Validating an Optimisation Technique for Estimating the Focus of a Listener

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

Motivation:

Listeners perceive structure in music. There are two important processes:

- Segmentation (perception of boundaries between sections)
- Grouping (categorisation of sections)

The analysis a listener prefers is affected by which "feature" of the music (e.g., harmony, melody, rhythm, timbre) they focus on. However, there is no way to observe a listener's auditory focus.

Question:

Can we instead estimate what someone was paying attention to based on their analysis?

To answer this we need to do three things:

1. Propose an algorithm;
2. Obtain a dataset of listeners' analyses paired with what they were paying attention to;
3. Test whether the algorithm can correctly estimate a listener's focus from their analysis.

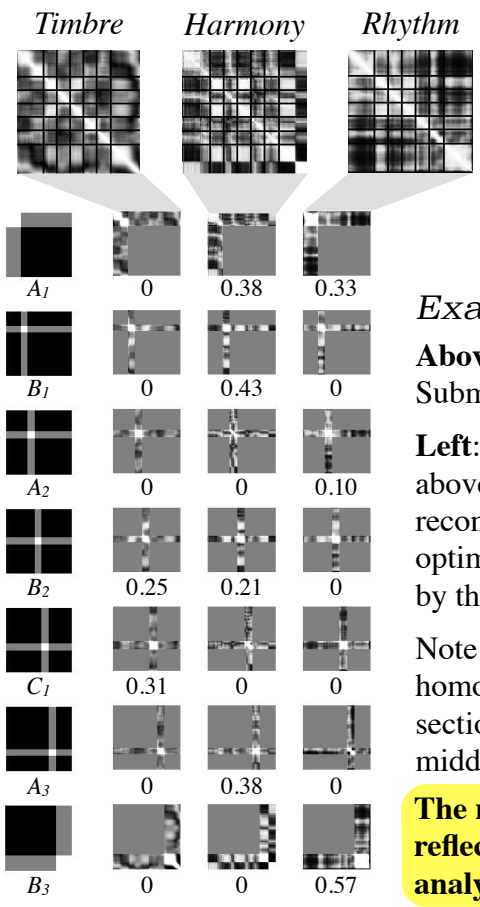
Part 1. An algorithm for estimating a listener's focus

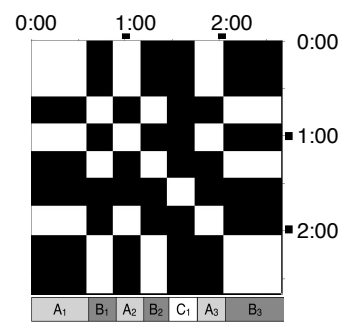
Goal:
By looking at a listener's structural analysis, estimate what musical features they focused on.

Premise:
Listeners focus on different musical features of a song at different parts. We want to find, for each part, which audio feature best justifies their analysis.

A self-similarity matrix (SSM) shows the similarity between all points in a song, revealing its structure. SSMs computed from audio features may strongly resemble the annotation at some parts, but not others.

Method:
1. Compute SSM for each audio feature;
2. Divide SSMs into components according to annotated boundaries;
3. Use quadratic programming (QP) to find the sum of SSM components that most closely resembles the annotation.





Example:

Above: Annotation-derived SSM for "Yellow Submarine" by the Beatles.

Left: Audio-derived SSMs for three features, above the SSM components used to reconstruct the annotation. Numbers are the optimal reconstruction coefficients estimated by the algorithm.

Note that the rhythm SSM captures the homogeneity of the opening and closing sections well, but the other features capture the middle sections better.

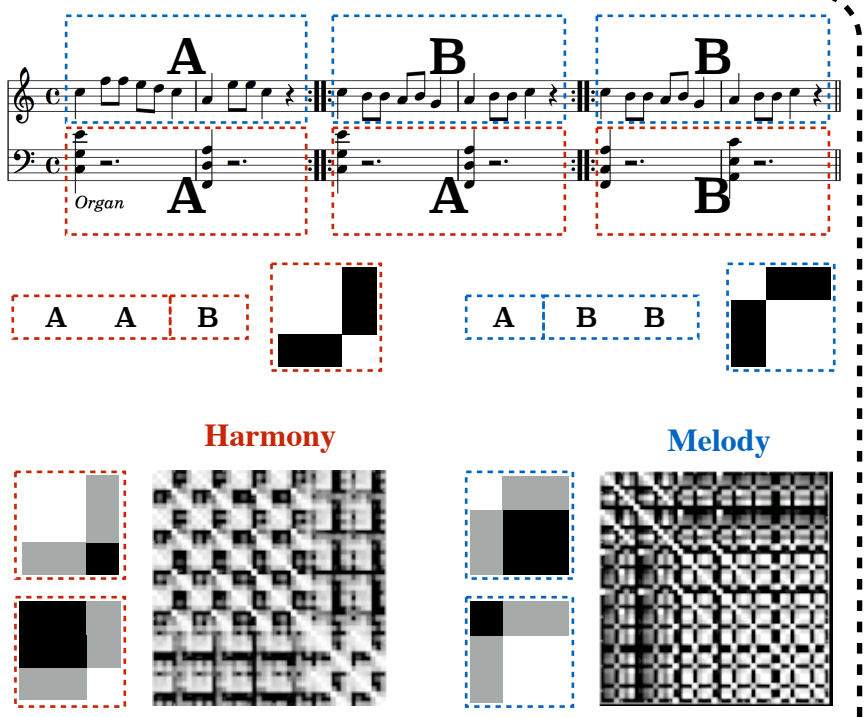
The reconstruction coefficients seem to reflect the importance of each feature to the analysis, but the algorithm needs validation.

Example

This music has ambiguous form:

Which audio feature best explains each analysis?

Pick an analysis, then divide each feature matrix into two components accordingly.



Harmony components match AAB *analysis best.*

For ABB, *melody is better.*

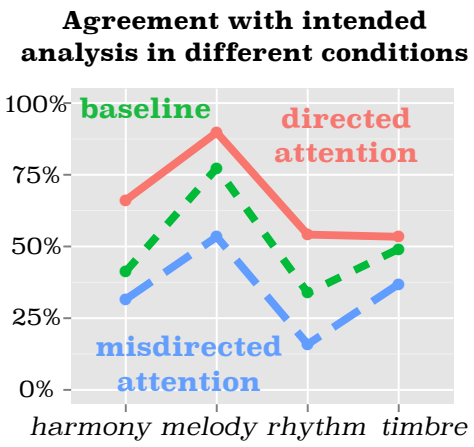
Part 2. A dataset of analyses and attentional targets

Goal:
Obtain a set of analyses of audio recordings where we know what the annotators focused on while listening.

Method:
We composed stimuli in which four features were manipulated: harmony, melody, rhythm and timbre. Each stimulus has form *AAB* with respect to one feature and *ABB* to another.

We composed 3 separate sets; in each, some features are "convolved" since they are varied within the same voice. For example, in the "HT-MR" set, there are four melodies (2 contours *x* 2 rhythms) and 4 harmonies (2 chord progressions *x* 2 timbres).

Experiment:
Participants heard the ambiguous excerpts and indicated the analysis they preferred: *AAB* or *ABB*. But first, we steered their attention with a distractor task, asking if a pattern (e.g., a melody, rhythm or chord progression) existed in the excerpt.



Above: Chance that a listener's analysis agreed with a given feature when their attention was directed toward or away from it, or not manipulated at all.

We found that listeners' analyses aligned with the target of their attention more often than chance (65% instead of 50%). Thus, attention affects the perception of grouping.

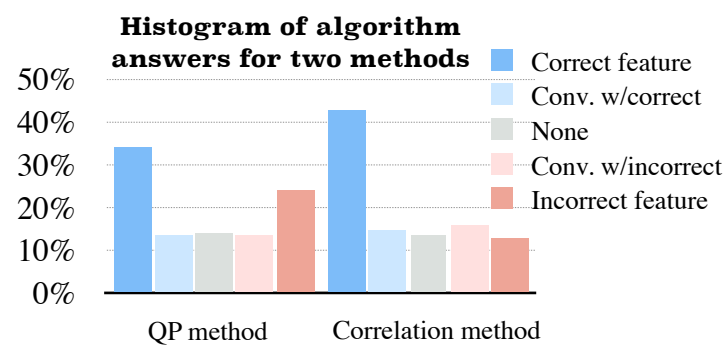
Part 3. Testing the algorithm

Goal:
Use the algorithm from Part 1 to predict the focus of the listeners in Part 2 based on their preference for *AAB* or *ABB*.

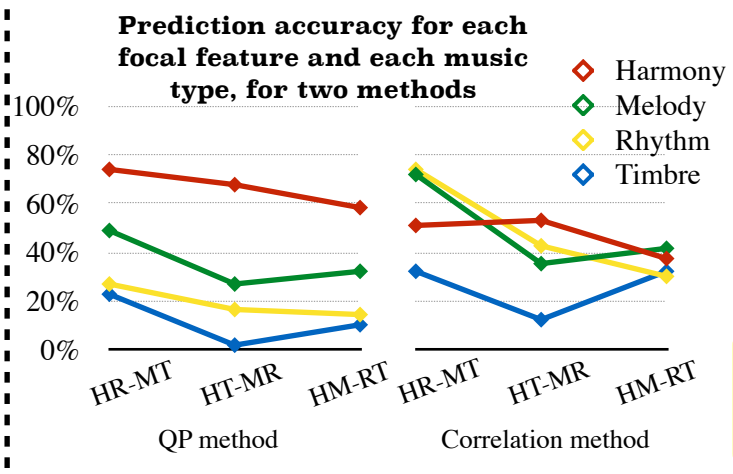
Method:
For each stimulus, we computed 8 standard audio features (2 per musical feature). We ran the QP algorithm on each stimulus twice, to consider both possible analyses (*AAB* and *ABB*).

When analysing the example in the middle with structure *AAB*, harmony is the "correct feature" and melody the "incorrect feature"; each of these is convolved with another feature, or with each other.

Q: Does the largest coefficient output by the algorithm identify the focal feature?
Yes, but not very strongly: it is correct in 33% of trials, above the chance level of 25%. (See histogram at left.)



Q: Does the prediction accuracy vary among the features?
A great deal! The lowest accuracy was for timbre, the greatest for harmony. Clearly, the audio features vary in their fidelity to the musical features.



Q: Do other, simpler methods work better?
It turns out, yes! Instead of using a QP approach, we can obtain relevance estimates by taking the correlation between the SSM components and the annotation. Using this approach improves overall performance from 33% to 40%.

Our method seems to work, but only modestly better than chance. Simple tweaks have yielded better results, so there is much room for improvement!

We are grateful for our support from:
Fonds de recherche Société et culture Québec
Queen Mary University of London
centre for digital music