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

Geoffroy Peeters and Emmanuel Deruty

IRCAM Sound Analysis/Synthesis Team - CNRS STMS,
geoffroy.peeters@ircam.fr,
WWW home page: http://www.ircam.fr

Abstract. Considering that M.I.R. content-extraction algorithms are evaluated over annotated test-sets, it is worth discussing the robustness of the concepts used for these annotations. In this paper we discuss the robustness of local music annotations, more specifically "Music Structure" annotation. We define four conditions to be fulfilled by an annotation method to provide robust local annotation. We propose mathematical formulations of two of them. We then measure these criteria on existing "Music Structure" test-sets and discuss the pro's and con's of each test-set. From these, we derive a robust set of concepts which form a "multi-dimensional" description of the "Music Structure". We then apply this description to a set of 300 tracks representing various music genres and discuss the results.

1 Introduction

A large part of present-day "Music Structure" research is devoted to the improvement of algorithms, through the improvement of recognition scores or to the definition of new measures of performances. But a question that should also be asked is "how pertinent is the structure annotation that is used for those evaluations?" This involves a precise definition of the annotation process. This question is important. Before thinking about how precise is the annotation, before thinking about how much it corresponds to the initial definitions, before thinking about how good is the automatic estimation, one should think about the annotation's relevance. This question arises directly when comparing annotations from the same tracks (as "The Beatles") coming from two different "Music Structure" test-sets made by different research teams. Actually "Music structure" is a notion that has never been clearly defined. Therefore, the amount of work concerning its automatic estimation or its evaluation is surprisingly large compared to the amount of work dedicated to providing a precise definition of the music structure the algorithms try to estimate.

Paper organization: The goal of this paper is to define a robust definition of "Music Structure" annotation. We start by defining a set of rules for a robust local music annotation (part 2). We then discuss the pro and cons of

previously existing "Music Structure" test-sets (part 3) along these rules of robustness. From this discussion and from the set of rules, we propose a robust multi-dimensional definition of "Music Structure" annotation (part 4). The validity of the proposed approach is then tested (part 4.4). 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.

2 Requirements for a robust annotation definition

It is possible to divide the notion of "local music annotation" into two categories: "information extraction" and "imitation".

"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. "Imitation" or "reduction" would include: - note / chord / melody annotation.

2.1 Information extraction: conditions

We have established the existence of several conditions on the annotation criteria for which "Information Extraction" will work on a given corpus. 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. A "chorus", for instance, is such an object, or indeed a descriptor. "Information Extraction" conditions concern those objects.

Four conditions can be identified:

1. **Definition:** an object, a descriptor, must be properly defined.
2. **Certainty:** in a given corpus, the object should be recognized without doubt.
3. **Concision:** the range of available descriptors should be limited.
4. **Universality:** a given descriptor should be used reasonably often.

2.2 Measuring "Certainty": Perceptive Recognition Rate (PRR) and Algorithmic Recognition Rate (ARR)

The condition ("Certainty") corresponds to a quantitative notion we name the "Perceptive Recognition Rate" or PRR. It can be measured by checking, on a given corpus, how many times a given object is recognized without doubt. It is capital to understand that this "PRR" is a key factor in annotation.

We note "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.

Indeed, if a given object is not easily recognized by ear, references to this object during annotation will be inaccurate (low PRR), and algorithm recognition experiments on this object will be invalidated. One could answer that ARRs are usually calculated on corpuses where PRR=1. However, this is not true, annotation in any field leads to uncertainty. We give below an example of this.

Application to the "chorus" case: We take here as example the case of the "chorus", which seems at first sight a very clear concept. Traditionally, a "chorus" is defined as: *"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".* We apply this definition on a first set of 112 songs (those songs are not particularly main-stream, neither are they particularly recent, their style is quite varied). Using this traditional "chorus" definition, the PRR is very low: less than 50%. It means that for this 112 songs test-set, we cannot tell if there is a chorus or not for 56 of them!

2.3 Measuring "Concision" and "Universality"

We measure the "Concision" and "Universality" of a given annotated test-set using the following measures:

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. A good annotation should have a small number L of labels.

$N(l)$**:** is the "N"umber of tracks using a specific label l, divided by the total number of tracks T. For a specific label l, a large value of $N(l)$ indicates that the concept of the label is usable over many tracks. The concept of the label is said to be universal. A small value of $N(l)$ is not necessary bad, it simply shows that the concept of the label l is not universal over track and is only applicable to a few tracks. It should be noted that $N(l)$ is 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). For a specific label l, a large value of $U(l)$ indicates that the concept is usable many times inside a track, it has by itself a structural role through its repetition inside the track. It should be noted that $U(l)$ is 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. A large value of mS indicates many segments in a song. Note that this value depends on the duration of the tracks annotated and the kind of music annotated.

mL**:** is average (over the tracks) number of different labels used for a specific track. A large value of mL indicates that many different labels are used for the description of a specific song. If mL is close to mS, it means that the label are only used once inside a track.

3 Related works

3.1 Existing test-sets

There has never been a clear definition of what "Music Structure" is. However several "Music Structure" test-sets have been proposed so far. We review them here and discuss them. In the following, we describe the existing "Music Structure" test-set and discuss their "Concision" and "Universality" using the proposed measures: T, L, $N(l)$, $U(l)$, mS and mL. Note that we cannot measure the "Definition" which is unfortunately not provided, neither the "Certainty" which requires to be present at the annotation time. We summarize these values in Table 1.

		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)**

Table 1. *Existing "Music Structure" test-sets and corresponding "Concision" and "Universality" measures.*

MPEG-7-Audio test-set [5]: The first test-set (2001) was realized by Ircam in the framework of an experimental scenario during the development of MPEG-7-Audio. T =25 tracks have been annotated in so-called structure (state and sequence structures). The annotations have been cross-checked by the other MPEG-7-Audio participants. L =50 different labels were used to label the segments: "bass", "break drum", "break guitar", "chorus 1", "chorus instru", "chorus variante", "verse glockenspiel". Clearly the terms used were a 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. The average number of segments per track is high mS=17.57 as the number of different labels used in a track mL=7.64. Most of the labels appear only for the description of a single track ("break 2", "break drum", "break guitar", "break piano", "intro synth", "intro voice") and only appear once in the track. The exception 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).

Comment: The number of labels is far too important, and their use very restricted.

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. For example, in the MPEG-7-Audio test-set, a large part of the tracks of the "state" corpus are made of "grunge" music, a music genre for which the instrumentation changes very significantly between the "verse" and "chorus" making the transition hence the annotation very clear. The second part of this corpus, named "sequence" corpus, is made of music where the instrumentation does not change over the track (early Beatles-like music), the structure is hence made by variations of the melodic line, which fit the sequence annotation definition.

QMUL test-set [1]: The Queen Mary University of London (QMUL) test-set starts from the MPEG-7-Audio "state" test-set and extend it a lot. It includes T =107 tracks of various pop-rock songs and many Beatles songs. It uses a total of L=107 different labels. The average number of segments per track is mS=12.33 and mL=6 different labels are used on average for a given track. Most of the labels appear only for the description of a single track ("crash", "fill", "drop", "crash") and only appear once in this track. The exception 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) .

Comment: The number of labels is far too important, and their use very restricted.

Beatles test-set [6]: The Beatles test-set has been 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). It describes T =174 tracks. All tracks are coming from The Beatles. L=55 different labels are used. The average number of segments per track is low mS=9.21 as the number of different labels used on average for a given track mL=5.23. Most of the labels appear only for the description of a single track ("close", "closing", "improv interlude") and only appear once in this track. 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).

Comment: Most of the labels are again used very few times. The most-used labels refer to the "musical role" of the part ("intro", "outro", "bridge", "verse", "chorus").

TUT07 Structure test-set [7]: The Tampere University of Technology (TUT) developed the largest test-set so far. The "TUTstructure07 musical structure database" contains T =557 Western popular music pieces (pop, rock, jazz, blues and "schlager" music) annotated into structure. This test-set seems annotated into "musical role" ("intro", "verse", "chorus") or "acoustical similarity" ("A", "B", "solo"). Unfortunately, since this test-set or its detailed description (except for the track list) is not available, we cannot provide detailed figures of it.

TU Vienna test-set [8]: The IFS TU Vienna uses a test-set of 109 tracks annotated into structure. Part of the tracks are coming from the QMUL (hence MPEG-7-Audio), RWC and Beatles test-sets. Because of that we do not give specific figures for this test-set. However, an interesting idea of this test-set is to allow several simultaneous descriptions of the same segment (describing a given part as a single segment or as a set of sub-segments) through the use of a hierarchical XML schema.

RWC test-set [3]: The RWC test-sets comes with the annotation of T=285 tracks into structure or chorus. The number of labels is restricted to L=17. The average number of segments per track is mS=15.73 (which is high but subdivision of segments are considered in this case) and mL=6.68 different labels are used on average for a given track. 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.

Comment: The 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"). Because of the restricted number of labels, their good coverage and the double "musical role"/ "acoustical similarity" this annotation is the best so far. However, the decision between the predominance of "musical role" over "acoustical similarity" is not always appropriate (some "intro" are in fact identical to "verse-a"). This highlights the necessity to separate both view-points.

3.2 Discussions

Main problems of existing "Music Structure" annotations: As one can see, each "Music Structure" test-set tends to use different rules and different set of labels. We summarize the main problems of these annotations here.

Number and coverage of labels: Most test-sets (except RWC) use a very large number of labels with a very-low usage.

Merging orthogonal view-points: Most test-sets merge descriptions related to - "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: Given two similarly different parts, the difference is sometimes interpreted as parts being identical and sometimes as parts being different. There is often a lack of consistency of the annotation process over a given test-set.

Describing the structure of the music, the melody, the instrumentation? It is not clear on which instruments the structure bases itself on. 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: The definition of the boundaries of the segments is often not coherent from track to track.

Segment sub-division: The definition of the "sub-division of a part A into sub-part a" is not coherent over a given test-set. If a chord succession is repeated over and over the verse, is this part a single "A" ? or a succession of repetition of "a" ?

Various possible definitions for a "Music Structure" annotation: In this part we propose several possibilities for the definition of "Music Structure". It is important to note that any choice of definition can be done if the appropriate test-set is chosen. For example it is possible to choose a "verse/chorus" description if the test-set contains only tracks with obvious "verses" end "choruses". Conversely 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.

It is also important to note that whatever choice is made for the description, it is important to avoid mixing various view-points. In the previous description of the test-sets, the labels which are used often merge various view-points, such as instrumentation and "musical role" ("break guitar", "break piano") or "acoustical similarity" and "musical role" (sometimes the part indicated as "introduction" is actually 100% similar to the part indicated as "verse").

Music Structure based on "musical role": One can rely on the choice of assigning labels according to the "musical role" that a part plays in a song ("introduction", "verse", "chorus", "bridge", "ending"). The same label is therefore used for parts playing the same role ("chorus", "verse"). However, in this case we also merge several notions. "Intro" and "Outro" refer to positions in the time axis of the track (sometimes the start or the end of the song is actually the "chorus"). Also they can be several versions of the "verse", "chorus" and "bridge" (hence the use in the previous test-sets of "verse A", "verse B"). We have already mentioned the problem of defining the "chorus"; now what is the definition of a "verse"? When one tries to annotate R'n'B music (one of the most popular and sold music today) there is often only a long verse, or several

parts which are all eligible to be verses, but which are called "hooks", "vamps" or indeed "verses".

Music Structure based on "acoustical similarity": One can rely on the choice of assigning labels according to the acoustical similarity between parts. In this description, the same label is assigned to parts with the same acoustical content. But the use of similarity however poses problems. Two parts are similar if they are identical (this is the case for example when using samples such as in the track Moby "Natural Blues"). But what about if there are small variations? "They are similar at 90%." What is the criterion used to say it is 90%? Is the timbre similarity more important than the harmonic or rhythmic ones? Is an instrument more important than another? This poses the problem of the point-of-view used to define the acoustical similarity. Then how do we go from the similarity between parts of 90% to the binary decision "they have the same (different) label(s)". This poses the problem of the choice of a threshold to make the binary decision. From this choice depends the number of labels used inside a track.

Music Structure based on "instrument role": One can rely on the choice of assigning labels to the description of the instrumentation of the track. In this, we describe the location of the lead singer parts, the (solo) guitar parts, ... This description is interesting but provides few insights into the global structure of the track. Furthermore providing the identity of the instrument will require a huge number of labels (guitar= classical guitar, folk-guitar, 12-string-guitar, electric guitar, wah-wah guitar ...). In this case it is more useful to describe the "role" plays by this specific instrument in the track, such as "Primary Lead" (the obvious front-men singer or instrument), "Secondary Lead" (the backing singer or side-man more generally). We call it "instrument role" in the following.

Music Structure based on the final application: One can also rely on the use of the structure in the final application. For example, if the final application is to create an audio summary or audio-thumbnail which must represent the most memorable part of the track (as used in [2]), it is maybe not necessary to spend time annotating the tracks in chorus location but only annotate the most repeated segments in them. The drawback of this approach is that the annotation can only be used to validate the target application and cannot be used for other applications.

Music Structure based on perceptual tests: One can also rely on perceptual tests to find the average human perception of the musical structure (as did [4] for tempo and beat perceptual annotation). Apart from the fact that this approach is very costly, another problem comes from the fact that, in the "Music Structure" case, the labels used by people to describe the structure of a track are usually not shared.

Proposed Music Structure, multi-dimensional representation: The main idea of the proposed description is to use simultaneously (but independently) the various view-points: "acoustical similarity", "musical role" and "instrument role". The idea is based on the way modern music is created through multi-track recording: a set of main patterns are repeated over time, with vari-

ations of those, with instrumentation super-imposed (singing voice), and which play a "musical role" such as introduction, transition, chorus, solo and ending. The proposed annotation method is based on "Constitutive Solid Loops", which are constitutive blocks whose limits and labels are derived from the various elements' synthetic perception. The criteria proposed in the following encompass the usual structure criteria such as "chorus" and "verses", but are much more powerful. Unlike the "chorus/ verse" approach, our method of structure annotation makes it possible to properly describe the structure for many different styles. The first example would be being able to annotate pieces that don't include any choruses, which are much more common than one would spontaneously think[1].

4 Proposed method: multi-dimensional music structure annotation

4.1 Overall explanation

The whole idea is that a track is formed:

- by a set of Constitutive Solid Loops (CSLoop) which represent a "musical phrase" or a "musical exposition" (a succession of chords). CSLoops with similar ID represent the same "musical phrase" although large variation can occur between them. Two CSLoops with the same ID can follow each other if the "musical phrase" is repeated twice successively.
- over which are super-imposed variations of the CSLoops ID. For example the same CSLoop which occurs in a lighter version (for example without the drum or without the bass) is indicated by "–"; if it is in a stronger version (for example with an extra second guitar) by "++".
- over which are superimposed important "instrument roles": such as presence of the "primary leads" (lead singer in popular music, lead instrument in jazz or electro music), "other leads" (choir, other lead instruments or melodic sample) or "solo mode" (electric-guitar solo, jazz chorus solo, ...)
- and which plays a "musical role" (intro, outro, transition, obvious chorus, or solo).

The track is therefore decomposed simultaneously on these various viewpoints. When a part is too complex to be described, it is annotated as "ComplexMode".

The mandatory decomposition is the CSLoop description. 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 can still be annotated as the repetition of the occurrence of a specific CSLoop, with PrimaryLead and OtherLead (Choir) which are distinctive elements.

[1] For instance, there is not a single chorus in the "Dark Side of The Moon" album from Pink Floyd, which sold 40 millions units, making it the 6th best selling album of all time.

In order to solve the 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.

The temporal boundaries of segments and markers are defined as the closest downbeat to the start or end of the respective described object.

4.2 Detailed description

In Table 2, we give the detailed specification and definition of the proposed annotation into "Music Structure".

Trans: indicates transitions which are structurally outside the CSL.

IO: indicates intro, outro parts or exotic parts (parts which have nothing to do with the rest of the song).

CSLoop 1, 2, 3, 4, 5, A, B : indicates a musical phrase, idea, or subject. The equality rule applies to the CSLoops 1 2 3 4 5, i.e. two CSLoops with the same ID represent the same thing. It does not apply for the CSLoop A and B. CSLoop A and B are used either when the track contains too many CSLoop to be annotated, or when the annotator cannot reliably decide about the equivalence between CSLoop but still want to mark a segment.

– **(++):** when applied to a CSLoop it indicates that this occurrence of CSLoop has a much lower (higher) loudness than the rest of the song, or a part in which two of the three references (rhythmic, melodic, harmonic) disappear (are added).

Cplx: indicates a very complex non-periodic part (such as in Frank Zappa free improvisation parts).

SMode: indicates a Solo Mode, whatever the instrument playing the solo (vocal, guitar, sax, piano), it can be super-imposed to a CSLoop to indicates the part over which the solo is performed.

PLead1: indicates the presence of the main melodic referent, which is usually the main singer or the instrument playing the theme in jazz music.

PLead2: indicates the presence of a second (side-man), third melodic referent in case of duo, trio, ...

Olead1, OLead2: indicates a second melodic referent which is not the main one (backing vocals, instrument interacting with the singer melody).

Chorus 1, 2: indicates "obvious chorus", i.e. when it is without any doubt the chorus. Note that two chorus ID are possible.

V1 (V2): is a marker (as opposed to the pervious descriptions which are segments), it indicates a sub-division inside a specific CSLoop with the part on the left of V1 (V2) being similar (not similar) to the one on the right.

The "Exclusion" column of the table indicates the exclusion rules of the labels. For example a segment cannot be "–" and "++" at the same time.

# id	Name	Full name	Type	Description	Equality rule	Exclusion	N(l)	U(l)
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

Table 2. *Definition of multi-dimensional "Music Structure" annotation labels and associated $N(l)$ and $U(l)$ over the 300 track test-set.*

4.3 Examples

In Figure 1 and Figure 2, we give two examples of the application of the proposed method for two tracks which are also describes in the test-sets of part 3.

Figure 1 represents the annotated structure of The Cranberries "Zombie". As one can see, the annotation is multi-dimensional (several criteria described at the same time). The main structure of the track relies on two CSLoops: "CSL1" and "CSL2". CLS1 is used as the introduction ("IO") in a lighter form ("–"). It is followed by "CSL2" in a strong version ("++"). Then back to "CSL1" in normal form which acts here as a transition ("trans"). The end of this part has an Other Lead ("OL1") (which is the guitar melody of "Zombie"). Then the CSL1 is repeated twice with singing voice ("PL1"). This part would be named "verse" in the previous test-sets, however naming it "verse" does not tell it is actually the same part as the transitions ("trans") and as the solo ("SMode"). "CLS2" follows in a strong version ("++") with singing voice ("PL1") and is obviously a chorus ("Chorus 1"). The rest of the track can be interpreted in the same way until the end of the track which is again a "CSL1" in light form ("–") acting as an outro ("IO"). Note also the added comma separation in the CSLoop ("V1", "V2") indicating sub-repetition ("V1") or sub-division ("V2"). Notice how concise is this representation, and the amount of information it contains.

Figure 2 represents the annotated structure of The Beatles "Come Together". The track is formed by 4 different CSLoop. It starts with the "CSL1" in normal form acting as an introduction ("IO"). The second "CSL1" has singing voice ("PL1") and ends with a lighter version of the "CSL1" ("–"). The next "CSL1" acts as a transition ("trans"). "CSL2" acts as the obvious chorus ("Chorus 1"). Around time 125s, a new part, "CSL3", starts with an Other Lead (OL1) which

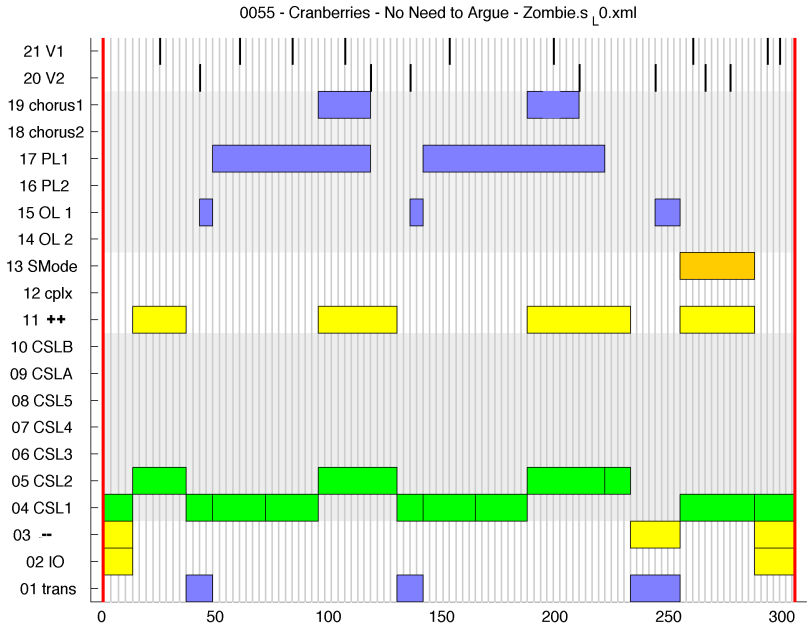

Fig. 1. *Example 1: Cranberries "Zombie".*

is the guitar, and the Other Lead acts as a Solo ("SoloMode" or "SMode"). The end of track is a "CSL4" with interlaced "Primary Lead 1" (the singer) and "Other Lead 1" (the guitar melody). Again, the description is quite simple for a complex structure.

4.4 Testing over a large variety of music genre

The applicability of the proposed description has been tested over a large 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 ...).

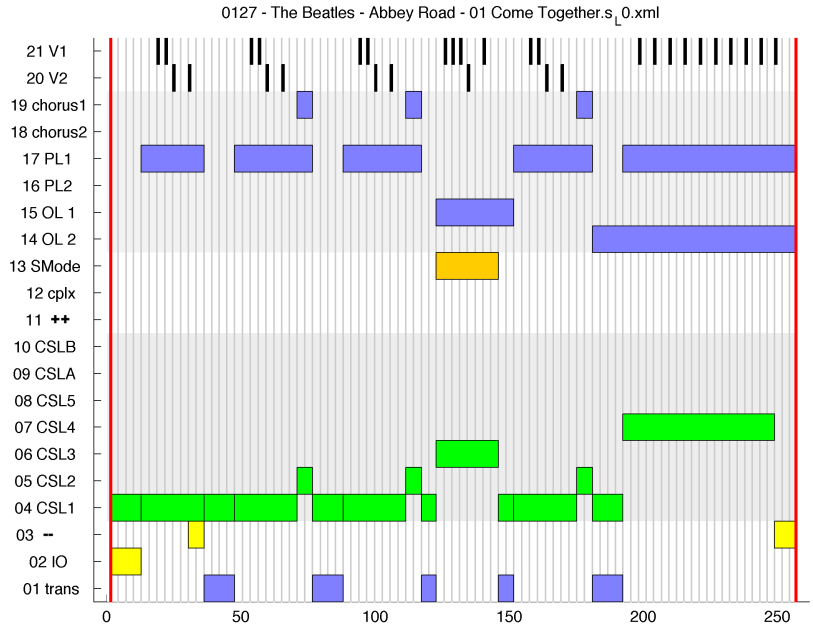

Fig. 2. *Example 2: The Beatles "Come Together".*

4.5 Information extraction conditions applied to the proposed multi-dimensional music structure annotation

Given that the first conditions "Definition" is fulfilled, we measure the other conditions "Certainty" (PRR), "Concision" and "Universality" (T, L, $N(l)$, $U(l)$, mS and mL values) of the proposed multi-dimensional music structure annotation on this 300 tracks test-set.

PRR: Observations made on a three-month period show that our multi-dimensional annotation method shows reliable results over time, and, more importantly these results don't depend on the annotator. The proposed method permits a much higher annotator agreement than previously existing method (tested on the same tracks). This indicates a high PRR.

L**:** 21 different labels (19 when omitting the subdivision with comma V1 and V2). As for the RWC, the total number of labels is small.

mS**:** on average a track is divided into 38.93 segments (22.80 when omitting the sub-division with comma V1 and V2). This mS is very high. Indeed, to measure mS we have considered that the appearance of each new labels super-imposed in the middle of another one (as PLead1 appearing in the middle of a CSLoop) creates a new segment. Considering only the *CSLoop*

segmentation will decrease a lot mS. This makes somehow our description a scalable description.

mL: on average a track uses 9.80 different labels (8.11 when omitting the subdivision with comma V1 and V2). This high value comes also from the multi-dimensionality of our description. Because of the simultaneous use of various view-points, several labels co-exist at the same time (such as "CSLoop1" with "–" with "Plead1"), the number of labels used inside a track is therefore larger.

$N(l)$ and $U(l)$: The detailed results of $N(l)$ and $U(l)$ are given in the last two columns of table 2. The mean (over labels) $N(l)$ is 0.47 (0.39 for RWC) which is very high, the mean (over labels) $U(l)$ is 3.21 (2.16 for RWC). This shows that the concept used are quite universal across music tracks (high $N(l)$) (can be used for many different tracks), and plays a structural role inside a track (high $U(l)$). The fact that these values are higher than for RWC and the fact that our test-set has much more music genres is very promising for our approach. Only the "CplxMode" and "Chorus2" are used in few tracks which is coherent with their functionality ("too complex to be annotated" and "there exist two different chorus").

Examples of the annotated test-set are accessible at the following address `http://recherche.ircam.fr/ equipes/ analyse-synthese/ peeters/ pub/ 2009_LSAS/`.

4.6 Use of the proposed multi-dimensional Music Structure annotation in M.I.R.

The starting point of this research on Music Structure annotation was the creation of a test-set to evaluate the performances of an algorithm for music structure estimation. Since this algorithm estimates a mono-dimensional structure, studies have been done on the development of a methodology to reduce the multi-dimensional structure annotation to a mono-dimensional one. A 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. The other criteria (PrimaryLead, OtherLead, –, ++ ...) are then considered as additional descriptions of the constitutive CSLoops and are used to find equivalence between them hence repetitions of parts over time. From the 300 music tracks test-set, only 200 music tracks could be reduced to a mono-dimensional structure. The structure of the remaining 100 music tracks did not fulfill the requirements of repetitive parts (whatever it is "acoustical similarity"-based "musical role"-based or "instrumentation"-based) over time. Their reduction would have lead to a very low PRR.

This problem of non-possible reduction to mono-dimensional structure should be addressed by the development of multi-dimensional structure estimation algorithms.

Apart from this evaluation use, the multi-dimensional annotations provide very rich information about the construction of music tracks. It allows highlighting the temporal relationship between the various dimensions (such as the 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").

5 Conclusion and Future works

In this work, we have proposed a set of conditions to define robust concepts to be used for local music annotation. We have used these conditions for the creation of a robust "Music Structure" annotation system. For this we have proposed the use of a multi-dimensional description of "Music Structure" which uses simultaneously various super-imposed view-points: "musical role", "acoustical similarity" and "instrument role". We have tested our description in an annotation experiment on a collection of 300 tracks coming from various music-genres. The four measures (Definition, Concision, Universality, and Certainty) were all above the results obtained with previous test-sets. Especially the proposed method permits a much higher agreement among annotators.

Further works will concentrate on defining a quantitative measure for the Perceptual Recognition Rate (PRR) that was used during the experiment. This quantity 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.

Further works will also concentrate on applying the same approach to other well-known local music annotation tasks, such as singing voice, chord or melody description.

6 Acknowledgments

This work was partly supported by "Quaero" Programme, funded by OSEO, French State agency for innovation? This work couldn't have been realized without the great work of Jean-Francois Rousse and Maxence Riffault. We would like to thanks the three anonymous reviewers for their fruitful comments which help on the improvement of this paper.

References

1. S. Abdallah, K. Nolan, M. Sandler, M. Casey, and C. Rhodes. Theory and evaluation of a bayesian music structure extractor. In *Proc. of ISMIR*, pages 420–425, London, UK, 2005.
2. M. Cooper and J. Foote. Automatic music summarization via similarity analysis. In *Proc. of ISMIR*, pages 81–85, Paris, France, 2002.
3. M. Goto. Rwc (real world computing) music database, 2005.
4. D. Moelants and M. McKinney. Tempo perception and musical content: What makes a piece slow, fast, or temporally ambiguous? In *International Conference on Music Perception and Cognition*. Evanston, IL, 2004.

5. MPEG-7. Information technology - multimedia content description interface - part 4: Audio, 2002.
6. B. Ong and P. Herrera. Semantic segmentation of music audio contents. In *Proc. of ICMC*, pages 61–64, Barcelona, Spain, 2005.
7. J. Paulus and A. Klapuri. Labelling the structural parts of a music piece with markov models. In *Proc. of CMMR*, Copenhagen, Denmark, 2008.
8. E. Peiszer, T. Lidy, and A. Rauber. Automatic audio segmentation: Segment boundary and structure detection in popular music. In *Proc. of LSAS*, Paris, France, 2008.
9. A. Pollack. "notes on ..." series. the official rec.music.beatles, 1989-2001.