A Perceptually Based Evaluation of Music Boundaries

Oriol Nieto
Morwaread M. Farbood
Juan P. Bello

Toronto, Canada
August 7th, 2013

MARL
NYU Music and Audio Research Laboratory

CogMIR™

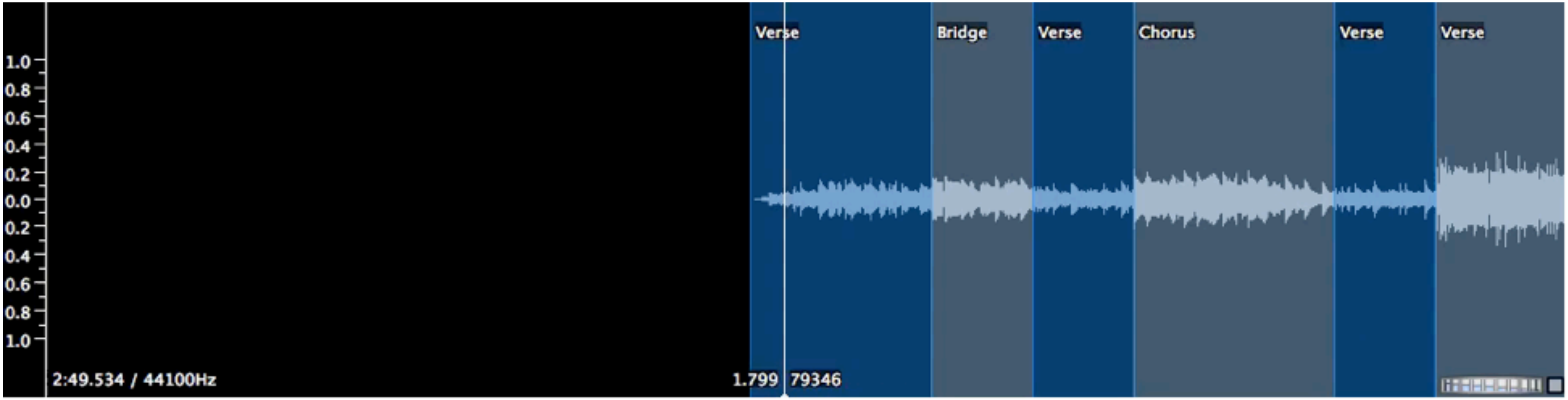
Overview

- Music Structure Analysis in MIR
- F-measure for Boundary Evaluation
- Experiments
- Redefining the F-measure
- Results and conclusions

MARL

Music Structure Analysis Overview

- Automatically get this:



(Trains by Porcupine Tree)

MARL

Music Structure Analysis Evaluation

- Typically, compare estimated results against ground truth annotations (e.g. The Beatles dataset, SALAMI dataset).
- Make use of the F-measure (or F1-score):
  - Quantizes the similarity between the annotations and the estimated results.
  - Commonly used to evaluate machine learning algorithms (e.g. search, document classification).
  - Is it appropriate in the framework of music structural analysis? Does it align with humans' perception of the structure in music?
- In this work we aim to perceptually redefine the F-measure for evaluating **music boundaries**.

MARL

F-measure for Boundary Evaluation

- Find intersection between annotations and estimated results:
  - Estimated boundaries are correct (hits) if they are within 3 seconds from the annotated one.
- **Precision**: Ratio between hits and the total number of estimated elements.
- **Recall**: Ratio between hits and the total number of annotated elements.

$$P = \frac{|\text{hits}|}{|\text{bounds}_e|} \qquad R = \frac{|\text{hits}|}{|\text{bounds}_a|}$$

- **F-measure**: Harmonic mean between P and R.
  - Weights both values equally.
  - Penalizes outliers.
  - Mitigates impact of large values.

$$\boxed{F = 2\frac{P \cdot R}{P + R}}$$

MARL

F-measure for Boundary Evaluation

- Higher Precision represents less false positives.
- Higher Recall represents less false negatives.
- When listening to estimated results of music structural analysis, it becomes apparent that these two values are perceptually very different.
- We decided to assess the relative effect that these differences had on human evaluations in order to redefine the F-measure.
  - Two Experiments

MARL

Experiment I

- Goal:
  - Investigate whether Precision or Recall is more perceptually relevant than the other.
  - Ensure that findings are robust across a relatively large set of subjects.
- Design:
  - Obtain track excerpts from the Levy dataset (Levy & Sandler, 2008) by finding the 1-minute segments containing the highest amount of boundaries.
  - Reduce the time of the experiment while maintaining participants' attention as high as possible.

MARL

Experiment 1

- Track Selection:
  - 5 tracks from the Levy dataset.
  - For each track excerpt, we synthesized 3 different segmentations:
    - Ground-truth boundaries (i.e. F-measure 100%)
    - High Precision (HP): Precision of 100% and Recall of ~65%
    - High Recall (HR): Recall of 100% and Precision of ~65%

Experiment 1 Excerpt List						
Song Name (Artist)	HP			HR		
	F	P	R	F	P	R
Black & White (Michael Jackson)	.809	1	.68	.794	.658	1
Drive (R.E.M.)	.785	1	.647	.791	.654	1
Intergalactic (Beastie Boys)	.764	1	.619	.792	.656	1
Suds And Soda (Deus)	.782	1	.653	.8	.666	1
Tubthumping (Chumbawamba)	.744	1	.593	.794	.659	1
Average	.777	1	.636	.794	.659	1

MARL

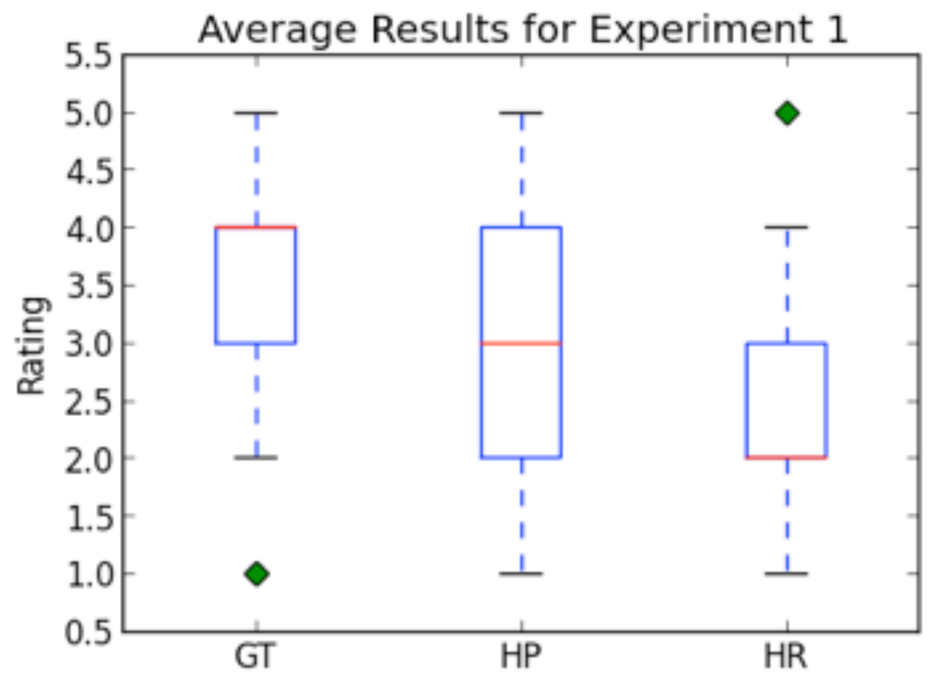
Experiment 1- Subjects

- Choose accuracy of each version of the excerpts.
- Rate them with a discrete value from 1 to 5.
- Total time of the experiment: **15 minutes** = 5 tracks x 3 excerpts/track x 1 minute/track
- A total of **48 subjects** took the experiment.

MARL

Experiment 1 - Results

- Higher accuracy ratings were assigned to Ground Truth, then HP, and finally HR.



- ANOVA was performed on the accuracy of the ratings with type (GT, HP, HR).
  - Main effect of type [$F(2,94)=90.74$, $p<.001$] is significant.

Experiment II

- Goal/Main Inquiries:
  - Experiment 1 shows the relative importance of Precision over Recall.
  - Can the F-measure, Precision and Recall predict subject's preference?
  - How this information can be used to design a perceptually-relevant evaluation metric?

- Design:
  - Obtain track excerpts sampled from a larger dataset.
  - Use three state-of-the-art algorithms:
    - Structural Features (Serrà et al. 2012): Best reported results.
    - Convex NMF (Nieto & Jehan 2013): Tends to oversegment.
    - SI-PLCA (Weiss and Bello 2011): Tends to undersegment (depending on params).

MARL

Experiment II - Excerpts

- Sampled from the union of three datasets:
  - The Beatles TUT dataset
  - Levy catalogue
  - SALAMI dataset (only the freely available on-line)
  - Total of 463 tracks.

- Run the three algorithms on the 463 tracks, and filter as follows:
  - (i) There are at least 2 algorithms that have similar F-measure (with a max 5% diff)
  - (ii) F-measure must be at least 45%
  - (iii) There is at least 10% difference between P and R.

MARL

Experiment II -Excerpts

- 41 out of 463 tracks met these criteria. Qualitatively select 20 (e.g. some SALAMI recordings have very poor sound quality).
- We kept both algorithmic outputs maximizing the difference between P and R to create two versions for each track: High Precision (HP) and High Recall (HR).
- The F-measure was almost identical for both versions.

Boundaries Version	F	P	R
HP	.65	.82	.56
HR	.65	.54	.83

MARL

Experiment II - Subjects

- Each subject was presented with 5 excerpts randomly selected from the 20.
- Each subject had to choose the version they found most accurate (instead of discretely measuring accuracy).
- Total of **23** participants took the experiment.

MARL

Experiment II - Results

- **72.88%** of the times subjects chose HP version over HR.

- Binary logistic regression analysis on the results in order to understand what values of the F-measure are useful to predict participants' preferences.

- Used three predictors:
  - The F-measure
  - The difference between P and R
  - The absolute difference between P and R

- As the table shows, the P-R is the only value that can predict participants' preferences in a significantly statistical way.

Logistic Regression Analysis of Experiment 2						
Predictor	β	S.E. β	Wald's χ^2	df	p	e^β
F-measure	-.012	1.155	.000	1	.992	.988
$P - R$	2.268	.471	23.226	1	.000	1.023
$\|P - R\|$	-.669	.951	.495	1	.482	.512
k	.190	.838	.051	1	.821	1.209

MARL

Perceptually Redefining the F-measure

- The generic form of the F-measure is:

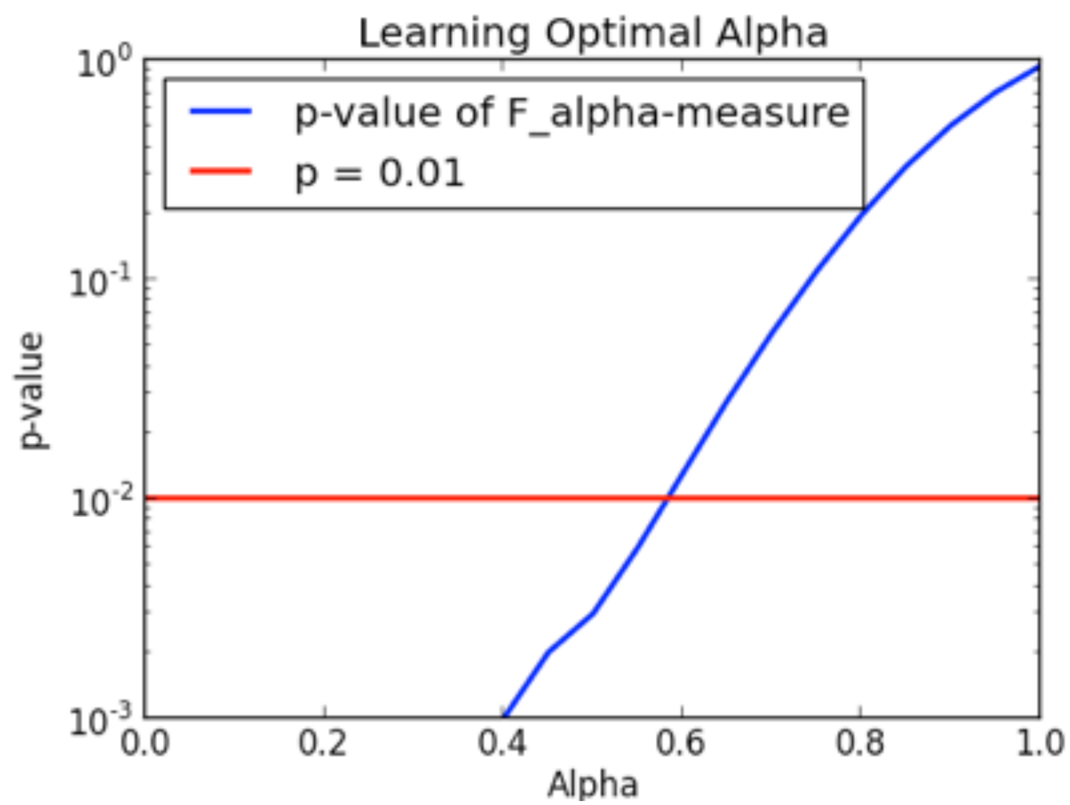
$$F_\alpha = (1 + \alpha^2)\frac{P \cdot R}{\alpha^2 P + R}$$

- If alpha = 1: R and P have the same weight (F1-score)
- If alpha > 1: more importance to R
- If alpha < 1: more importance to P

MARL

Perceptually Redefining the F-measure

- To learn alpha, we sweep alpha from 0 to 1 (hop size of 0.05), and compute the binary logistic regression analysis at every step. We pick the alpha that obtains statistically significant results (p=0.01)



- This results in **alpha=0.58**
- If we apply this alpha to evaluate our experiments, the new F-measure aligns well with participants' preferences.

MARL

Conclusions and Future Work

- Precision tends to have more perceptual relevance than Recall when evaluating music boundaries.
- Experiment 2 showed that the difference between P and R has a high predictive power to perceptually choose a set of boundaries.
- We have proposed a new method to evaluate music boundaries that better aligns with subject preferences.
- As P-R increases, we expect participants' preferences to decrease. A new experiment should be performed to confirm this.

MARL

Thank you!

MARL

References

- Levy, M., & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 318–326. doi:10.1109/TASL.2007.910781
- Nieto, O., & Jehan, T. (2013). Convex Non-Negative Matrix Factorization For Automatic Music Structure Identification. In Proc. of the 38th IEEE International Conference on Acoustics Speech and Signal Processing (pp. 236–240). Vancouver, Canada.
- Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2012). Unsupervised Detection of Music Boundaries by Time Series Structure Features. In Proc. of the 26th AAAI Conference on Artificial Intelligence (pp. 1613–1619). Toronto, Canada.
- Smith, J. B., Burgoyne, J. A., Fujinaga, I., De Roure, D., & Downie, J. S. (2011). Design and Creation of a Large-Scale Database of Structural Annotations. In Proc. of the 12th International Society of Music Information Retrieval (pp. 555–560). Miami, FL, USA.
- Weiss, R., & Bello, J. P. (2011). Unsupervised Discovery of Temporal Structure in Music. IEEE Journal of Selected Topics in Signal Processing, 5(6), 1240–1251.

MARL