Template-based chord recognition from audio signals

Laurent Oudre

Supervisors: Yves Grenier
Cédric Févotte

November 3rd, 2010

Contents

Contents

What is a chord ?

Definition

Chord : aggregate of musical pitches played simultaneously

Characterization of a chord

What is a chord ?

Definition

Chord : aggregate of musical pitches played simultaneously

Characterization of a chord

root note upon which the chord is perceived

What is a chord ?

Definition

Chord : aggregate of musical pitches played simultaneously

Characterization of a chord

root note upon which the chord is perceived

type harmonic structure of the chord

What is a chord ?

Definition

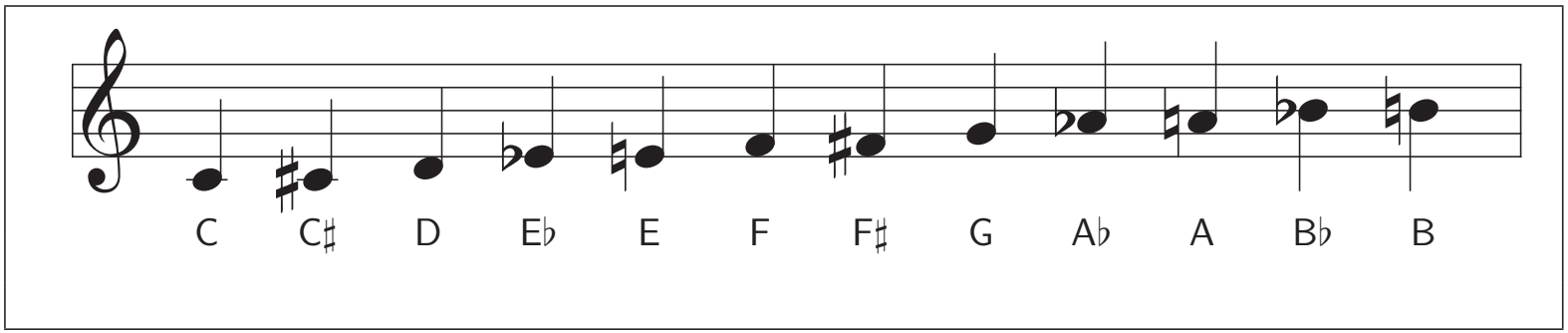
Chord : aggregate of musical pitches played simultaneously

Characterization of a chord

root	note upon which the chord is perceived
type	harmonic structure of the chord
inversion	relationship of the bass note to the other notes in the chord

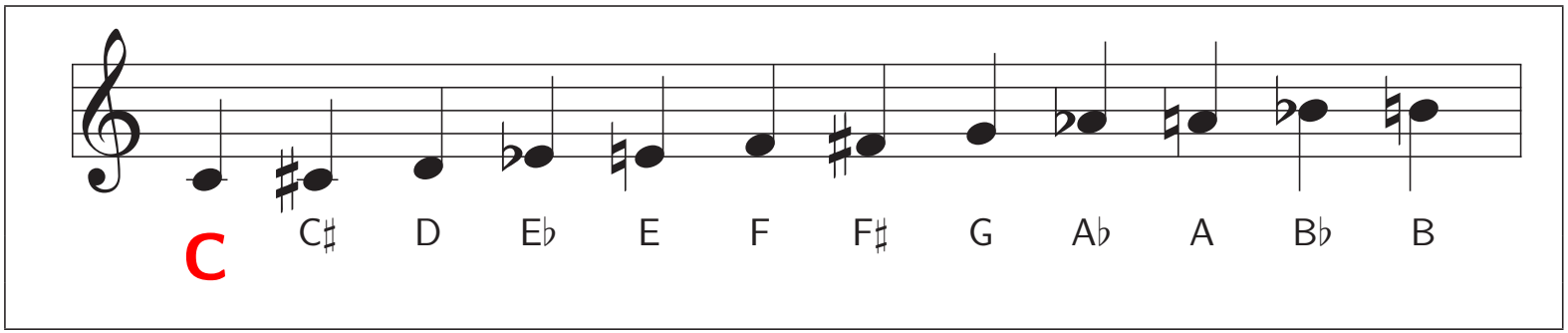
Construction of a C minor chord

C minor

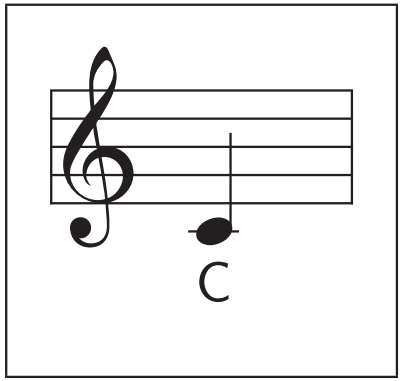


Construction of a C minor chord

C minor

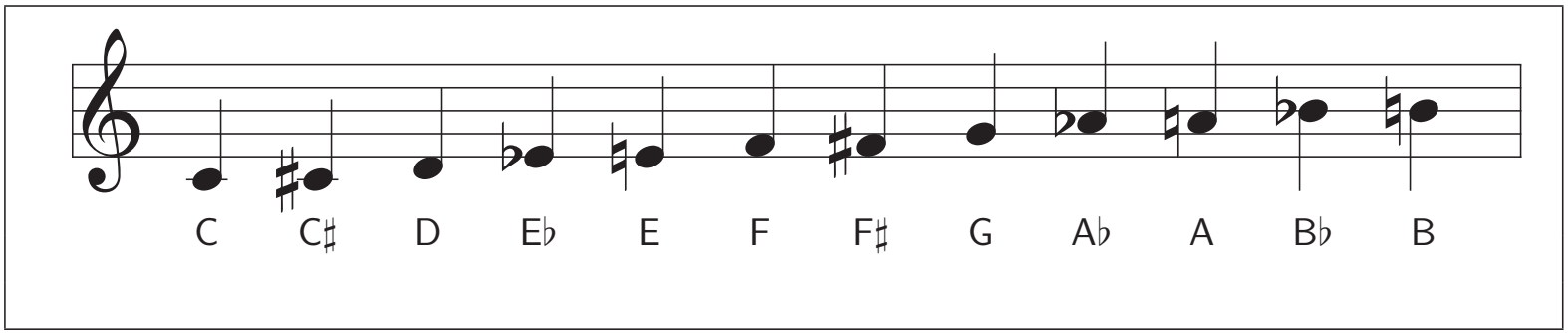


root C



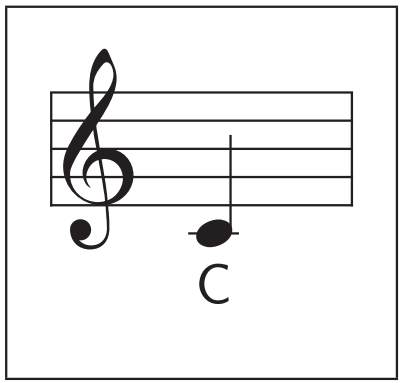
Construction of a C minor chord

C minor



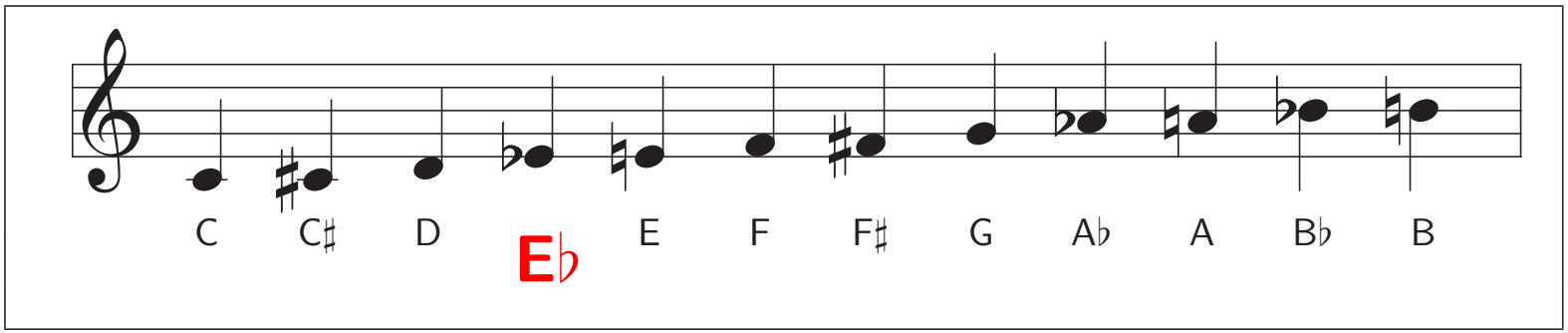
root C

type minor : minor third (E♭) + fifth (G)



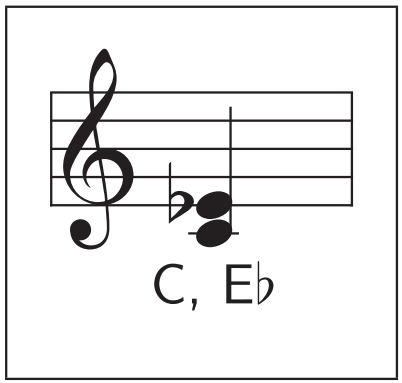
Construction of a C minor chord

C minor



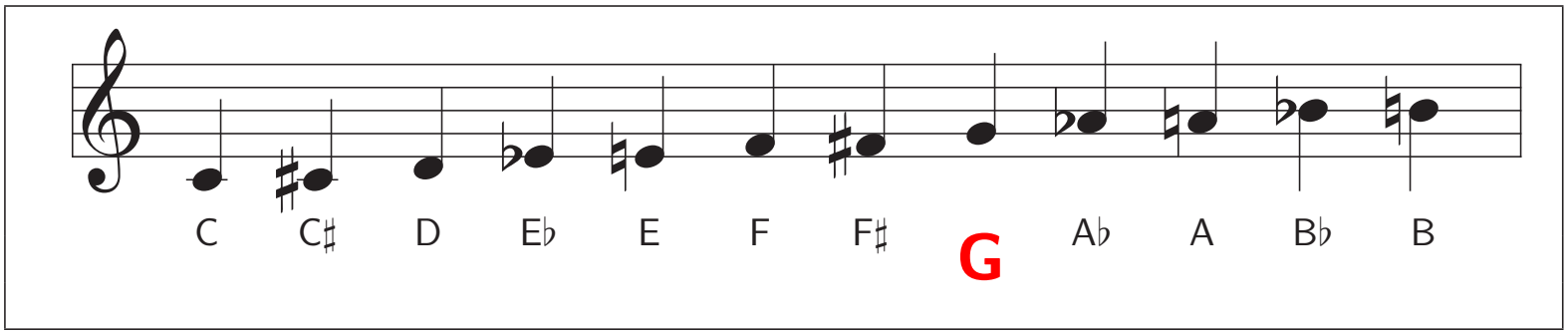
root C

type minor : minor third (E♭) + fifth (G)



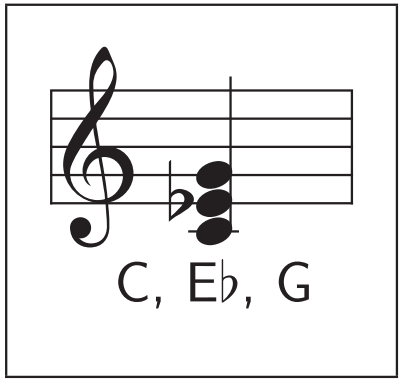
Construction of a C minor chord

C minor



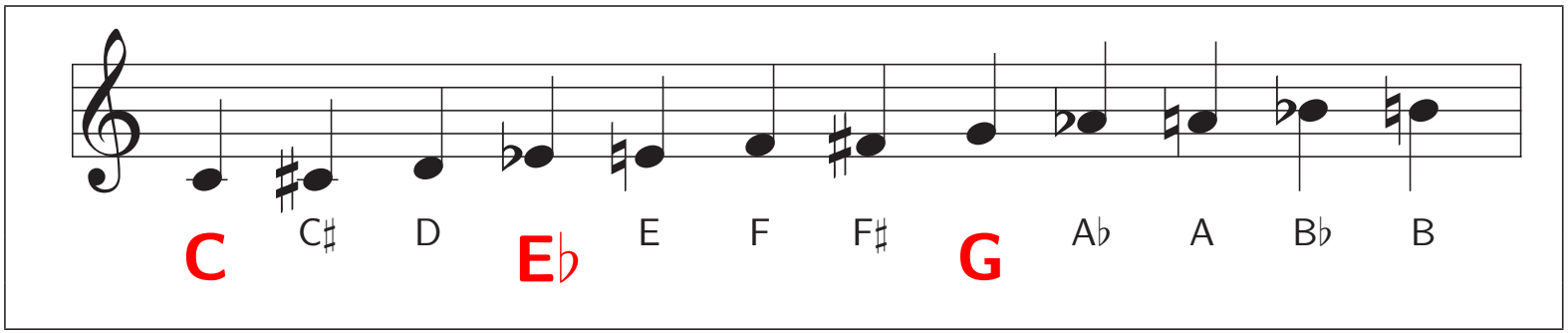
root C

type minor : minor third (E♭) + fifth (G)



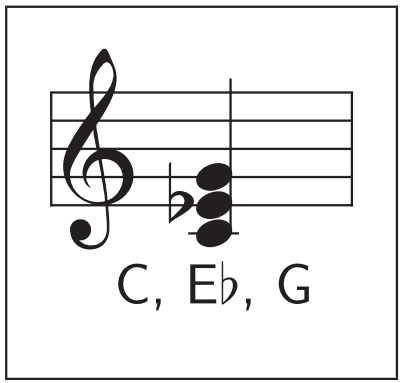
Construction of a C minor chord

C minor



root C

type minor : minor third (E♭) + fifth (G)



Applications

What is a chord transcription ?

Sequence of chords with their respective start and end times

Some applications

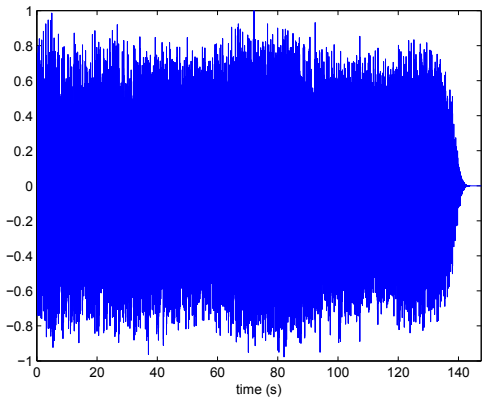
- Song playback
- Key, rhythm, structure
- Song identification
- Query by similarity
- Music classification

Contents

Overview

2 phases

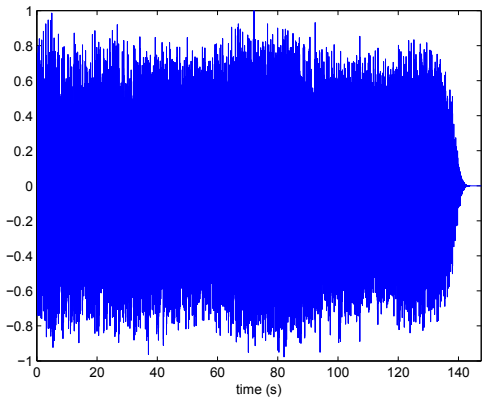
Overview



waveform

2 phases

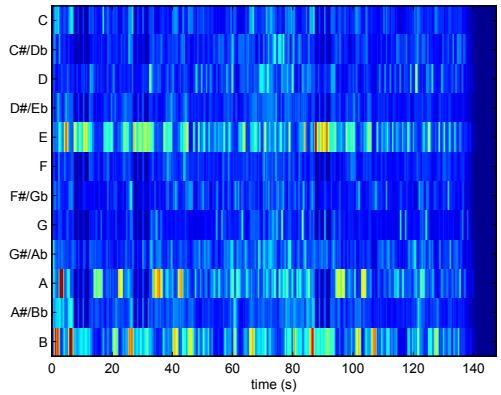
Overview



waveform

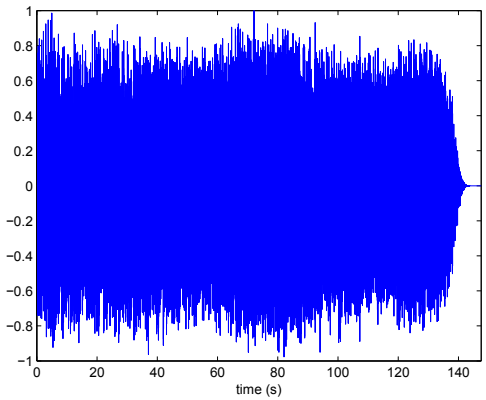
2 phases

Input features calculation



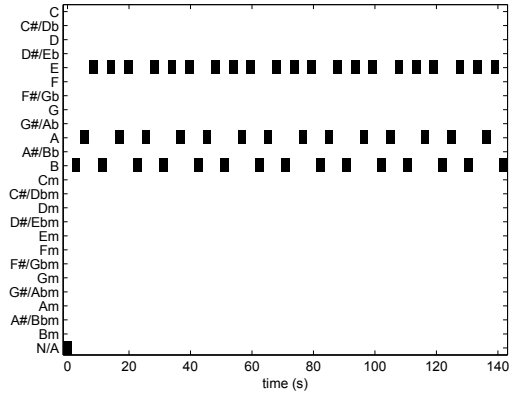
chromagram

Overview



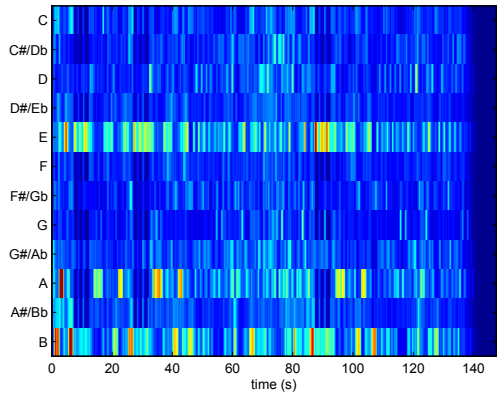
waveform

2 phases



chord transcription

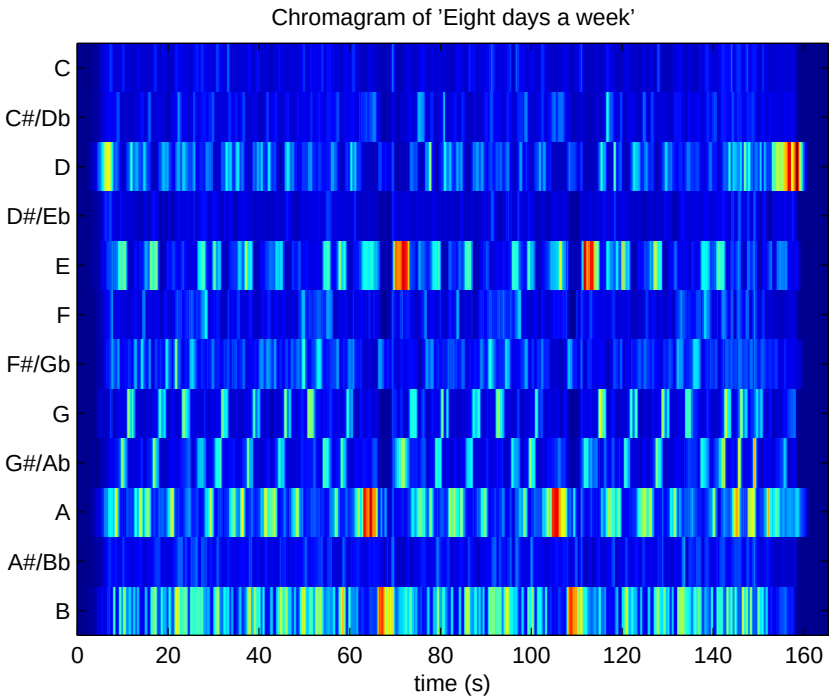
Input features calculation



chromagram

Chord recognition

Chroma vectors



- 12-dimensional vectors
- Every component : spectral energy or salience of a semi-tone within the chromatic scale (regardless of the octave)
- Chromagram : succession of chroma vectors over time
- Calculated with STFT or CQT

Chord recognition

How can we perform chord recognition ?

Chord recognition

How can we perform chord recognition ?

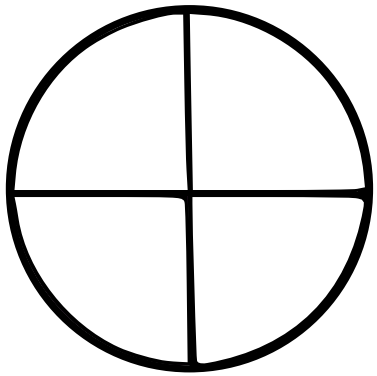
- music theory :
  - key
  - rhythm
  - possible chord transitions

Chord recognition

How can we perform chord recognition ?

- music theory :
  - key
  - rhythm
  - possible chord transitions
- training

Four main categories

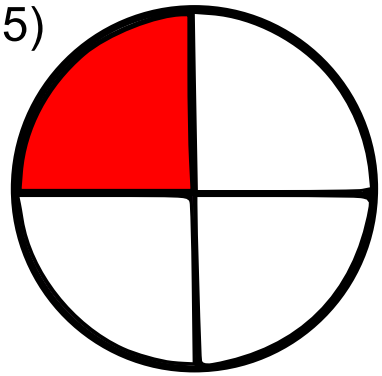


→ four main categories of chord recognition methods

Four main categories

Template-based

Fujishima (1999)
Harte & Sandler (2005)



$\longrightarrow$ only the chord definition is necessary

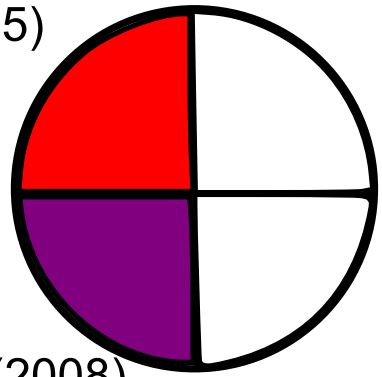
MUSIC THEORY : NO TRAINING : NO

Four main categories

Template-based

Fujishima (1999)

Harte & Sandler (2005)



Training-based

Sheh & Ellis (2003)

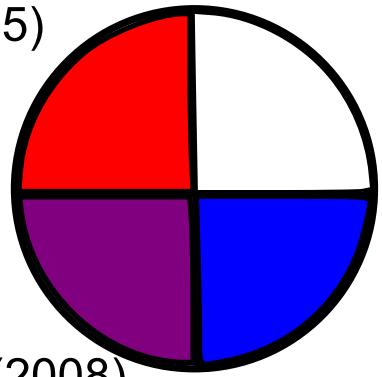
Ryynänen & Klapuri (2008)

$\longrightarrow$ only training and annotated data are necessary

MUSIC THEORY : NO TRAINING : YES

Four main categories

Template-based
Fujishima (1999)
Harte & Sandler (2005)



Training-based
Sheh & Ellis (2003)
Ryynänen & Klapuri (2008)

Music-driven
Bello & Pickens (2005)
Papadopoulos & Peeters (2008)
Mauch & Dixon (2010)

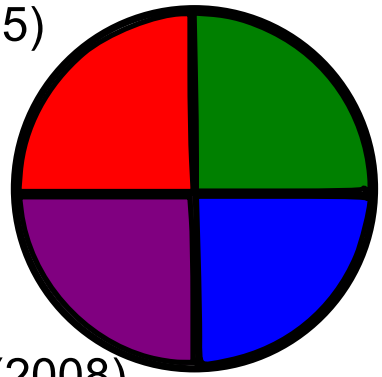
$\longrightarrow$ only extensive musical knowledge is necessary

MUSIC THEORY : YES TRAINING : NO

Four main categories

Template-based
Fujishima (1999)
Harte & Sandler (2005)

Hybrid
Yoshioka et al. (2004)
Lee & Slaney (2008)
Khadkevich & Omologo (2009)



Training-based
Sheh & Ellis (2003)
Ryynänen & Klapuri (2008)

Music-driven
Bello & Pickens (2005)
Papadopoulos & Peeters (2008)
Mauch & Dixon (2010)

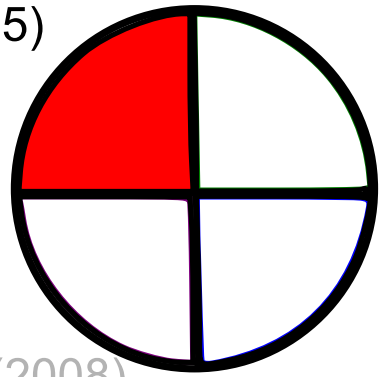
$\longrightarrow$ training and musical knowledge are necessary

MUSIC THEORY : YES TRAINING : YES

Four main categories

Template-based
Fujishima (1999)
Harte & Sandler (2005)

Hybrid
Yoshioka et al. (2004)
Lee & Slaney (2008)
Khadkevich & Omologo (2009)



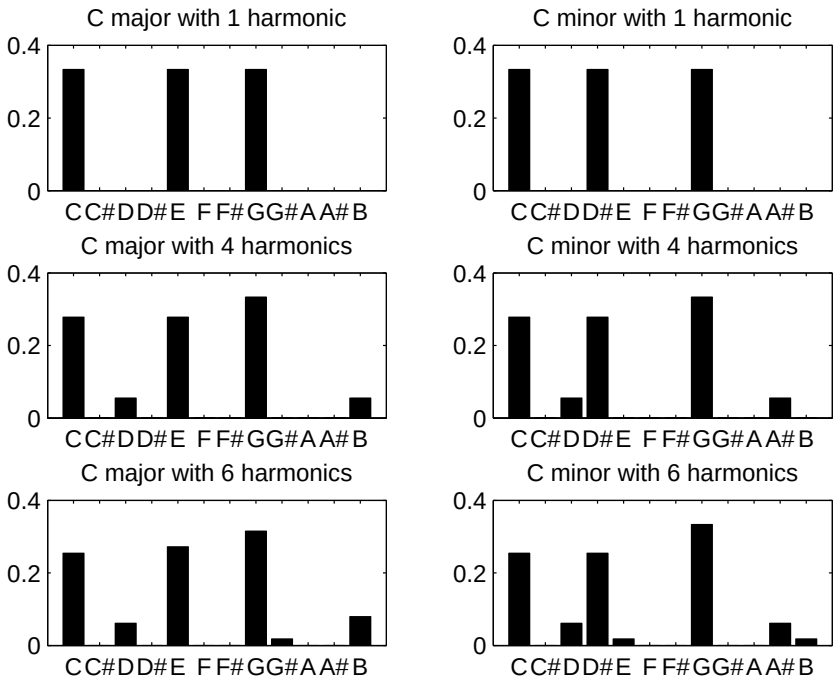
Training-based
Sheh & Ellis (2003)
Ryynänen & Klapuri (2008)

Music-driven
Bello & Pickens (2005)
Papadopoulos & Peeters (2008)
Mauch & Dixon (2010)

$\longrightarrow$ our methods

Contents

Chord templates



- 12-dimensional vectors
- Give the theoretical amplitudes of the notes in the chord
- Taking into account 1, 4 or 6 harmonics with exponentially decreasing spectral profile for the amplitudes of the partials (Gomez, 2006)

$\rightarrow$ Chord dictionary composed of K chord templates $\mathbf{w}_k$

Overview

- Given a set of chord templates $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$, a set of chroma vectors $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ and a measure of fit $D(\,.\,;\,.\,)$
- Fit a scale parameter $h_{k,n} = \underset{h}{\operatorname{argmin}}\ D(h\,\mathbf{c}_n; \mathbf{w}_k)$
- Compute $d_{k,n} = D(h_{k,n}\,\mathbf{c}_n; \mathbf{w}_k)$
- Detect chord $\hat{k}_n$ on frame n as :

$$\hat{\gamma}_n = \underset{k}{\operatorname{argmin}}\ d_{k,n}$$

Measures of fit

- Euclidean *(EUC)* :

$$D_{EUC}(\mathbf{x}|\mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Itakura-Saito *(2 variants IS1 & IS2)* :

$$D_{IS}(\mathbf{x}|\mathbf{y}) = \sum_i \frac{x_i}{y_i} - \log\left(\frac{x_i}{y_i}\right) - 1$$

- Kullback-Leibler *(2 variants KL1 & KL2)* :

$$D_{KL}(\mathbf{x}|\mathbf{y}) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right) - x_i + y_i$$

Filtering methods

Exploiting persistence by introducing a post-processing step

- So far : chord detection done frame by frame
- In practice : unlikely for a chord to last only one frame

Filtering

- Filtering upstream on the calculated measures $d_{k,n}$ and not on the sequence of detected chords
- 2 filtering methods :
  - low-pass : $\tilde{d}_{k,n} = \frac{1}{L} \sum_{n'=n-\frac{L-1}{2}}^{n+\frac{L-1}{2}} d_{k,n'}$
  - median : $\tilde{d}_{k,n} = \text{med}\left\{d_{k,n'}\right\}_{n-\frac{L-1}{2} \leq n' \leq n+\frac{L-1}{2}}$

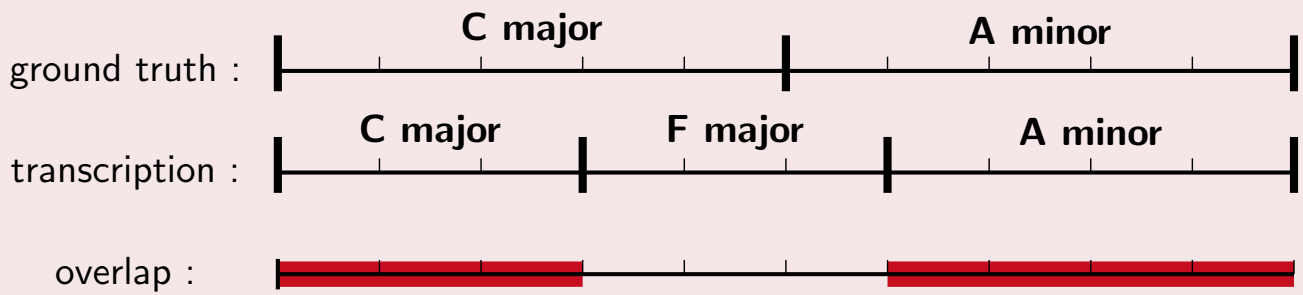
$$\hat{\gamma}_n = \underset{k}{\text{argmin}}\ \tilde{d}_{k,n}$$

Evaluation

Corpus

- 180 songs by the Beatles annotated by Christopher Harte
- All chord types mapped to major/minor for the evaluation

Overlap Score (MIREX)



$$\text{Example : Overlap Score} = \frac{3+4}{10} = 0.70$$

Experiments

Parameters

- 5 measures of fit (EUC, IS1, IS2, KL1, KL2)
- 2 filtering strategies (low-pass & median) with various step sizes (from 3 to 25)
- 3 chord models (1, 4 & 6 harmonics)

$\rightarrow$ 375 chord recognition systems
$\rightarrow$ 375 Average Overlap Scores (AOS)

Detection of major and minor chords

	no filtering			filtering		
	1 harm.	4 harm.	6 harm.	1 harm.	4 harm.	6 harm.
EUC	0.665	0.636	0.588	0.710	0.684	0.646
IS1	0.665	0.441	0.399	0.706	0.465	0.422
IS2	0.657	0.667	0.170	0.704	0.714	0.178
KL1	0.665	0.487	0.140	0.700	0.532	0.151
KL2	0.667	0.672	0.612	0.714	0.718	0.656

Conclusions

Detection of major and minor chords

	no filtering			filtering		
	1 harm.	4 harm.	6 harm.	1 harm.	4 harm.	6 harm.
EUC	0.665	0.636	0.588	0.710	0.684	0.646
IS1	0.665	0.441	0.399	0.706	0.465	0.422
IS2	0.657	0.667	0.170	0.704	0.714	0.178
KL1	0.665	0.487	0.140	0.700	0.532	0.151
KL2	0.667	0.672	0.612	0.714	0.718	0.656

Conclusions

- KL2 tends to give best results

Detection of major and minor chords

	no filtering			filtering		
	1 harm.	4 harm.	6 harm.	1 harm.	4 harm.	6 harm.
EUC	0.665	0.636	0.588	0.710	0.684	0.646
IS1	0.665	0.441	0.399	0.706	0.465	0.422
IS2	0.657	0.667	0.170	0.704	0.714	0.178
KL1	0.665	0.487	0.140	0.700	0.532	0.151
KL2	0.667	0.672	0.612	0.714	0.718	0.656

Conclusions

- KL2 tends to give best results
- Taking into account harmonics in the model is not really useful

Detection of major and minor chords

	no filtering			filtering		
	1 harm.	4 harm.	6 harm.	1 harm.	4 harm.	6 harm.
EUC	0.665	0.636	0.588	0.710	0.684	0.646
IS1	0.665	0.441	0.399	0.706	0.465	0.422
IS2	0.657	0.667	0.170	0.704	0.714	0.178
KL1	0.665	0.487	0.140	0.700	0.532	0.151
KL2	0.667	0.672	0.612	0.714	0.718	0.656

Conclusions

- KL2 tends to give best results
- Taking into account harmonics in the model is not really useful
- Filtering is definitely useful

Detection of major and minor chords

	no filtering			filtering		
	1 harm.	4 harm.	6 harm.	1 harm.	4 harm.	6 harm.
EUC	0.665	0.636	0.588	0.710	0.684	0.646
IS1	0.665	0.441	0.399	0.706	0.465	0.422
IS2	0.657	0.667	0.170	0.704	0.714	0.178
KL1	0.665	0.487	0.140	0.700	0.532	0.151
KL2	0.667	0.672	0.612	0.714	0.718	0.656

Conclusions

- KL2 tends to give best results
- Taking into account harmonics in the model is not really useful
- Filtering is definitely useful

→ Best system **OGF1 (maj-min)** : KL2, 4 harmonics, median filtering on 15 frames (2.04s)

Detection of other chord types

Detection of other chord types

- Introduction in priority of the most common chord types

Detection of other chord types

- Introduction in priority of the most common chord types
- Once the chords have been detected with their appropriate model, they are mapped to the major or minor type

Detection of other chord types

- Introduction in priority of the most common chord types
- Once the chords have been detected with their appropriate model, they are mapped to the major or minor type

Chord types	AOS	Optimal parameters
maj-min	0.718	KL2, 4 harm, median, L=15 (2.04s)
maj-min + 7	0.724	KL2, 1 harm, median, L=17 (2.23s)
maj-min + 7 + min7	0.706	IS1, 1 harm, low-pass, L=13 (1.86s)

Detection of other chord types

- Introduction in priority of the most common chord types
- Once the chords have been detected with their appropriate model, they are mapped to the major or minor type

Chord types	AOS	Optimal parameters
maj-min	0.718	KL2, 4 harm, median, L=15 (2.04s)
maj-min + 7	0.724	KL2, 1 harm, median, L=17 (2.23s)
maj-min + 7 + min7	0.706	IS1, 1 harm, low-pass, L=13 (1.86s)

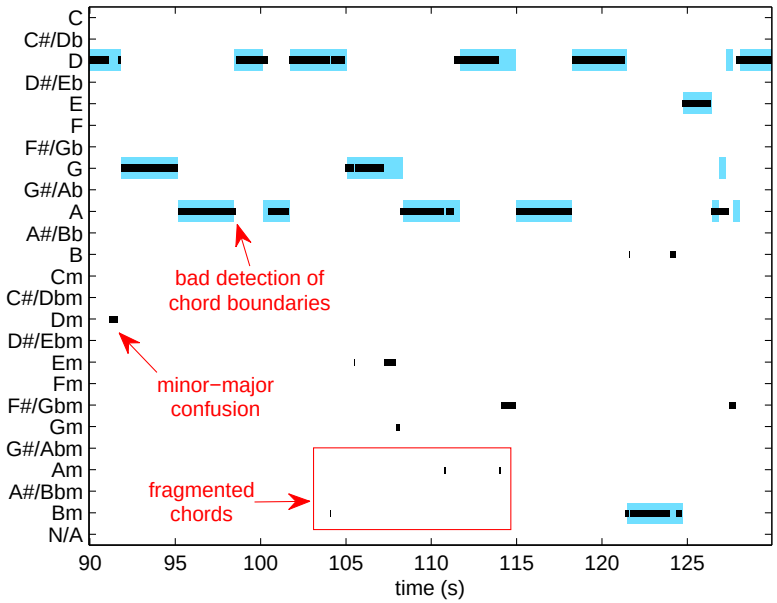
Conclusions

→ Best system **OGF2 (maj-min-7)** : KL2, 1 harmonic, median filtering on 17 frames (2.23s)

Waterloo by ABBA : OGF1 (maj-min), OS = 0.855

Play Resume Pause Stop

Conclusions



Satisfactory results BUT :

- Over-fragmentation of the piece : unsatisfying play-back of the song
- Bad detection of chord boundaries
- Major-minor confusions
- Over-estimation of the chord vocabulary (set of different chords played in the song)

Contents

Normalization vs Generative model

Notations

- Until now, scale parameter $h_{k,n}$ used for normalization :

$$h_{k,n}\ \mathbf{c}_n \approx \mathbf{w}_k$$

- We now define an amplitude parameter $a_{k,n}$, defining a generative model :

$$\mathbf{c}_n \approx a_{k,n}\ \mathbf{w}_k$$

Generative model

Parallel between measures of fit and probabilistic models

$$-\log\, p\left(\mathbf{c}_n | a_{k,n}, \mathbf{w}_k\right) = \varphi_1\, D\left(\mathbf{c}_n | a_{k,n}\, \mathbf{w}_k\right) + \varphi_2$$

- Euclidean distance $\leftrightarrow$ additive Gaussian noise
- Itakura-Saito divergence $\leftrightarrow$ multiplicative Gamma noise
- Kullback-Leibler divergence $\leftrightarrow$ Poisson noise

Generative model

Variables

- $\gamma_n \in [1, \ldots, K]$: discrete random state indicating the chord present on frame n
- α_k : probability for chord k to appear in the song

$$P(\gamma_n = k) = \alpha_k$$

Mixture model

$$p\left(\mathbf{c}_n | \boldsymbol{\alpha}, \mathbf{a}_n\right) = \sum_{k=1}^{K} \alpha_k \, p\left(\mathbf{c}_n | a_{k,n}, \mathbf{w}_k\right)$$

$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ and $\mathbf{A} = \{a_{k,n}\}_{kn}$ are estimated through an EM algorithm

Generative model

Chord recognition

Given estimates of these parameters, we choose for frame n the chord with highest state posterior probability :

$$\hat{\gamma}_n = \underset{k}{\text{argmax}}\ p\left(\gamma_n = k | \mathbf{c}_n, \boldsymbol{\alpha}, \mathbf{a}_n\right)$$

Filtering

Just like in the deterministic approach : post-processing filtering now applied on the state posterior probabilities

Experiments

	no filtering	low-pass	median
Gaussian	0.714	0.748	0.749
Gamma	0.730	0.758	0.758
Poisson	0.727	0.742	0.744

Best systems

Experiments

	no filtering	low-pass	median
Gaussian	0.714	0.748	0.749
Gamma	0.730	0.758	0.758
Poisson	0.727	0.742	0.744

Best systems

- **PCR/Gaussian** : median filtering on 17 frames (2.23s)

Experiments

	no filtering	low-pass	median
Gaussian	0.714	0.748	0.749
Gamma	0.730	0.758	0.758
Poisson	0.727	0.742	0.744

Best systems

- **PCR/Gaussian** : median filtering on 17 frames (2.23s)
- **PCR/Gamma** : low-pass filtering on 15 frames (2.04s)

Experiments

	no filtering	low-pass	median
Gaussian	0.714	0.748	0.749
Gamma	0.730	0.758	0.758
Poisson	0.727	0.742	0.744

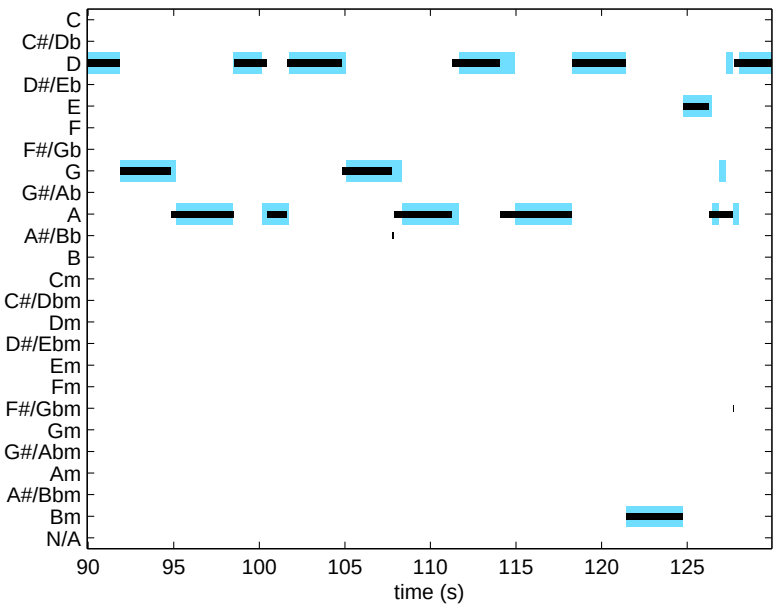
Best systems

- **PCR/Gaussian** : median filtering on 17 frames (2.23s)
- **PCR/Gamma** : low-pass filtering on 15 frames (2.04s)
- **PCR/Poisson** : median filtering on 13 frames (1.86s)

Waterloo by ABBA : PCR/Gamma, OS = 0.893

Play Resume Pause Stop

Conclusions



- Deletion of some fragmented chords
- Better estimation of the chord vocabulary through the estimation of vector $\boldsymbol{\alpha}$
- Still some temporal issues due to the filtering process

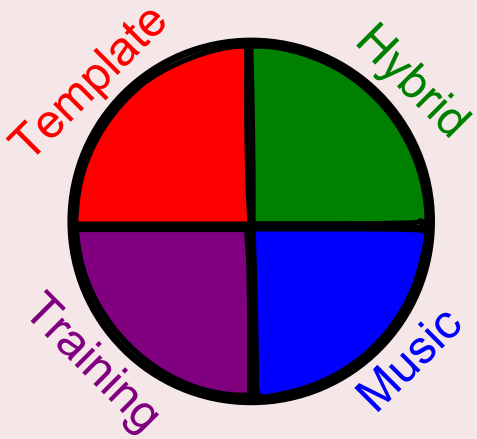
Contents

State-of-the-art

MIREX 2008 & 2009

Methods whose codes were available during the development of our systems

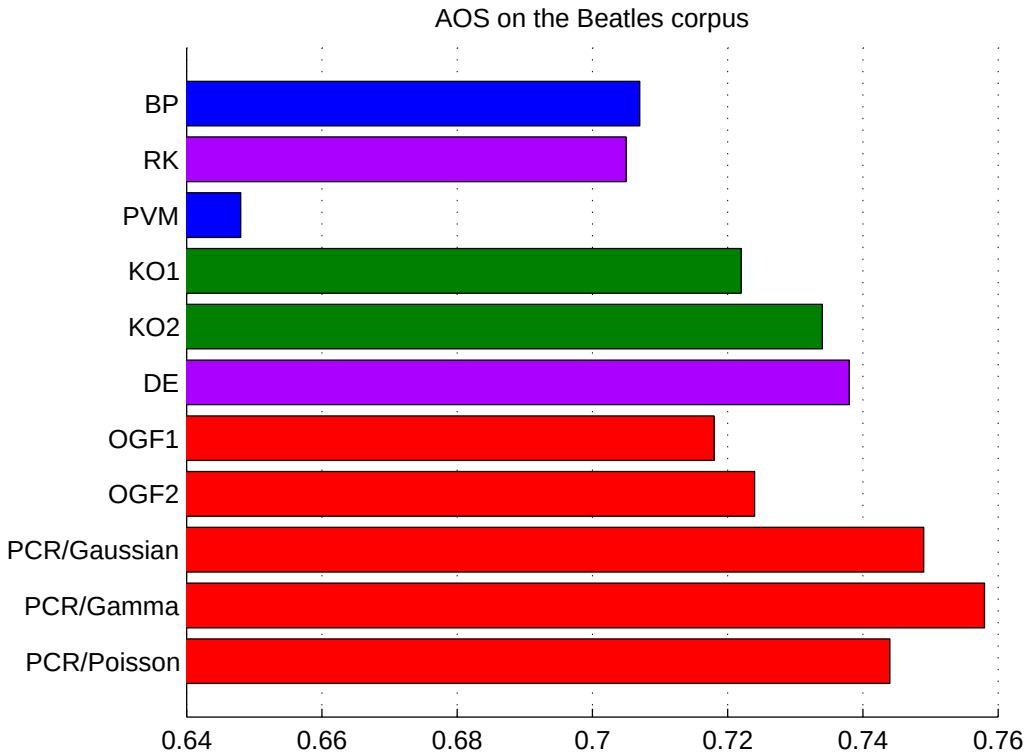
- MIREX 2008 :
  - BP : Bello & Pickens (2005)
  - RK : Ryynänen & Klapuri (2008)
  - PVM : Pauwels et al. (2008)
- MIREX 2009 :
  - KO1 & KO2 : Khadkevich & Omologo (2009)
  - DE : Ellis (2009)



+ OGF1 *(maj-min)*, OGF2 *(maj-min-7)*, PCR/Gaussian, PCR/Gamma & PCR/Poisson

Results on the Beatles corpus (AOS)

BP	0.707
RK	0.705
PVM	0.648
KO1	0.722
KO2	0.734
DE	0.738
OGF1	0.718
OGF2	0.724
PCR/Gaussian	0.749
PCR/Gamma	0.758
PCR/Poisson	0.744



Average Overlap Score (AOS) $\rightarrow$ as high as possible

Other corpus, other metrics

Other corpus

QUAERO corpus : 20 songs by various artists and music styles

Other metrics

Other corpus, other metrics

Other corpus

QUAERO corpus : 20 songs by various artists and music styles

Other metrics

- Segmentation : Average Hamming Distance (AHD)
  $\rightarrow$ as low as possible

Other corpus, other metrics

Other corpus

QUAERO corpus : 20 songs by various artists and music styles

Other metrics

- Segmentation : Average Hamming Distance (AHD)
  $\rightarrow$ as low as possible
- Fragmentation : Average Chord Length (ACL)
  $\rightarrow$ as close to 1 as possible

Other corpus, other metrics

Other corpus

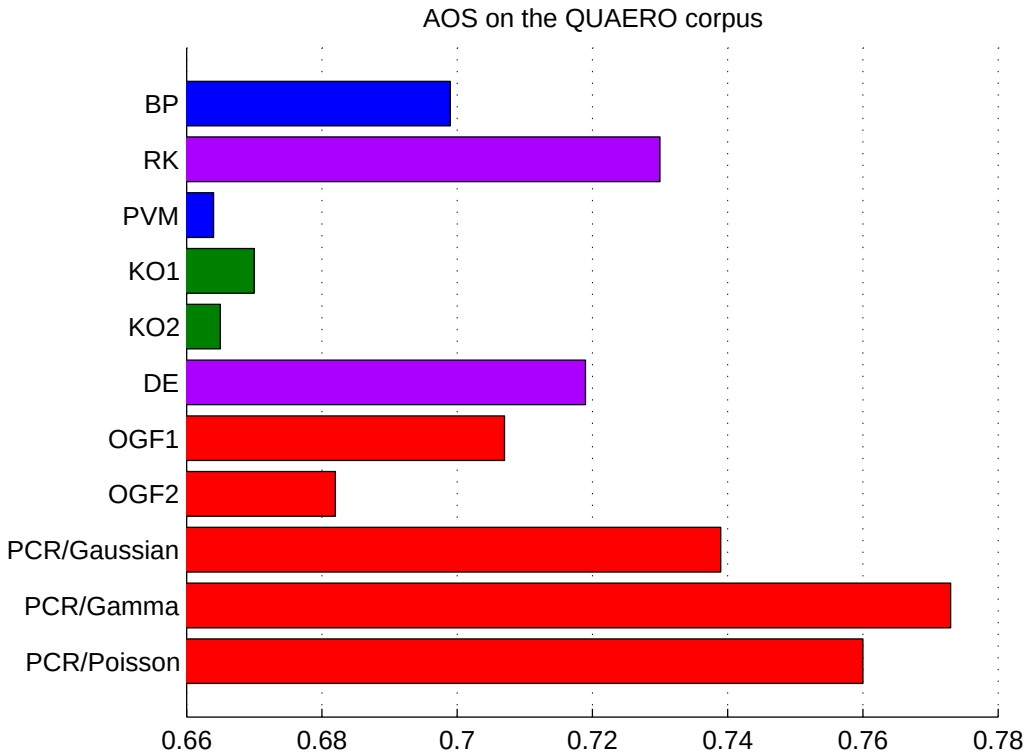
QUAERO corpus : 20 songs by various artists and music styles

Other metrics

- Segmentation : Average Hamming Distance (AHD)
  $\rightarrow$ as low as possible
- Fragmentation : Average Chord Length (ACL)
  $\rightarrow$ as close to 1 as possible
- Chord vocabulary : Average Chord Number (ACN)
  $\rightarrow$ as close to 1 as possible

Results on the QUAERO corpus (AOS)

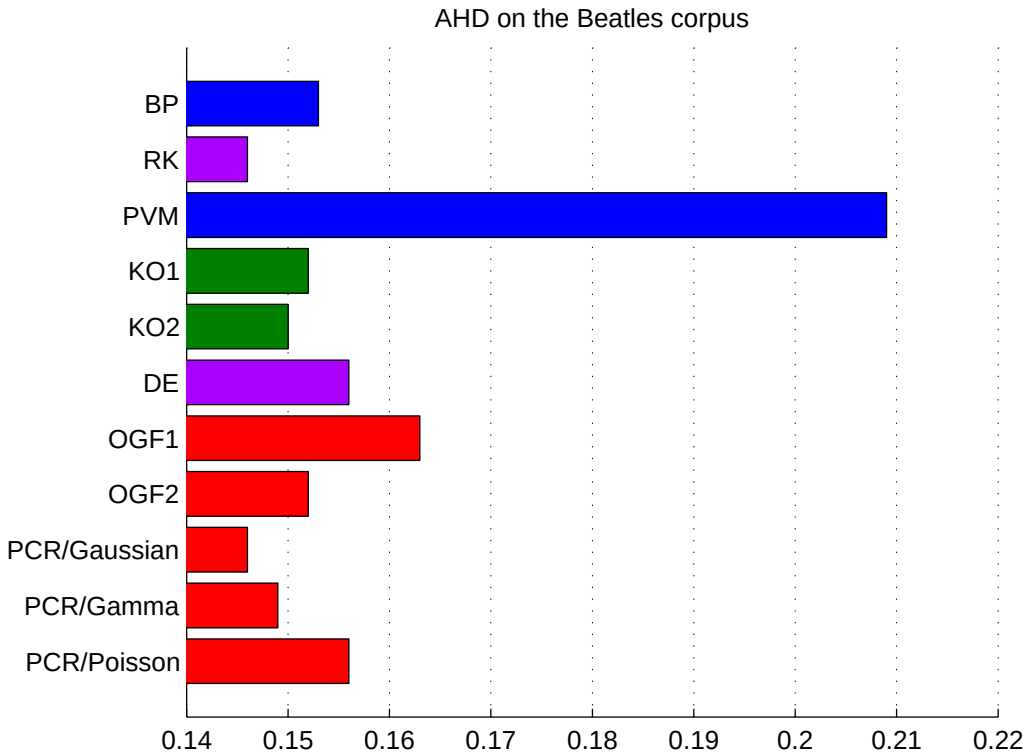
BP	0.699
RK	0.730
PVM	0.664
KO1	0.670
KO2	0.665
DE	0.719
OGF1	0.707
OGF2	0.682
PCR/Gaussian	0.739
PCR/Gamma	0.773
PCR/Poisson	0.760



Average Overlap Score (AOS) $\rightarrow$ as high as possible

Results on the Beatles corpus : Segmentation (AHD)

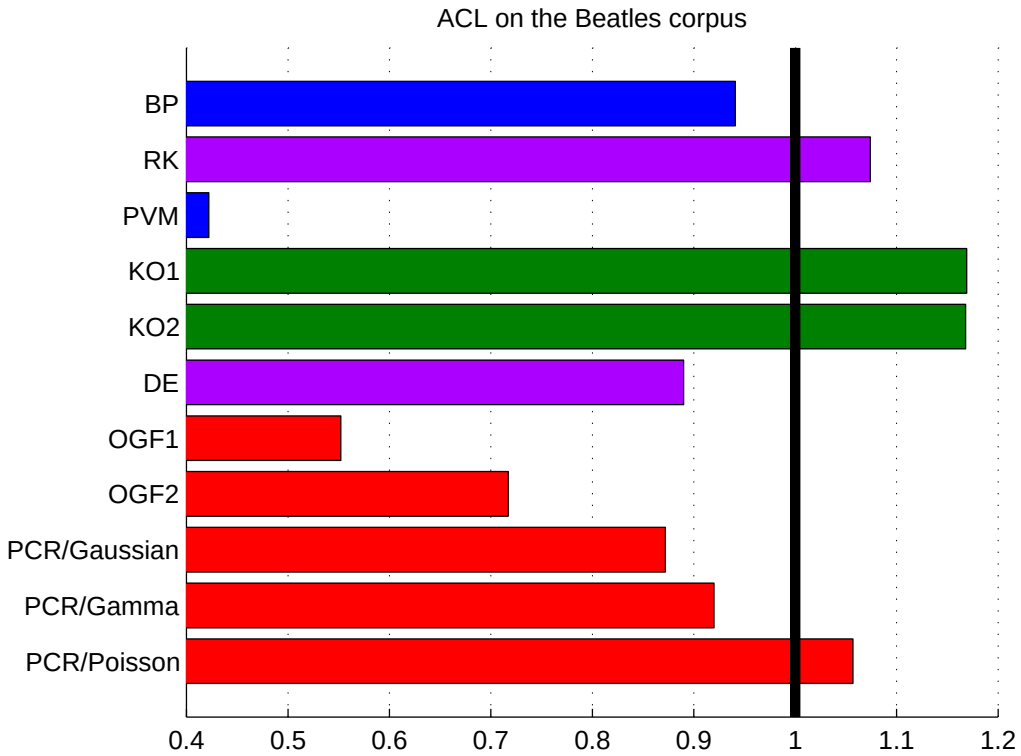
BP	0.153
RK	0.146
PVM	0.209
KO1	0.152
KO2	0.150
DE	0.156
OGF1	0.163
OGF2	0.152
PCR/Gaussian	0.146
PCR/Gamma	0.149
PCR/Poisson	0.156



Average Hamming Distance (AHD) $\rightarrow$ as low as possible

Results on the Beatles corpus : Fragmentation (ACL)

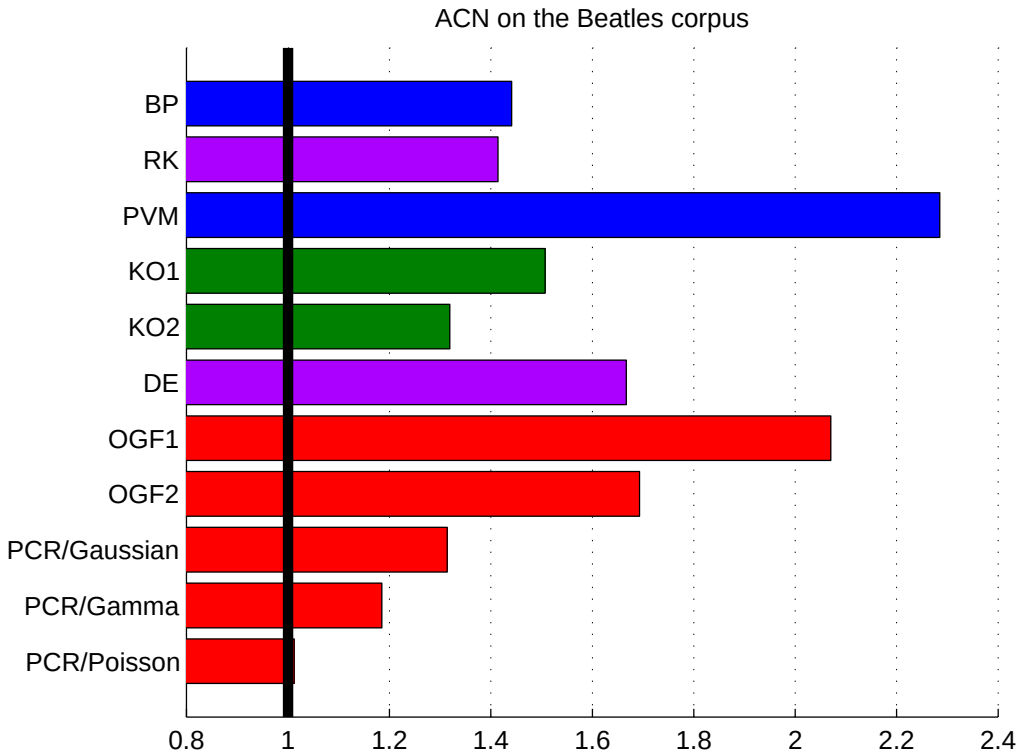
BP	0.941
RK	1.074
PVM	0.422
KO1	1.169
KO2	1.168
DE	0.890
OGF1	0.552
OGF2	0.717
PCR/Gaussian	0.872
PCR/Gamma	0.920
PCR/Poisson	1.057



Average Chord Length (ACL) $\rightarrow$ as close to 1 as possible

Results on the Beatles corpus : Chord vocabulary (ACN)

BP	1.441
RK	1.414
PVM	2.285
KO1	1.507
KO2	1.319
DE	1.667
OGF1	2.070
OGF2	1.693
PCR/Gaussian	1.314
PCR/Gamma	1.185
PCR/Poisson	1.012



Average Chord Number (ACN) $\rightarrow$ as close to 1 as possible

Contents

Contributions

Two approaches for template-based chord recognition which do not need training nor extensive musical knowledge

- Satisfactory performances on several corpus
- Reduction of the fragmentation effect
- Precise estimation of the chord vocabulary

Perspectives

Perspectives

- Temporal context taken into account within the probabilistic model

Perspectives

- Temporal context taken into account within the probabilistic model
- Adaptive estimation of the parameters

Perspectives

- Temporal context taken into account within the probabilistic model
- Adaptive estimation of the parameters
- Improvement of the front end

Perspectives

- Temporal context taken into account within the probabilistic model
- Adaptive estimation of the parameters
- Improvement of the front end
- Use of vector α to retrieve the key