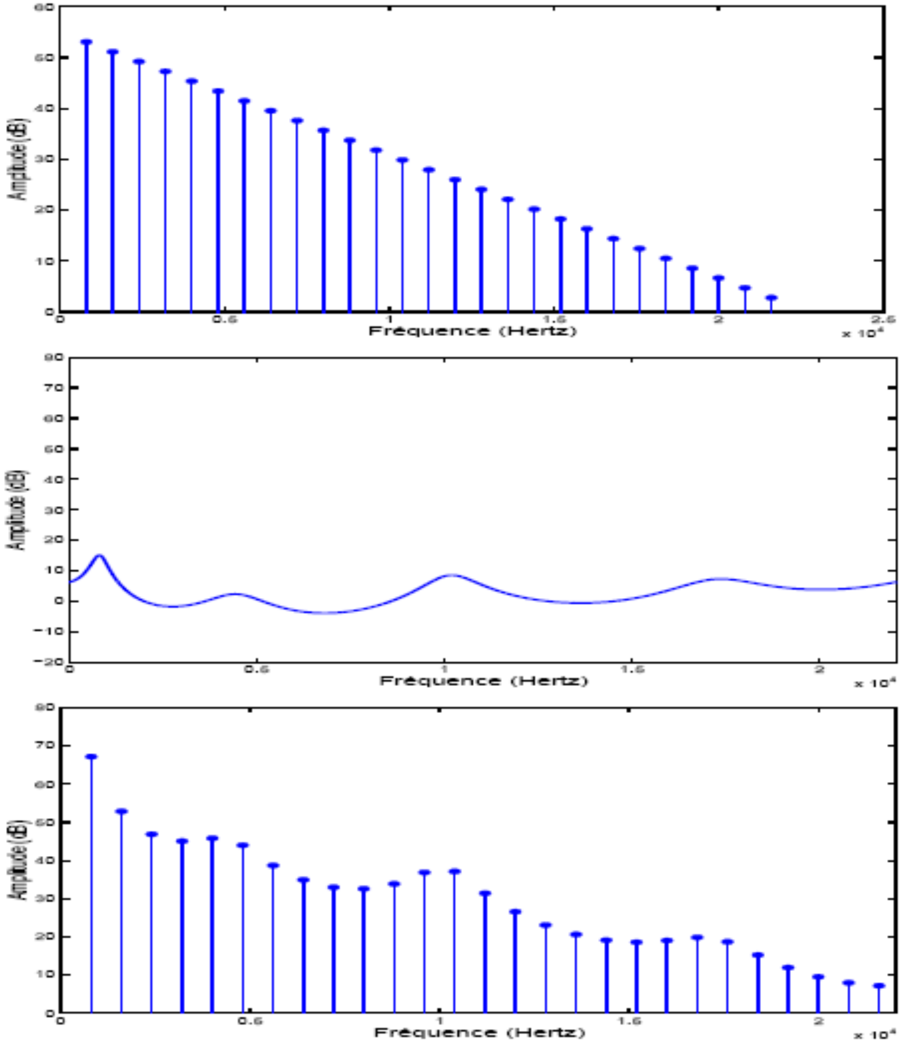
- The speech signal is produced when the air flow coming from the lungs go through the vocal chords and the vocal tract.
  - The size and the shape of the vocal tract as well as the vocal chords excitations are changing relatively slowly
  - The speech signal can therefore be considered as quasi-stationary over short period of about 20 ms.
- Type of speech production
  - Voiced: <a>, <e>, …
  - Unvoiced: <s>, <ch>,
  - Plosives: <pe>, <ke>



Source / Filter Model

- In the case of an idealized **voiced** speech signal, the vocal chords are producing a perfectly periodic harmonic signal

- The influence of the vocal tract can be considered as a filtering with a given frequency response whose maximas are called formants.



Source / Filter Coding

- Algorithm :
  - Voiced / Unvoiced detection;
  - Voiced case: the source signal is approximated with a Dirac comb:
    - a Dirac comb whose successive Diracs are respectively T spaced by T as a spectrum which is a Dirac comb whose successive combs are 1/T spaced.
    - Parameters : T, gain
  - Unvoiced: the source signal is approximated by a stochastic signal:
    - Parameter : gain.
  - The Source signal is next filtered.
    - Parameters : filter coefficients.

« Code-Excited Linear Predictive » (CELP)

For each frame of 20 ms :

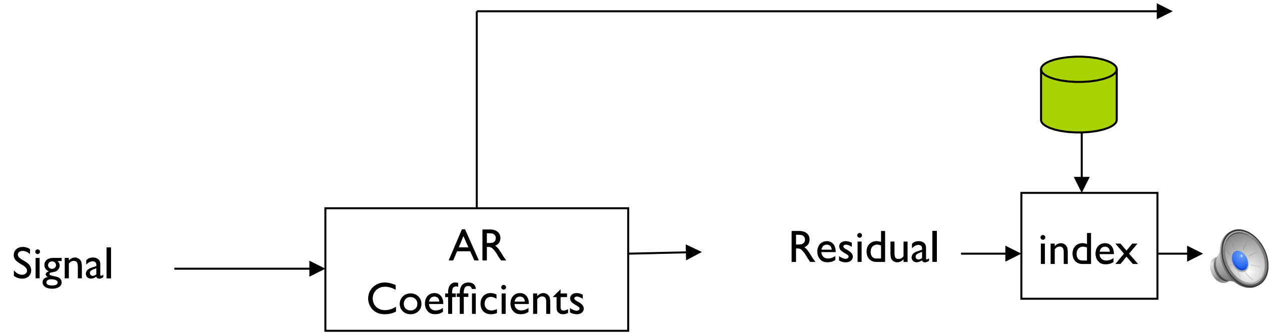
- Auto-Regressive coefficients are computed such that the prediction error is minimized over the entire duration of the frame:

$$\hat{s}[n] = \sum_{i=0}^{d} a[i]s[n-i]$$

$$E = \sum_{n=0}^{N} (s[n] - \hat{s}[n])^2$$

- Quantified coefficients and an index encoding the error signal are transmitted.

« Code-Excited Linear Predictive » (CELP)

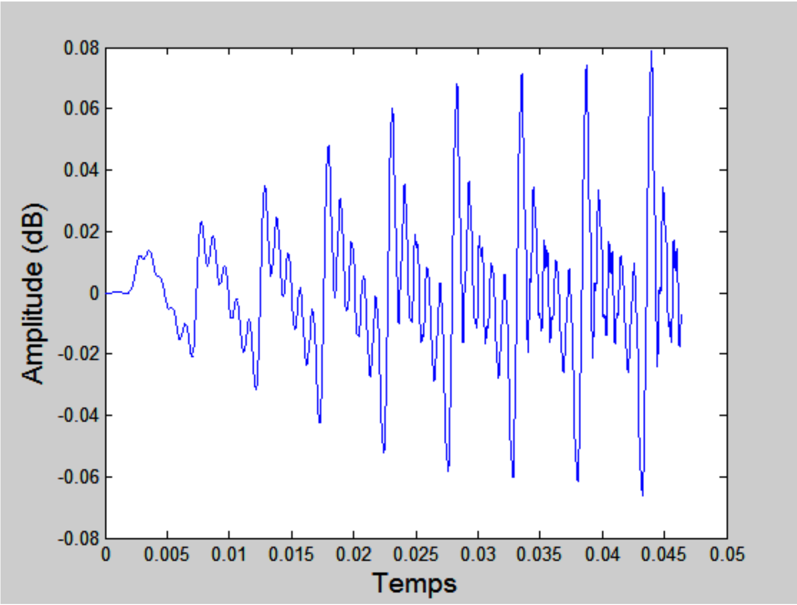


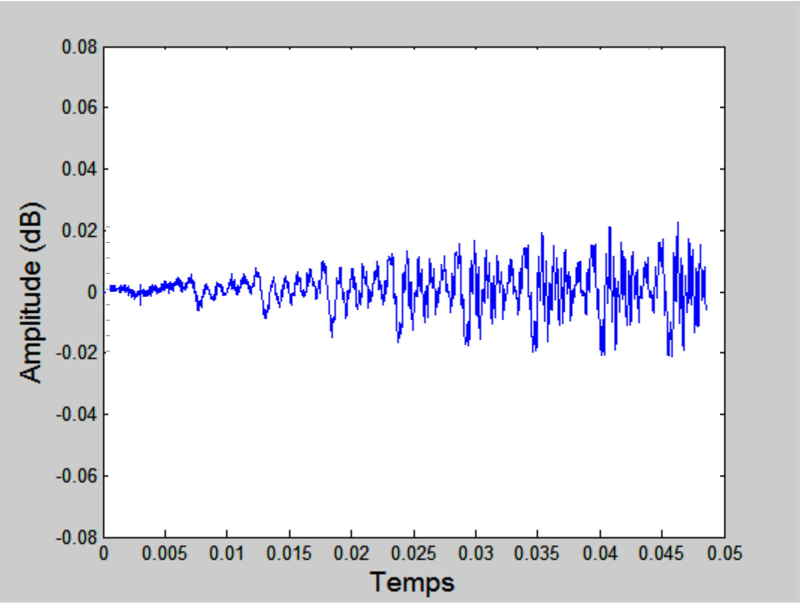
Signal

AR Coefficients

Residual

index





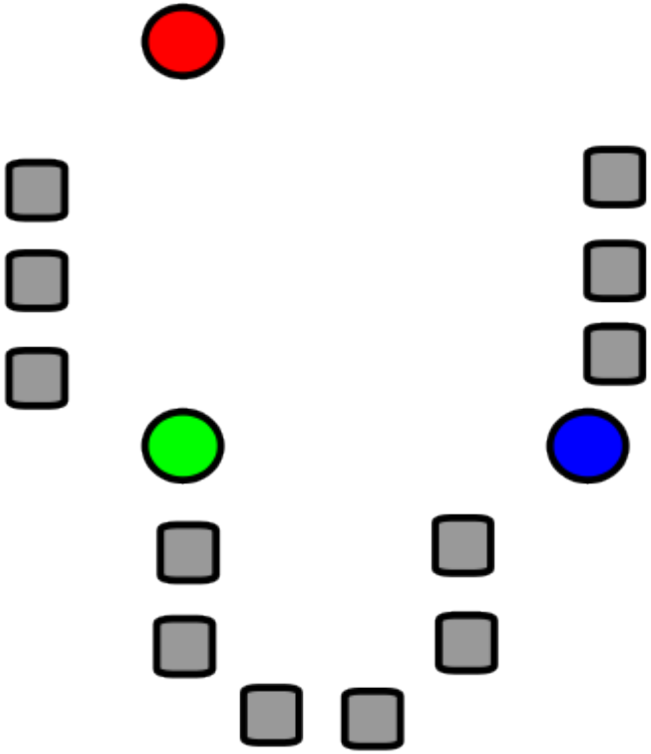
Vector Quantization

- Works by dividing a large set of points (vectors in the feature space) into groups
  - Groups are represented by their centroid
  - Use of standard k-means algorithms to jointly determine the groups and the centroid
  - The set of centroids from the codebook
- At the encoding stage
  - For each residual frame, the closet centroid is determined
  - The index is transmitted
- At the decoding stage
  - The centroid is retrieved using the index value

K-means

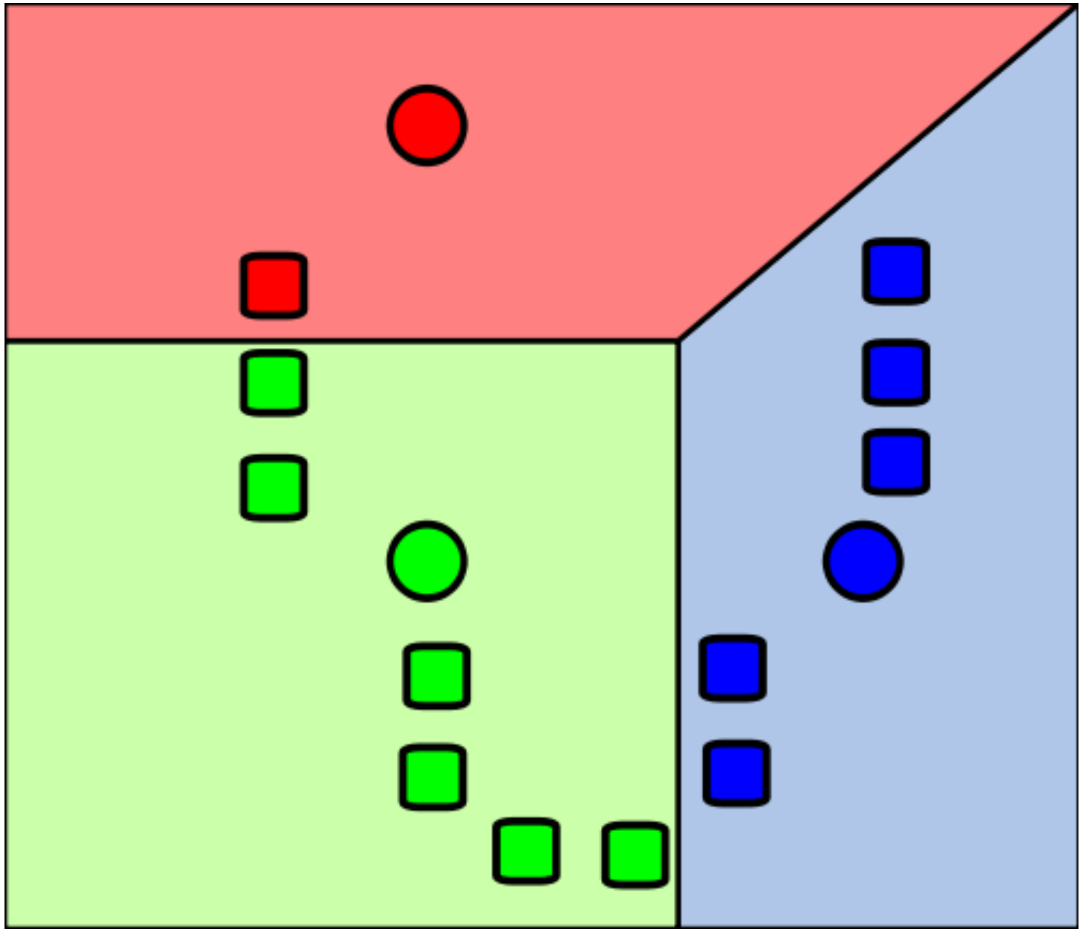
- Place k-centroids at random
- Iterate until stabilisation
    - Determine assignement
    - Compute centroid

K-means



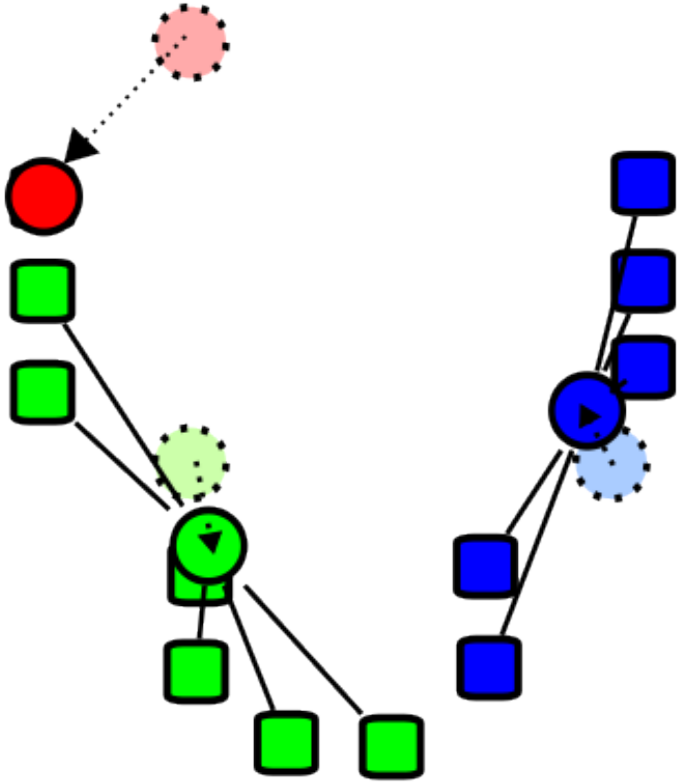
(Fig. Wikipedia)

K-means



(Fig. Wikipedia)

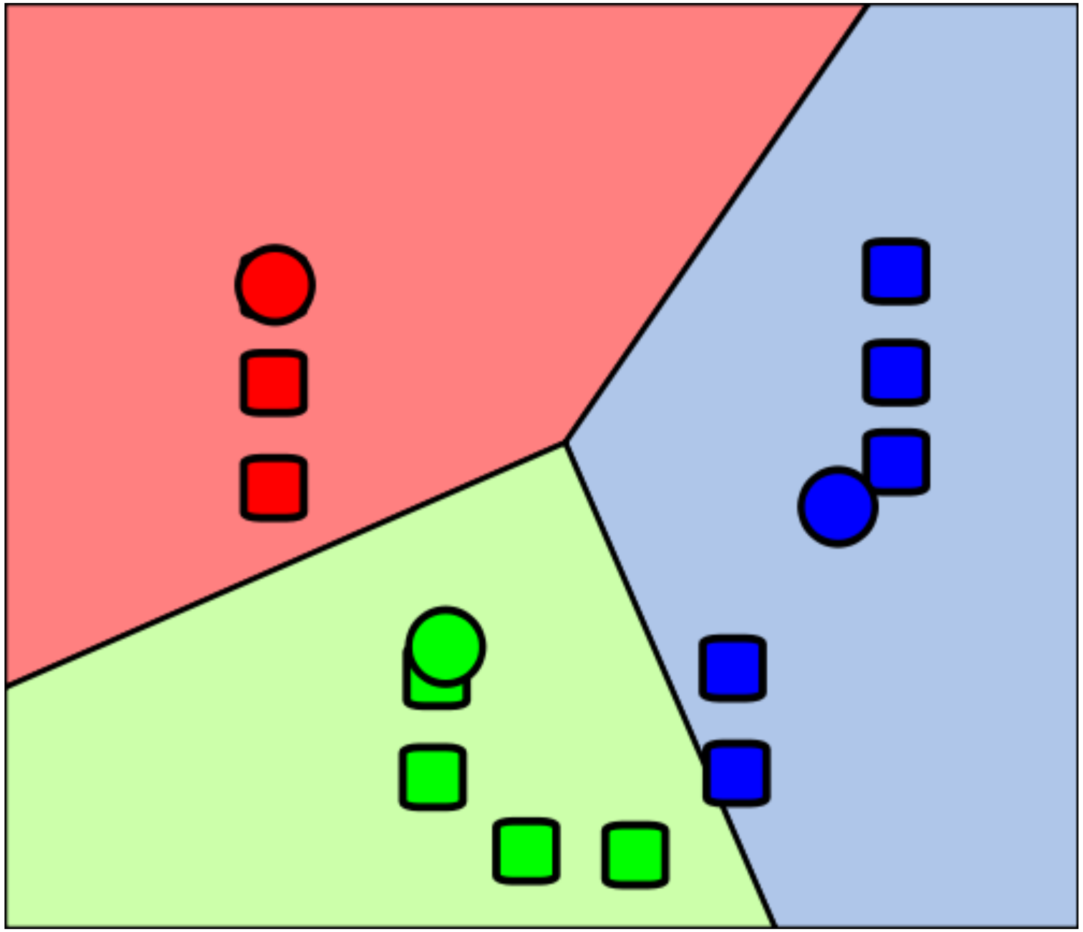
K-means



(Fig. Wikipedia)

K-means

- Video



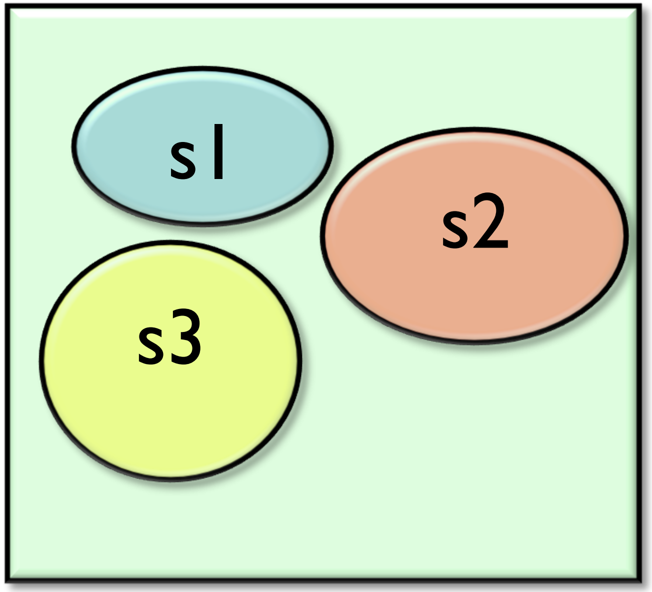
(Fig. Wikipedia)

K-means

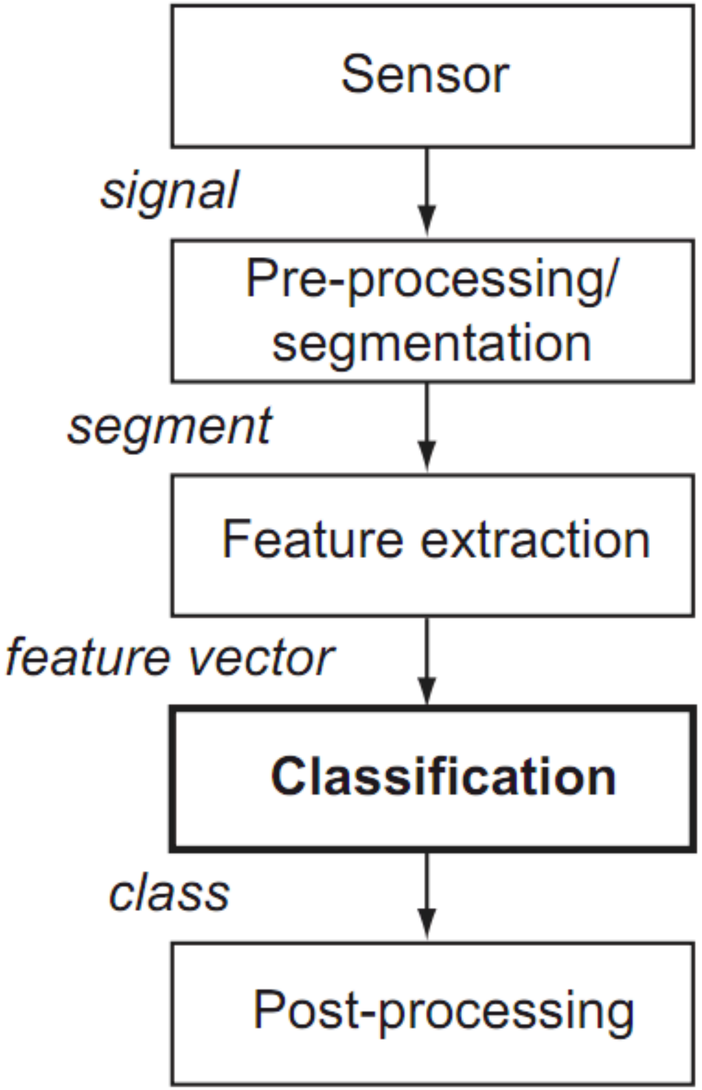
- Place k-centroids at random
- Iterate until stabilisation
  - Determine assignement
  - Compute centroid

Speaker Recognition

- Classical pattern recognition problem
- Specific problems:
  - Open Set / Closed Set: rejection problem
  - Identification / Verification
  - Text Dependency
- Method
  - Feature extraction: model each speech with Mel-Frequency Cepstral Coefficients (MFCCs) and their derivatives.
  - Classification
    - Text independent: Vector Quantization Codebooks or Gaussian Mixture Models (GMMs)
    - Text dependent: Dynamic Time Warping (DTW) or Hidden Markov Model (HMM)



Classification scheme

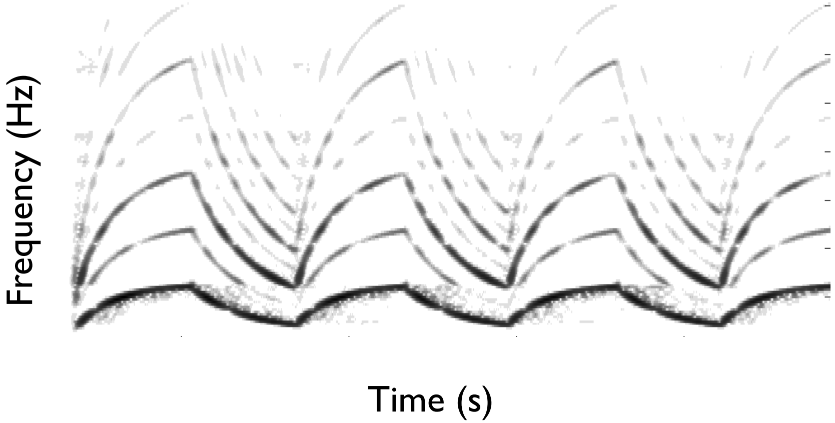


Features

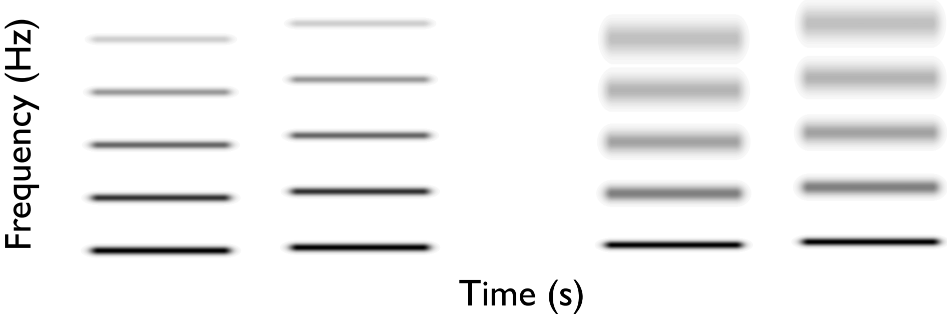
- The decision system is not usually fed directly with the sound signal
- Infer a reduced set of features
  - Smaller `feature space' (fewer dimensions)
  - Simpler models (fewer parameters)
  - Less training data needed
- Expert knowledge is necessary for efficient inference of meaningful descriptors or features.
  - Meaningful means here that the features
    - Extract from the signal interesting properties for the task at hand
    - Invariance under irrelevant modification
    - have to be nicely handled by the decision system

Audio Features

- Frequency-dependant information
  - Fourier transform

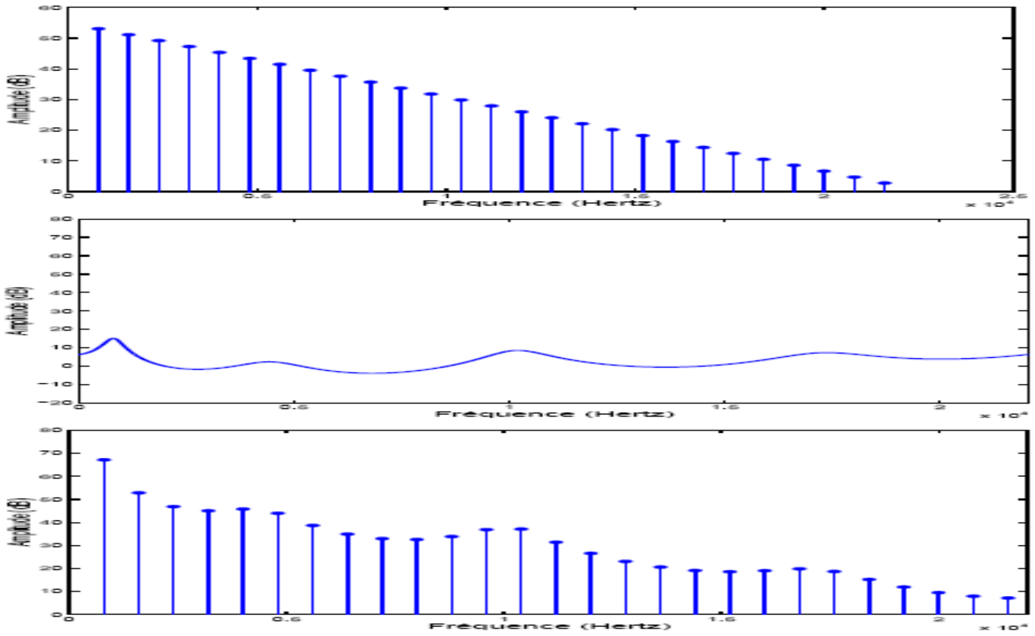


- Mel-Spectrum
  - Robustness to harmonic translation



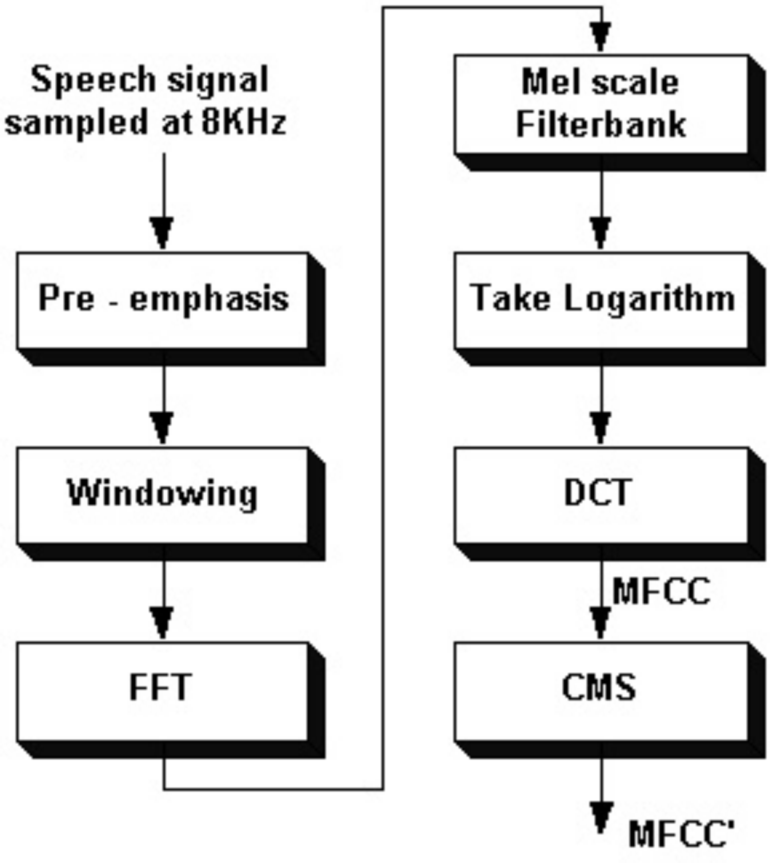
Audio Features

- For most classification tasks, we put the focus on the spectral envelope
  - Speech: formant
  - Music: genre



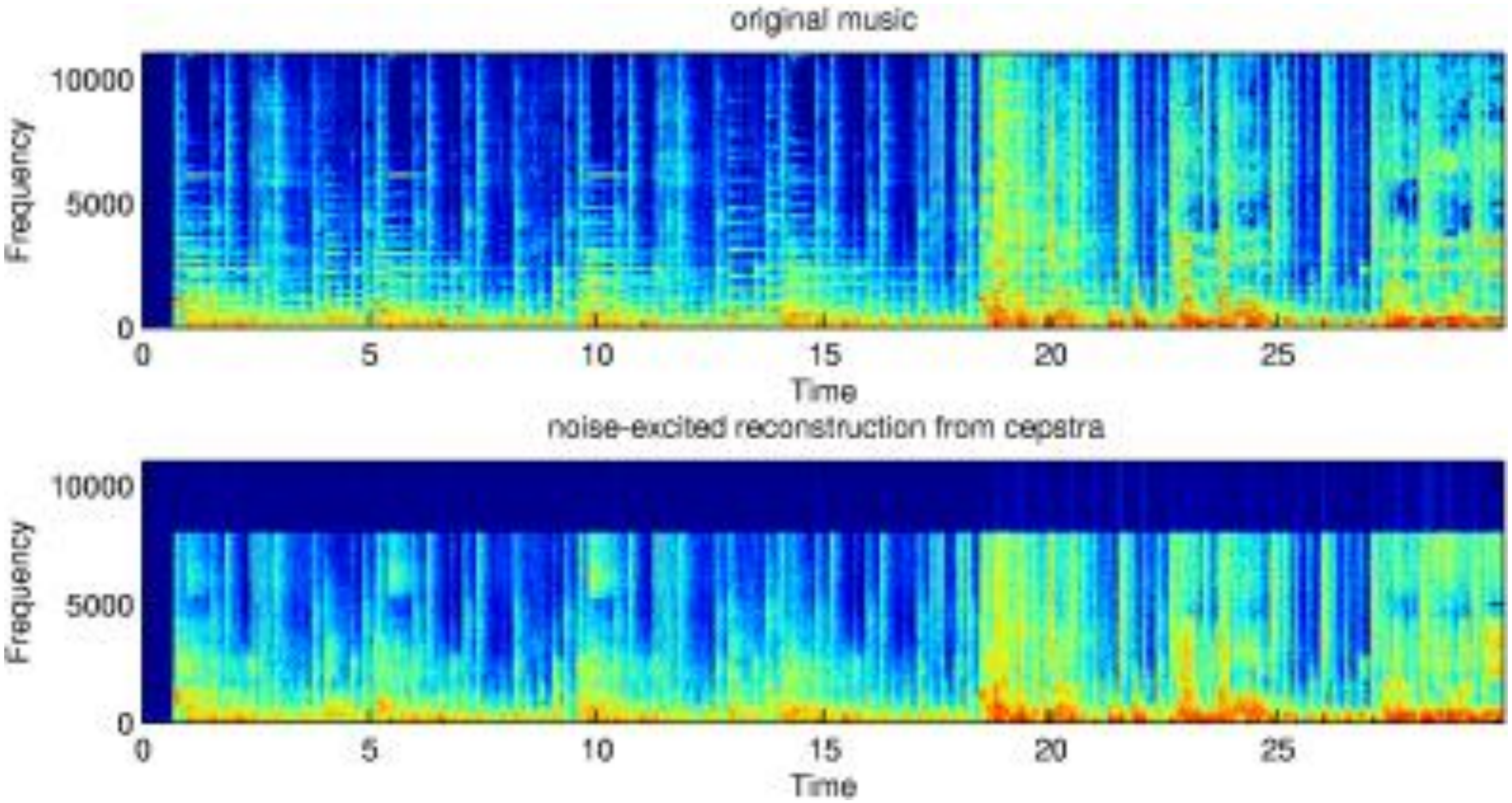
MFCCs rules ?

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform (DCT)
5. The MFCCs are the amplitudes of the resulting spectrum.



Example

- Audio



Potentials of the DCT step

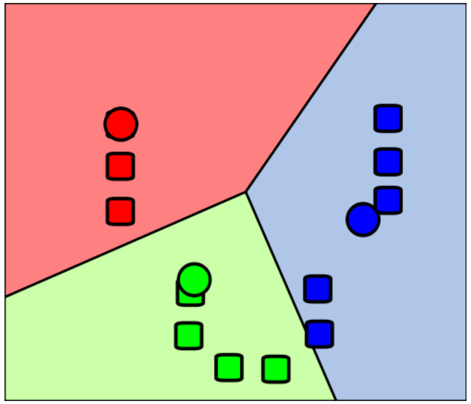
- Observation of Pols that the main components capture most of the variance using a few smooth basis functions, smoothing away the pitch ripples
- Principal components of vowel spectra on a warped frequency scale aren't so far from the cosine basis functions
- Decorrelates the features.
  - This is important because the MFCC are in most cases modelled by Gaussians with diagonal covariance matrices

Issues

- The MEL frequency wraping:
  - highly criticized form a perceptual point of view (Greenwood)
  - conceptually: periodicity analysis over data that are not periodic anymore (Camacho)
- The Cepstral Coefficients are COSINE coefficients:
  - cannot shift with speaker size to capture the shift in formant frequencies that occurs as children grow up and their vocal tracts get longer
- Not a sound representation:
  - no way to provide enhancements such as speaker and channel adaptation, background noise suppression, source separation

Decision System

- From hard assignment to soft assignment
  - K-means:



Gaussian Mixture Models (GMMs)

- The data is modeled as a weighted sum of Gaussians

$$g(\boldsymbol{x}, \boldsymbol{\Phi}) = \sum_{k=1}^{g} \pi_k f(\boldsymbol{x}, \boldsymbol{\theta}_k),$$

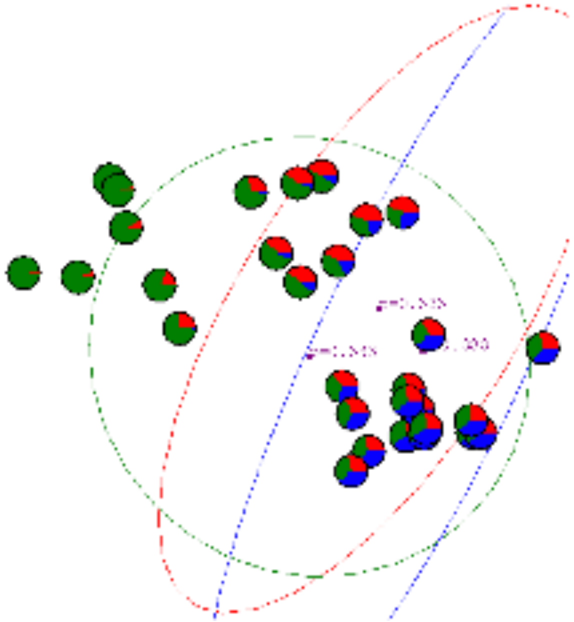
- Estimation of the weights, means and variances of the Gaussians can be done by maximizing the log-likelihood

$$L(\mathbf{x}; \boldsymbol{\Phi}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{g} \pi_k f(\boldsymbol{x}_i, \boldsymbol{\theta}_k) \right).$$

- Usually done with the E-M algorithm
  - E-step : expectation
  - M-step : maximisation

E-M example

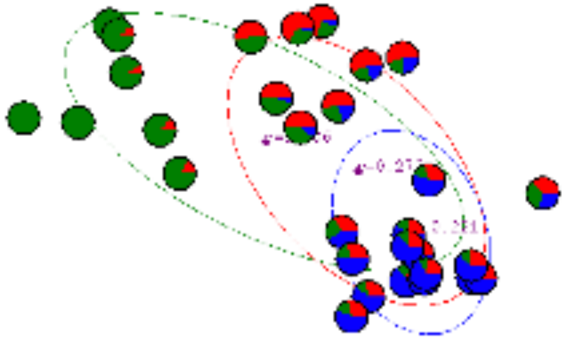
- Start



(Fig. From A. Moore's Tutorial)

E-M example

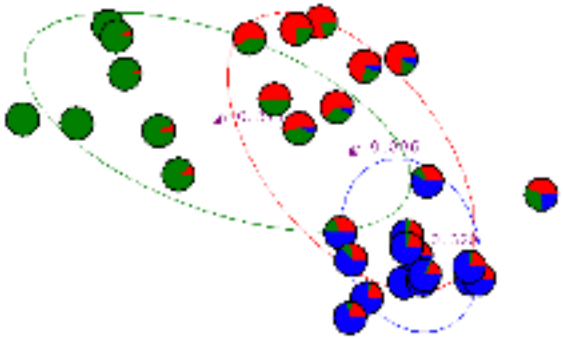
- 1-st iteration



(Fig. From A. Moore's Tutorial)

E-M example

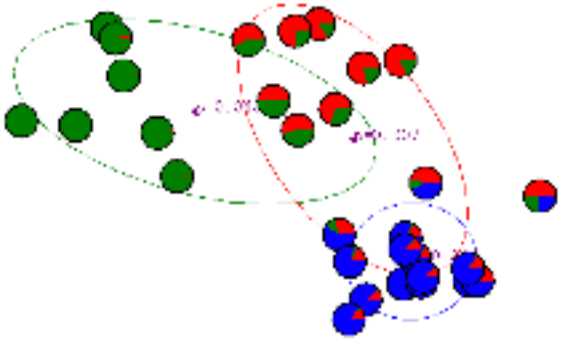
- 2-nd iteration



(Fig. From A. Moore's Tutorial)

E-M example

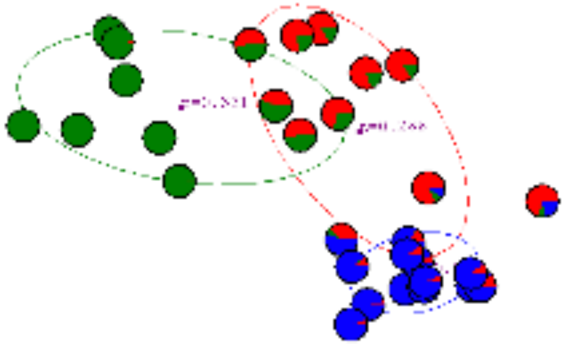
- 3-rd iteration



(Fig. From A. Moore's Tutorial)

E-M example

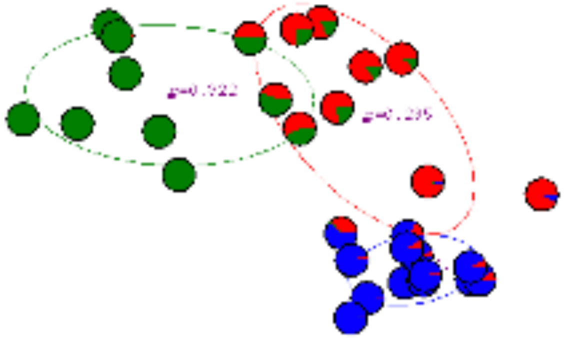
- 4-th iteration



(Fig. From A. Moore's Tutorial)

E-M example

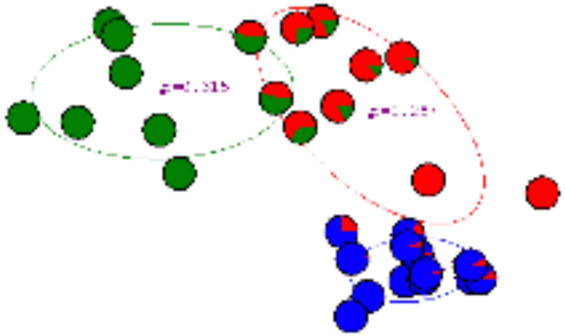
- 5-th iteration



(Fig. From A. Moore's Tutorial)

E-M example

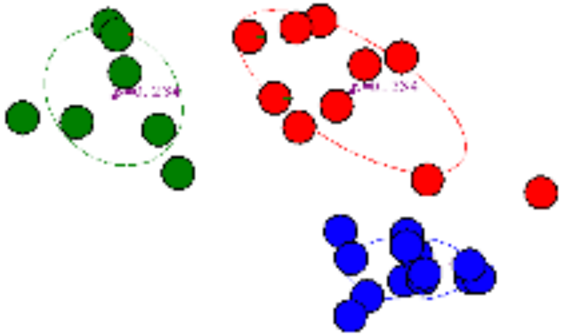
- 6-th iteration



(Fig. From A. Moore's Tutorial)

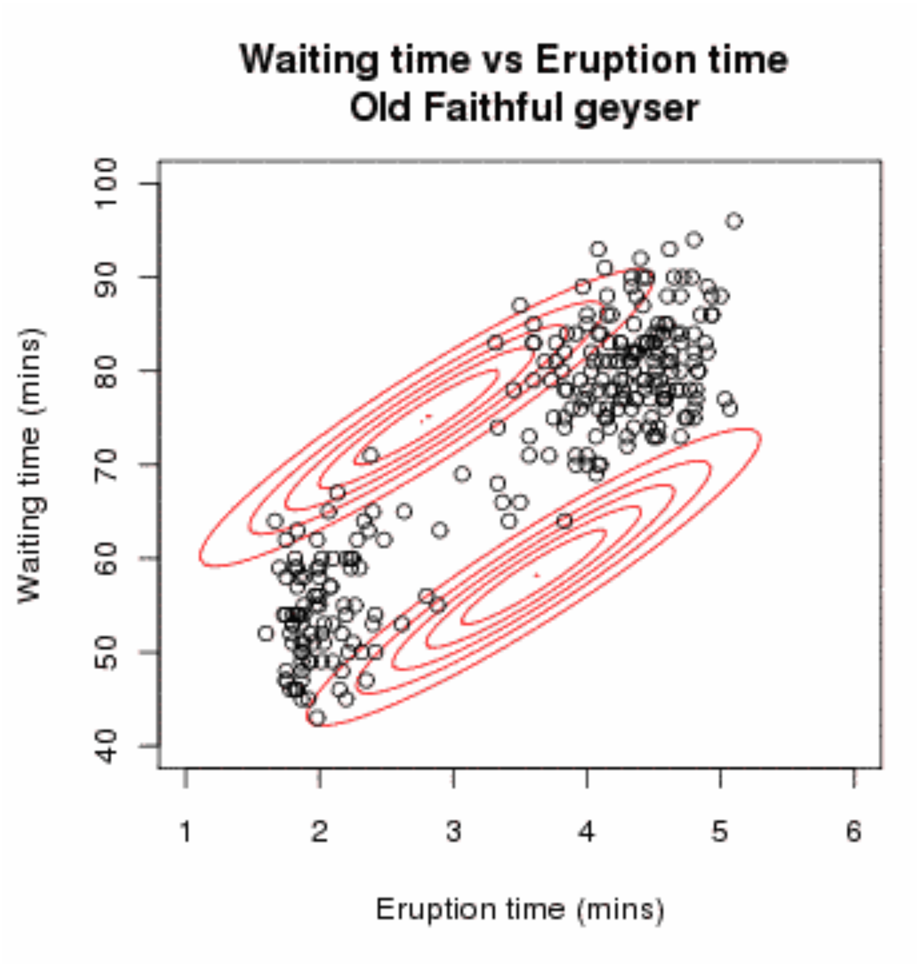
E-M example

- 20-th iteration



(Fig. From A. Moore's Tutorial)

Density Estimation



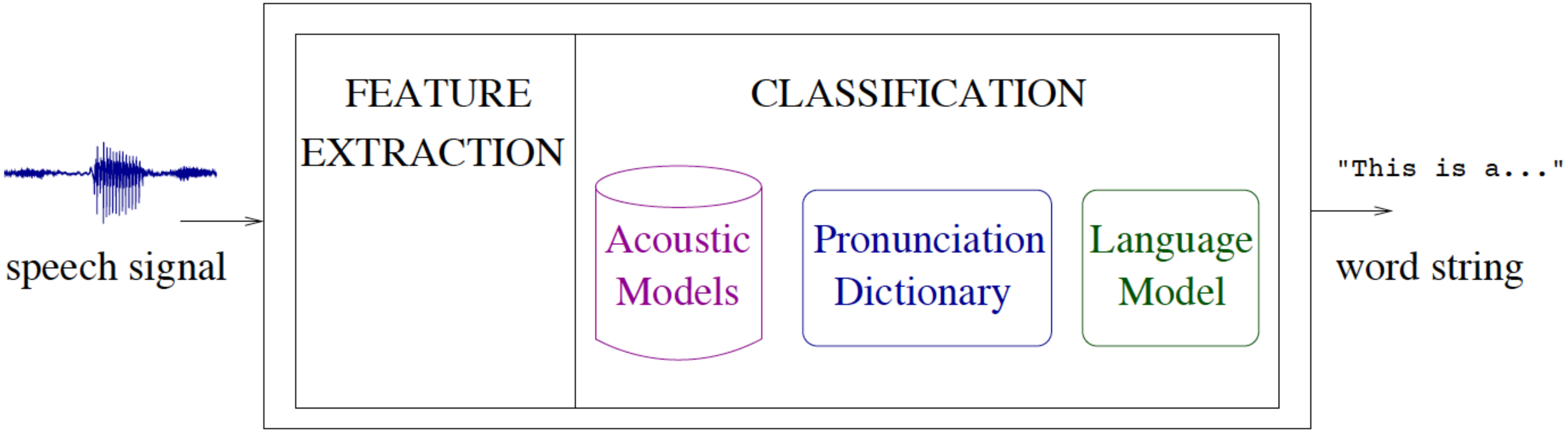
(Fig. Wikipedia)

GMMs for the speaker recognition task

- Given a density of probability estimated for each speaker
  - Search the one that best explains the observed features
- Recent systems are more complex than this
  - Universal Background Model (UBM)
  - Nuisance Attribute Projection (NAP)

Speech recognition

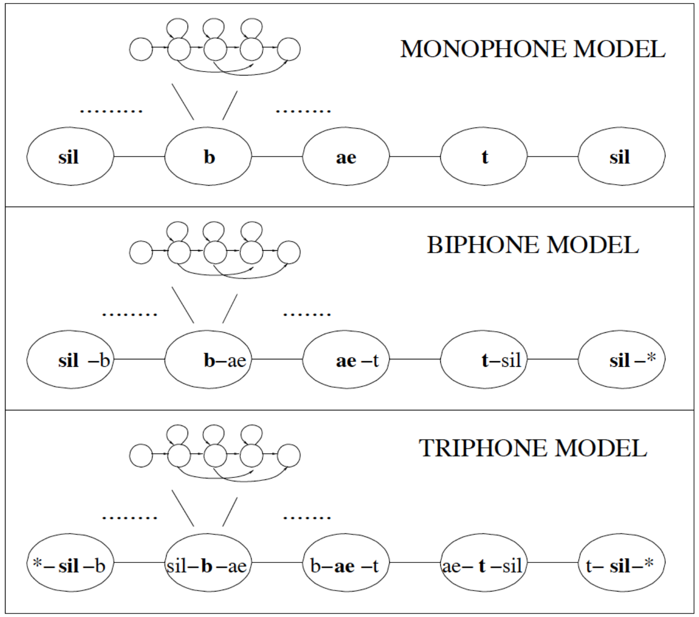
- The aim of an Automatic Speech Recognizer (ASR) is to
    - Output the spoken words
    - Using the speech signal only.



(Fig. from HTK documentation)

Speech recognition

- An Automatic Speech Recognition System is typically decomposed into:
  - Feature Extraction: MFCCs
  - Acoustic Models: HMMs trained for set of phones
    - Each phone is modelled with 3 states
  - Pronunciation dictionary: convert a series of phones into a word
  - Language Model: predict the likelihood of specific words occurring one after another with n-grams



(Fig. from HTK documentation)

Summary

- A convenient way of modeling sound is to split the sampled signal into overlapping frames within which the signal is considered as stationary
- Speech encoding:
  - reduced set of parameters that is necessary to synthesize a perceptually similar signal
  - Non parametric: Vector Quantization (K-means)
- Speaker recognition
  - Need to abstract the signal into a meaningful set of parameters: MFCCs
- Speech recognition
  - Sequentiality is important: from GMMs to HMMs

Outline

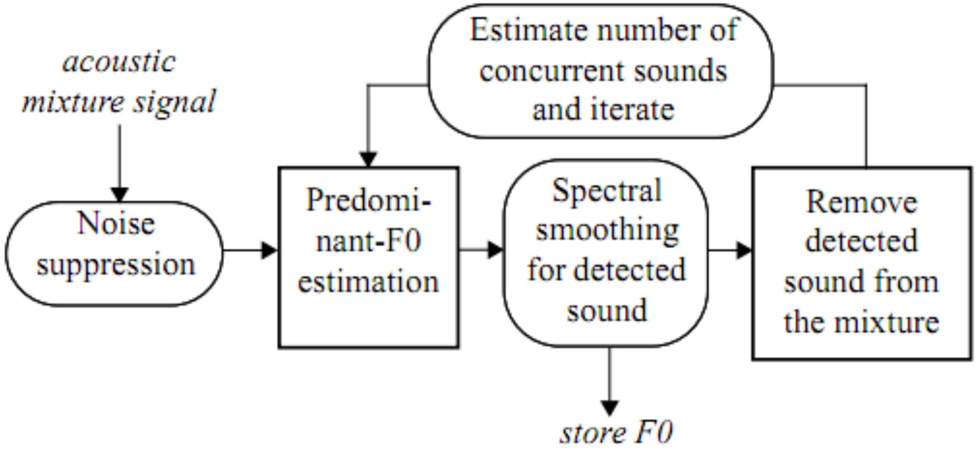
1. Introduction
   1. Context and challenges
2. Past and Present
   1. Speech
      1. Model
      2. Applications (coding, speaker recognition, speech recognition)
   2. Audio (Music)
      1. Sound models
      2. Retrieving information within songs
      3. Retrieving information across songs

Retrieving information within songs

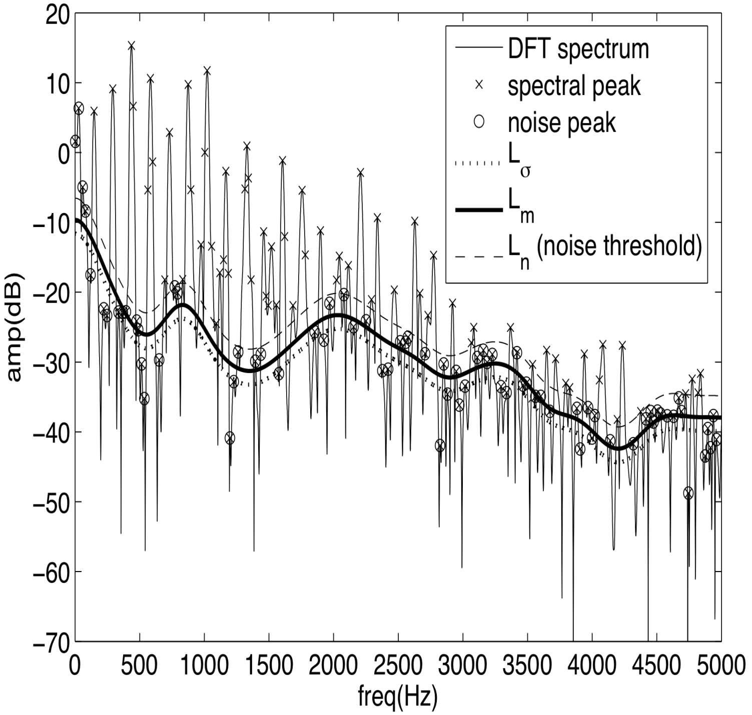
- Harmony
  - Pitch Tracking
  - Melody estimation
  - Multi-F0 estimation
  - Chord estimation
- Rhythm
  - Onset detection
  - Tempo estimation
  - Beat tracking
- Orchestration
  - Instrument recognition

Multi-F0 Estimation

- From an observed spectrum, we want to estimate the fundamental frequency (f0) of each note.
  - Most algorithms perform an iterative search:
    - Estimate the dominant f0
    - Remove its contribution

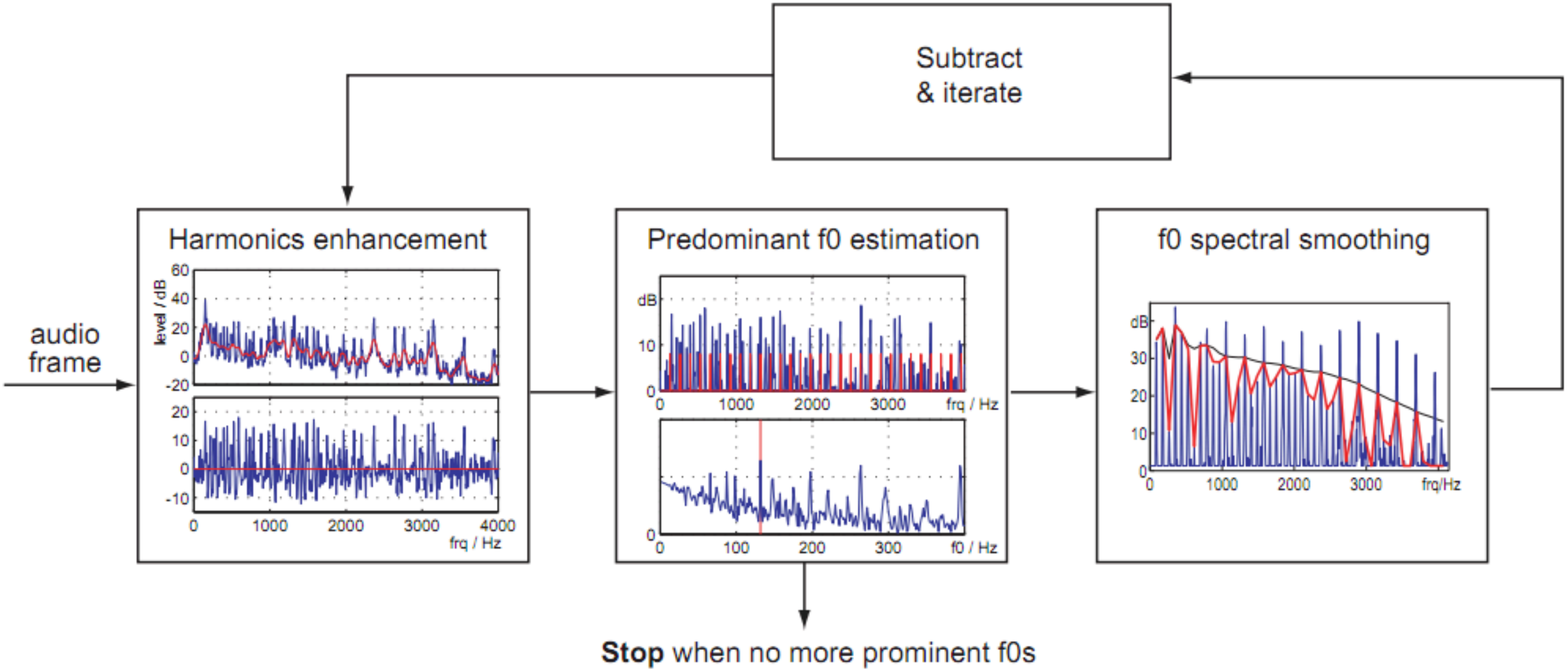

acoustic mixture signal
Noise suppression
Predominant-F0 estimation
Spectral smoothing for detected sound
Remove detected sound from the mixture
Estimate number of concurrent sounds and iterate
store F0

(Fig. from Klapuri 2004)



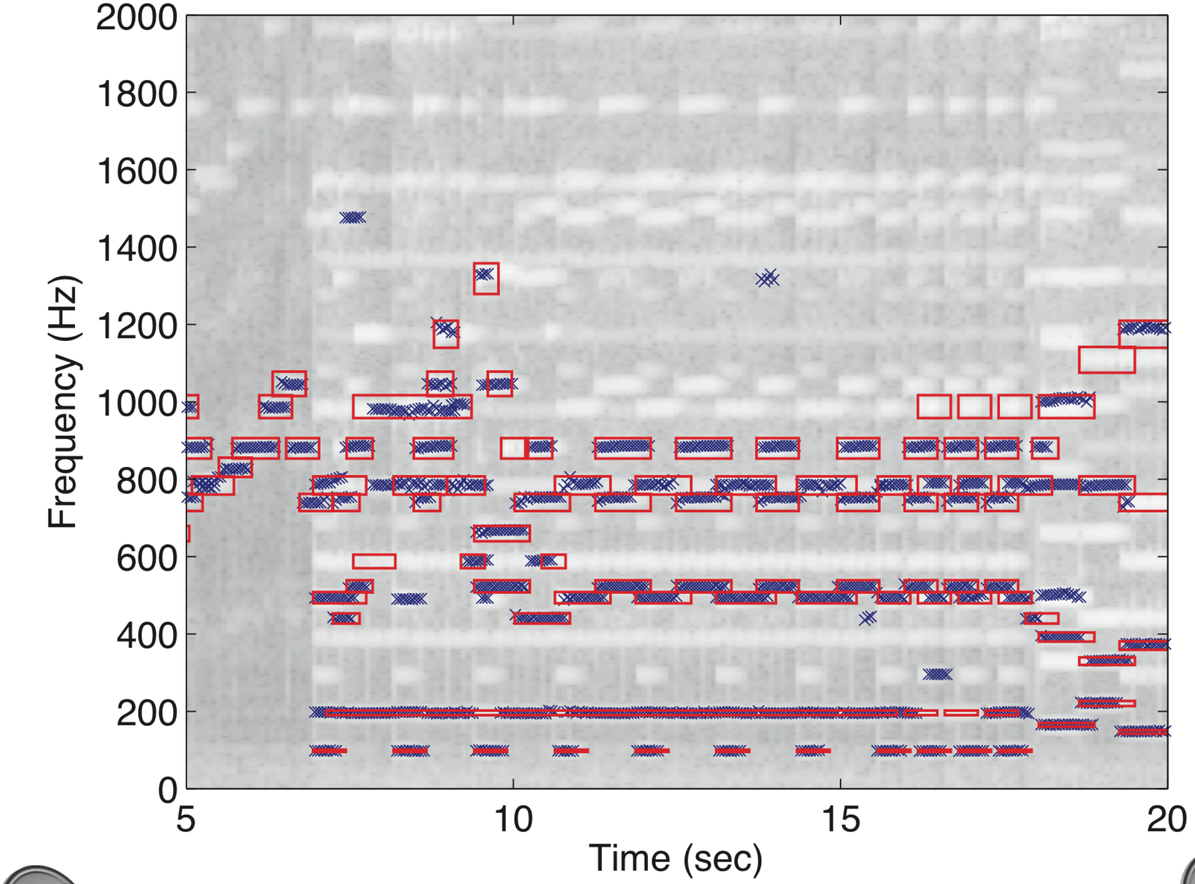
(Fig. from Yeh 2010)

Multi-F0 Estimation



Stop when no more prominent f0s

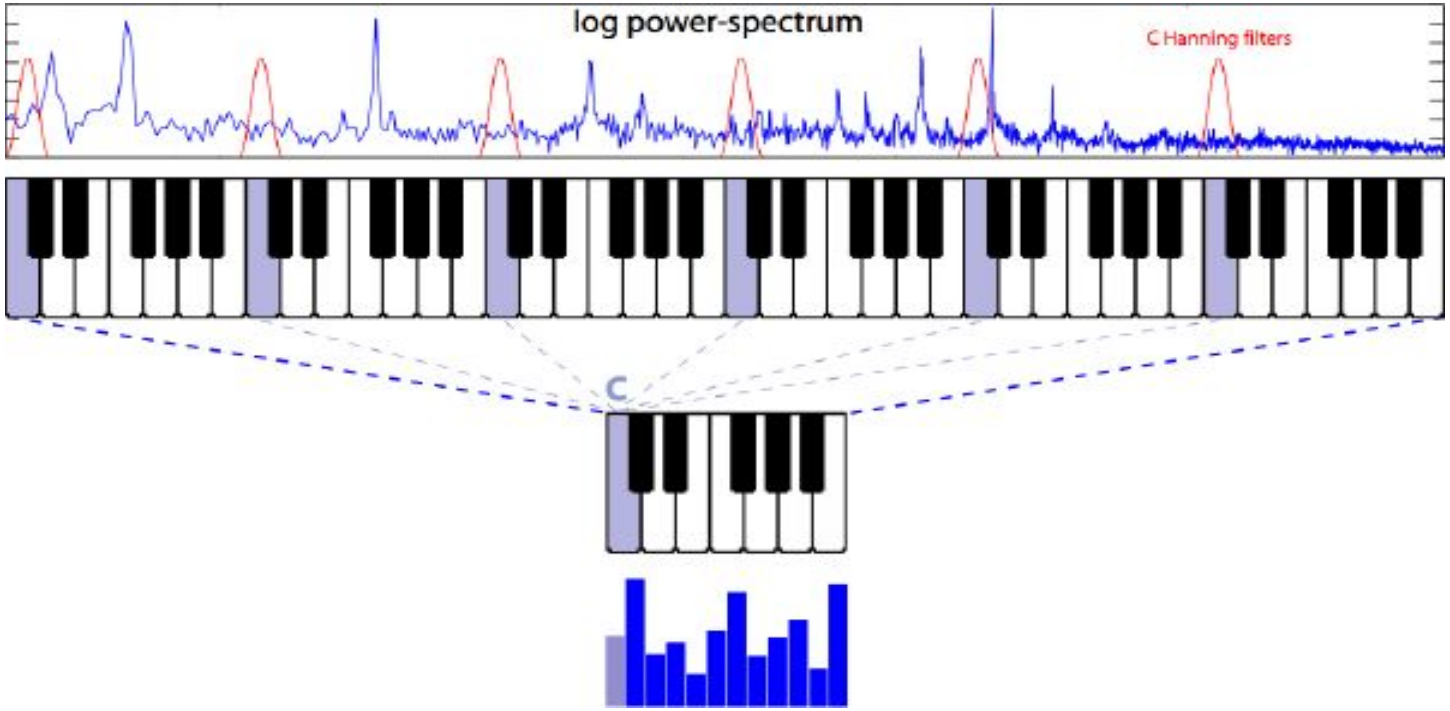
Multi-F0 Estimation



(Fig. from Yeh 2010)

Chord Estimation

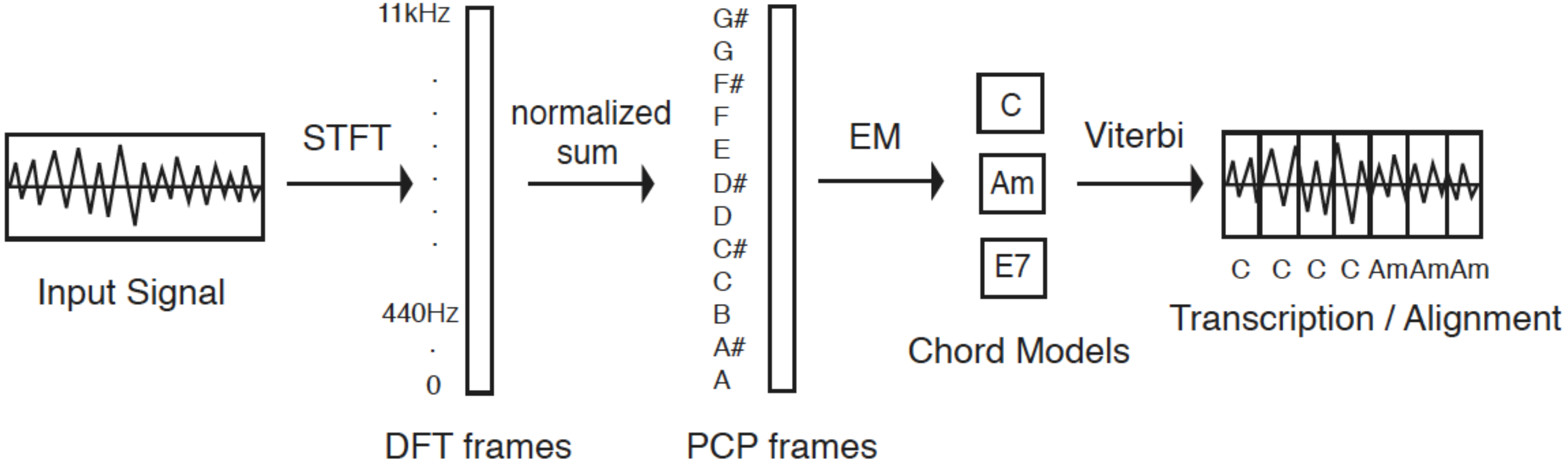
- Relies on a Chroma representation



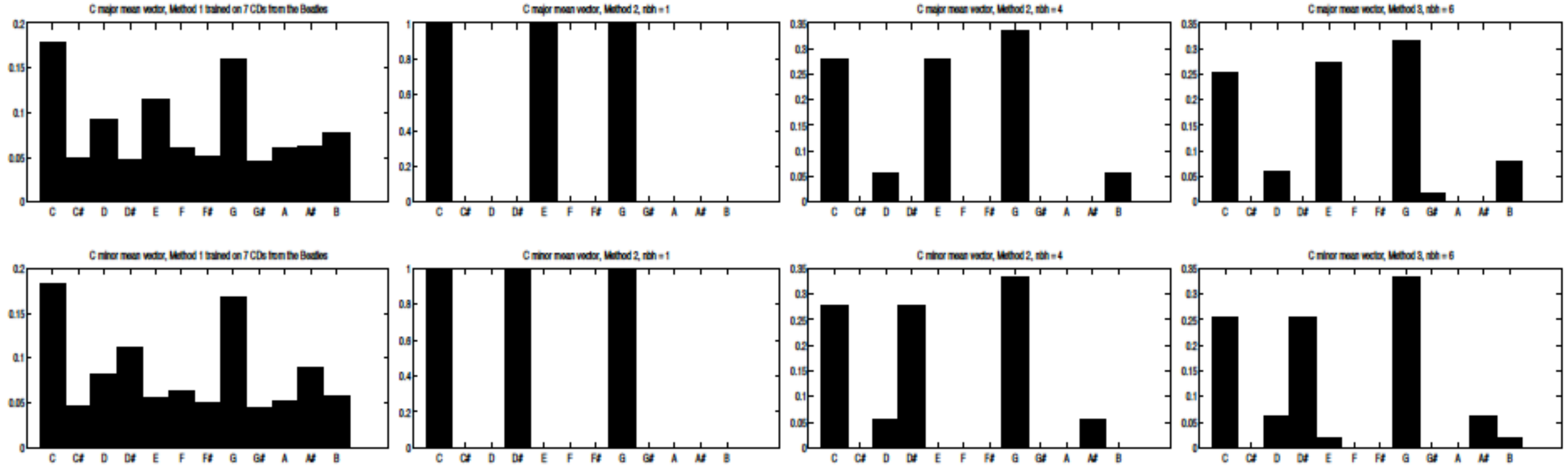
(Fig. from Padapopoulos 2007)

Chord Estimation

- Matches observed chromas to chords templates



Issues with template models



- World is dirty, so are our models
- Our models are clean, let us clean the world

Structure Analysis

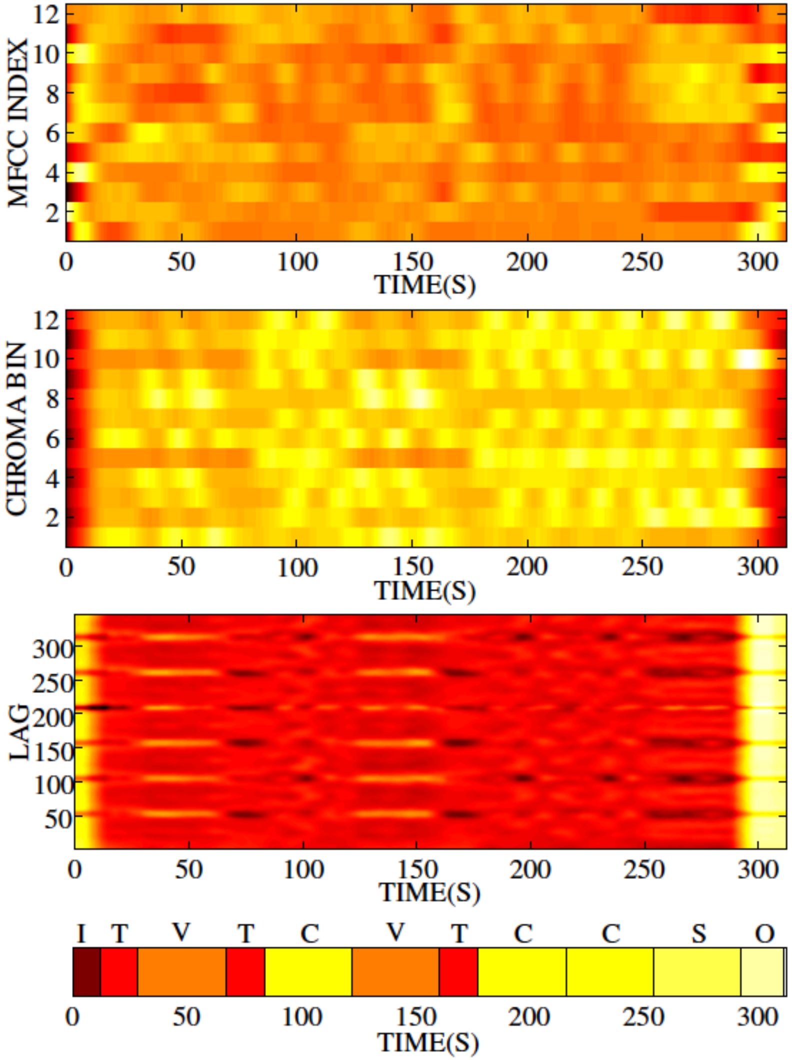
- Aims at estimating the musical structure
  - Ex.: Intro, Verse, Chorus, Verse, Chorus, Chorus, Outro
- Method:
  - Compute features
  - Compute the similarity between every features
  - Perform segmentation based on
    - Novelty
    - Homogeneity
    - Repetition

Paulus'10

Pauls J., Muller M. and Klapuri A.. AUDIO-BASED MUSIC STRUCTURE ANALYSIS
11th International Society for Music Information Retrieval Conference (ISMIR 2010)

Structure Analysis

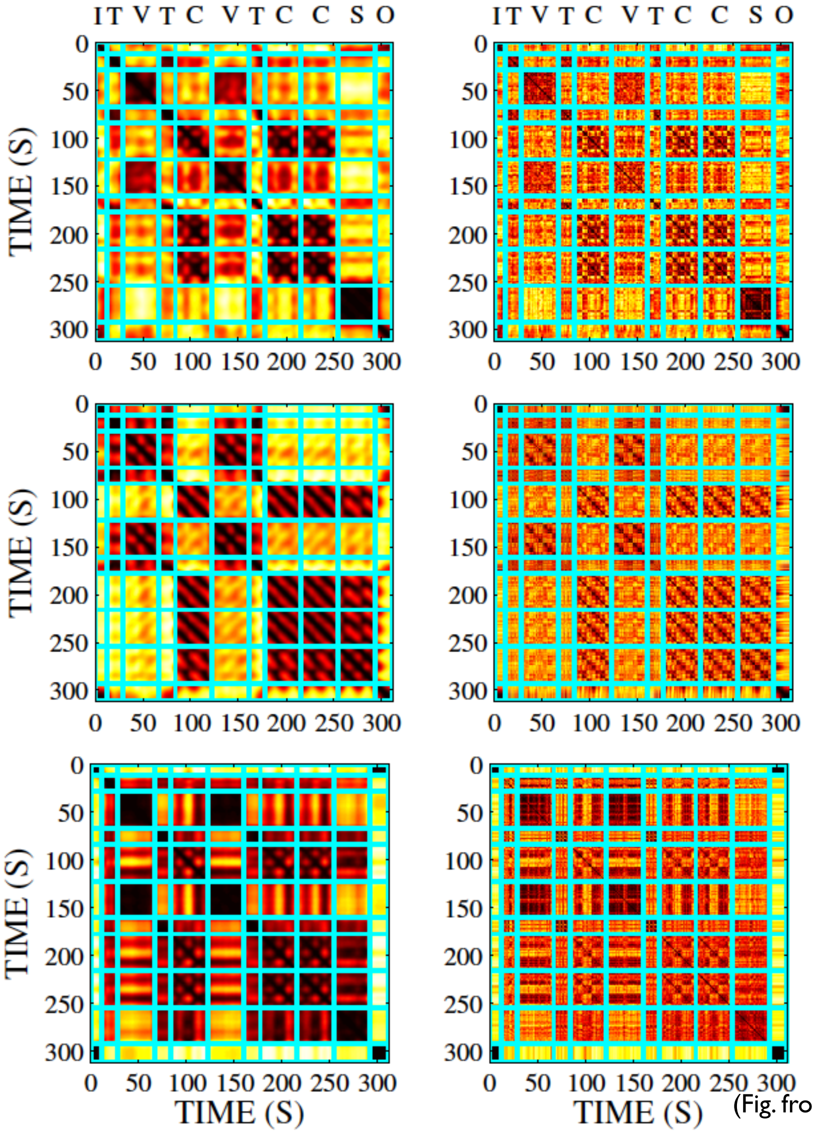
- Feature extraction
  - Timbre: MFCCs
  - Harmony: Chromas
  - Rhythm.



(Fig. from Paulus 2010)

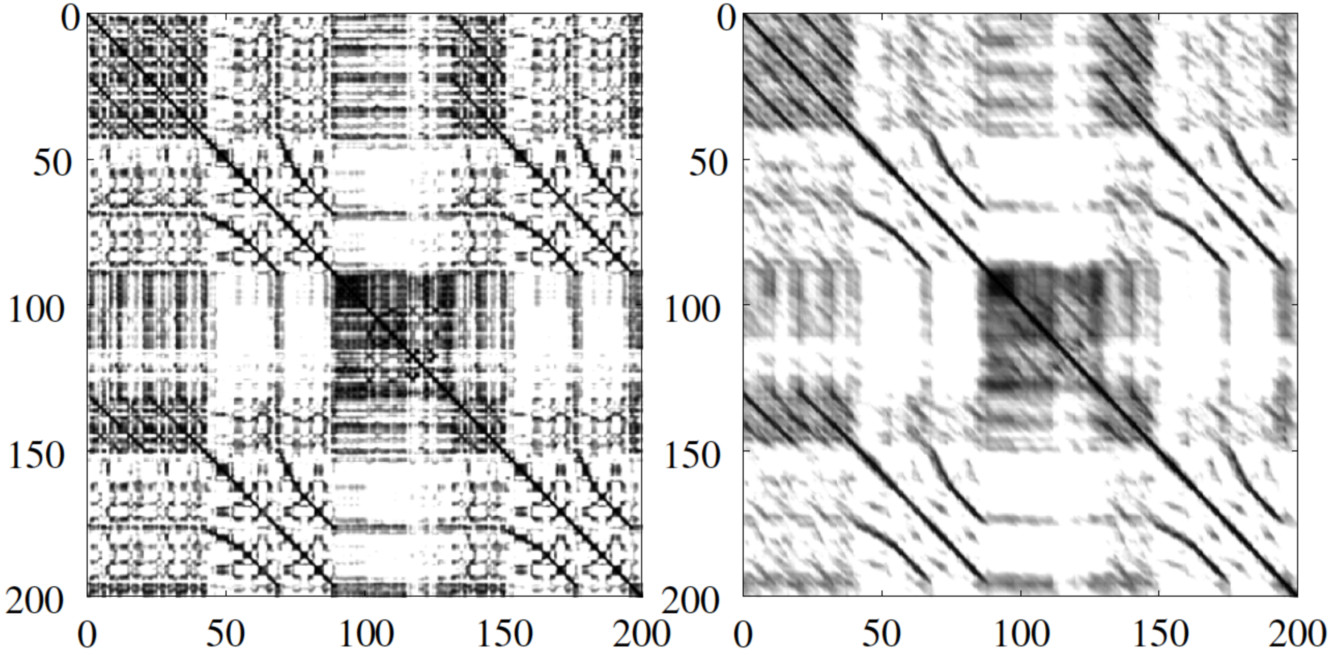
Structure Analysis

- Self-Similarity Matrix
  - For the 3 features
  - At different granularity



(Fig. from Paulus 2010)

Structure Analysis

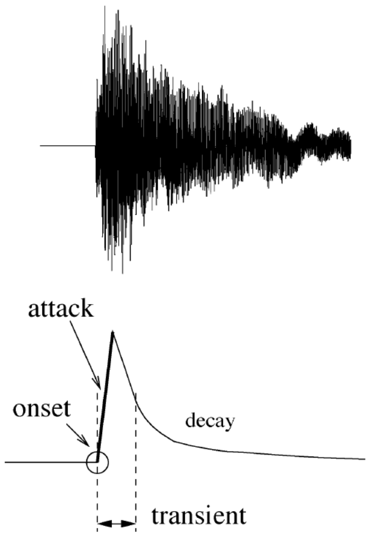


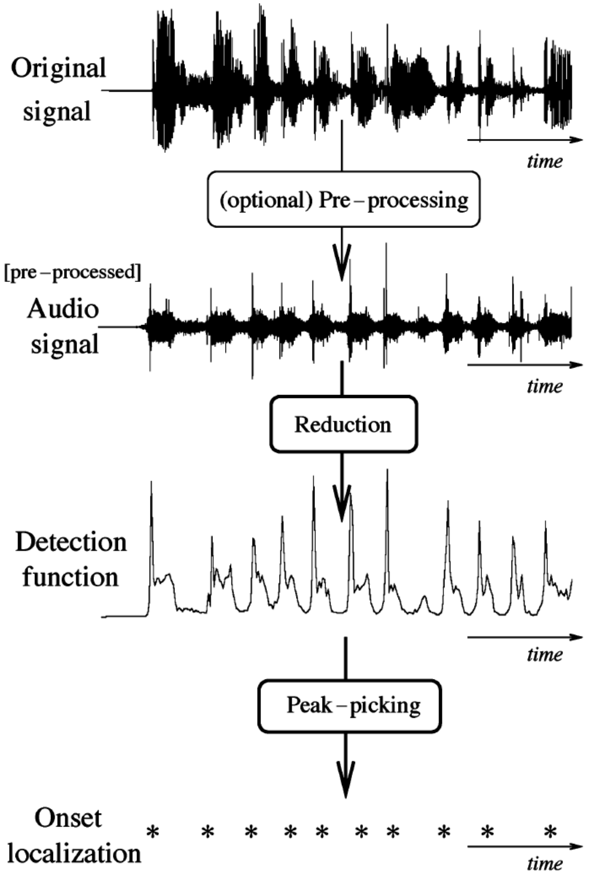
(Fig. from Paulus 2010)

- segmentation based on
  - Novelty
  - Homogeneity
  - Repetition

Onset Detection

- What is an onset ?
- How do we extract it ?

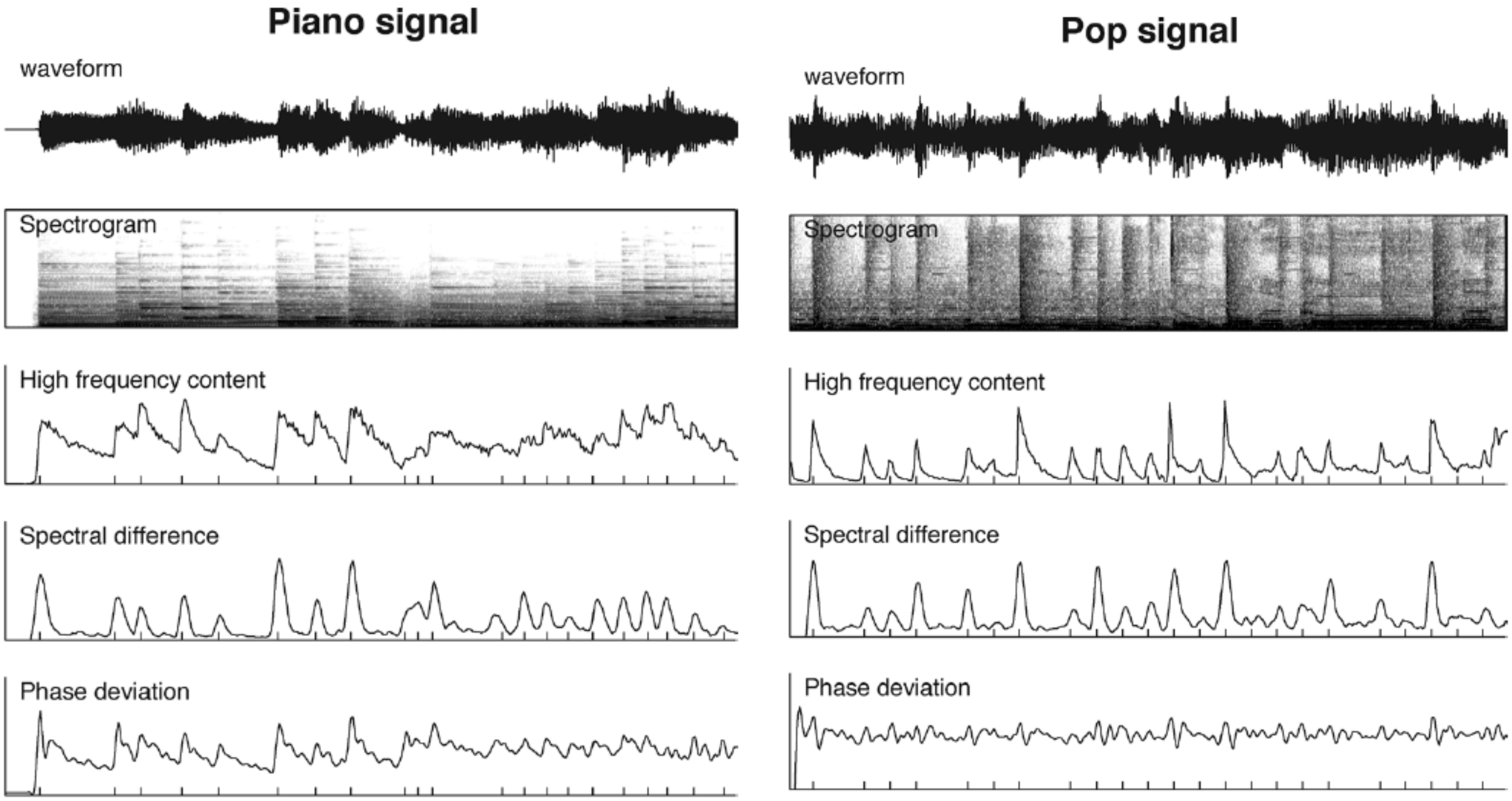




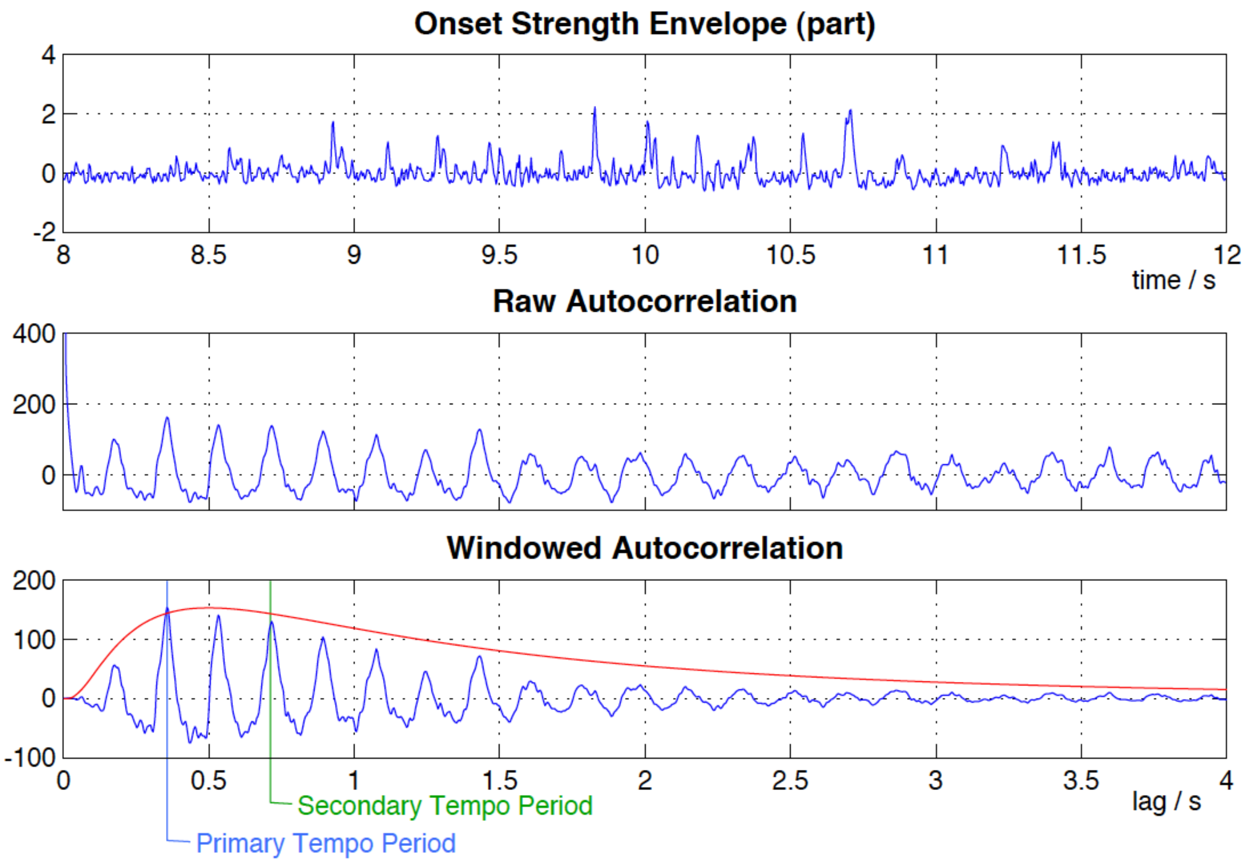
Bello & al

J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions on, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.*

Onset detection

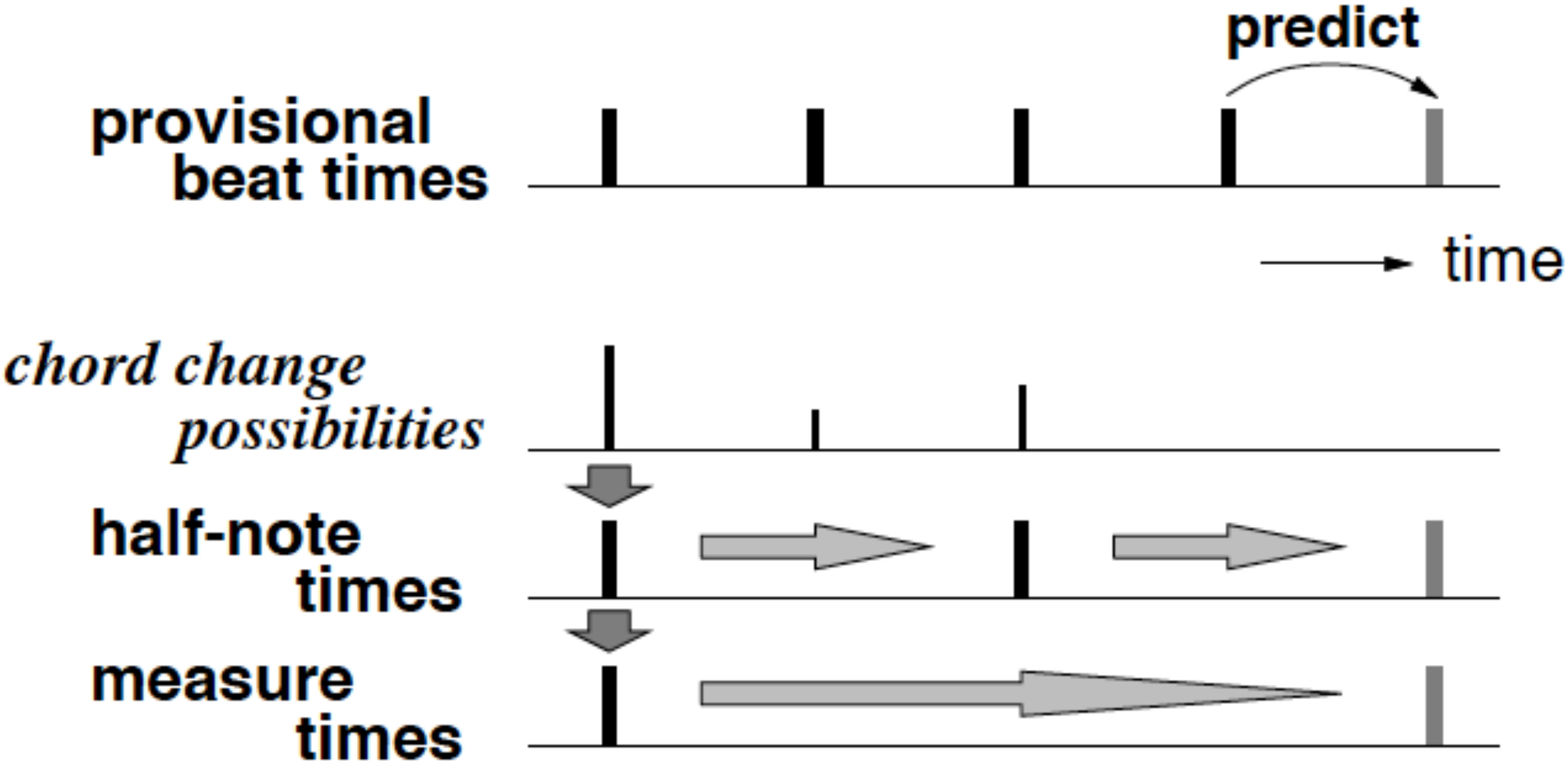


Tempo estimation



Beat tracking

- From an estimate of the tempo
  - Infer the beat position
  - Possibly the down beat



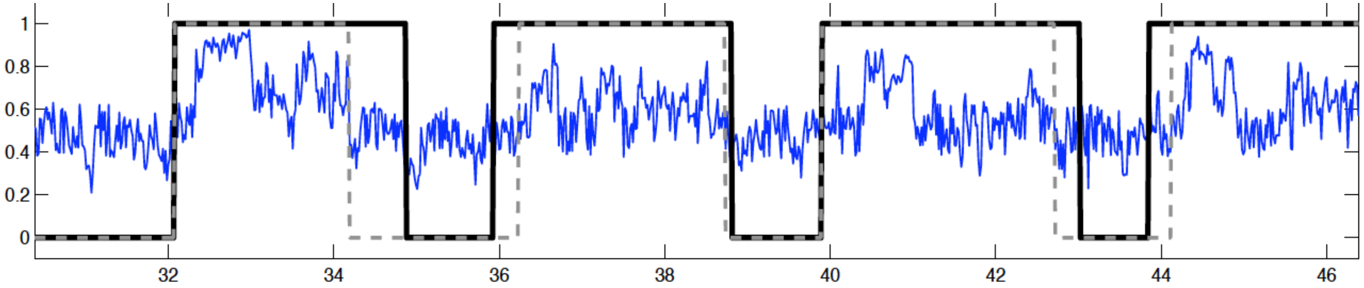
Retrieving information across songs

- Classification
  - Genre recognition
  - Tag inference
- Similarity
  - Music similarity
  - Cover detection

Classification

- Method: [Tzanetakis'02]
  - Agree on mutually exclusive set of tags (the ontology)
  - Extract features from audio (MFCCs and variations)
  - Train statistical models:
    - Due to the high dimensionality of the feature vectors discriminatives approaches are prefered (SVMs)

- Segmentation
  - Smoothing decision using dynamic programming (DP)

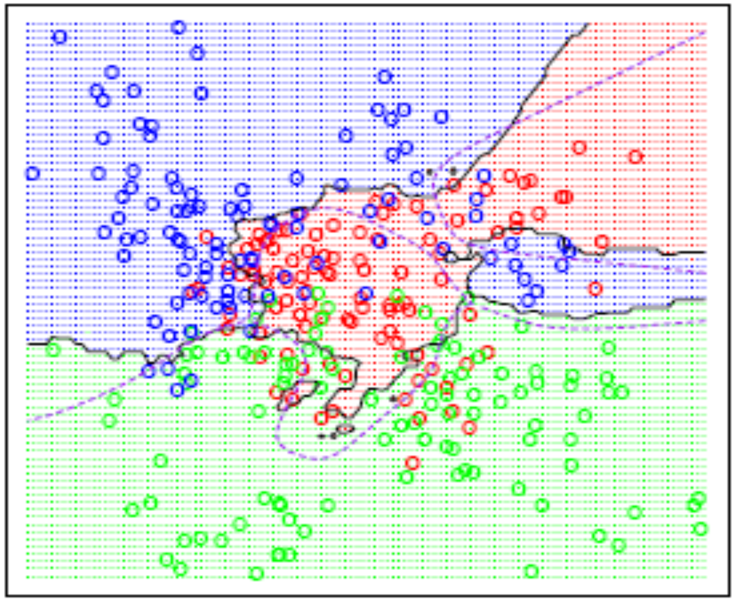


(Fig. from [Ramona07])

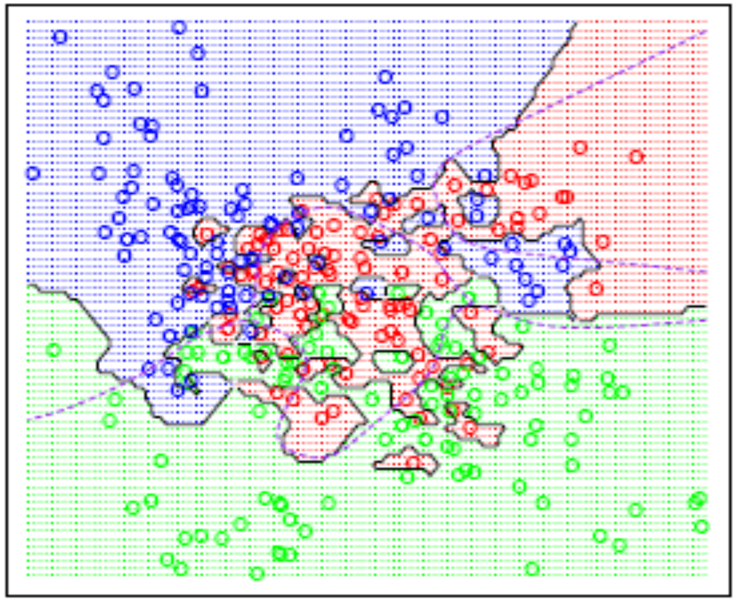
Tzanetakis'02 Tzanetakis, G. Cook, P. Musical Genre Classification of Audio Signals
IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING 2002

From parametric to non-parametric

- K-Nearest Neighbors: simple but effective non parametric approach to classification
- Assuming given
  - A metric that defines the similarity of 2 items
  - Some labeled items



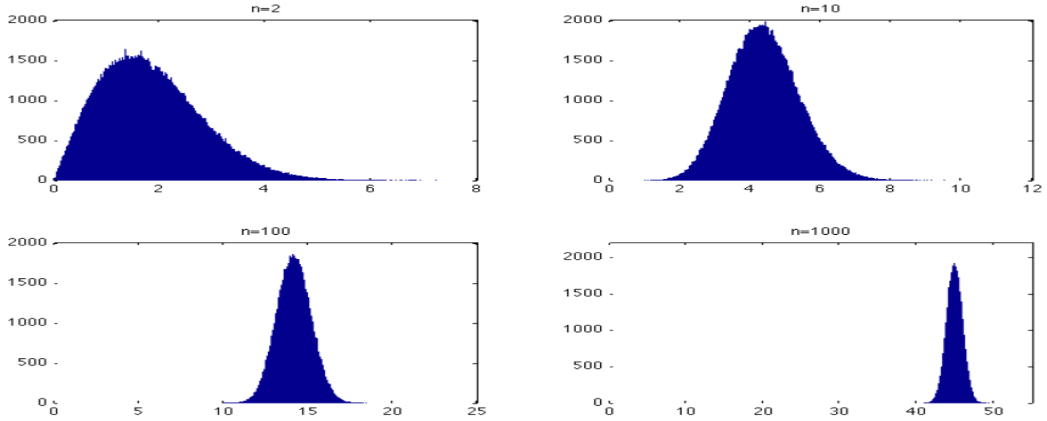
k=15



k=1

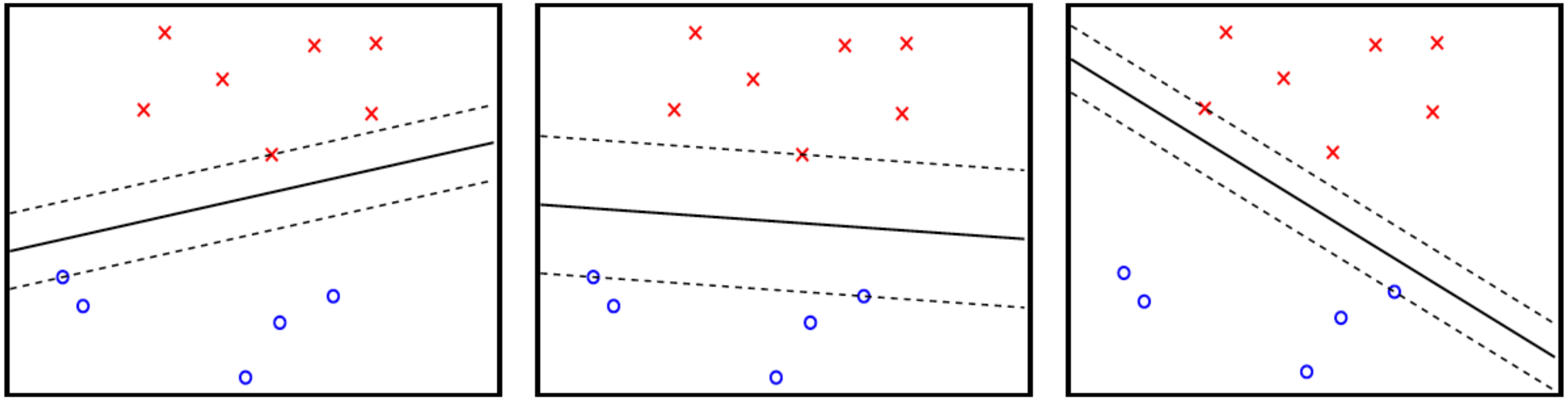
Issues with k-NN

- Do not scale to large problems (lots of items)
- In high dimensions (lots of features),
  - due to the curse of dimensionality,
  - items tend to be equally far from each others
  - neighbors tend to be non local and meaningless



Support Vector Machines (SVMs)

- Discriminative approach towards classification
  - In the linear case, SVMs aim at maximizing the distance between margin hyperplanes (dashed lines), called the margin M.

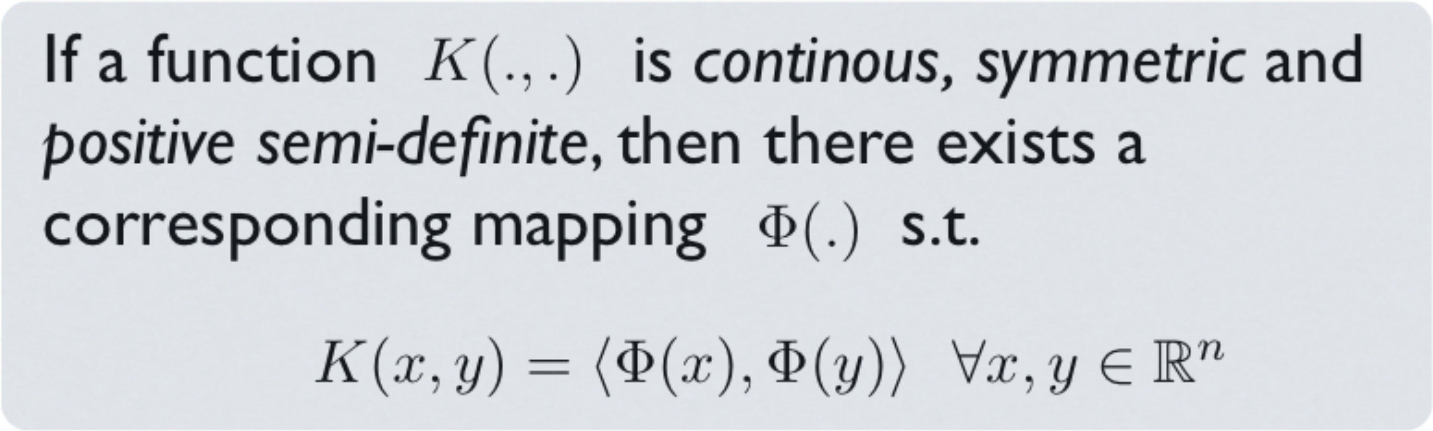


(Fig. from [Ramona Phd])

  - Allows to minimize the structural risk by jointly minimize the
    - Empirical risk
    - Dimension de Vapnik et Chervonenkis

Kernel-based SVMs

- Data is usually non-linearly separable wich lead to the use of some kernel function to project the data into higher dimensional space



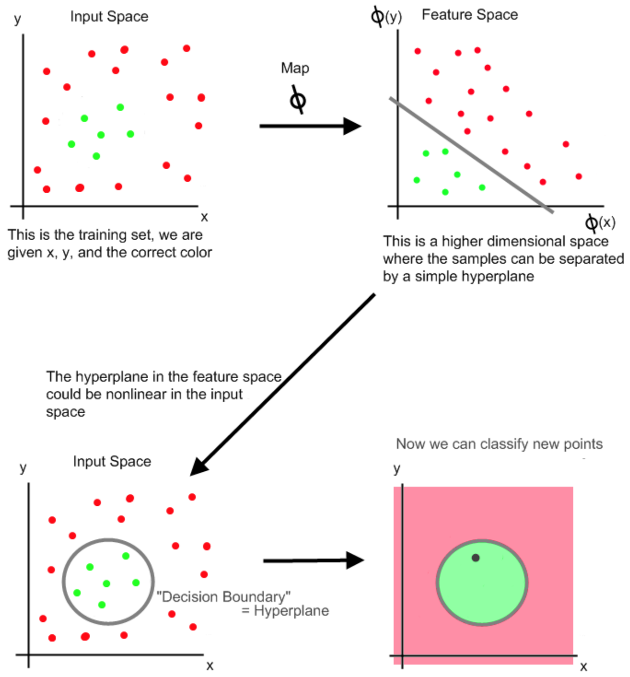
If a function $K(.,.)$ is *continous, symmetric* and *positive semi-definite*, then there exists a corresponding mapping $\Phi(.)$ s.t.

$$K(x,y) = \langle \Phi(x), \Phi(y) \rangle \quad \forall x, y \in \mathbb{R}^n$$

Kernel-based SVMs

Let us consider a polynomial kernel (Video) such as

K([x1, x2], [y1, y2]) = x1x2+y1y2+(x1^2+y1^2)(x2^2+y2^-2)


y
Input Space
x
This is the training set, we are given x, y, and the correct color
Map
ϕ
ϕ(y)
Feature Space
ϕ(x)
This is a higher dimensional space where the samples can be separated by a simple hyperplane
The hyperplane in the feature space could be nonlinear in the input space
y
Input Space
"Decision Boundary" = Hyperplane
x
Now we can classify new points
y
x

Multi-Class Discriminative Classification

- Usually performed by combining binary classifiers

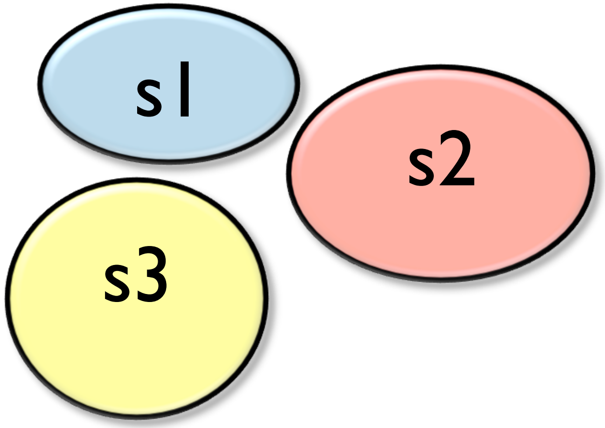
- Two approaches:
  - **One-vs-all**: For each class build a classifier for that class versus the rest

    $$f(\mathbf{x}) = \arg\max_i f_i(\mathbf{x}).$$

    - Often very imbalanced classifiers (use asymmetric regularization)
  - **All-vs-all** Build a classifier for each couple of class

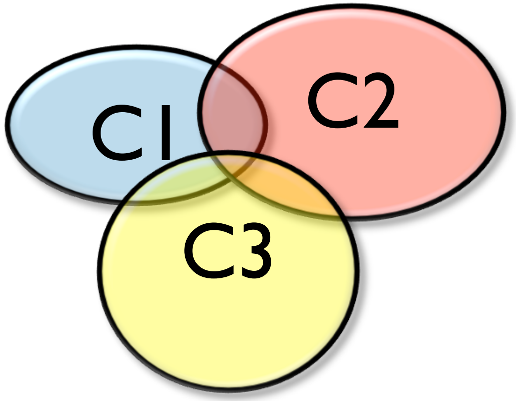
    $$f(\mathbf{x}) = \arg\max_i \left(\sum_j f_{ij}(\mathbf{x}) \right).$$

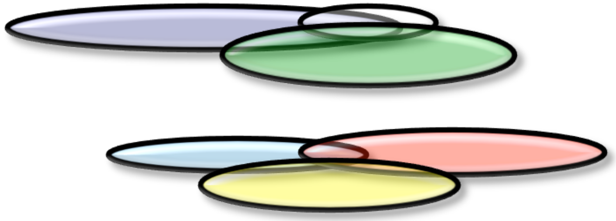
    - A priori a large number of classifiers to build ***but*** the pairwise classification are faster and the classifications are balanced (easier to find the best regularization)



Multi-Label Discriminative Classification

- Each object may be tagged using several labels
- Computational approaches
  - Power Sets
  - Binary Relevance (equivalent to one-vs-all)
- Multiple criteria:
  - « Flattening » the ontology
  - Research trend: considering the ontology structure to benefit from co-occurrence labels of different semantic criterion





Music Similarity

- Question to solve: « Given a seed song, provide us with the entries of the database which are the most similar »
- Annotation type: Artist / Album
- Method:
  - Songs are modeled as a Bag of Features (BoF), usually Gaussian models of MFCCs
  - proximity of GMMs are considered as similiarity measure
    - Diagonal covariance GMMs [Aucouturier'04]:
      - Likelihood (requires access to the MFCCs)
      - Monte carlo sampling
    - Full covariance Gaussian:
      - KL divergence

[Aucouturier'04] J.-J. Aucouturier and F. Pachet. Improving Timbre Similarity: How High is the Sky? *Journal of Negative Results in Speech and Audio Sciences,* 1 (1), 2004.

Cover Version Detection

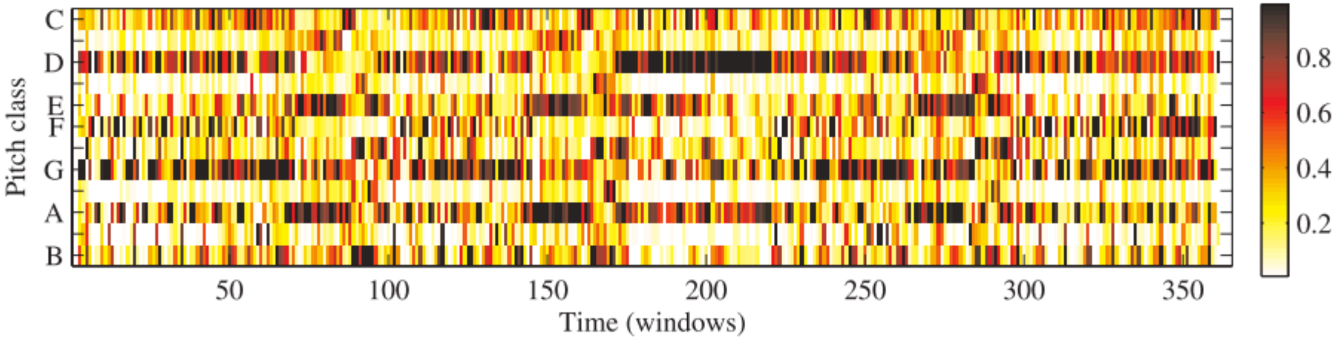
- Question to solve: « Given a seed song, provide us with the entries of the database which are cover versions »
- Annotation: canonical song
- Method: [Serra'08]
    - Songs are modeled as a time series of Chromas
    - Computation of the similarity matrix between the two time series
    - Similarity is measured using Dynamic Programming Local Alignment

(Fig. from [Serr08])

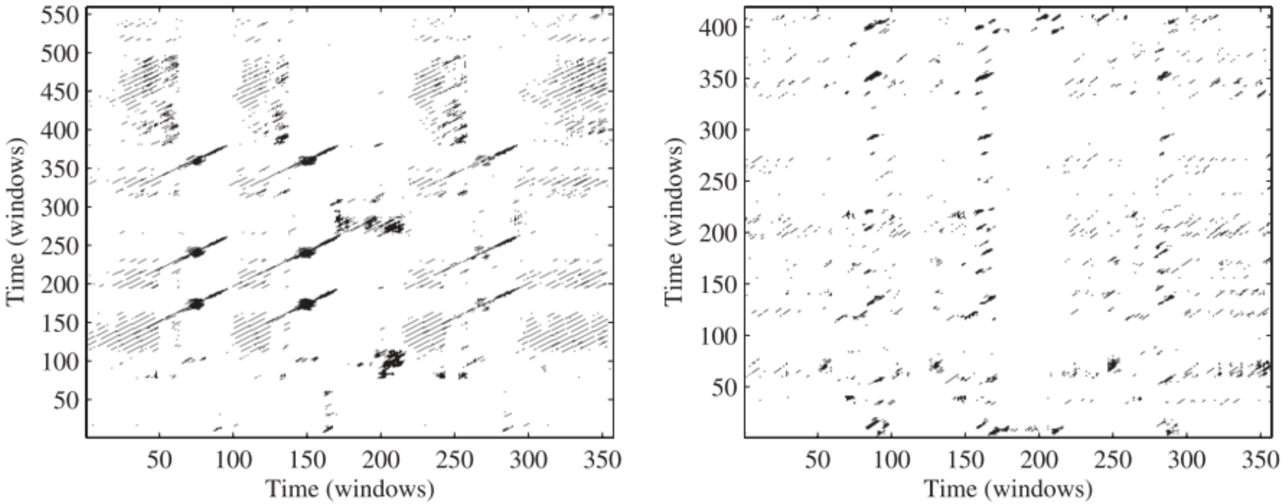
[Serra'08] Chroma Binary Similarity and Local AlignmentApplied to Cover Song Identification, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2008

Cover Song Detection

- Chromagram of “Day Tripper”



- Chroma similarity of “Day Tripper” with a cover (left) not a cover (right)



Future

1. Issues
    1. Description of audio and music
        1. Polyphonic
        2. Multiple shapes varying in various ways
    2. Statistical Modeling
        1. Curse of dimensionality
        2. Sense of structure relevant at multiple levels of temporality
2. Research Trends
    1. Polyphony handling by Source separation
    2. Building and using priors by performing Auditory Scene Analysis (ASA)
    3. Joint estimation of several musical parameters

Vocabulary ?

- STFT
- MFCCs
- Chromas
- K-means
- GMMs
- HMMs

References

- Music information Retrieval (MIR) is an emerging field
  - Browse ISMIR proceedings
    - using Google Scholar and use citation index
  - Stay tuned via Music-IR mailing list
  - Go to the MIREX webpage to see how things work and are evaluated

Live coding in Matlab

- You can find the source here:
  - http://recherche.ircam.fr/equipes/analyse-synthese/lagrange/teaching/atiam11/coursAtiam2011Intro.m
- You will need some external dependencies, web locations are provided in the code
- The code uses cell mode, please look at the Matlab documentation for usage