Chord Label Personalization through Deep Learning of Integrated Harmonic Interval-based Representations

Hendrik Vincent Koops*[1], W. Bas de Haas†[2], Jeroen Bransen‡[2], and Anja Volk§ [1]

[1]Utrecht University, Utrecht, the Netherlands
[2]Chordify, Utrecht, the Netherlands

œ

The increasing accuracy of automatic chord estimation systems, the availability of vast amounts of heterogeneous reference annotations, and insights from annotator subjectivity research make chord label personalization increasingly important. Nevertheless, automatic chord estimation systems are historically exclusively trained and evaluated on a single reference annotation. We introduce a first approach to automatic chord label personalization by modeling subjectivity through deep learning of a harmonic interval-based chord label representation. After integrating these representations from multiple annotators, we can accurately personalize chord labels for individual annotators from a single model and the annotators' chord label vocabulary. Furthermore, we show that chord personalization using multiple reference annotations outperforms using a single reference annotation.

Keywords: Automatic Chord Estimation, Annotator Subjectivity, Deep Learning

1 Introduction

Annotator subjectivity makes it hard to derive one-size-fits-all chord labels. Annotators transcribing chords from a recording by ear can disagree because of personal preference, bias towards a particular instrument, and because harmony can be ambiguous perceptually as well as theoretically by definition [Schoenberg, 1978, Meyer, 1957]. These reasons contributed to annotators creating large amounts of heterogeneous chord label reference annotations. For example, on-line repositories for popular songs often contain multiple, heterogeneous versions.

*h.v.koops@uu.nl
†bas@chordify.net
‡jeroen@chordify.net
§a.volk@uu.nl

H.V. Koops, W.B. de Haas, J. Bransen and A. Volk. 2017. Chord Label Personalization through Deep Learning of Integrated Harmonic Interval-based Representations. Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN. Anchorage, US. May, 2017. 1(1). pp 19-25

One approach to the problem of finding the appropriate chord labels in a large number of heterogeneous chord label sequences for the same song is data fusion. Data fusion research shows that knowledge shared between sources can be integrated to produce a unified view that can outperform individual sources [Dong et al., 2009]. In a musical application, it was found that integrating the output of multiple Automatic Chord Estimation (ACE) algorithms results in chord label sequences that outperform the individual sequences when compared to a single ground truth [Koops et al., 2016]. Nevertheless, this approach is built on the intuition that one single correct annotation exists that is best for everybody, on which ACE systems are almost exclusively trained. Such reference annotation is either compiled by a single person [Mauch et al., 2009], or unified from multiple opinions [Burgoyne et al., 2011]. Although most of the creators of these datasets warn for subjectivity and ambiguity, they are in practice used as the *de facto* ground truth in MIR chord research and tasks (e.g. MIREX ACE).

On the other hand, it can also be argued that there is no single best reference annotation, and that chord labels are correct with varying degrees of "goodness-of-fit" depending on the target audience [Ni et al., 2013]. In particular for richly orchestrated, harmonically complex music, different chord labels can be chosen for a part, depending on the instrument, voicing or the annotators' chord label vocabulary.

In this paper, we propose a solution to the problem of finding appropriate chord labels in multiple, subjective heterogeneous reference annotations for the same song. We propose an automatic audio chord label estimation and personalization technique using the harmonic content shared between annotators. From deep learned shared harmonic interval profiles, we can create chord labels that match a particular *annotator vocabulary*, thereby providing an annotator with familiar, and personal chord labels. We test our approach on a 20-song dataset with multiple reference annotations, created by annotators who use different chord label vocabularies. We show that by taking into account annotator subjectivity while training our ACE model, we can provide personalized chord labels for each annotator.

Contribution. The contribution of this paper is twofold. First, we introduce an approach to automatic chord label personalization by taking into account annotator subjectivity. Through this end, we introduce a harmonic interval-based mid-level representation that captures harmonic intervals found in chord labels. Secondly, we show that after integrating these features from multiple annotators and deep learning, we can accurately personalize chord labels for individual annotators. Finally, we show that chord label personalization using integrated features outperforms personalization from a commonly used reference annotation.

2 Deep Learning Harmonic Interval Subjectivity

For the goal of chord label personalization, we create an harmonic bird's-eye view from different reference annotations, by integrating their chord labels. More specifically, we introduce a new feature that captures the shared harmonic interval profile of multiple chord labels, which we deep learn from audio. First, we extract Constant Q (CQT) features from audio, then we calculate Shared Harmonic Interval Profile (SHIP) representations from multiple chord label reference annotations corresponding to the CQT frames. Finally, we train a deep neural network to associate a context window of CQT to SHIP features.

From audio, we calculate a time-frequency representation where the frequency bins are geometrically spaced and ratios of the center frequencies to bandwidths of all bins are equal, called a Constant Q (CQT) spectral transform [Schörkhuber and Klapuri, 2010]. We calculate

H.V. Koops, W.B. de Haas, J. Bransen and A. Volk. 2017. Chord Label Personalization through Deep Learning of Integrated Harmonic Interval-based Representations. Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN. Anchorage, US. May, 2017. 1(1). pp 19-25

	C	C#	D	D#	E	F	F#	G	G#	A	A#	B	N	♯3	♭3	⋆3	♯7	♭7	⋆7
G:maj7	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0
G:maj	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1
G:maj7	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0
G:minmaj7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0
SHIP	0	0	0	0	0	0	0	1	0	0	0	0	0	0.75	0.25	0	0.75	0	0.25

Table 1: Interval profiles from root notes of HIPs of different chord labels and their SHIP

these CQT features with a hop length of 4096 samples, a minimum frequency of ≈ 32.7 Hz (C1 note), $24 \times 8 = 192$ bins, 24 bins per octave. This way we can capture pitches spanning from low notes to 8 octaves above C1. Two bins per semitone allows for slight tuning variations.

To personalize chord labels from an arbitrarily sized vocabulary for an arbitrary number of annotators, we need a chord representation that (i) is robust against label sparsity, and (ii) captures an integrated view of all annotators. We propose a new representation that captures a harmonic interval profile (HIP) of chord labels, instead of directly learning a chord label classifier. The rationale behind the HIP is that most chords can be reduced to the root note and the stacked triadic intervals, where the amount and combination of triadic interval determines the chord quality and possible extensions. The HIP captures this intuition by reducing a chord label to its root and harmonic interval profile. HIP is a concatenation of multiple one-hot vectors that denote a root note and additional harmonic intervals relative to the root that are expressed in the chord label.

In this paper, we use a concatenation of three one-hot vectors: roots, thirds and sevenths. The first vector is of size 13 and denotes the 12 chromatic root notes (C...B) + a "no chord" (N) bin. The second vector is of size 3 and denotes if the chord denoted by the chord label contains a major third (♯3), minor third (♭3), or no third (⋆3) relative to the root note. The third vector, also of size 3, denotes the same, but for the seventh interval (♯7, ♭7, ⋆7). The HIP can be extended to include other intervals as well. In Table 1 we show example chord labels and their HIP equivalent. The last row shows the SHIP created from the HIPs above it.

2.1 Deep Learning Shared Harmonic Interval Profiles

We use a deep neural network to learn SHIP from CQT. Based on preliminary experiments, a funnel-shaped architecture with three hidden rectifier unit layers of sizes 1024, 512, and 256 is chosen. Research in audio content analysis has shown that better prediction accuracies can be achieved by aggregating information over several frames instead of using a single frame [Sigtia et al., 2015, Bergstra et al., 2006]. Therefore, the input for our DNN is a window of CQT features from which we learn the SHIP. Preliminary experiments found an optimal window size of 15 frames, that is: 7 frames left and right directly adjacent to a frame. Consequently, our neural network has input layer size of $192 \times 15 = 2880$. The output layer consists of 19 units corresponding with the SHIP features as explained above.

We train the DNN using stochastic gradient descent by minimizing the cross- entropy between the output of the DNN with the desired SHIP (computed by considering the chord labels from all annotators for that audio frame). We train the hyper-parameters of the network using mini-batch (size 512) training using the ADAM update rule [Kingma and Ba, 2014]. Early stopping is applied when validation accuracy does not increase after 20 epochs. After training the DNN, we can create chord labels from the learned SHIP features.

3 Annotator Vocabulary-based Chord Label Estimation

The SHIP features are used to associate probabilities to chord labels from a given vocabulary. For a chord label **L** the HIP **h** contains exactly three ones, corresponding to the root, thirds and sevenths of the label **L**. From the SHIP **A** of a particular audio frame, we project out three values for which **h** contains ones (**h**(**A**)). The product of these values is then interpreted as the combined probability **CP** (= Π **h**(**A**)) of the intervals in **L** given **A**. Given a vocabulary of chord labels, we normalize the **CP**s to obtain a probability density function over all chord labels in the vocabulary given **A**. The chord label with the highest probability is chosen as the chord label for the audio frame associated to **A**.

For the chord label examples in Table 1, the products of the non-zero values of the point-wise multiplications $\approx$ 0.56, 0.19, and 0.19 for G:maj7, G:maj, and G:minmaj7 respectively. If we consider these chord labels to be a vocabulary, and normalize the values, we obtain probabilities $\approx$ 0.6, 0.2, 0.2, respectively. Given extracted SHIP from multiple annotators providing reference annotations and chord label vocabularies, we can now generate annotator specific chords labels.

4 Evaluation

SHIP models multiple (related) chords for a single frame, e.g., the SHIP in Table 1 models different flavors of a G and a C chord. For the purpose of personalization, we want to present the annotator with only the chords they understand and prefer, thereby producing a high chord label accuracy for each annotator. For example, if an annotator does not know a G:maj7 but does know an G, and both are probable from an SHIP, we like to present the latter. In this paper, we evaluate our DNN ACE personalization approach, and the SHIP representation, for each individual annotator and their vocabulary.

In an experiment we compare training of our chord label personalization system on multiple reference annotations with training on a commonly used single reference annotation. In the first case we train a DNN (DNN_{SHIP}) on SHIPs derived from a dataset introduced by Ni et al. [2013] containing 20 popular songs annotated by five annotators with varying degrees of musical proficiency. In the second case, we train a DNN (DNN_{ISO}) on the HIP of the *Isophonics* (ISO) single reference annotation [Mauch et al., 2009]. ISO is a peer-reviewed, and *de facto* standard training reference annotation used in numerous ACE systems. From the (s)HIP the annotator chord labels are derived and we evaluate the systems on every individual annotator. We hypothesize that training a system on SHIP based on multiple reference annotations captures the annotator subjectivity of these annotations and leads to better personalization than training the same system on a single (ISO) reference annotation.

It could be argued that the system trained on five reference annotations has more data to learn from than a system trained on the single ISO reference annotation. To eliminate this possible training bias, we evaluate the annotators' chord labels directly on the chord labels from ISO ($\text{ANN}|_{\text{ISO}}$). This evaluation reveals the similarity between the SHIP and the ISO and puts the results from DNN_{ISO} in perspective. If DNN_{SHIP} is better at personalizing chords (i.e. provides chord labels with a higher accuracy per annotator) than DNN_{ISO} while the annotator's annotations and the ISO are similar, then we can argue that using multiple reference annotations and SHIP is better for chord label personalization than using just the ISO. In a final baseline evaluation, we also test ISO on DNN_{ISO} to measure how well it models the ISO.

Ignoring inversions, the complete dataset from Ni et al. [2013] contains 161 unique chord

	Annotator 1			Annotator 2			Annotator 3			Annotator 4			Annotator 5			ISO
	DNN_{SHIP}	$ANN\|_{ISO}$	DNN_{ISO}	DNN_{SHIP}	$ANN\|_{ISO}$	DNN_{ISO}	DNN_{SHIP}	$ANN\|_{ISO}$	DNN_{ISO}	DNN_{SHIP}	$ANN\|_{ISO}$	DNN_{ISO}	DNN_{SHIP}	$ANN\|_{ISO}$	DNN_{ISO}	DNN_{ISO}
ROOT	0.85	0.73	0.66	0.82	0.74	0.67	0.80	0.72	0.65	0.80	0.73	0.65	0.77	0.67	0.60	0.86
MAJMIN	0.82	0.69	0.61	0.69	0.67	0.53	0.67	0.69	0.53	0.73	0.67	0.55	0.72	0.61	0.55	0.69
MIREX	0.82	0.70	0.61	0.69	0.68	0.54	0.66	0.69	0.54	0.73	0.68	0.56	0.72	0.62	0.55	0.69
THIRDS	0.82	0.70	0.62	0.75	0.67	0.59	0.79	0.69	0.62	0.76	0.68	0.61	0.72	0.62	0.55	0.83
7THS	0.77	0.56	0.50	0.64	0.53	0.42	0.64	0.56	0.43	0.53	0.48	0.40	0.72	0.53	0.55	0.65

Table 2: Chord label personalization accuracies for the five annotators

labels, comprised of five annotators using 87, 74, 62, 81 and 26 unique chord labels respectively. The intersection of the chord labels of all annotators contains just 21 chord labels meaning that each annotator uses a quite distinct vocabulary of chord labels. For each song in the dataset, we calculate CQT and SHIP features. We divide our CQT and SHIP dataset frame-wise into 65% training (28158 frames), 10% evaluation (4332 frames) and 25% testing (10830 frames) sets. For the testing set, for each annotator, we create chord labels from the deep learned SHIP based on the annotators' vocabulary.

We use the standard MIREX chord label evaluation methods to compare the output of our system with the reference annotation from an annotator [Raffel et al., 2014]. We use evaluations at different chord granularity levels. ROOT only compares the root of the chords. MAJMIN only compares major, minor, and "no chord" labels. MIREX considers a chord label correct if it shares at least three pitch classes with the reference label. THIRDS compares chords at the level of root and major or minor third. 7THS compares all above plus the seventh notes.

5 Results

The DNN_{SHIP} columns of Table 2 for each annotator show average accuracies of 0.72 ($\sigma = 0.08$). For each chord granularity level, our DNN_{SHIP} system provides personalized chord labels that are trained on multiple annotations, but are comparable with a system that was trained an evaluated on a single reference annotation (ISO column of Tab. 2). Comparable high accuracy scores for each annotator show that the system is able to learn a SHIP representation that (i) is meaningful for all annotators (ii) from which chord labels can be accurately personalized for each annotator. The low scores for annotator 4 for SEVENTHS form an exception. An analysis by Ni et al. [2013] revealed that between annotators, annotator 4 was on average the most different from the consensus. Equal scores for annotator 5 for all evaluations except ROOT are explained by annotator 5 being an amateur musician using only major and minor chords.

Comparing the DNN_{SHIP} and DNN_{ISO} columns, we see that for each annotator DNN_{SHIP} models the annotator better than DNN_{ISO}. With an average accuracy of 0.55 ($\sigma = 0.07$), DNN_{ISO}'s accuracy is on average 0.17 lower than DNN_{SHIP}, showing that for these annotators, ISO is not able to accurately model chord label personalization. Nevertheless, the last column shows that the system trained on ISO modeled the ISO quite well. The results of $ANN|_{ISO}$ show that the annotators in general agree with ISO, but the lower score in DNN_{ISO} shows that the agreement is not good enough for personalization. Overall, these results show that our system is able to personalize chord labels from multiple reference annotations, while personalization using a commonly used single reference annotation yields significantly worse results.

H.V. Koops, W.B. de Haas, J. Bransen and A. Volk. 2017. Chord Label Personalization through Deep Learning of Integrated Harmonic Interval-based Representations. Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN. Anchorage, US. May, 2017. 1(1). pp 19-25

6 Conclusions and Discussion

We presented a system that provides personalized chord labels from multiple reference annotations from audio, based on the annotators' specific chord label vocabulary and an interval-based chord label representation that captures the shared subjectivity between annotators. To test the scalability of our system, our experiment needs to be repeated on a larger dataset, with more songs and more annotators. Furthermore, a similar experiment on a dataset with instrument/proficiency/cultural-specific annotations from different annotators would shed light on whether our system generalizes to providing chord label annotations in different contexts. From the results presented in this paper, we believe chord label personalization is the next step in the evolution of ACE systems.

Acknowledgments

We thank Y. Ni, M. McVicar, R. Santos-Rodriguez and T. De Bie for providing their dataset.

References

J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate features and adaboost for music classification. *Machine learning*, 65(2-3):473–484, 2006.

J.A. Burgoyne, J. Wild, and I. Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proc. of the 12th International Society for Music Information Retrieval Conference, ISMIR*, volume 11, pages 633–638, 2011.

X.L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proc. of the VLDB Endowment*, 2(1):550–561, 2009.

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. of the 3rd International Conference on Learning Representations, ICLR*, 2014.

H.V. Koops, W.B. de Haas, D. Bountouridis, and A. Volk. Integration and quality assessment of heterogeneous chord sequences using data fusion. In *Proc. of the 17th International Society for Music Information Retrieval Conference, ISMIR, New York, USA*, pages 178–184, 2016.

M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler. Omras2 metadata project 2009. In *Late-breaking demo session at 10th International Society for Music Information Retrieval Conference, ISMIR*, 2009.

L.B. Meyer. Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, 15(4):412–424, 1957.

Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie. Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2607–2615, 2013.

C. Raffel, B. McFee, E.J. Humphrey, J. Salamon, O. Nieto, D. Liang, D.P.W. Ellis, and C. Raffel. mir_eval: A transparent implementation of common mir metrics. In *Proc. of the 15th International Society for Music Information Retrieval Conference, ISMIR*, pages 367–372, 2014.

A. Schoenberg. *Theory of harmony*. University of California Press, 1978.

C. Schörkhuber and A. Klapuri. Constant-q transform toolbox for music processing. In *Proc. of the 7th Sound and Music Computing Conference, Barcelona, Spain*, 2010.

S. Sigtia, N. Boulanger-Lewandowski, and S. Dixon. Audio chord recognition with a hybrid recurrent neural network. In *Proc. of the 16th International Society for Music Information Retrieval Conference, ISMIR*, pages 127–133, 2015.

H.V. Koops, W.B. de Haas, J. Bransen and A. Volk. 2017. Chord Label Personalization through Deep Learning of Integrated Harmonic Interval-based Representations. Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN. Anchorage, US. May, 2017. 1(1). pp 19-25