Perceptually-informed Chord Label Evaluation

Andrew McLeod,[1] Xavier Suermondt,[1] Steffen A. Herff,[1] Martin Rohrmeier[1]

[1]*Digital and Cognitive Musicology Lab, EPFL, Switzerland*
[1]{andrew.mcleod, xavier.suermondt, steffen.herff, martin.rohrmeier@epfl.ch}

Background

Automatic chord detection, the segmenting and labeling of a musical piece with chord symbols, is a fundamental step in the harmonic analysis of music. However, comparatively little focus has been given to the evaluation of such labels. Usually, evaluation is based on a binary correct/incorrect accuracy (Chord Symbol Recall; CSR), either of the full chord symbol, or specific features of the chord symbols (e.g., root, chord type, and inversion), without taking the perceptual similarity between label and chord into account (e.g., Harte, 2010).

Aims

We propose and implement a new perceptually-informed evaluation metric for chord label accuracy evaluation, based on Spectral Pitch Similarity (SPS) (Milne, Laney, & Sharp, 2015). SPS is calculated as the cosine similarity between the spectrograms of two sounds, and has been shown to correlate well with perceptual similarity. SPS is similar to simply measuring shared pitch classes between two chords, but also takes into account the harmonic content of the resulting audio. Our goal with this work is to demonstrate how evaluation differs when taking this perceptual information into account.

Method

Given two chord symbols (root, chord type, and inversion), we generate a MIDI file of each, where its bass note is in octave 4, and other notes are initially stacked thirds. Inversions are generated by shifting the lowest note(s) up an octave. We synthesize piano audio for each, and calculate SPS Accuracy as one minus the SPS of the resulting spectrograms. We then compare performance of a harmonic analysis model (McLeod & Rohrmeier, 2020) with different parameter settings, and evaluate performance using conventional CSR and SPS Accuracy against an expert-annotated ground truth.

Results

The model achieves an overall CSR of 0.472, and a root-only CSR of 0.650. Meanwhile, it achieves an SPS Accuracy of 0.730. Systematically changing the model's parameters and reevaluating the results reveal that the CSR and SPS accuracy are overall highly correlated. However, specific parameters affect the measures differently. For example, varying the model's "KSM-exponent"—which was previously identified as a large contributor to model performance using CSR (McLeod & Rohrmeier, 2020)—only marginally affects SPS accuracy. Consequently, the decision of whether CSR or SPS is being used to evaluate the model greatly influences model optimisation and outcome interpretation.

Conclusions

In many regards, SPS accuracy and CSR produce similar results, suggesting that SPS accuracy captures a lot of the information previously evaluated by CSR. However, in some instances, evaluation of model performance can vary dramatically, depending on whether perceptual similarity is accounted for. We believe that—depending on the research question—SPS Accuracy may be more representative of label quality. For example, chords such as G7 and Bdim: they do not match at all for CSR metrics, yet they function as substitutes of one another in many musical contexts, which is reflected in relatively high SPS Accuracy. However, chords such as Cmaj7 and Emin (which have the same number of overlapping pitches as G7 and Bdim, and a very similar SPS Accuracy) do not function similarly (in the key of C major).

This highlights the need for different evaluation metrics depending on the desired task: our proposed SPS Accuracy is well-suited for tasks in which listener perception is key, whereas more purely music theoretical tasks (like automatic functional analysis) might require even more sophisticated methods. Overall, we argue that great care should be given to the means of evaluation in chord detection models.

References

Harte, C. (2010). Towards Automatic Extraction of Harmony Information from Music Signals (PhD Thesis). Queen Mary University of London.

McLeod, A. & Rohrmeier, M. (2020). A Modular System for Harmonic Structure Analysis of Music. DMRN+15: Digital Music Research Network One-day Workshop, 6.

Milne, A. J., Laney, R., & Sharp, D. B. (2015). A spectral pitch class model of the probe tone data and scalic tonality. Music Perception: An Interdisciplinary Journal, 32(4), 364-393.

Keywords: harmonic analysis, chord detection, evaluation metrics, Spectral Pitch Similarity.