Labelling the Structural Parts of a Music Piece with Markov Models

Jouni Paulus, Anssi Klapuri

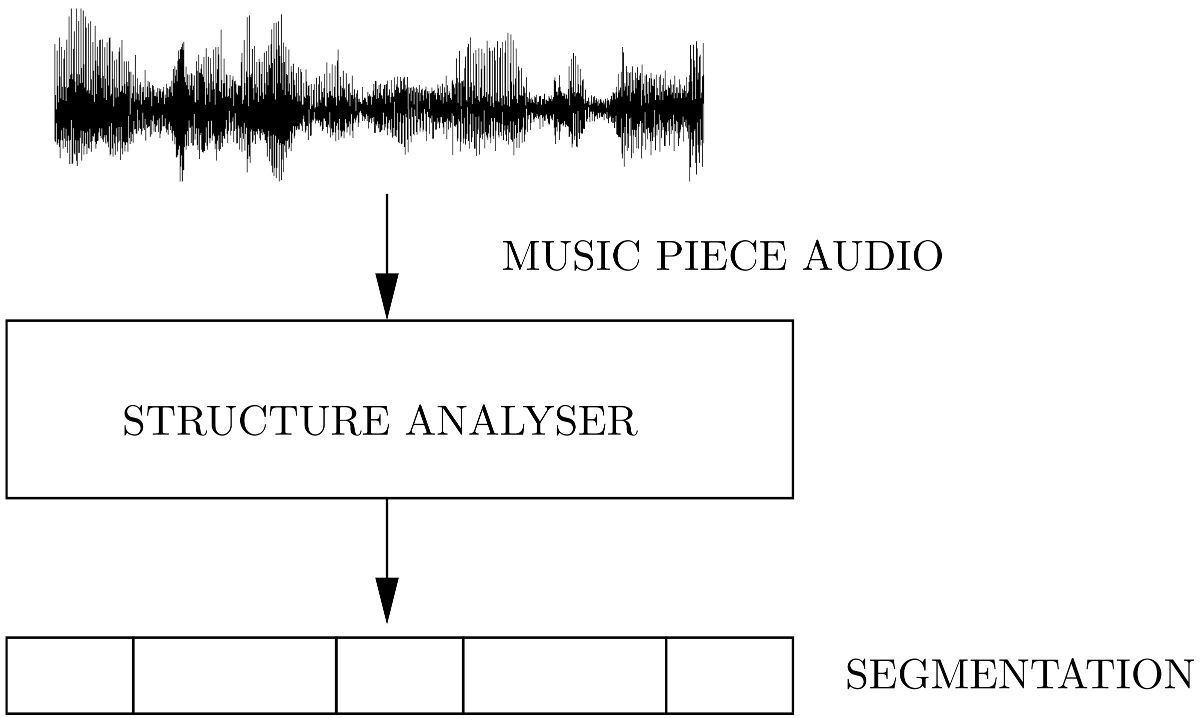
Audio Research Group
Department of Signal Processing
Tampere University of Technology
Tampere, Finland

2008 Computer Music Modeling and Retrieval
20.5.2008

Music structure analysis

- In general: recover sectional form of musical piece.
- Different levels for the task:
  - Piece segmentation.
    - Logan & Chu, 2000, ICASSP
    - Foote, 2000, ICME
    - Goodwin & Laroche, 2004, ICASSP
    - Jensen, 2007, JASP
    - Turnbull et al, 2007, ISMIR

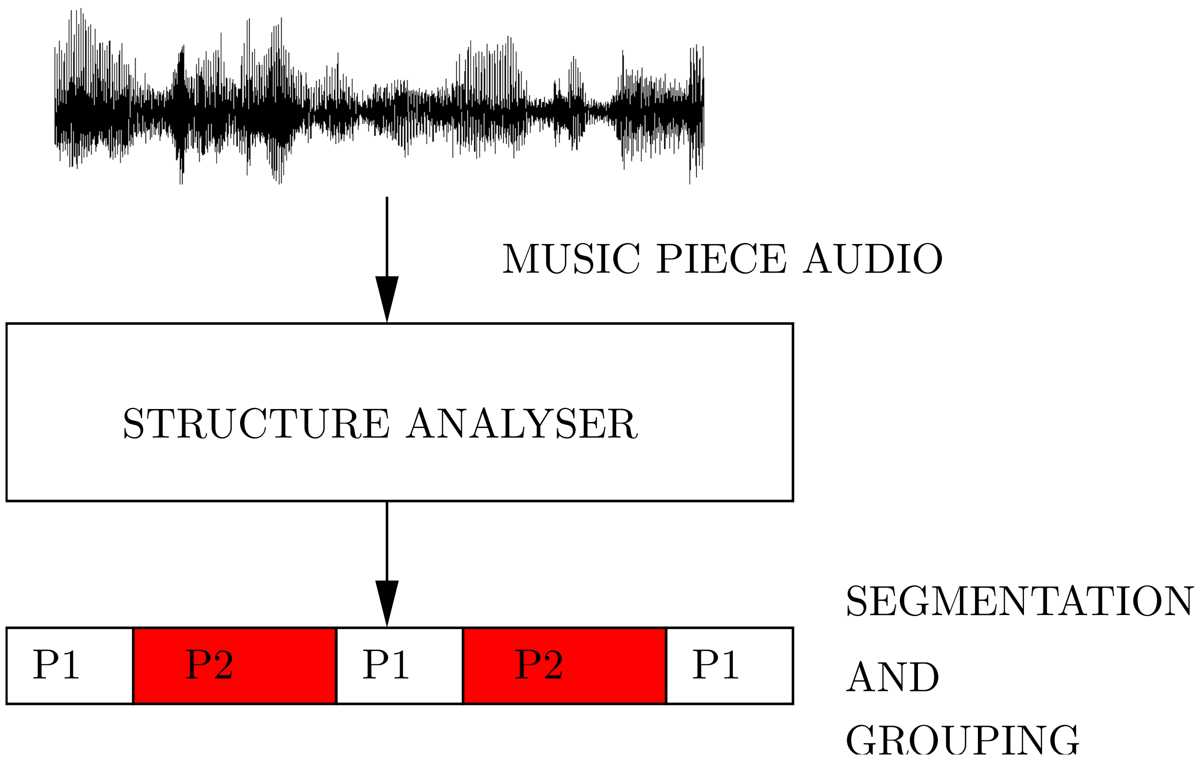
Piece segmentation



Music structure analysis

- In general: recover sectional form of musical piece.
- Different levels for the task:
  - Piece segmentation.
  - Piece segmentation and grouping of segments representing same musical part.
    - Cooper & Foote, 2003, WASPAA
    - Peeters, 2003, CMMR
    - Aucouturier et al, 2005, IEEE-TMM
    - Chai, 2005, PhD
    - Ong, 2006, PhD
    - Rhodes & Casey, 2007, ISMIR
    - Levy & Sandler, 2008, IEEE-TALSP

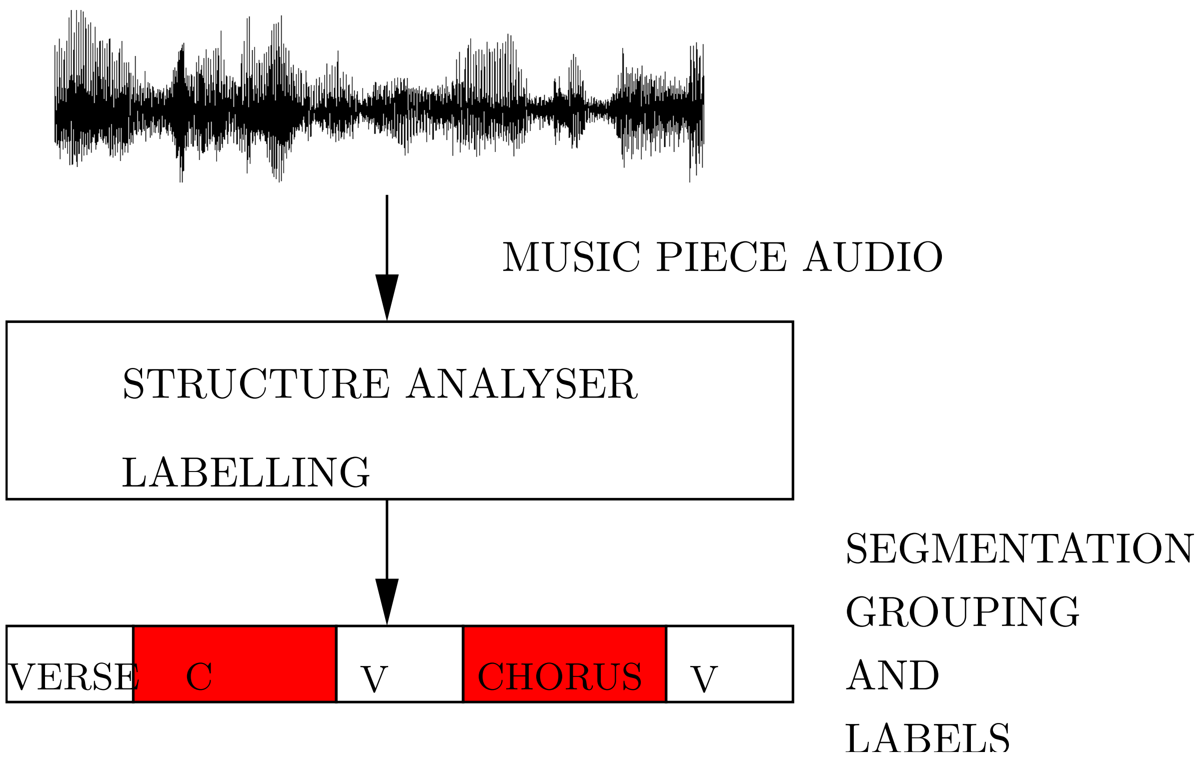
Piece segmentation and grouping



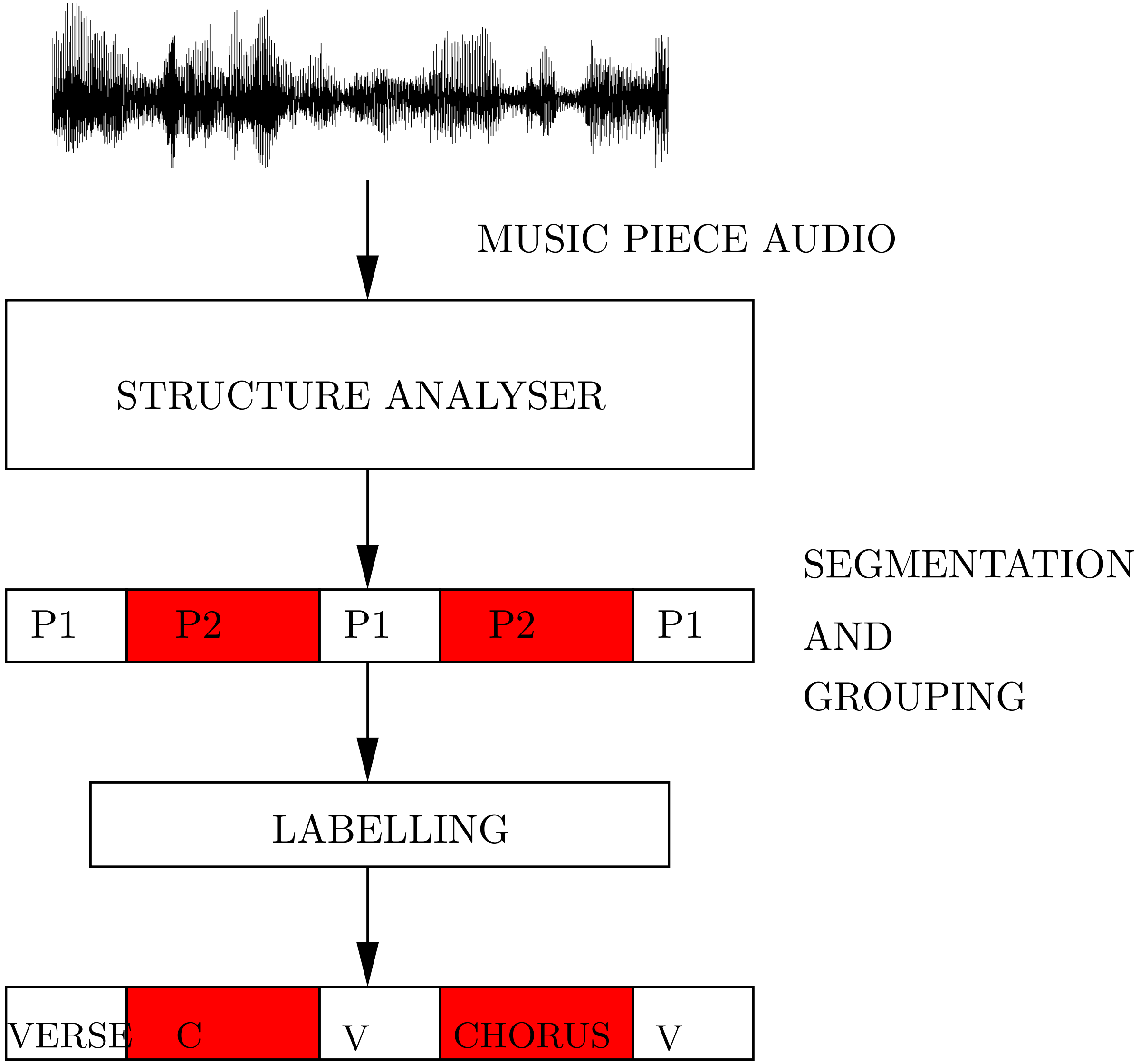
Music structure analysis

- In general: recover sectional form of musical piece.
- Different levels for the task:
  - Piece segmentation.
  - Piece segmentation and grouping of segments representing same musical part.
  - Piece segmentation, grouping of segments of same part, and labelling of the groups.
    - Shiu et al, 2005, SPIE
    - Maddage, 2006, IEEE-MM

Piece segmentation, grouping and labelling



Piece segmentation, grouping and labelling as post-processing

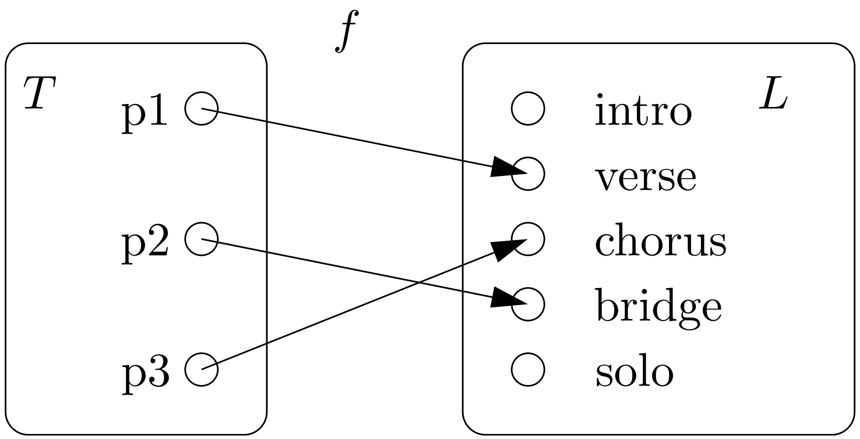


Motivation for labels

- Add *musically meaningful* labels to the output of existing structure analysis systems. (Labelling task)
- Test N-grams for modelling musical structure. (Modelling task)
- Abstract segment names given in order (p1, p2, a, b, c, . . .) not practical for modelling.
  - Same “identifier” has different meaning in different pieces.
  - N-gram -like models too random.
  - Still, less than 10% of abstract sequences occurred more than once in TUTstructure07.

Task

- Input: musical piece structure as a string of symbols describing the grouping of segments, e.g. “p1, p3, p1, p3, p2, p3”.
- Output: sequence of musically meaningful labels, e.g., “verse, chorus, verse, chorus, bridge, chorus”.
- Find: an injective mapping function $f : T \to L$.



- Any injective mapping is valid, but which is the best one?

Scoring the mappings

- Model musical structures with N-grams of musical labels. N-gram model (Markov assumption)

$$p(S_{i+1}|S_{1:i}) = p(S_{i+1}|S_{(i-N+1):i}) \quad (1)$$

- Probability over a whole sequence is

$$p(S_{1:K}) = \prod_{k=1}^{K} p(S_k|S_{(k-N):(k-1)}) \quad (2)$$

- Find the mapping f_{OPT} maximising the total probability

$$f_{OPT} = \operatorname*{argmax}_{\mathrm{f}} \{p\,(f(R_{1:K}))\},\ f : T \rightarrow L \text{ injective} \quad (3)$$

Optimisation algorithm

- Exhaustive testing of all combinations laborous.
- Employ a variant of N-best token passing.
  - Greedy, beam search -like algorithm.
  - State space from possible labels.
  - At each iteration search proceeds to next segment.
  - Transition probabilities between labels from N-grams.
  - Prohibit transitions breaking mapping injectivity.
- Result compared with exhaustive search: almost always the same result is found, computationally much lighter.

Data

- Segmentation of pieces and labelling of parts.
  $\rightarrow$ Possible to treat as sequences.
- Two separate data sets:
  1. TUTstructure07
     - 557 pieces.
     - Mainly pop/rock, some jazz, blues, schlager etc.
     - Annotated in-house by non-professionals.
  2. UPF Beatles
     - 174 pieces by The Beatles.
     - Piece forms by Alan W. Pollack.
     - Segment time stamps annotated in Universitat Pompeu Fabra.

Data diversity

- Stereotypical assumption "every pop piece follows the same form..." was shown to be invalid.
  - Top-3 in TUTstructure07, each with 4 occurrences.
    - "intro", "verse", "chorus", "verse", "chorus", "C", "chorus", "outro"
    - "intro", "A", "A", "B", "A", "solo", "B", "A", "outro"
    - "intro", "verse", "chorus", "verse", "chorus", "chorus", "outro"
  - 524 (94%) of the label sequences in TUTstructure07 are unique.
  - UPF Beatles has a clear winner occurring 17 times:
    - "intro", "verse", "verse", "bridge", "verse", "bridge", "verse", "outro"
  - 135 (78%) of the label sequences in UPF Beatles are unique.

Data diversity, cont'd

- Small amount of sequences and large vocabulary.
- Reduce vocabulary size by retaining only the labels covering 90% of all part occurrences.
  - TUT: from 80 to 13 labels
  - UPF: from 50 to 8 labels
- Different N-gram model orders. (Number of parameters to be estimated A^N.)
- Variable-order Markov models (VMMs).
  - Try to determine best model order automatically based on data.
- Maximise usable training data in evaluations with leave-one-out cross-validations.

TUTstructure07, hit-%

label	N=1	N=2	N=3	N=4	N=5	VMM
chorus	68.1	76.3	**80.8**	76.6	74.9	78.5
verse	42.3	62.4	64.4	64.9	**66.0**	66.0
bridge	17.7	38.6	45.6	**47.4**	44.4	43.7
intro	27.6	97.6	**98.2**	97.8	97.8	96.4
pre-verse	4.2	40.7	**46.3**	43.3	41.7	43.3
outro	13.9	98.3	**98.6**	97.8	92.1	98.3
c	0.0	38.0	42.1	47.4	**54.8**	49.3
theme	0.0	0.0	2.7	**4.4**	3.3	3.3
solo	0.0	4.4	7.2	16.0	**18.2**	14.9
chorus_a	0.0	0.0	7.5	**15.7**	11.2	3.0
a	0.0	0.0	**32.5**	31.7	27.0	29.4
chorus_b	0.0	0.9	5.3	**12.4**	7.1	2.7
MISC	12.6	29.5	38.3	37.1	**40.3**	38.3
average	30.9	55.6	**60.3**	59.9	59.5	59.8

UPF Beatles, hit-%

label	N=1	N=2	N=3	N=4	N=5	VMM
verse	72.4	79.9	86.7	85.7	83.7	**87.5**
refrain	30.1	32.1	62.2	66.3	68.7	**70.7**
bridge	36.7	40.7	**78.0**	74.0	74.0	70.6
intro	0.0	93.2	88.9	92.0	**93.8**	93.2
outro	0.0	99.3	**99.3**	97.2	93.0	97.9
verses	0.0	16.1	48.2	**50.0**	44.6	44.6
versea	0.0	5.9	7.8	17.6	**21.6**	5.9
MISC	0.0	15.9	22.3	**25.5**	23.6	22.3
average	33.5	58.9	72.1	72.8	72.1	**73.0**

Conclusions

- Assigning *musically meaningful* labels to structural parts only based on sequential dependencies between them works surprisingly well.
  - True labelling would require the use of other information sources, too.
    - Acoustic cues...
    - Lyrical information...
- Low model order sufficient (due to data sparsity?)
- Simple model contains much musical information.
  $\rightarrow$ *How to utilise it in the structural analysis process?*

Raw bi-gram transition probs in UPF Beatles