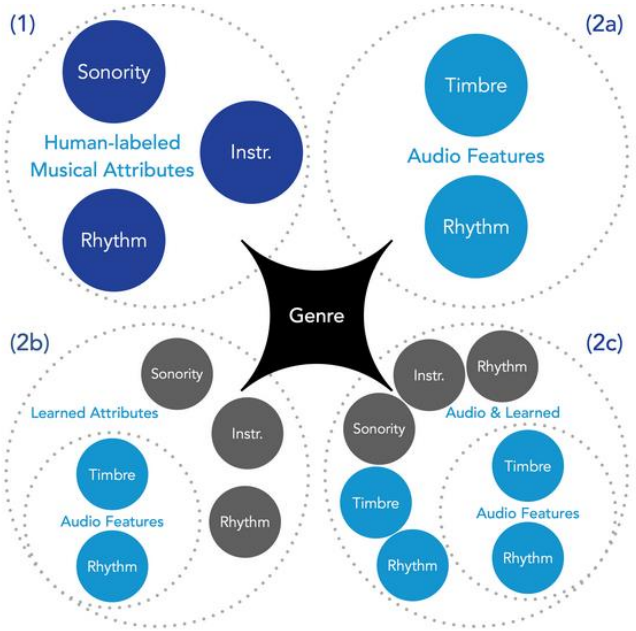
Audio similarity

its evaluation, and meta-analyses

Meta-analysis of MIREX data

- **Music Information Retrieval Exchange (MIREX)**
  - Sets tasks each year for researchers to test new algorithms
  - Works better in some areas than others
  - Results announced at **ISMIR conference**
- Tasks focus on ten or so areas of MIR
  - Grading done by volunteers
  - Two important meta-analyses of results



Alan Marsden meta-analysis (JNMR 2012)

- Looked at MIREX 2002-2006, with emphasis on 2005
- Similarity may be in the ear (or eye) of the beholder. [Credit = A/ Tversky]
- **Reductive approaches** produce *inconsistent results*.

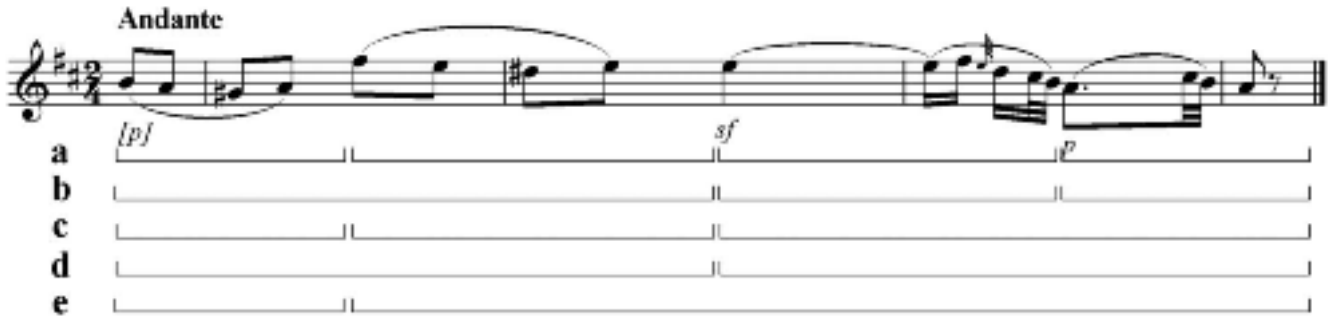


Fig. 6. Alternative segmentations of the second phrase of the theme of the third movement of Mozart's string quartet in A major, K. 464.

Mozart, K. 464, II

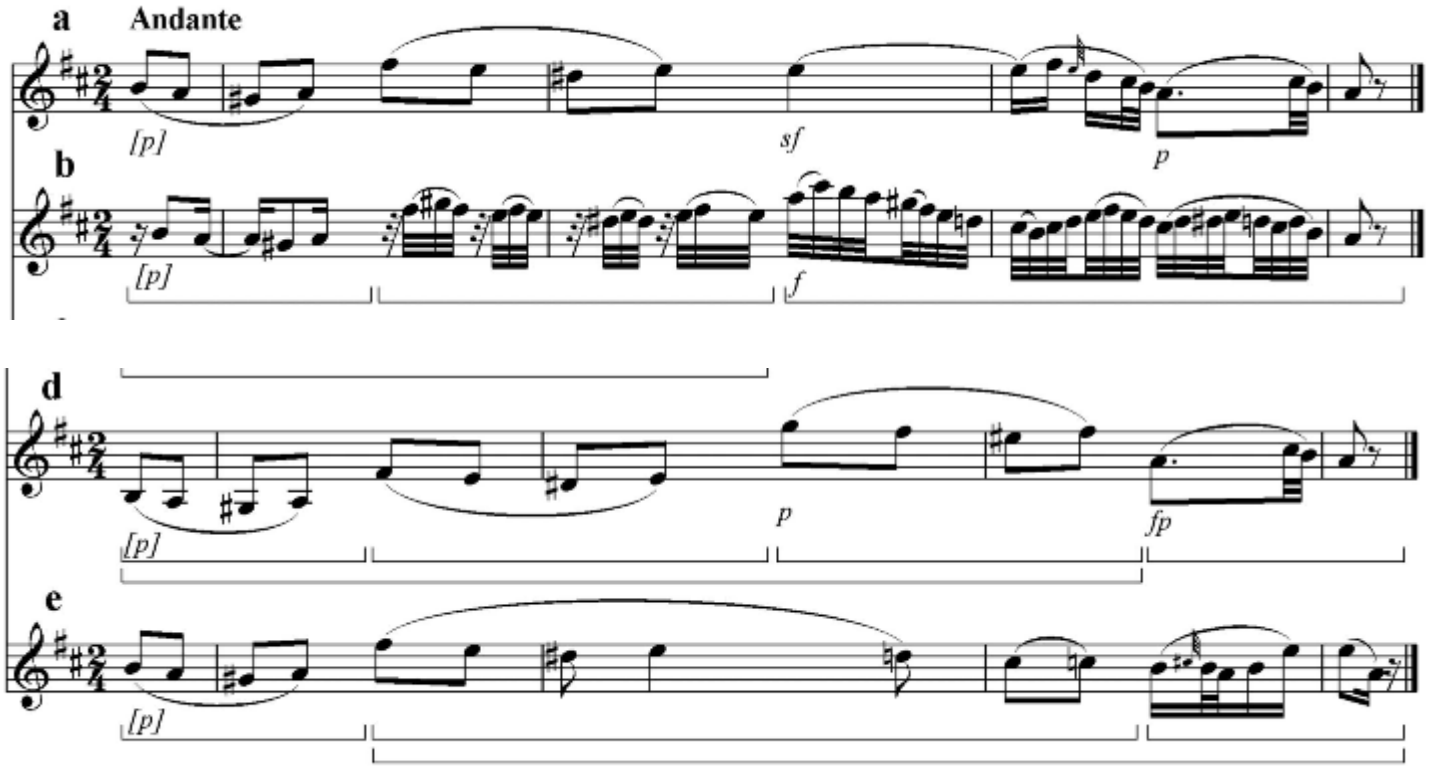


Fig. 7. Different segmentations found in variations in Mozart's K. 464 of the theme from Figure 6.

Citations include…

Arthur Flexer[(1)] et al. meta-analyses:
MIREX 2006-2014 plus own data [Soundpark]

- With Thomas Grill (no picture), Markus Schedl (2), and Julián Urbano (3)
- Stage 1: re-examination of MIREX analysis related to similarity
- Stage 2: independent user studies with off-label Austrian pop

Austrian pop used by Flexer et al.

Austrian Center for AI (OFAI)

- Set up FM4 Soundpark
- Allow artists to upload own works
- Holdings used for research
  - Sound-processing
  - **Human-response studies**

Flexer et al. 2014

- **Inter-rater agreement** in audio music similarity (ISMIR2014)
- In 2006-2015, performance peaked in 2009.

	grader1	grader2	grader3
grader1	1.00	0.43	0.37
grader2		1.00	0.40
grader3			1.00

- Why lack of inter-rater agreement?
  - Concept of music similarity is too "coarse"
  - Upper bounds can be achieved by algorithms
  - Performance in 2009 cannot be exceeded without changes of approach

Flexer et al., 2014

Observations

- **Musical similarity is complex** and depends *on individual exposure* and *experience*.
- **Human judgments** will therefore *vary* from person to person
- "Any evaluation of MIR systems…based on 'ground truth' [as] annotated by human beings"…has the same limitations.

Flexer and Grill (2016)

"The problem of limiting inter-rater agreement in modelling music similarity", *Journal of New Music Research* 45/3 (2016), 239-251.

http://dx.doi.org/10.1080/09298215.2016.1200631

- Quantitative relations should **mirror human perception** of similarity…

  but they don't.

- Computational models that *exceed limits* of human perception are useless.

Tests used by Flexer and Grill

Performance comparisons

1. Modeling music similarity *between pieces*.
2. Modeling structural analyses [i.e. segmentation, of pop] *within pieces*.

Set-up: Three graders for each task

Highlights from **findings**

1. In Task 1**, timbre and rhythm** were *most influential features*.
2. **Same algorithms** *did not perform consistently* from year to year.
3. Algorithms performed almost as well as people [cf. Haydn/Mozart QQ]
4. **Classical and world music** more difficult to model than popular music.

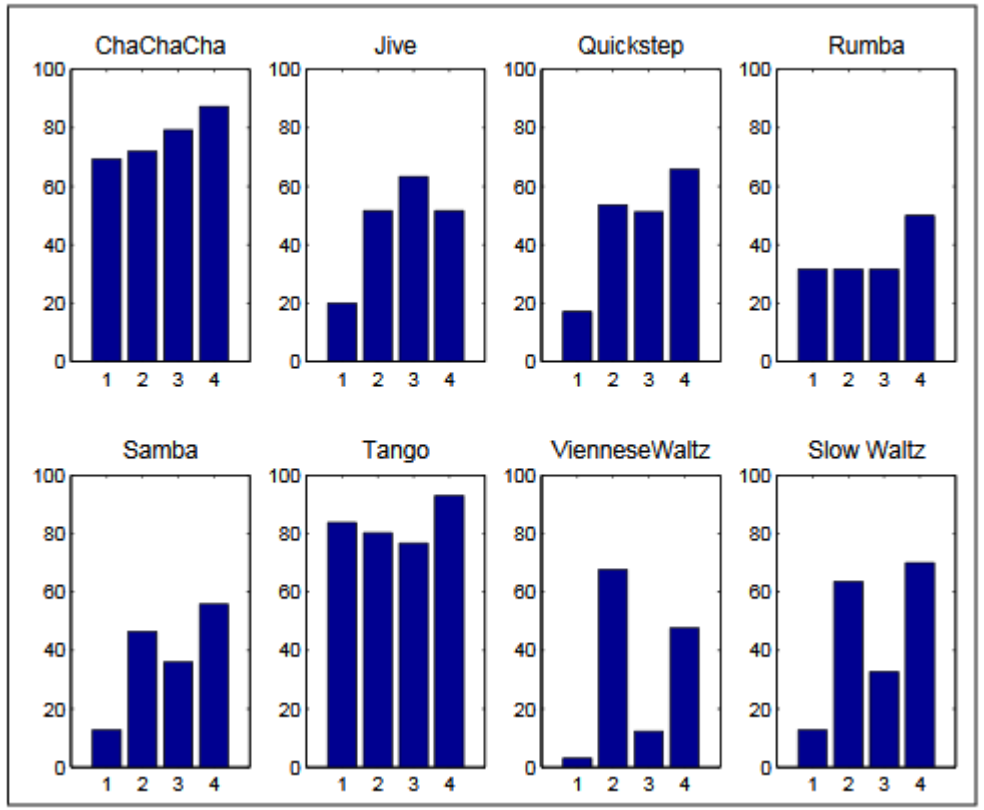
Flexer and Grill, Tasks 1 and 2

Questions:

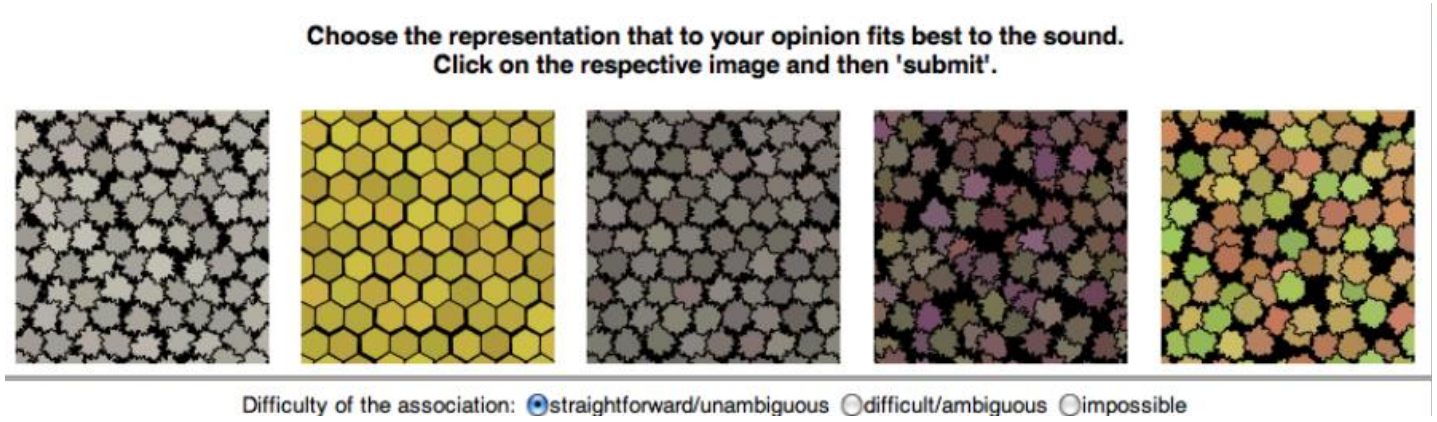
- 1. Should MIR *evaluate whole systems* instead of individual items?
- 2. Should we *refocus on a core set of* better-defined *tasks*? (MIREX)

Other work by Flexer (now Linz)

- With Markus Schedl, 2012:
  - ~~Genre~~ is *too fuzzy* a concept to model. Use similarity instead.
  - Make personalized systems.
- Probablistic combination of features for music classification (2006)
  - Rhythmic similarity in dance-music data



Sensory mapping (for intuitive searches); Timbral-visual mapping



2012 (WWTF)

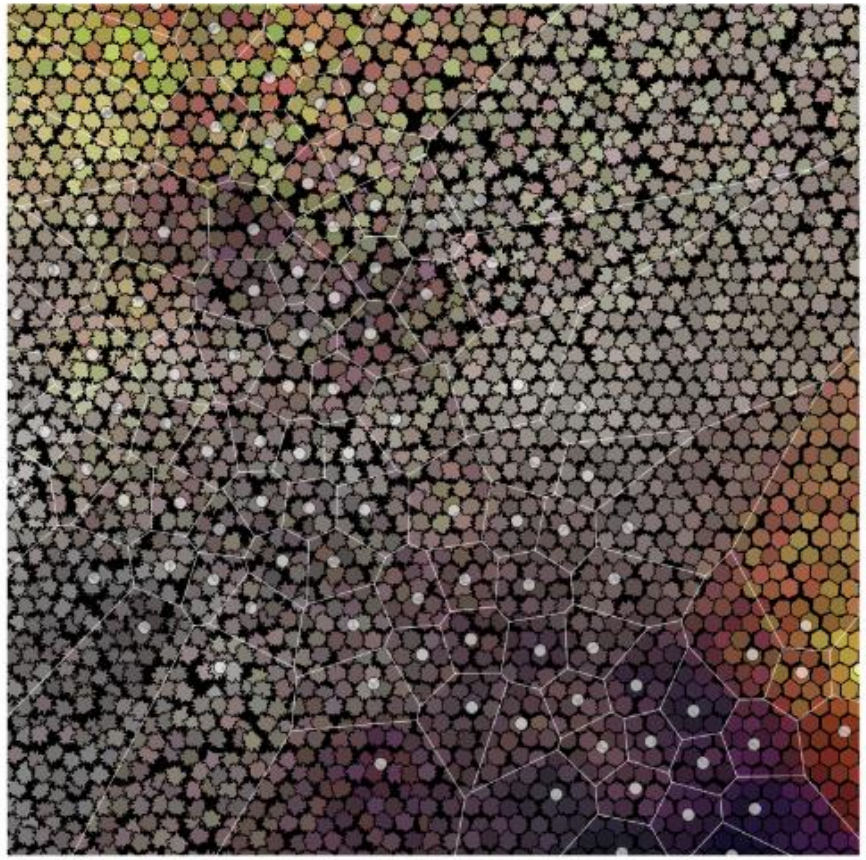
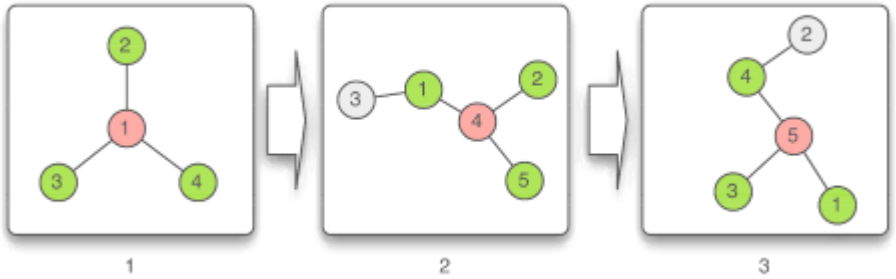


Figure 7. Web-browser based interface for browsing textural sounds building on the perceptually informed visualization strategy. The tiling is interpolated between the individual sound positions for a clearer appearance.

Limitations of audio-based recommendation systems (ACM 2010)

- Series of experiments under title **FM4 Soundpark**
- *Main focus*: Why some songs in a large database are never recommended?
- User builds similarity net
- Focus on **hub**



- Performance evaluated by **actual listening**, not mere downloading
- **Songs similar to many others more likely to be listened to**

Work of Julián Urbano et al. (now Delft)

- Geometric models of melodic similarity (symbolic data): ISMIR 2013
  - Transposition invariant
  - Time-scale invariant (CMMR 2010; Springer Verlag)
- Evaluation of MIR systems

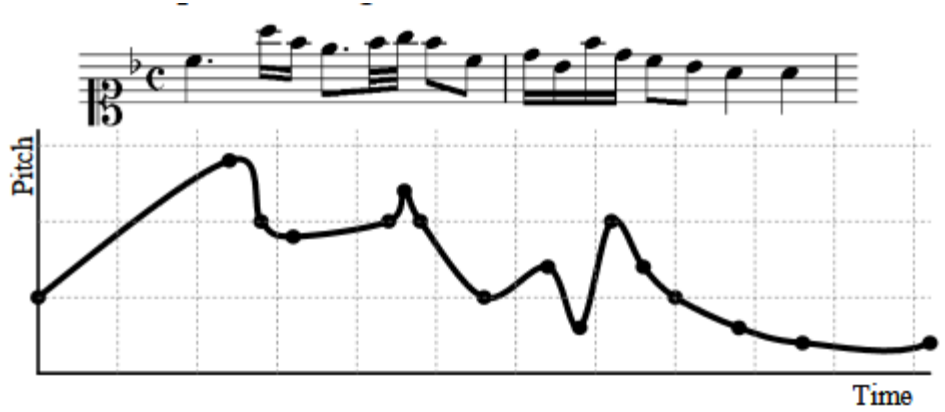


Figure 1. Melody represented as a curve in the pitch-time plane.

5. RESULTS

Table 1 shows and excerpt of the official MIREX results [5], with the overall figures for the systems described. Notably, all our four systems ranked in the top 5 for all 10 effectiveness measures (5th only in 4 of the 40 cases).

	JU1	JU2	JU3	JU4
ADR	0.307 (5)	0.309 (3)	0.317 (2)	0.371 (1)
NRGB	0.297 (3)	0.294 (4)	0.288 (5)	0.328 (1)
AP	0.300 (3)	0.299 (4)	0.301 (2)	0.349 (1)
PND	0.373 (2*)	0.373 (2*)	0.368 (4)	0.399 (1)
Fine	0.579 (5)	0.583 (2)	0.581 (3)	0.606 (1)
Psum	0.613 (4)	0.620 (2)	0.615 (3)	0.642 (1)
WCsum	0.559 (3*)	0.563 (2)	0.559 (3*)	0.580 (1)
SDsum	0.532 (3)	0.535 (2)	0.531 (4)	0.549 (1)
Greater0	0.777 (5*)	0.790 (3)	0.783 (4)	0.827 (1)
Greater1	0.450 (2*)	0.450 (2*)	0.447 (4)	0.457 (1)
Median Rank	3	2	3.5	1

Table 1. MIREX overall results for our four systems. Ranks per effectiveness measure appear in parentheses. * for tied ranks.

Accommodation of variation (Urbano)

- Use of interpolated splines
- Experimental results

Graphical edit distance

- Insertion
- Deletion
- Substitution
- Match

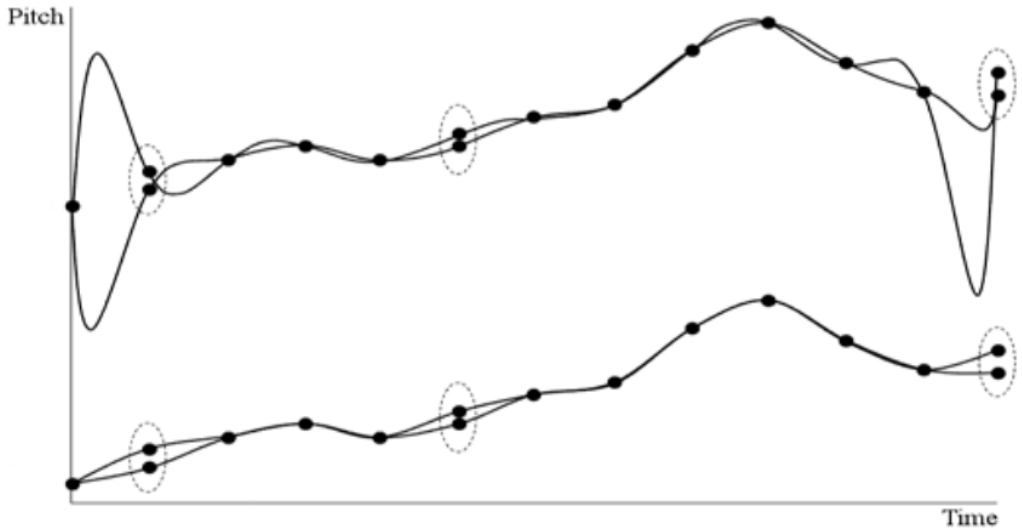


Fig. 11. Runge's Phenomenon

- **Findings**
  - Spans of 4 notes perform best; performance degrades with length
  - Model ignores rests, which are often missing in MIREX test sets

Urbano:

- https://link.springer.com/content/pdf/10.1007%2F978-3-642-23126-1.pdf
- Compares diverse concepts of “equality”
  - Octave equivalence
  - Degree (harmonic) equality
  - Note equality
  - Harmonic similarity
  - Time-signature equivalence
  - Tempo, duration
- Measures of dissimilarity

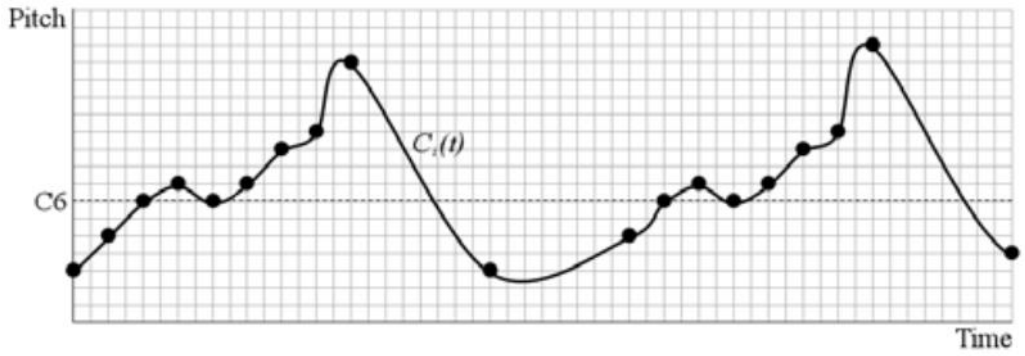


Fig. 10. Melody represented as a curve in a pitch-time plane

Urbano: vertical features in “matching”

- Octave equivalence: allow (perceptually, I’d say disallow)
- Scale-degree (melodic) equivalence: if key irrelevant
- “Note equivalence”: same as transposition/scale degree
- Pitch variation: allowance of approximate matches (no discussion of accent)
- Harmonic similarity: rank
- Voice separation: problem of working with composite and single voices

Urbano: horizontal features in matching

- Time-signature equivalence: 2/4, 4/4
- Tempo equivalence: gets into metronome markings
- Duration equivalence: quality of performance
- *Duration variation*: or, *Privilege accented notes*?

Solutions to equivalence problems (Urbano)

- Nos. 1-3: use **scale-degree differences**, not exact pitch differences
- Horizontal requirements:
  - Time signature difference not important when equivalent
  - Duration can be measured two ways:
    - *Elapsed time* in performance
    - *Implied time* in score: he gets into pitch-time splines here
- Then: *measure dissimilarity* in splines (*oscillation*)
- Finally: a [new] model for transposition and *time-scale invariant comparison.*