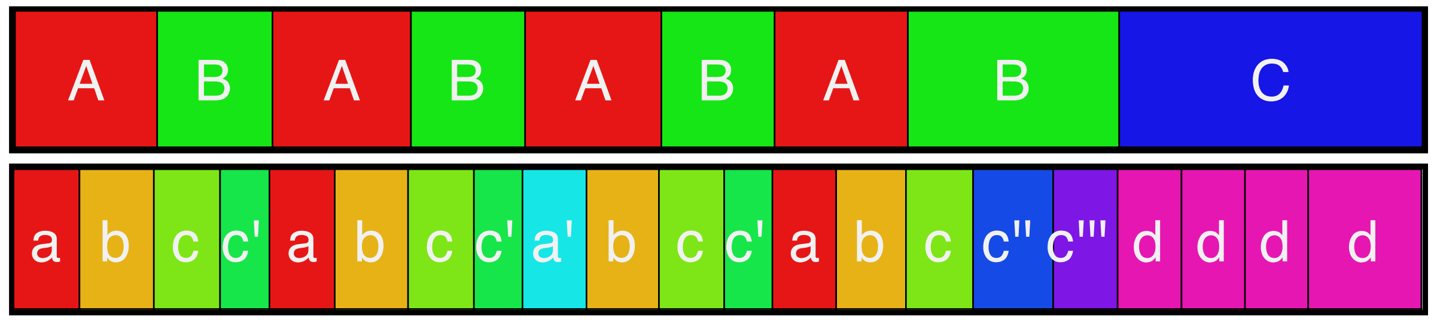
Automatic interpretation of music structure analyses

A validated technique for post-hoc estimation of the rationale for an annotation

Jordan B. L. Smith, National Institute of Advanced Industrial Science and Technology (AIST), Japan
Elaine Chew, Queen Mary University of London

1. Motivation

Structural descriptions are usually single-dimensional, or perhaps hierarchical:



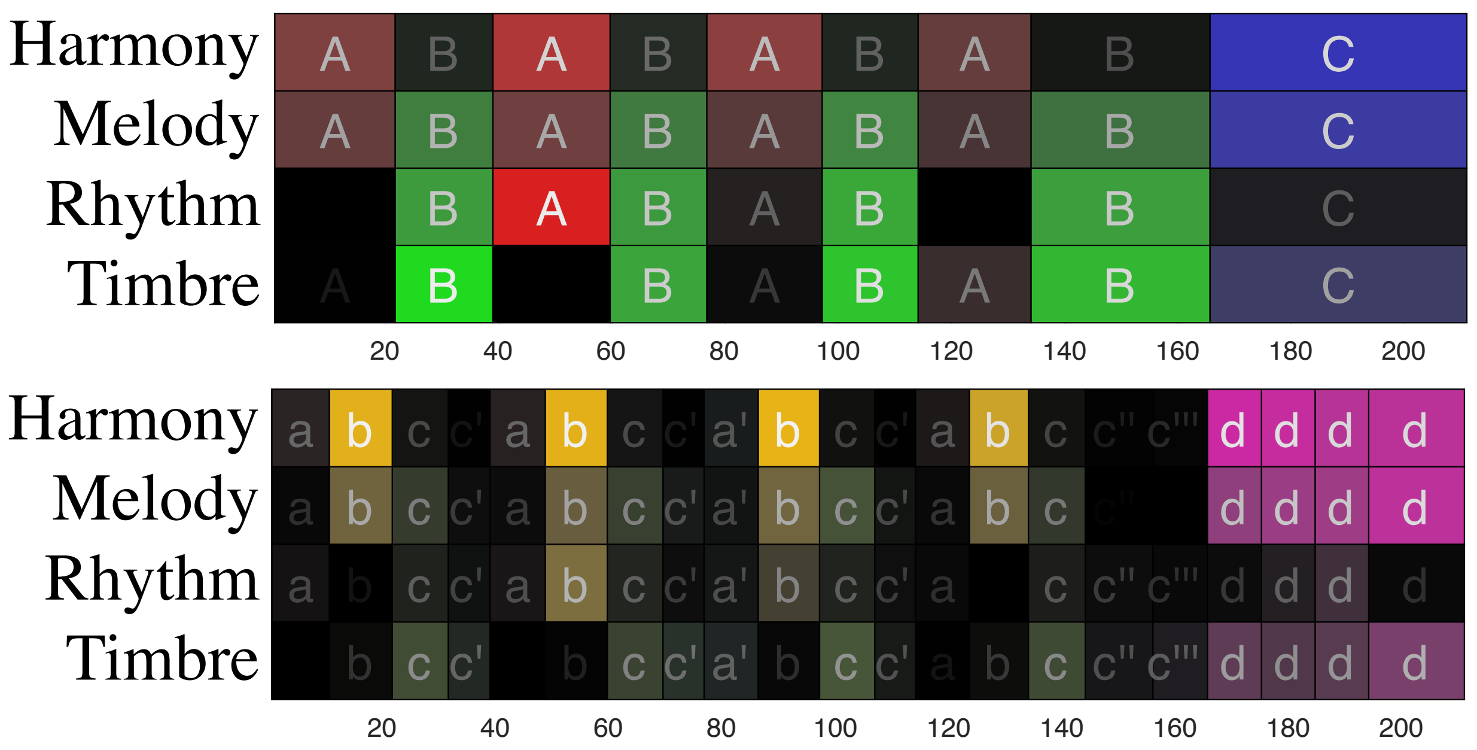
"Hello, Goodbye", by The Beatles

This annotation tells us that sections **A** and **B** are different—but what makes them different? Do listeners think **B** is defined by a harmonic or melodic progression, or by a timbre? What was the listener's **rationale**?

Collecting this information from listeners is onerous, and the introspection required is difficult. Instead, we aim to **automatically interpret existing annotations** by comparing them to the audio.

If successful, we could visualize structure to see which musical attributes characterize each section:

"Hello, Goodbye", by The Beatles



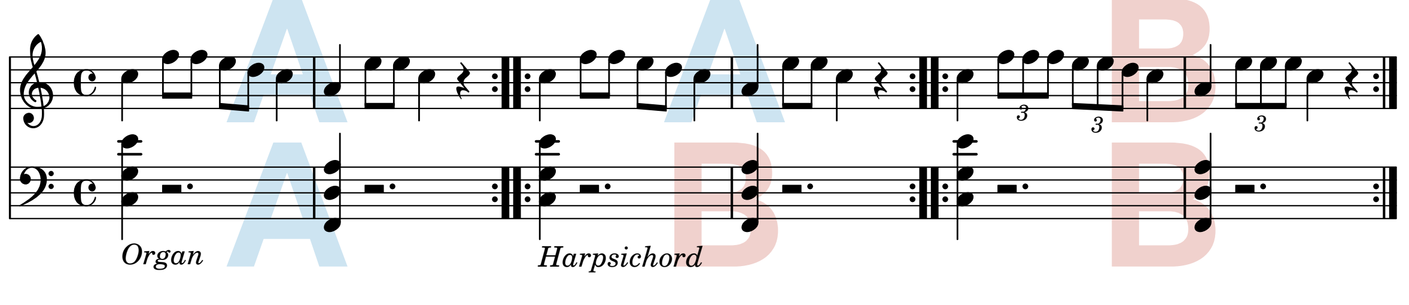
How to read: cells are brightest when a feature is:

1. homogenous throughout that section;
2. similar in other sections with the same label;
3. different in other sections with different labels.

2. Data

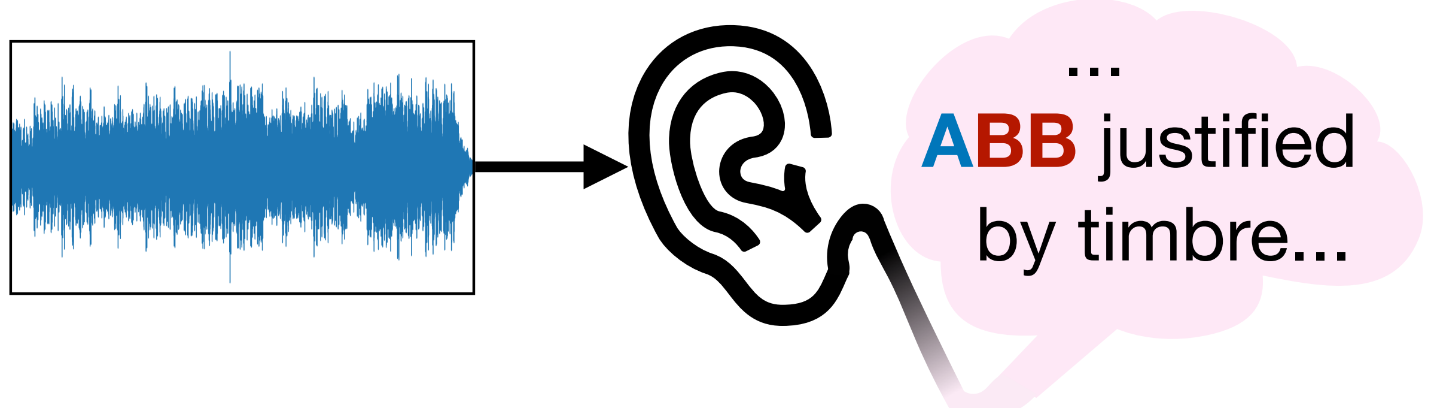
Finding appropriate data is not trivial! To validate the algorithm, we need structural annotations **paired** with listener rationales.

We obtained the data in a music perception study: we composed stimuli with *intended* forms, each suited to *intended* rationales:



AAB justified by rhythm
ABB justified by timbre

We also confirmed that listeners perceived these structure with the same rationales:



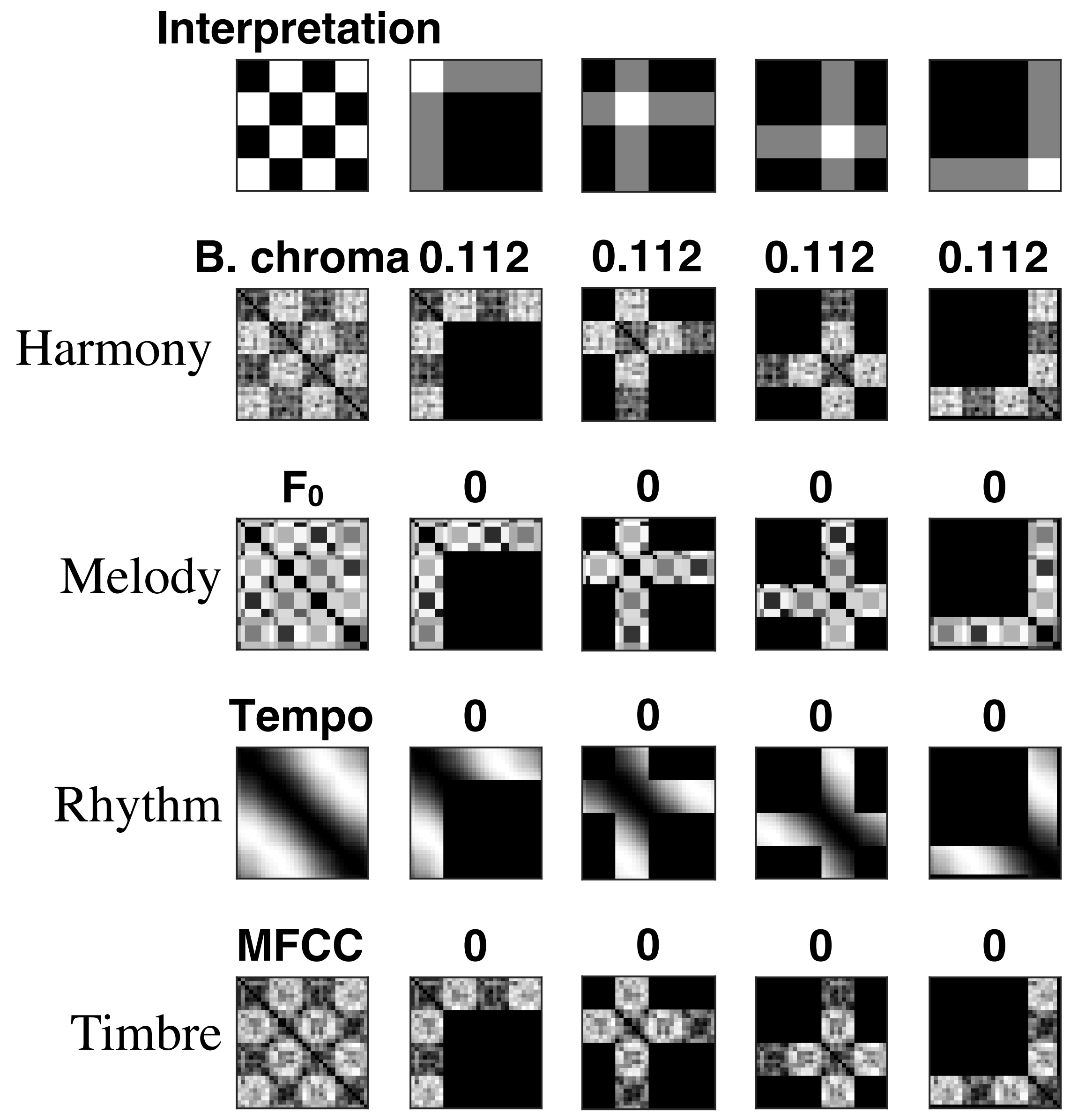
We have a large number of stimuli, in three styles, with either 3 parts (**AAB** vs. **ABB**) or 4 parts (**AABB** vs. **ABAB** vs. **ABBA**).

3. Algorithm

We compute self-similarity matrices (SSMs) from several **audio features**, each of which is *assumed* to correlate with a relevant **musical attribute**.

We generate **masked SSM segments**, each revealing the relationship of a segment to the rest of the piece.

Then, a **quadratic program** (QP) estimates coefficients to recreate the ground truth SSM from the masked segments. E.g.:

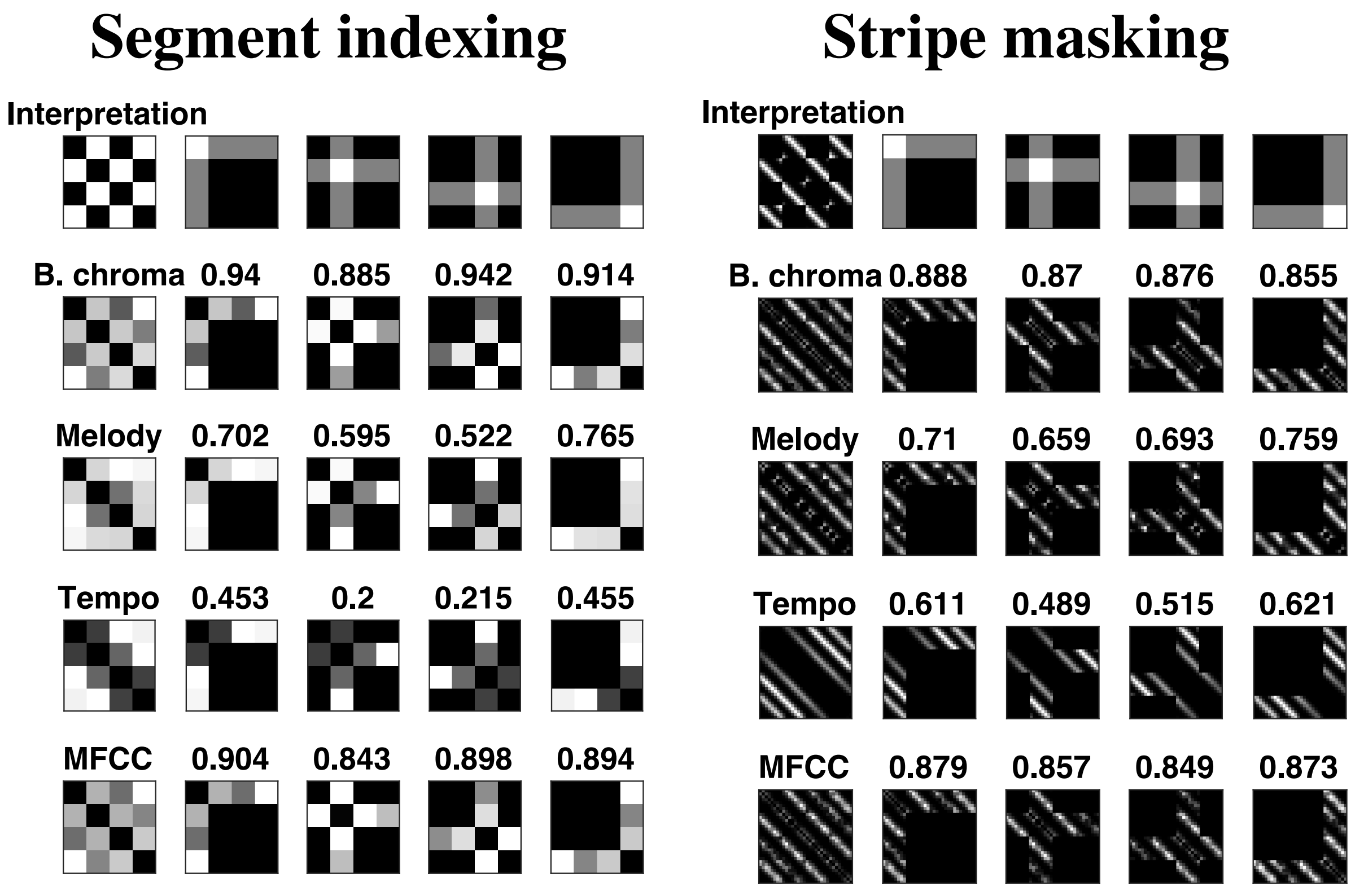


This piece has structure: **ABBA** justified by timbre
AABB justified by rhythm
ABAB justified by harmony

The QP reconstructs the **ABAB** interpretation using only bass chroma.

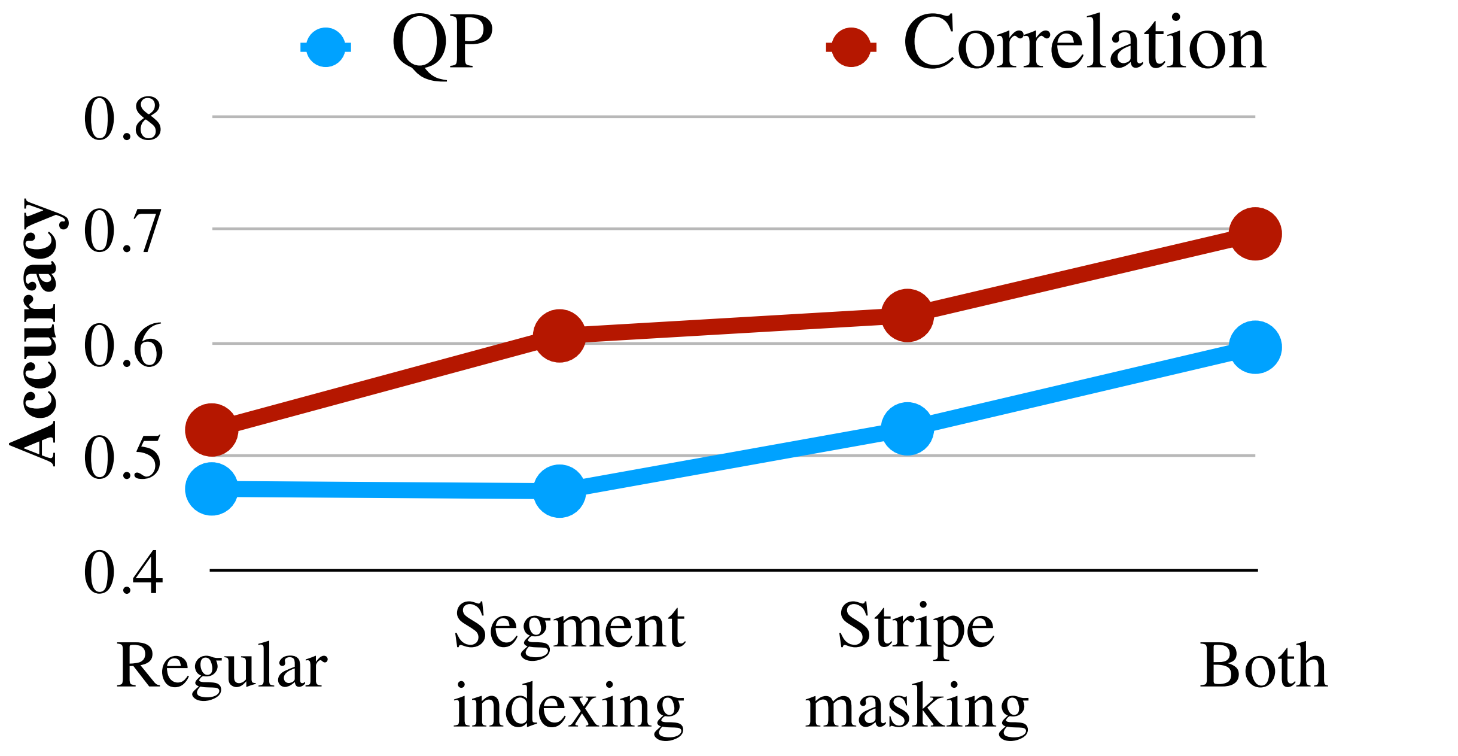
The QP approach has clear limitations:

- If two musical attributes explain a section equally, the QP might only point to one. Instead, we can measure **correlation**.
- Sequences that are repeated but non-homogenous may be overlooked in a point-wise SSM comparison. Instead, we can use **segment-indexed** SSMs, or apply additional **stripe masking**.


Segment indexing
Interpretation
B. chroma 0.94 0.885 0.942 0.914
Melody 0.702 0.595 0.522 0.765
Tempo 0.453 0.2 0.215 0.455
MFCC 0.904 0.843 0.898 0.894
Stripe masking
Interpretation
B. chroma 0.888 0.87 0.876 0.855
Melody 0.71 0.659 0.693 0.759
Tempo 0.611 0.489 0.515 0.621
MFCC 0.879 0.857 0.849 0.873

4. Validation

The suggested improvements all had a positive impact: the best algorithm used the stripe-masked SSMs, indexing by segment, and correlation instead of the QP output.



But accuracy varied among musical styles and features, as these confusion plots show:

	Style 1 H	Style 1 M	Style 1 R	Style 1 T		Style 2 H	Style 2 M	Style 2 R	Style 2 T	
B. chroma	**0**	13	5	0	0	**78**	0	0	63	.55
Chord	**384**	0	0	0	1	**256**	74	4	8	.75
Melody	0	**202**	10	39	.80	4	**238**	8	19	.88
T. chroma	0	**65**	1	0	.98	10	**2**	1	16	.07
Tempo	0	48	**235**	4	.82	25	17	**73**	171	.26
Onset	0	20	**133**	0	.87	1	0	**294**	25	.92
MFCC	0	23	0	**9**	.28	8	41	0	**58**	.54
Low level	0	13	0	**332**	.96	2	12	4	**24**	.57
	1	.70	.96	.89		.87	.63	.96	.21	

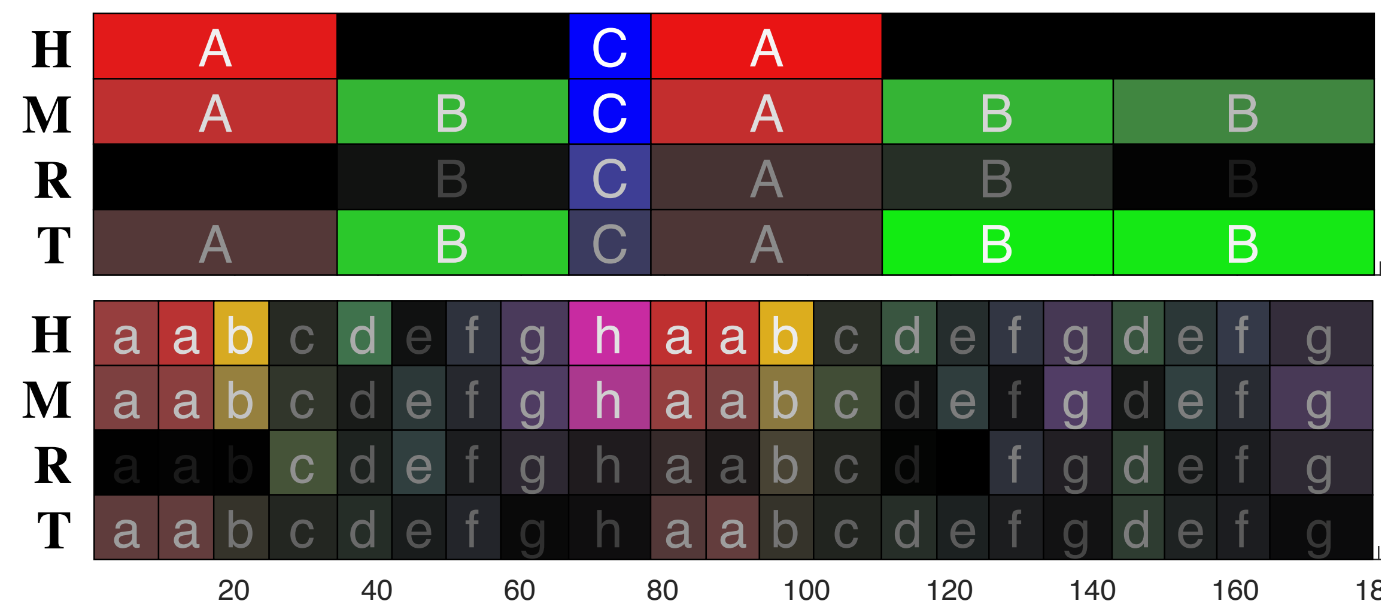
5. Application

We can use the validated approach to analyze SALAMI annotations:

"We Are The Champions", by Queen

A: Harmonies stable, orchestration builds up;
→ harmonies in **a** and **b** are unique across the piece.

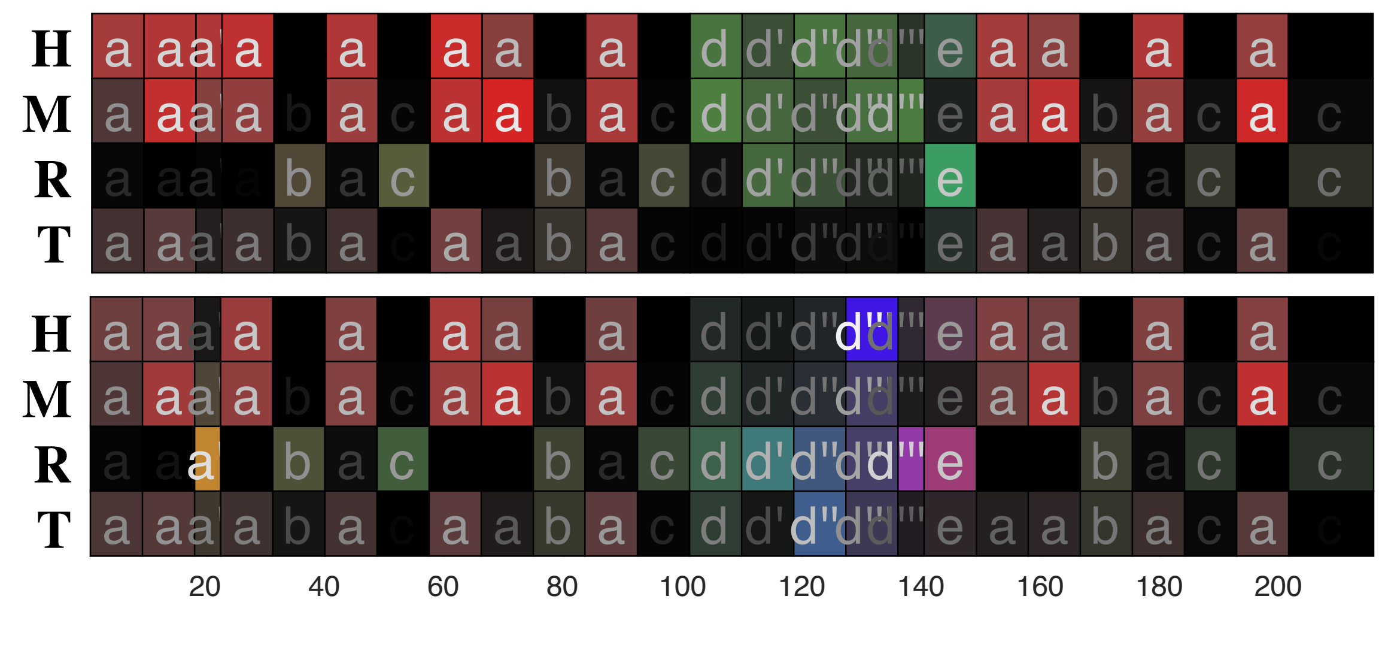
B: Complex chord sequence, stable timbre;
→ timbre cannot explain individuated subsections.



Some analyses have prime markers. If we consider primed sections to be similar or different changes the interpretation.

"Another One Bites The Dust", by Queen

d=d': Stable, stripped-down harmony throughout.
d≠d': Sections feature odd, varying sound effects.


H
M
R
T
a a a a b a c a a b a c d d' d'' d d''' e a a b a c a c
20
40
60
80
100
120
140
160
180
200