CoCoPops is a dataset of expert transcriptions of (currently) 20th century popular music, extending a preexisting dataset, and combining with a second into a common format. CoCoPops currently has two main parts:

- The **Billboard** subset, consisting of new melodic transcriptions added to 214 harmonic transcriptions from the existing McGill Billboard dataset (Burgoyne et al., 2011).
- The **Rolling Stone** subset, a conversion of Temperley’s and de Clercq’s (2011) RS200 corpus to a common tabular (humdrum) format.

**Rationale**

Harmony and melody are two perceptually salient musical features, yet most datasets consist of either harmonic annotations, or melody for classical music. We lack expert-quality melodic data for popular music. Adding (and unifying) melodic as well as perceptual (emotion) data to existing corpora facilitates research in computational musicology and perception.

Computational research is frequently burdened by datasets across:

- Multiple encoding formats (e.g., musicXML, kern, MIDI)
- Different representation schemes (e.g., RN, leadsheet, Harte, etc.)
- Low or unknown quality or data scarcity
- Aimed at symbolic or audio

The CoCoPops dataset, released under a CC-BY-4.0 license, aims to facilitate cross-corpus study by extending and integrating expert data using widely-used representations in a common format (humdrum) supported by all modern computational musicology software (e.g., humdrumR, music21) with a tabular format (i.e., csv) that can support audio and symbolic data.

**New Data Includes:**

- 214 expert melodic transcriptions
- Continuous valence & arousal ratings (100 song subset)
- Root Mean Square (RMS) (100 song subset)
- Linked MusicBrainz IDs
- All Billboard and RS200 data to common format

**Summary Statistics Comparison**

Despite differences in their sampling and transcription procedures, basic musical features are quite consistent between the two datasets, which suggests that treating them like a single large dataset is appropriate.

---

**Sample CoCoPops File**

**Distribution of Pitches in CoCoPops**

The distributions of absolute (vocal) pitch in the two datasets are similar.

**Distribution of Scale Degrees in CoCoPops**

The ranked frequency of scale degrees in the two datasets are nearly identical.

**Distribution of Harmonies in CoCoPops**

Nine of the ten most frequent harmonies in are the same between the two datasets.

Only the fourth and fifth spots (VI and I) and the tenth and eleventh spots (IV and ii) differ slightly in relative frequency.

---

https://github.com/Computational-Cognitive-Musicology-Lab/CoCoPops

Search GitHub for #humdrum #digital-scores to find other humdrum datasets.