EVALUATING A CHORD-LABELLING ALGORITHM

Daniel Müllensiefen, David Lewis, Christophe Rhodes, Geraint Wiggins
Department of Computing, Goldsmiths, University of London, London, SE14 6NW
{d.mullensiefen,d.lewis,c.rhodes,g.wiggins}@gold.ac.uk

ABSTRACT

This paper outlines a method for evaluating a new chord-labelling algorithm using symbolic data as input. Excerpts from full-score transcriptions of 40 pop songs are used. The accuracy of the algorithm's output is compared with that of chord labels from published song books, as assessed by experts in pop music theory. We are interested not only in the accuracy of the two sets of labels but also in the question of potential harmonic ambiguity as reflected the judges' assessments. We focus, in this short paper, on outlining the general approach of this research project.

1 INTRODUCTION AND BACKGROUND

Our motivation comes from the need to derive sequences of chord labels from transcriptions of pop songs for a current project [1] hosted at Goldsmiths College, one subtask of which is to provide a summary of the harmonic structure of a song in the form of a sequence of chord labels, of the sort used in lead sheet notation. Several algorithms have been proposed that assign chord labels to points in time, based on note events in a score-like data structure (see *e.g.* [6]), but none of these algorithms proved fully suited for our purpose. We require the algorithm not only to give the chord root and chord type, functional bass note and extensions for the note events in a time window but also to decide on the optimal width of the time window itself and, furthermore, deal with music where the structures of classical harmony may apply to only a limited extent.

We have proposed [5] using Bayesian model selection to tackle segmentation into appropriate time windows and chord label assignment simultaneously. An initial evaluation using manually-generated ground truth showed an accuracy of around 75% for root and type of the chord at each beat of the test set. This preliminary evaluation raised some concerns and questions that motivated this paper; chief among these was the way in which the ambiguity of the task is not considered in ground-truth-based evaluation.

[1] http://doc.gold.ac.uk/isms/m4s

1.1 Inherent ambiguity of the chord-labelling task

From the classical and pop music analysis literature [4, 1, 3] there are several well-known cases where harmonic ambiguity has no simple solution and the choice of chord label depends on what the analyst wishes to convey. These sources of ambiguity include: the ambiguities in chord root assignment; the incompatibility of a musical style with classical harmony; and the relative autonomy of bass note and harmony. They pose a challenge for the approach of evaluating an algorithm against a ground truth that strictly allows only for a single correct chord label assigned to a particular set of pitch classes. In many situations there is no single correct answer, or there may be several acceptable ones. This is true not only for chord labelling but also for other tasks in MIR, such as genre classification, song segmentation or chorus finding.

We try to answer two questions regarding the evaluation of ambiguous data. Firstly, to what degree do experts agree (or disagree) about a given chord-labelling solution? If there is high agreement among human experts that a chord label is wrong and they propose identical or similar corrections, this may be taken to indicate that there is generally one correct solution to chord labelling. In this case, the traditional approach of working with definitive, context-free ground-truth data can be justified. The second question relates to the performance of our labelling: does the agreement between the labelling and the human experts differ significantly from the agreement between the experts? Or, in other words, are the computer-generated labels significantly worse than the baseline as given by the experts' responses?

2 METHOD

2.1 The chord labelling algorithm

Our chord-labelling algorithm consists of three modules that determine chord-type and root, functional bass note, and chord extensions. The core module is the Bayesian chord-type and root model that also decides on the appropriate window size for labelling. The window size is then fed as an input to the bass note and extensions modules. The details of the novel Bayesian core module are described in [5]; it consists of two essential parts. The first is a class of models for pitch-class contributions to a window given a triadic chord (we currently consider the chord types major, minor, sus4, sus9, augmented and

diminished), modeled using the Dirichlet distribution for proportions. The second component of the chord root and type labelling scheme decides what regions to treat as a unified whole. For this, we use Bayesian Model Selection (see *e.g.* [2]), and currently consider all possible beat-wise subdivisions of a bar. For determining the functional bass note of a time window we use a rule system that generally favours longer and more prominent pitch classes sounding in the lower register. As our core model currently only takes a set of pitch classes and no voice-leading information into account, we restrict ourselves to labelling extensions as notes that are not part of the model-derived triad and that have a significant duration in the chord.

2.2 The evaluation method

We obtained detailed feedback from four highly-qualified experts on chord labellings of excerpts of 40 pop songs as labelled for song books and by our algorithm; each expert is an academic musician with substantial experience in score reading and pop harmony analysis. Their task was to compare chord labels in a lead sheet-like representation with a score of the song (taken from MIDI, via Sibelius). They were asked to indicate whether the chord labelling was correct, for each beat of each bar on the lead sheet, evaluating separately the correctness of chord symbol (chord-type and root), bass note, and chord extensions.

Audio realisations of the songs were provided to the experts on CD, and ten excerpts (different for each expert) were provided without scores. In these cases, the experts were instructed to perform their assessment of the chord labellings by ear only.

2.3 Coherence of experts' judgements

Our evaluation can be divided into two steps that correspond to the two main questions discussed in 1.1 above. As there are no time constraints or other sources of distraction in the task, and as we believe all four experts to be sufficiently qualified, we treat all of their responses as necessarily true data and not as approximate judgements with a measurement error attached. As a performance measure, we use the relative number of beats for which an expert disagrees with the labelling on the leadsheet (the beat error rate). We assess the coherence in three ways, answering slightly different questions. In an initial parametric test we ask whether the experts made roughly the same number of corrections to their scores, using as a very simple indicator the potential overlap between the ranges of $\pm$ 3 standard deviations around the mean beat error rate of each pair of judges. A lack of overlap indicates that two experts assess the accuracy of the algorithm differently.

Regardless of whether the overall level of agreement with the chord labellings is comparable between experts, we ask whether they agree on a rank level of which excerpts are considered accurate and which are not. An item-based non-parametric correlation, Spearman's ρ, is used. We also compare the corrections that the experts provide for instances of disagreement with the algorithmic labelling. We therefore select all instances where two experts provide a correction at a point of disagreement and count the relative number of instances where the correction is identical. The binary variable which reflects identity or difference in the corrections of two judges is tested against a minimum threshold of agreement required with a binomial test.

2.4 Performance of chord-labelling algorithm

For assessing the performance of our chord labelling itself and the labels from song books, we use the median, range and inter-quartile range of the beat error rates. As there is potential mutual disagreement between judges, we treat the feedback provided by each expert as an individual chord-labelling solution, and evaluate it just like the algorithmic labelling, again using the percentiles for the labelling from each expert with regard to the other three experts, resulting in an error baseline.

3 CONCLUSION

With this paper we tackle two problems that frequently arise when dealing with ground-truth data in MIR tasks. The first concerns evaluating the feasibility of working with a single ground truth, *i.e.* a data set where for every instance of multivariate musical data there is a predefined datum to be predicted for a correct answer to be predicted. The second problem requires a method for situations in which there is ambiguity that allows for several correct answers for a given set of musical data, and thus the data can take multiple equally or variably valid values.

4 ACKNOWLEDGEMENTS

The authors are supported by EPSRC grants EP/D038855 and GR/S84750. MIDI transcriptions used in this project were provided by Geerdes midimusic [2] .

5 REFERENCES

[1] D. de la Motte. *Harmonielehre*. dtv / Bärenreiter, 1976.

[2] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[3] A. Moore. *Rock: The primary Text*. Open University, 1993.

[4] W. Piston. *Harmony*. Victor Gollancz, 1948.

[5] C. Rhodes, D. Lewis, and D. Müllensiefen. Bayesian Model Selection for Harmonic Labelling. In *Mathematics and Computation in Music*, Berlin, 2007.

[6] D. Temperley. *The Cognition of Basic Musical Structures*. MIT Press, Cambridge, MA, 2001.

[2] `http://www.midimusic.de/`