Is Music Structure Annotation Multi-Dimensional? A Proposal for Robust Local Music Annotation

Geoffroy Peeters and Emmanuel Deruty
IRCAM – CNRS/STMS

with the help of Jean-Francois Rousse and Maxence Riffault

quaero

ircam Centre Pompidou

Local Music Annotation

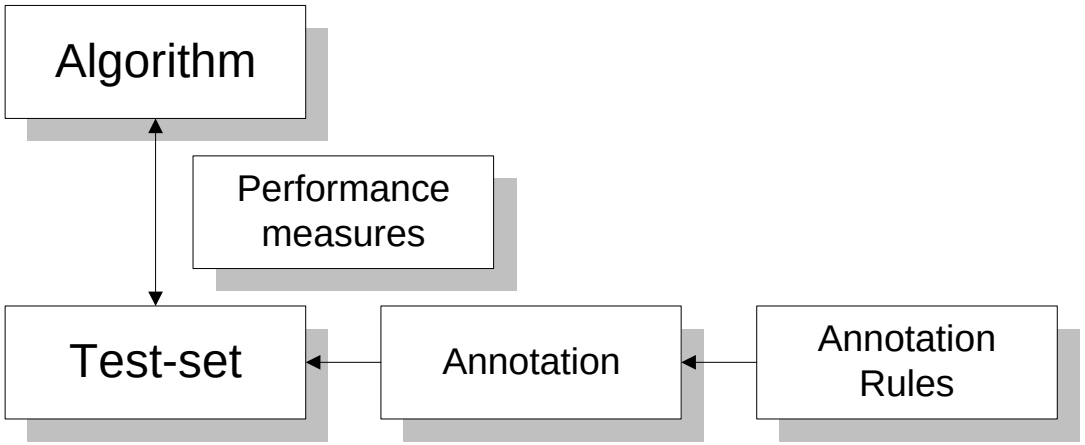
Criteria to define a robust annotation

Music Structure

Multi-dimensional annotation

Introduction

- Large part of present-day “Music Structure” research devoted to
  - Improvement of algorithms
  - Improvement of recognition scores
  - Definition of new performance measures



- But **“how pertinent is the structure annotation that is used for those evaluations?”**
  - Require a precise definition of the annotation process
  - Example: comparison of annotations from the same tracks (“The Beatles”) coming from two different “Music Structure” test-sets made by different research teams

Introduction

- Goal of this paper:
  - Define a set of rules to define what is a robust annotation
  - Use it to define a robust definition of “Music Structure” annotation
    - Multi-dimensional
  - This work comes from a one year long experiment of testing thinking and validating made by three professional musicians playing the role of computer-annotators.

Presentation overview

- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Requirements for a robust annotation definition

- Divide the notion of “local music annotation” into two categories

  - **“Information extraction”:**
    - consists in mapping a piece of music to extract information which describes aspects of the piece. “Information extraction”
    - would include: - structure annotation, - beat annotation, - singing voice annotation.

  - **“Imitation” or “reduction”:**
    - consists in finding audio objects that sound like the original piece. Those audio objects can then be compared to the original.
    - would include: - note / chord / melody annotation.

Presentation overview

- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Requirements for a robust annotation definition

Information extraction: conditions

- When doing “Information Extraction”, we look at the piece of music from a certain point of view, and then connect certain aspects of the music to an abstract object.
  - Example of abstract object: “chorus”

- **Definition:**
  - an object, a descriptor, must be properly defined

- **Certainty:**
  - in a given corpus, the object should be recognized without doubt
  - Perceptual Recognition Rate (use of common MIR measures between annotation)

- **Concision:**
  - the range of available descriptors should be limited.

- **Universality:**
  - a given descriptor should be used reasonably often.

Measuring “Certainty”: Perceptive Recognition Rate (PRR)

Certainty

- “Perceptive Recognition Rate” or PRR
  - checking, on a given corpus, how many times a given object is recognized without doubt

- “Algorithmic Recognition Rate” or ARR
  - the recognition rate traditionally used in the M.I.R. field to evaluate algorithms

  - If PRR=1 (perfect case), then the notion of ARR is justified.
  - If PRR=0 (worst case), then any result including ARR do not make any sense

Application to the “chorus” case

Application

- Chorus definition:
  - “A part of the track which includes the lead vocalist, - a part in which the lyrics contain the song title, - a recurrent part which happens at least 2 times during the song”
- Application
  - set of 112 songs (those songs are not particularly main-stream, neither are they particularly recent, their style is quite varied).
- PRR is very low: less than 50%
  - For this 112 songs test-set, we cannot tell if there is a chorus or not for 56 of them!

Requirements for a robust annotation definition

Information extraction: conditions

- When doing “Information Extraction”, we look at the piece of music from a certain point of view, and then connect certain aspects of the music to an abstract object.
  - Example of abstract object: “chorus”
- **Definition:**
  - an object, a descriptor, must be properly defined
- **Certainty:**
  - in a given corpus, the object should be recognized without doubt
  - Perceptual Recognition Rate (use of common MIR measures between annotation)
- **Concision:**
  - the range of available descriptors should be limited.
- **Universality:**
  - a given descriptor should be used reasonably often.

Measuring Concision and Universality

Concision/ Universality

- **T**: is the total number of tracks in the given test-set.
- **L**: is the total number of different labels l used over the given test-set.

- **N(l)**: is the “N”umber of tracks using a specific label l, divided by the total number of tracks T.
  - close to the “document frequency” measure used in Information retrieval.
- **U(l)**: is the average (over tracks) “U”se of a specific label l in a specific track (when the label is used at least once in this specific track).
  - close to the “term frequency” measure used in Information retrieval.

- **mS**: is the average (over the tracks) number of different segments used for a specific track.
- **mL**: is average (over the tracks) number of different labels used for a specific track.

Presentation overview

- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Existing Music Structure corpuses test-set

- **IRCAM MPEG-7-Audio test-set**
  - 2001, Ircam
  - cross-checked by MPEG-7 Audio partners
- T = 25 tracks (state and sequence structures)
- L = 50 different labels: “bass”, “break drum”, “break guitar”, “chorus 1”, “chorus instru”, “chorus variante”, “verse glockenspiel”.
- Average number of segments per track is mS=17.57
- Number of different labels used in a track mL=7.64
- Remark: Often, for the development of a “Music Structure” test-set, the list of music tracks are chosen to fit the definition of the annotation system used

- Label ? blend between
  - a description of the “musical role” that a part plays inside a track (“intro”, “verse”, “chorus”) and
  - a description of the specific instrumentation used in it
- Most of the labels appear only for the description of a single track and only appear once in the track
  - “break 2”,
  - “break drum”,
  - “break guitar”,
  - “break piano”,
  - “intro synth”,
  - “intro voice”.
- The exceptions are
  - “break” with N(l)=0.43 (with a mean-use inside a track of U(l)=2.16),
  - “chorus” 0.93 (4.38),
  - “intro” 0.86 (1.25),
  - “verse” 0.93 (3.92).

Existing Music Structure corpuses test-set

- Queen Mary University of London
  - 2003, QMUL
  - starts from the MPEG-7-Audio "state" test-set and extend it a lot.
- T = 107 tracks of various pop-rock songs and many Beatles songs
- L = 107 different labels
- Average number of segments per track is mS=12.33
- Number of different labels used in a track mL=6

- Most of the labels appear only for the description of a single track and only appear once in this track
  - "crash",
  - "fill",
  - "drop",
  - "crash".
- The exceptions are
  - "break" which appears N(l)=0.22 (with a mean-use inside a track of U(l)=1.53),
  - "bridge" 0.55 (1.6),
  - "chorus" 0.43 (3.96),
  - "intro" 0.85 (1.27),
  - "outro" 0.38 (1),
  - "verse" 0.87 (3.30) .

Existing Music Structure corpuses test-set

- Beatles test-set
  - 2005, developed by Universitat Pompeu Fabra (UPF) based on the annotations made by the musicologist Alan W. Pollack [9]. It has been later modified by the Tampere University of Technology (TUT).
- T =174 tracks
- L = 55 different labels
- Average number of segments per track is mS=9.21
- Number of different labels used in a track mL=5.23

- Most of the labels appear only for the description of a single track and only appear once in this track.
  - “close”,
  - “closing”,
  - “improv interlude”
- The exception are
  - “bridge” N(l)=0.59 (U(l)=1.73),
  - “intro” 0.86 (1.08),
  - “outro” 0.82 (1),
  - “refrain” 0.42 (3.41),
  - “verse” 0.86 (3.33),
  - “verses” 0.28 (1.16)

Existing Music Structure corpuses test-set

- TUT07 Structure test-set
  - 2007, Tampere University of Technology
- T =557 Western popular music pieces (pop, rock, jazz, blues and “schlager” music) annotated into structure
- seems annotated into “musical role” (“intro”, “verse”, “chorus”) or “acoustical similarity” (“A”, “B”, “solo”)

Existing Music Structure corpuses test-set

- TU Vienna test-set
  - 2007, IFS TU Vienna
- Part of the tracks are coming from the QMUL (hence MPEG-7-Audio), RWC and Beatles test-sets
- allow several simultaneous descriptions of the same segment (describing a given part as a single segment or as a set of sub-segments): hierarchical XML schema

Existing Music Structure corpuses test-set

- RWC test-set
  - 2006, AIST
  - T = 255 tracks
  - L = 15 different labels
- Average number of segments per track is mS=15.73
- Number of different labels used in a track mL=6.68
- Comments: annotation mainly describes the “musical role” of the parts (“intro”, “ending”, “verse”, “chorus”, “bridge”). It however merge “acoustical similarity” with it (“verse-a”, “verse-b”, “verse-c”, “verse-d”).

- All labels are used at least for 10 tracks (“bridge-d”) and for most more than 50 times.
  - The mean (over labels) value of N(l) is therefore high: 0.39.
  - The mean (over labels) value of U(l) is 2.16.

Existing Music Structure corpuses test-set

		MPEG-7 Audio Test Set	QMUL test-set	TU Vienna test-set	Beatles test-set	TUT Structure Test-set	RWC test-set
T	Number of tracks	25	107	109	174	557	285
L	Number of Labels	50	107	close to QMUL	55	not available	17
N(l)	Number of track using a specific label					not available	0.39
U(l)	Average use of a label inside a track					not available	2.16
mS	mean Segment per track	17.57	12.33	close to QMUL	9.21	not available	15.73
mL	mean Label per Track	7.64	6	close to QMUL	5.23	not available	6.68

Presentation overview

- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Main problems of existing “Music Structure” annotations

- Number and coverage of labels
- Merging orthogonal view-points
  - “musical role”, i.e. the role that a part plays inside a track, such as verse, chorus ... (furthermore these concepts are not applicable to all kind of music), - “acoustical similarity” and - “instrumentation”
- Similarity boundary definition
  - difference is sometimes interpreted as parts being identical and sometimes as parts being different
- Describing the structure of the music, the melody, the instrumentation ?
  - If the accompaniment remains constant over the entire track, then the voice variations are described (The Beatles). If the voice remains constant over the entire track, then the accompaniment variations are described (Rap music)
- Temporal boundaries definition
  - boundaries of the segments is often not coherent from track to track
- Segment sub-division
  - definition of the “sub-division of a part A into sub-part a” is not coherent over a given test-set

Various possible definitions for a “Music Structure” annotation

- Preliminary remarks
  - Any choice of definition can be done if the appropriate test-set is chosen
  - Important to avoid mixing various view-points
  - It is also possible to start from a test-set and find the best-fitting description for this given test-set
  - The goal of this paper is to find a description of the “Music Structure” that can be applied to any kind of music

Various possible definitions for a “Music Structure” annotation

- **1) based on Musical Role**
  - Music role that a part plays in a song
  - Introduction, Verse, Chorus, Bridge, Ending
  - Problems:
    - Intro, Outro= time position, can be the Chorus
    - Several version of Chorus and Verse
    - Definition of Chorus and verse not clear (Rap, R’n’B)
- **2) based on Acoustic Similarity**
  - Acoustic similarity between parts
  - Problems:
    - Similar= identical, What about if small variation ?
    - How to quantify ?
    - How to put the threshold ?
    - What criteria for similarity ? Timbre ? Harmony ? Rhythm ? Lyrics ?
- **3) based on Instrument Role**
  - Location of lead singer, location of solo guitar
  - Problems:
    - Few insights into the global structure
    - Identify instrument: huge number of labels (guitar= classical ? Folk ? Electric ? WhaWha ?)
  - Solution: Describe the instrument role: Primary Lead, Secondary Lead
- **4) based on Perceptual Tests**
    - Average human perception of structure
  - Problems:
    - Very costly !
    - Labels used by people are usually not shared

Presentation overview

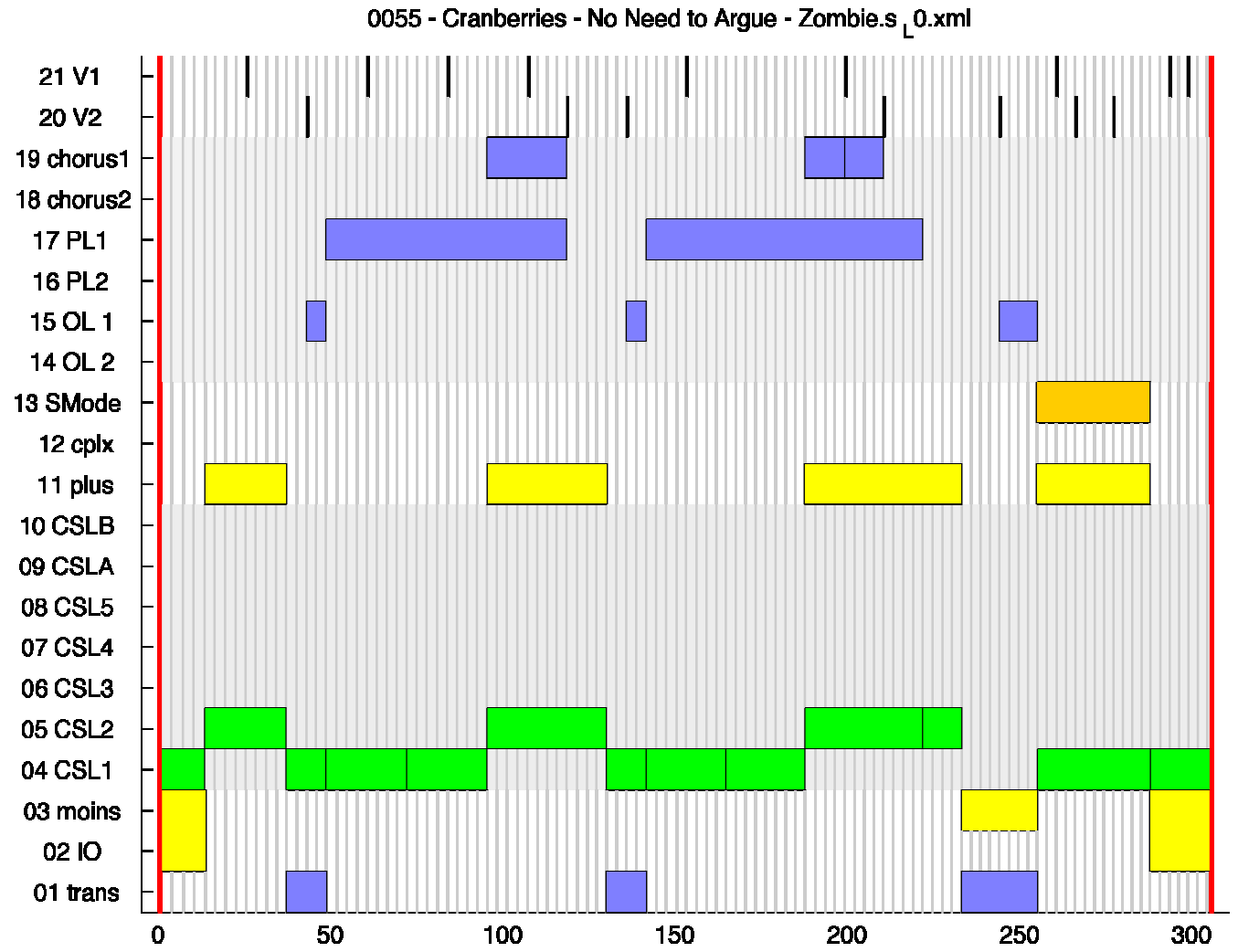
- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Proposed method
Multi-dimensional music structure annotation

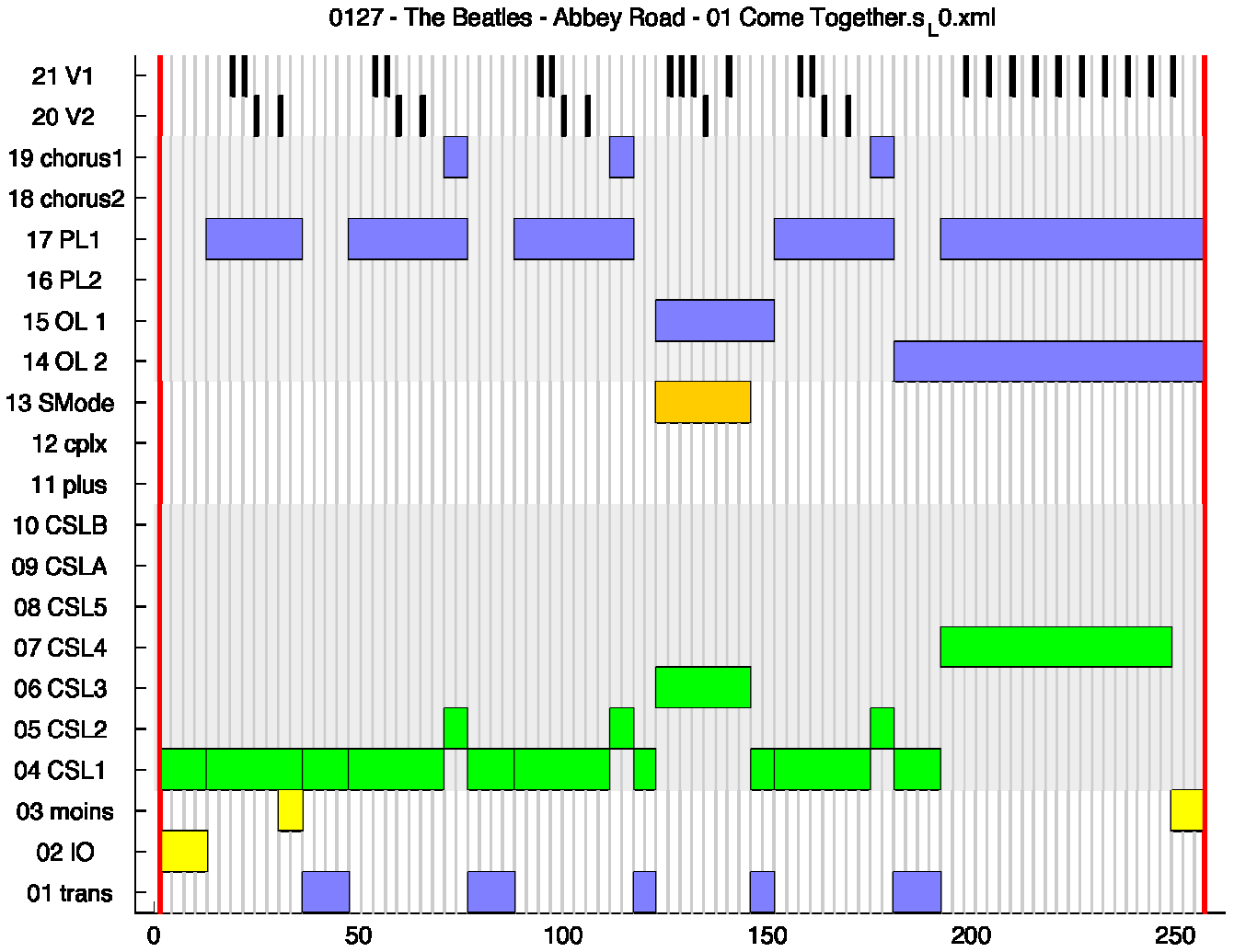
- Use simultaneously (but independently) the various view-points: “acoustical similarity”, “musical role” and “instrument role”
- A track is formed:
  - • by a set of Constitutive Solid Loops (CSLoop)
    - represent a ”musical phrase”/”musical exposition” (succession of chords).
    - CSLoop with similar ID represent the same ”musical phrase” although large variation can occur between them.
  - • over which are super-imposed variations of the CSLoops ID.
    - “--”: same CSLoop in a lighter version (for example without the drum or without the bass)
    - “++” if it is in a stronger version (for example with an extra second guitar)
  - • over which are superimposed important “instrument roles”:
    - such as primary leads (lead singer in pop., lead instr. in jazz/ electro)
    - other leads (choir, other lead instruments or melodic sample)
    - or solo mode (electric-guitar solo, jazz chorus solo, ...)
  - • which plays a “musical role” (intro, outro, transition, obvious chorus, solo).

- When a part is too complex to be described, it is annotated as ComplexMode.
- When a CSLoop is an obvious chorus it is annotated as “chorus”, when it is not obvious, it is not annotated as “chorus” but it is still annotated as the repetition of the occurrence of a specific CSLoop, with PrimaryLead and OtherLead (Choir) which are distinctive elements.
- Segment sub-division problem: markers can be placed inside a CSLoop segment to indicate further possible sub-divisions. Two types of markers can be placed (V1 and V2) indicating respectively similarity and dissimilarity between the parts on the left and on the right of the marker.

Proposed method
Examples



Proposed method
Examples



quaero

ircam Centre Pompidou

Proposed method
Multi-dimensional music structure annotation

# id	Name	Full name	Type	Description	Equality rule	Exclusion
21	V1	comma 1	marker	divides a CSL - means 'identical' or 'very similar'	(marker)	/
20	V2	comma 2	marker	divides a CSL - means 'different'	(marker)	/
19	Chorus1	obvious chorus 1	segment	chorus	yes	17
18	Chorus2	obvious chorus 2	segment	chorus	yes	18
17	PLead1	primary lead 1	segment	main melodic reference 1	yes	15
16	PLead2	primary lead 2	segment	main melodic reference 2, 3+...	no	16
15	OLead1	other lead 1	segment	secondary melodic reference / solo 1	yes	13
14	OLead2	other lead 2	segment	secondary melodic reference / solo 2, 3, 4+...	no	14
13	SMode	solo mode	segment	solo mode	no	/
12	Cplx	complex	segment	complex part, structure is not clear	no	all CSLs
11	++	"plus" part	segment	a part in which loudness is spectacularly higher than the rest of the song	no	3
10	CSLB	constitutive solid loop B	segment	a musical [phrase / idea]	no	04/05/06/07/08
9	CSLA	constitutive solid loop A	segment	a musical [phrase / idea]	no	04/05/06/07/09
8	CSL5	constitutive solid loop 5	segment	a singular musical [phrase / idea]		
7	CSL4	constitutive solid loop 4	segment	a singular musical [phrase / idea]	yes	04/05/06/08/09
6	CSL3	constitutive solid loop 3	segment	a singular musical [phrase / idea]	yes	04/05/07/08/09
5	CSL2	constitutive solid loop 2	segment	a singular musical [phrase / idea]	yes	04/06/07/08/09
4	CSL1	constitutive solid loop 1	segment	a singular musical [phrase / idea]	yes	05/06/07/08/09
3	--	"minus" part	segment	a part in which loudness is spectacularly lower than the rest of the song a part in which two of the three references (rhythmic, melodic, harmonic) disappear	no	10
2	IO	in / out / exotic	segment	intros, outros, and non I/O parts that have nothing to do with the rest of song	no	/
1	Trans	transition	segment	transitions	no	/

Presentation overview

- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Proposed method
Testing over a large variety of music genre

- Set of 300 music tracks coming from various music genres including:
  - Progressive-Rock (Pink Floyd, Queen, Frank Zappa ...),
  - World-Music (Ali Farka Toure, Buenavista Social Club, Stan Getz/ Gilberto Gil ...),
  - Electro-Music (The Chemical Brothers, Squarepusher, ...),
  - Rap-music (50 Cent, Outkast ...),
  - Mainstream-music (Michael Jackson, The Beatles, Eric Clapton, Nirvana, Cranberries, Bauhaus, The Cure ...).

- T= 300 tracks
- L= 21 different labels

- on average a track is divided into mS=38.93 segments (22.80 when omitting the sub-division with comma V1 and V2).
- on average a track uses mL=9.80 different labels

- mean (over labels) N(l) is 0.47 (0.39 for RWC) which is very high,
- mean (over labels) U(l) is 3.21 (2.16 for RWC)

Existing Music Structure corpuses test-set

		MPEG-7 Audio Test Set	QMUL test-set	TU Vienna test-set	Beatles test-set	TUT Structure Test-set	RWC test-set	Ircam 2009
T	Number of tracks	25	107	109	174	557	285	**300**
L	Number of Labels	50	107	close to QMUL	55	not available	17	**21**
N(l)	Number of track using a specific label					not available	0.39	**0.47**
U(l)	Average use of a label inside a track					not available	2.16	**3.21**
mS	mean Segment per track	17.57	12.33	close to QMUL	9.21	not available	15.73	**38.93 (22.80)**
mL	mean Label per Track	7.64	6	close to QMUL	5.23	not available	6.68	**9.80 (8.11)**

Proposed method
Information extraction conditions applied

# id	Name	Full name	Type	Description	Equality rule	Exclusion	N(I)	U(I)
21	V1	comma 1	marker	divides a CSL - means 'identical' or 'very similar'	(marker)	/	0,9	10
20	V2	comma 2	marker	divides a CSL - means 'different'	(marker)	/	0,8	8,7
19	Chorus1	obvious chorus 1	segment	chorus	yes	17	0,4	3,2
18	Chorus2	obvious chorus 2	segment	chorus	yes	18	0	5,3
17	PLead1	primary lead 1	segment	main melodic reference 1	yes	15	0,9	3,3
16	PLead2	primary lead 2	segment	main melodic reference 2, 3+...	no	16	0,3	3,5
15	OLead1	other lead 1	segment	secondary melodic reference / solo 1	yes	13	0,5	2,2
14	OLead2	other lead 2	segment	secondary melodic reference / solo 2, 3, 4+...	no	14	0,1	2,2
13	SMode	solo mode	segment	solo mode	no	/	0,3	1,3
12	Cplx	complex	segment	complex part, structure is not clear	no	all CSLs	0	1
11	++	"plus" part	segment	a part in which loudness is spectacularly higher than the rest of the song	no	3	0,1	1,6
10	CSLB	constitutive solid loop B	segment	a musical [phrase / idea]	no	04/05/06/07/08	0,1	2,1
9	CSLA	constitutive solid loop A	segment	a musical [phrase / idea]	no	04/05/06/07/09	0,1	1,9
8	CSL5	constitutive solid loop 5	segment	a singular musical [phrase / idea]			0,2	1,9
7	CSL4	constitutive solid loop 4	segment	a singular musical [phrase / idea]	yes	04/05/06/08/09	0,5	2,3
6	CSL3	constitutive solid loop 3	segment	a singular musical [phrase / idea]	yes	04/05/07/08/09	0,7	2,6
5	CSL2	constitutive solid loop 2	segment	a singular musical [phrase / idea]	yes	04/06/07/08/09	0,9	3,5
4	CSL1	constitutive solid loop 1	segment	a singular musical [phrase / idea]	yes	05/06/07/08/09	1	5,1
3	--	"minus" part	segment	a part in which loudness is spectacularly lower than the rest of the song a part in which two of the three references (rhythmic, melodic, harmonic) disappear	no	10	0,7	2
2	IO	in / out / exotic	segment	intros, outros, and non I/O parts that have nothing to do with the rest of song	no	/	0,9	1,6
1	Trans	transition	segment	transitions	no	/	0,4	2

Presentation overview

- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Proposed method

Use in M.I.R. ?

- Evaluate the performances of an algorithm for music structure estimation
  - mono-dimensional structure
  - Reduction from multi-dimensional to mono-dimensional:
    - set of rules based on weighting of the various dimensions have been created which allows deciding weither a CSLoop is “constitutive” of the music track structure or not
    - Additional descriptions of the constitutive CSLoops and are used to find equivalence
    - From the 300 music tracks test-set, only 200 music tracks could be reduced to a mono-dimensional structure
- Very rich information about the construction of music tracks.
  - allows highlighting the temporal relationship between the various dimensions
    - use of “++” over CSLoop before the entrance of PrimaryLead or stereotype used in specific music genre (such as the “chorus” based on the same CSLoop as the “verse”

Presentation overview

- Introduction
- Requirements for a robust annotation definition
  - “Information extraction”/ “Imitation” or “Reduction”
  - Definition/ Certainty/ Concision/ Universality
  - Perceptual Recognition Rate
- Existing Music Corpus Test-sets
  - Main problems of existing “Music Structure” annotations
  - Various possible definitions for a “Music Structure” annotation
- Proposal: Multi-dimensional music structure annotation
  - Examples
  - Evaluation
  - Use in Music Information Retrieval
- Conclusion and Future works

Conclusion and Future works

- Defined a set of conditions to create robust concepts to be used for local music annotation
- Application for the creation of a robust “Music Structure” annotation system
  - a multi-dimensional description of “Music Structure” which uses simultaneously various super-imposed view-points: “musical role”, “acoustical similarity” and “instrument role”
- Further works
  - defining a quantitative measure for the Perceptual Recognition Rate (PRR)
    - could actually be obtained using the performance measures (insertion, deletion, equivalence between labels) commonly used to evaluate M.I.R. algorithms but applied this time between annotations performed by different annotators
  - applying the same approach to other well-known local music annotation tasks, such as singing voice, chord or melody description