

# 7 Steps on the Stairway to Heaven

or

## How to evaluate algorithms for ambiguous MIR tasks

(and some examples from a chord labelling task)

Daniel Müllensiefen, David Lewis, Christophe Rhodes, Geraint Wiggins

ISMS Group, Department of Computing, Goldsmiths, University of London

### Background: Necessity of comprehensive evaluation

#### The common case: Evaluation against a single Ground Truth

- Use data set with finite number of items, each with a given value of *target*
- All *target* values are defined (assumed?) to be correct ('true')
- Count the number of items in the data set for which your algorithm can predict 'equivalent' values for *target*. ('equivalence' is a task-dependant concept).
- Use a measure based on the number of correctly predicted items as an indicator of algorithm performance

#### Common problems with single Ground Truth evaluation as only evaluation device

- Does not allow for multiple values of *target* (for same item)
- Does not usually consider item difficulty and item importance
- Does not consider item and task ambiguity
- Assumes metric for assessing (degree of) correctness of item prediction
- Does not question 'correctness' of Ground Truth data set

#### Relevance for MIR work

- Single Ground Truth based evaluation is highly desirable - less time intensive and much clearer than other methods.
- **But** Many popular MIR tasks are inherently ambiguous - two experts can validly disagree on value of *target*.
- Inherently ambiguous tasks include: Chord labelling, Genre classification, Similarity computation, Chorus finding, Segmentation, Key finding, Cover and remix detection, Mood classification

#### Ambiguous tasks: a more comprehensive approach

- Looks at algorithmic results from different perspectives
- Assesses task ambiguity
- Positions algorithm in environment of alternative solutions

### General Case

#### 1. Define task and output

##### Task:

- Is task something that humans can do?
- Why do humans do this task?
- What human skill/training is required to do the task?
- What are the advantages of a machine doing the task?
- Is the task part of an application?
- Are there critical performance limits for usefulness of algorithm?

##### Output:

- What is the appropriate scale / alphabet for target?
- Is an indicator of certainty computed along with the values of target variable?
- Should algorithm be considered cognitively adequate, i.e.
  - should output mirror human behaviour (including errors)?
  - should algorithm be fed same information that humans use?
  - should human and algorithmic processing time be related?

#### 2. Qualitative Analysis

##### Check the paradigmatic case

- Does algorithm give right answer in very easy cases?
- Does algorithm give right answer in very common situation?

##### Check the problematic case

- Are answers acceptable in difficult situations?
- Where are weaknesses? Where does the algorithm give nonsense answers?

#### 3. Determine task ambiguity (qualitatively)

Are there cases with more than 1 possible and acceptable answer?

Which factors enhance / diminish ambiguity?

#### 4. Quantify task ambiguity

Measure coherence between multiple GT data sets or human judges

=> If ambiguity is small enough, single GT testing is justified

Measure item difficulty or item-wise level of disagreement

#### 5. Test on Ground Truth data set

Test against single Ground Truth data set (the common case)

Test against multiple Ground Truth data sets

- Does algorithm match any human solution?
- Is number and pattern of algorithmic deviations significantly different from disagreement between humans?
- How well performs algorithm above baseline of overlap between multiple Ground Truth data sets?

#### 6. Compare to alternative solutions

Test rivaling algorithms using different approaches on same Ground Truth data set

What is performance level of trivial models, e.g. all values from most frequent class?

Is copying existing data an alternative (e.g. web spidering)? How good is this data compared to Ground Truth data?

#### 7. Let output be judged

Present algorithmic output to competent judges

Use output to predict data of psychological test (if cognitive adequacy is a desired feature)

Use algorithm in application and measure performance of application

### Specific Case: Chord Labelling

#### 1. Define task and output

##### Task:

- Ear-trained humans can perform chord labeling from audio and symbolic data.
- Chord labels are produced for play-along sheet music and analytical purposes.
- Humans are slow and for large amounts of music a computer is needed.
- The main goal is to find time windows of constant harmonic content and to summarise the content as a chord label including chord root, triad, bass, and extensions.
- The application is the identification of recurring harmonic patterns in pop music.
- For harmonic pattern recognition, chords need not to be labeled correctly, but similar harmonic situations need to get the same chord label.

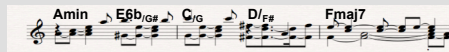
##### Output:

- The chord symbol alphabet should be comparable to those in more sophisticated song books.
- The algorithm (Rhodes et al., 2007) uses pitch class distributions over time windows derived from MIDI transcriptions.
- Humans use pitch class distributions as only one information source.

#### 2. Qualitative Analysis

##### Check the paradigmatic case

Detection of changes of simple chords at half and whole bars in *Stairway to Heaven*, bars 1-3



##### Check the problematic case

Vaguely meaningful labeling of opening bars of *California Dreaming* as sus4 and minor chords



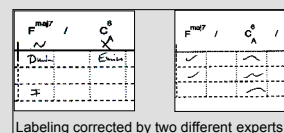
#### 3. Determine task ambiguity (qualitatively)

##### Ambiguous cases from music theory, e.g.:

- Diminished chords
- 6-4 chords
- Chords with identical pitch class sets

##### Factors influencing ambiguity:

- Knowledge about harmonic context
- Stylistic information
- Knowledge of harmony instruments



#### 4. Quantify task ambiguity

Coherence between 4 judges, measured by corrections of chord labels of 40 pop song excerpts

- Number of identical corrections applied in same place
- Number of corrections applied in same place
- Correlation between number of corrections over songs

Option: Exclusion of difficult items, i.e. items with strong disagreement regarding corrections

#### 5. Test on Ground Truth data set

Test against single Ground Truth data set

- 80% of 1178 beats from 12 songs with chord root and type identical to solution provided by 1 expert (Rhodes et al. 2007)

Test against multiple GTs

not yet assessed

#### 6. Compare to alternative solutions

Test with Temperley's (2001) chord labeling algorithm from Melisma package on same single GT

- 77% of beats correctly labeled
- Accuracy level of all chord labels being tonic triad of key on same single GT
- 34% of beats correctly labeled
- Accuracy of chord labels taken from official song books for 40 pop song excerpts
- 87% of beats with at least acceptable labels (partial result)

#### 7. Let output be judged

Chord labels of 40 song excerpts judged by experts as 'correct', 'acceptable' or 'wrong' (part. result)

- 77% of beats with correct chord labels
- 83% of beats with at least acceptable labels
- 82% of beats with correct bass note
- 86% of beats with correct chord extensions