A comparison and evaluation of approaches to the automatic formal analysis of musical audio

Jordan B. L. Smith
McGill University / USC
jordan.smith2@mail.mcgill.ca

Ichiro Fujinaga
McGill University
ich@music.mcgill.ca

1. Problem:

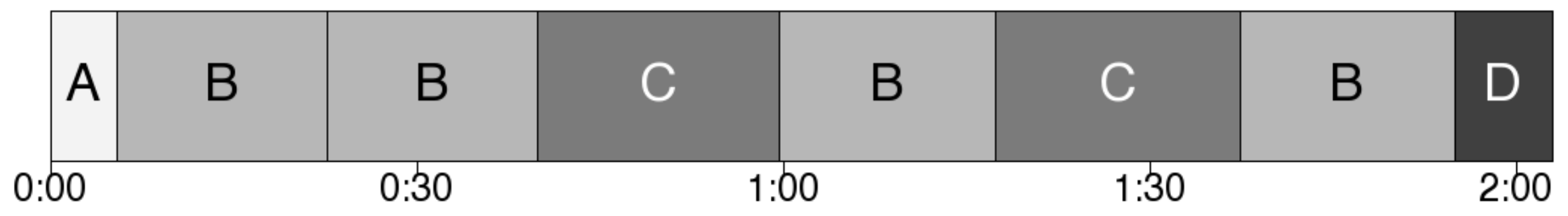
Many music informatics researchers have sought to develop algorithms that analyze the long-term structure or form of pieces of music, but evaluating these algorithms is difficult. In this research, we compare three such algorithms and assess their performance on three separate collections of music.

1.1. Operational Definition:

A large-scale structural description of a piece of music comprises:

1. **boundaries** between sections;
2. **labels** that identify which sections are similar to each other.

For example, here is the structure for "Yesterday" by The Beatles:



2. Motivation:

Effective structural analysis algorithms could be used to:

- Facilitate large-scale musicological studies;
- Identify different versions of a piece within a collection;
- Generate audio summaries;
- Search for music by form.

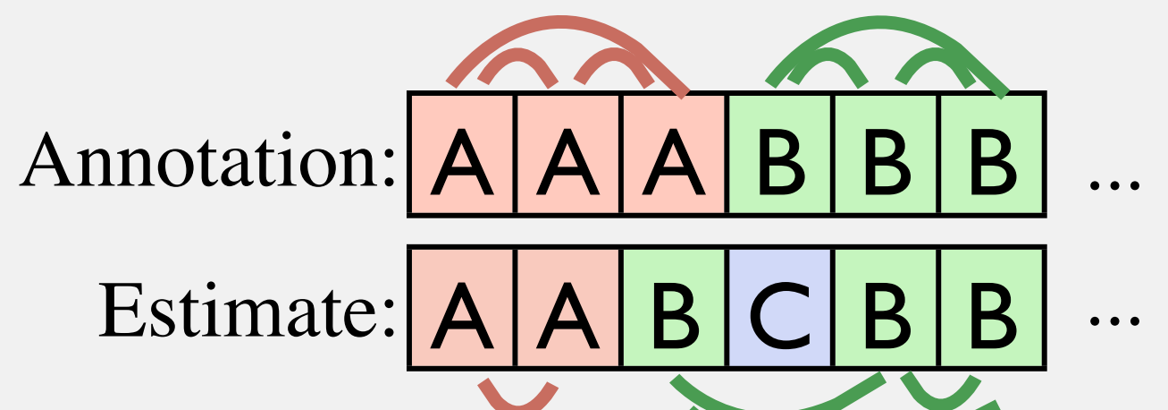
3. Experiment:

- **Three structure analysis algorithms**, described by Barrington et al. (2009), Levy and Sandler (2008) and Peiszer (2007), were used to analyze **three corpora** of different genres (see *Corpora* info box).
- **Naïve baseline algorithms** placed a boundary every 10, 15, 20, or 30 seconds, and either used random or uniform labelling. Uniform labelling tended to be ranked higher.
- **Algorithm input parameters** included the audio feature being used (pitch- or timbre-based) and the maximum number of unique labels (varied between 3 and 10).

3.1. Evaluation:

One set of common evaluation metrics includes pairwise *precision*, *recall* and f-*measure*.

- *Pairwise recall*: the percentage of pairwise matching segments in the annotated description that are contained in the estimated description.
- *Pairwise precision:* the percentage of estimated pairwise matches that are correct.
- *Pairwise* f-*measure:* harmonic mean of precision and recall.



In this example:
precision = 2/4 = 50%
recall = 2/6 = 33%
f-*measure* =
2·33·50/(33+50) = 0.4

3.2. Algorithms:

Barrington et al. (2009):

- Music modelled as mixture of time-changing "dynamic textures"; entire problem solved in a single, long, computation-intensive step.

Levy and Sandler (2008):

- Estimates low level "transcription-like" representation using a Hidden Markov model (HMM), then associates section labels with neighbourhoods of HMM states.

Peiszer (2007):

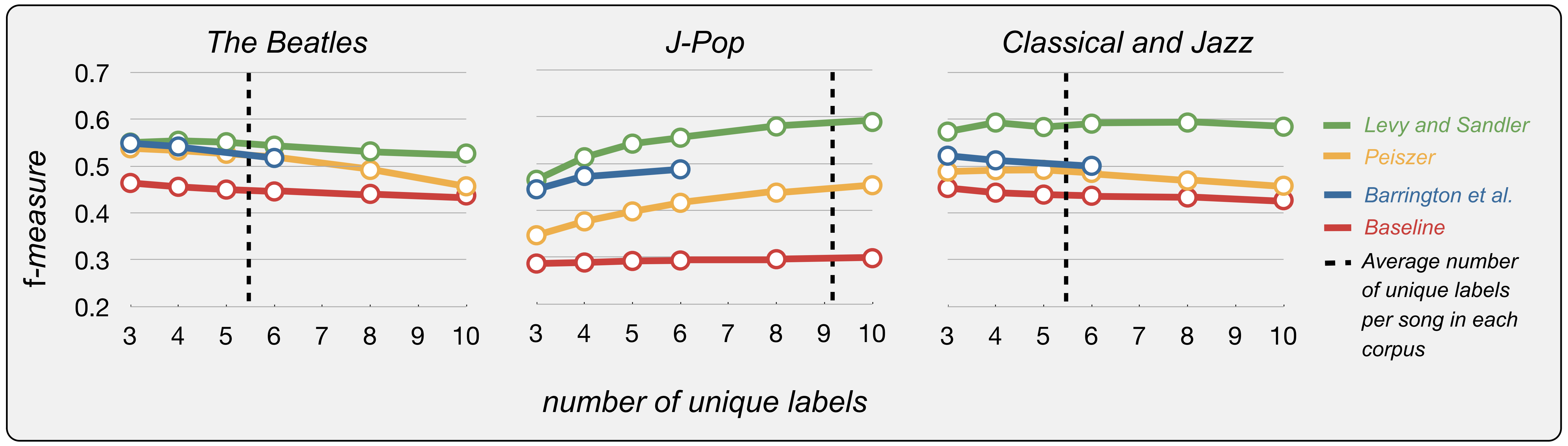
- Estimates boundaries by finding points of maximum novelty, then clusters resulting segments according to overall similarity.

4. Results:

- As seen at left, performance was comparable across corpora, though Levy and Sandler tended to do best.
- Overall, the algorithms were more successful using timbral features.
- When the prescribed number of section types was nearer to the true value, the *f*-measure tended to be higher.
- Examples of the algorithms' estimates are shown on the right.

Corpora:

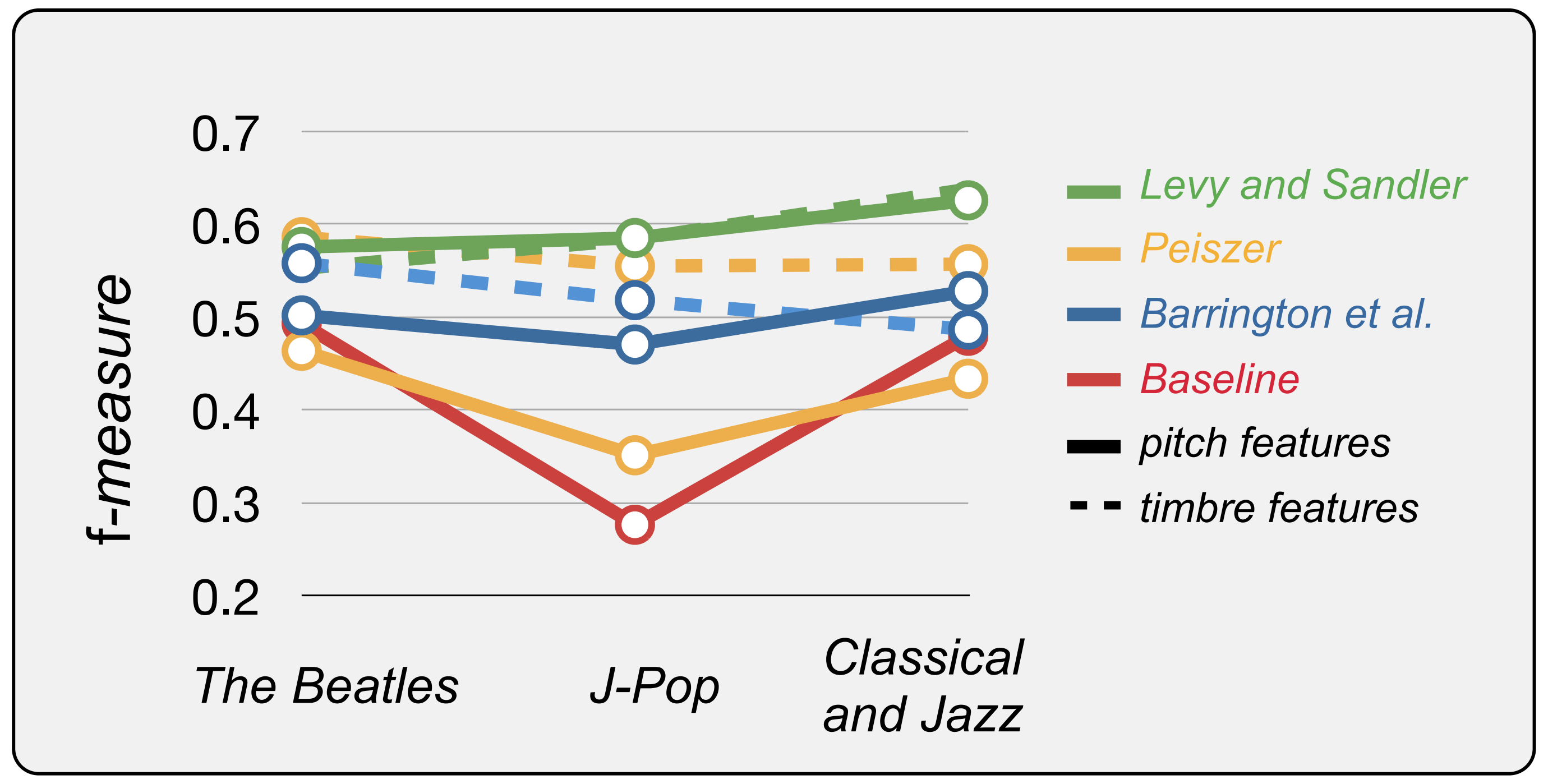
	The Beatles	J-Pop	Classical and Jazz
Contents	All 12 studio albums by The Beatles	Real-World Computing (RWC) Popular Music Database of Japanese pop songs	Public domain recordings, chosen for their structural simplicity and covering a variety of periods and styles
# of pieces	180	100	15 classical and 15 jazz
Annotations	Based on analyses by musicologist Alan W. Pollack	Annotations created by RWC	Produced by the author for this project



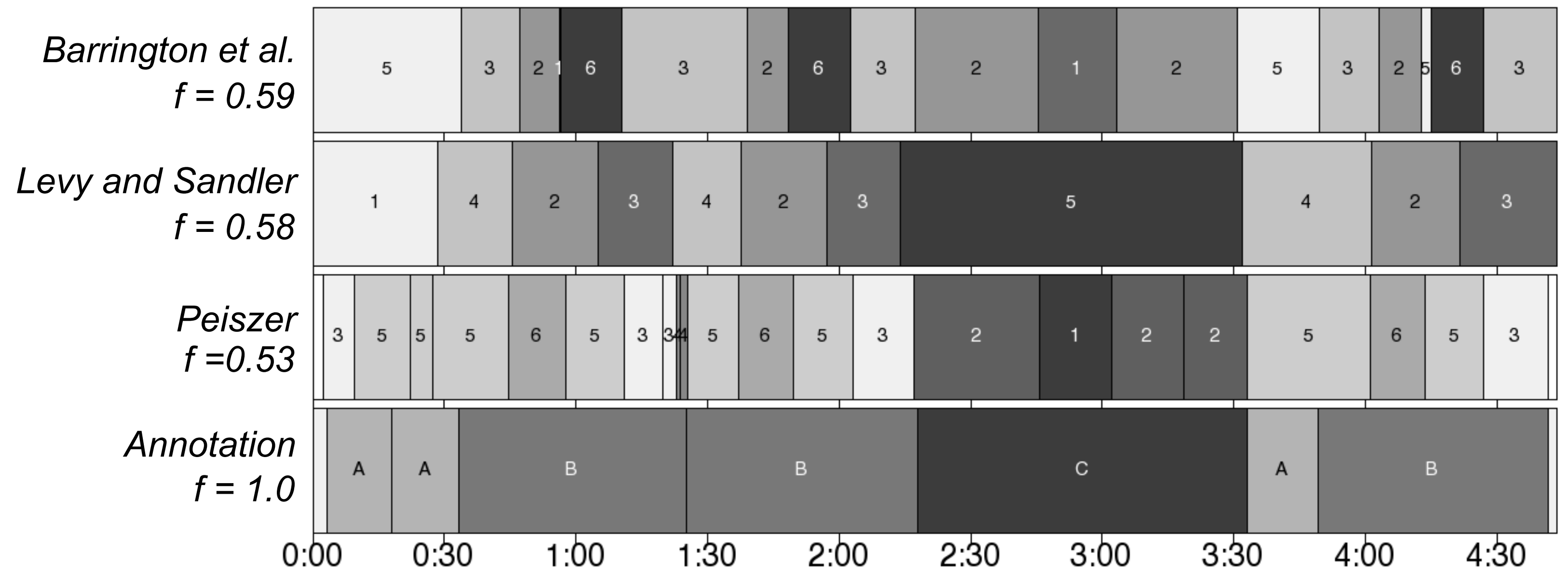
Above: when the input parameter "number of unique labels" was set near the average for that corpus (given by the dotted line), scores increased. Note that due to computing constraints, Barrington's algorithm was operated fewer times.

Right: algorithm success as a function of corpus and feature set. Note that overall, timbre features outperformed pitch features.

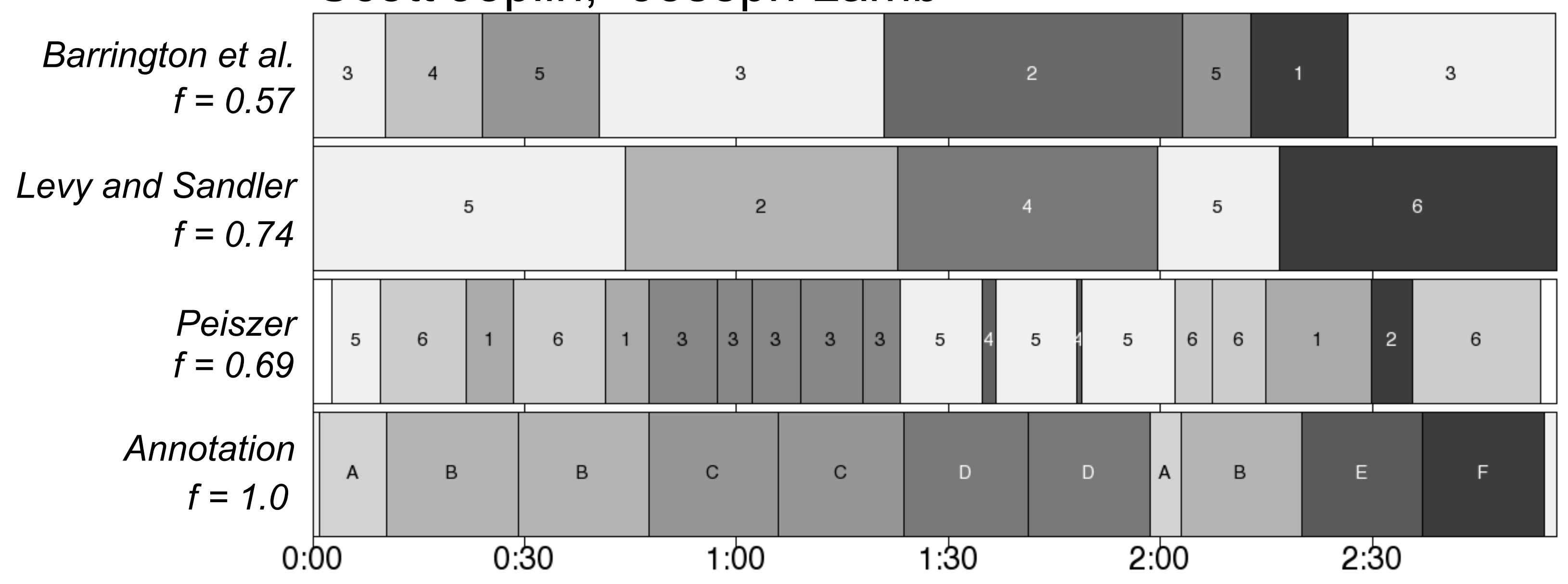
Below right: example output and f-measure from each algorithm for two pieces. The f-measure obtained by a single-label baseline was 0.57 for the Beethoven and 0.45 for the Joplin.



Beethoven, Symphony No. 8, Op. 93, iii



Scott Joplin, "Joseph Lamb"



5. Conclusion:

- The algorithms studied here performed about as well on classical and jazz pieces as on Beatles and J-Pop music.
- This indicates that the analytical models they are founded on are general and powerful.
- However, their performance does not greatly exceed that of the naïve baselines, and the results are plainly noisy.
- Fortunately, development on these algorithms is ongoing and all are available in some form online for use in your own research.

McGill
Schulich School of Music
École de musique Schulich
CIRMMT Centre for Interdisciplinary Research in Music Media and Technology
DDMAL DISTRIBUTED DIGITAL MUSIC ARCHIVES & LIBRARIES LAB
Fonds de recherche sur la société et la culture Québec
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada