Perceptual Analysis of the F-Measure for Evaluating Section Boundaries in Music

Oriol Nieto[1], Morwaread M. Farbood[1], Tristan Jehan[2], Juan P. Bello[1]
[1]Music and Audio Research Laboratory, New York University
[2]The Echo Nest / Spotify

{oriol, mfarbood, jpbello}@nyu.edu
tristan@echonest.com

Abstract

We aim to raise awareness of limitations of F-measure when evaluating segment boundaries:

- Multiple experiments with humans subjects to assess perceptual preferences.
- Results: *Precision* value of the F-measure is regarded as more relevant than the *Recall* value when the F-measure is sufficiently high.
- We propose an alternative evaluation to emphasize *Precision*.

F-measure for Boundaries

Standard metric to evaluate boundaries:

- ***hit*** is found every time an estimated boundary falls within a time window ***t*** from a reference one.
- Compute *Precision* and *Recall* based on |hits|.
- Formally:

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

F-measure **equally** weights *P* and *R*:

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

Preliminary Study

Choose the best estimated boundaries from three algorithms **A**, **B**, and **C**.

Results of the algorithms on the Levy dataset (***t***=3):

Algorithm	F	P	R
$\mathcal{A}$	49%	57%	47%
$\mathcal{B}$	44%	46%	46%
$\mathcal{C}$	51%	47%	64%

2 subjects: 68% chose the same algorithm.

58.5% chose **A**.

14.6% chose **C**.

A has the highest *P*, and **C** the highest *R*.

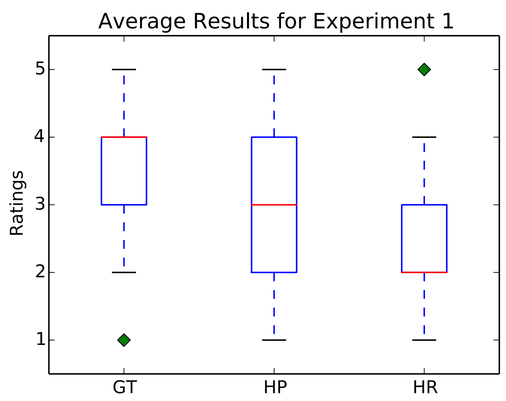
Experiment I

Explore hypothesis that *P* is perceptually more relevant than *R*.

- Choose five one-minute excerpts in the Levy dataset with highest number of boundaries.
- For each excerpt we synthesize three sets of boundaries,: high *P* (HP), high *R* (HR), and Ground-Truth (GT) (***t***=3).
- Subjects rated the "quality" of the boundaries for each version of the five tracks by choosing a discrete value between 1 and 5.

Results suggest perceptual preference for HP over HR.

- N = 48 participants.
- 2-way ANOVA on ratings using **type** and **excerpt** as effects.
- Effect on type of boundaries is significant: $F(2, 94) = 90.74$, $p < .001$.



Experiment II

Investigate actual algorithm outputs using a larger music collection:

- Run C-NMF, SI-PLCA, and SF algorithms over dataset of 463 tracks (Levy, Beatles, and free-SALAMI).
- Select 20 songs that have two estimated sets of boundaries (HP and HR) with similar F-measures and over 10% difference between *P* and *R* (***t***=3).

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

23 Subjects listened to both versions and chose the "best" one:

- **67% of the time subjects chose HP, 33% chose HR.**
- Logistic Regression to predict results based on numerous factors:

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
$\lvert P - R \rvert$	-.669	.951	.495	1	.482	.512
k	.190	.838	.051	1	.821	1.209

- F-measure does not sufficiently characterize the boundaries perception.
- ***P* more important than *R*** (its difference P-R can predict preference).

Enhancing the F-measure

Generic form of the F-measure:

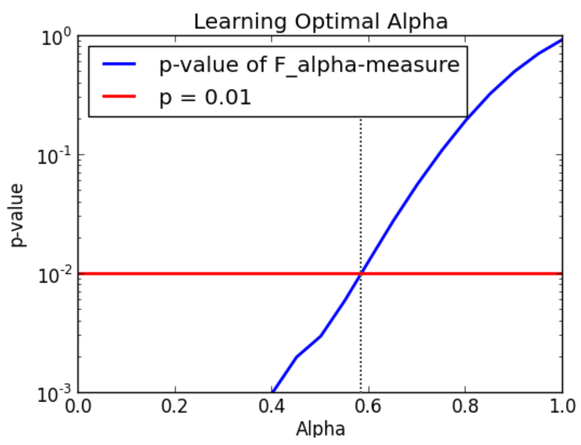
$$F_\alpha = (1 + \alpha^2)\frac{P \cdot R}{\alpha^2 P + R}$$

Where if:

- $\alpha = 1$ same weight for *P* and *R*.
- $\alpha > 1$ more weight for *R*.
- $\alpha < 1$ more weight for *P*.

Sweeping α and running logistic regression as in Exp2, for P - R we can estimated approximate value:

- $\alpha = 0.58$



Discussion

Our experiments suggest that *P* is perceptually more relevant than *R*. However:

- Need more data (more participants and larger dataset) in order to find a more generic α.
- What happens when the difference of P - R is too large?
  - (What happens when the F-measure is sufficiently low?).
- Experiments might be biased towards small variations of the sonified boundaries, therefore relying on the existing boundaries instead of the non-existing ones that were not extracted.
- A better way to evaluate boundaries: use a "salience" value associated to each boundary?
- Or to have a gaussian window centered in the boundary of size ***t*** in order to incorporate a weight?
  - i.e. an estimated boundary right on top of the reference boundary is a full hit, whereas a boundary deviated 0.5 seconds is less weighted hit.

NEW YORK UNIVERSITY

Experiments Data: www.urinieto.com/NYU/ISMIR14-BoundariesExperiment.zip