Neural Computing and Applications (2020) 32:929–939
https://doi.org/10.1007/s00521-018-3703-y

S.I. : DEEP LEARNING FOR MUSIC AND AUDIO

Automatic chord label personalization through deep learning of shared harmonic interval profiles

Hendrik Vincent Koops[1] · W. Bas de Haas[2] · Jeroen Bransen[2] · Anja Volk[1]

Received: 3 December 2017 / Accepted: 10 September 2018 / Published online: 21 September 2018

Abstract
Current automatic chord estimation systems are trained and tested using datasets that contain *single reference annotations*, i.e., for each corresponding musical segment (e.g., audio frame or section), the reference annotation contains a *single* chord label. Nevertheless, theoretical insights on harmonic ambiguity from harmony theory, experimental studies on annotator subjectivity in harmony annotations, and the availability of vast amounts of heterogeneous (subjective) harmony annotations in crowd-sourced repositories make the notion of a single-harmonic "ground truth" reference annotation a tenuous one. Recent studies suggest that subjectivity is intrinsic to harmonic reference annotations that should be embraced in automatic chord estimation rather than resolved. We introduce the first approach to automatic chord label personalization by modeling annotator subjectivity through harmonic interval-based chord representations. We integrate these representations from multiple annotators and deep learn them from audio. From a single trained model and the annotators' chord-label vocabulary, we can accurately personalize chord labels for individual annotators. Furthermore, we show that chord personalization using multiple reference annotations outperforms using just a single reference annotation. Our results show that annotator subjectivity should inform future research on automatic chord estimation to improve the state of the art.

Keywords Automatic chord estimation · Personalization · Harmony

1 Introduction

Extracting time-aligned sequences of chords from a given audio music signal, commonly referred to as automatic chord estimation (ACE), is a well-researched topic in music information retrieval (MIR). ACE systems consist of some variation of audio feature extraction followed by a pattern matching step in which the audio features are associated with chord labels [22]. Both feature extraction and pattern matching in modern ACE systems are commonly performed using machine learning technique; in current state-of-the-art ACE systems usually some flavor of deep learning [10, 17, 22]. Although current state-of-the-art ACE performance power allows them to be used in commercial products (e.g., Chordify,[1] Riffstation[2]), their performance nevertheless seems to be tapering off in recent years[3] [10]. One of the reasons, Humphrey et al. [10] argue, is that the perception of chords in recorded music can be highly subjective, which is problematic for deriving a single reference "ground truth" chord label annotation.

1.1 The problem of 'ground truth' in harmony transcriptions

Annotators transcribing chords from a recording by ear can disagree because of personal preference and bias toward a

✉ Hendrik Vincent Koops
h.v.koops@uu.nl

W. Bas de Haas
bas@chordify.net

Jeroen Bransen
jeroen@chordify.net

Anja Volk
a.volk@uu.nl

[1] Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

[2] Chordify, Boothstraat 2a, 3512 BW Utrecht, The Netherlands

[1] https://chordify.net/.

[2] https://www.riffstation.com/.

[3] http://www.music-ir.org/mirex/wiki/2017:Audio_Chord_Estimation.

particular instrument, and because harmony can be ambiguous perceptually as well as theoretically by definition [23, 27]. The harmonic content of an audio recording is often ambiguous and can result in annotators disagreeing about which chord label best describes a musical segment. For example, if in a recording the simultaneously sounding notes C,E, and G are combined with a melody touching a B, it is up to the annotator whether to include B in the chord label (C:maj7) or not (C:maj). Neither of these choices would be objectively wrong, but each expresses a subjective selection of the harmonic content of the audio signal.

Furthermore, reharmonization (altering an original harmony) is a common phenomenon in harmony transcriptions of popular music, which can happen implicitly because of perceptual differences between annotators, or explicitly to make a transcription more useful in a particular context. That is, there is an element of usefulness of chord labels that influences chord label choice. This relates to pedagogical aspects (e.g., the playability of a chord on a certain instrument) or proficiency of a performer (e.g., a music student of a certain level). Online chord-label repositories (e.g., Ultimate-Guitar,[4] E-Chords[5]) for popular songs often contain multiple, heterogeneous chord sheets of the same song. These chord sheets often contain reharmonizations in the form of "convenient" chord labels, which probably do not exactly describe the chords that the musicians of the original performance played. However, these different reharmonizations are useful, because playing a song by oneself might call for slightly different chords than in the context of a complete band. These kinds of disagreement should not be confused with error. Annotating chords by listening is a challenging task even for professional musicians and composers, which creates ample opportunity for annotation errors. For example, if the instruments in a recording play the notes C,E, and G, then the chord label C:min is not a valid subjective choice but simply wrong.

This so-called *annotator subjectivity* problem makes it hard to derive one-size-fits-all chord labels and contributed to annotators creating large amounts of heterogeneous chord label reference annotations. 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 some form of data integration, such as data fusion. Data fusion research shows that knowledge shared between sources can be integrated to produce a unified view that can outperform individual sources when compared to a single ground truth [6] In a musical application, it was found that integrating the output of multiple ACE algorithms results in chord label sequences that outperform the individual sequences when compared to a single ground truth [14]. However, this approach is built on the assumption that one single correct ground truth annotation exists that is best for everybody, on which ACE systems are exclusively built. Such a reference annotation is either compiled by a single person [19] or unified from multiple opinions [2]. Although most of the creators of single reference annotation datasets warn for subjectivity and ambiguity, they are in practice used as the *de facto* ground truth in MIR harmony research and related tasks, including but not limited to: ACE [22], analysis of harmonic trends over time [3, 8, 20], computational hook discovery [30], chorus analysis of popular music [31], data fusion of ACE algorithms [14], automatic structural segmentation [5], and computational creativity, such as automatic generation of harmonic accompaniment [4] and harmonic blending [11].

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 [24]. 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 an analysis, Humphrey et al. [10] found a high-degree variance between different "ground truth" annotations that are often used in ACE research. In an experimental study, Ni et al. [24] found that annotators transcribing the same music recordings disagree on roughly 10 percent of chord labels. Similarly, low inter-rater agreement was found in an analysis of human annotations in the MIREX audio similarity task [7]. Ni et al. found that that state-of-the-art ACE systems trained on single reference annotations perform worse on a consensus of annotators than on the single reference annotations. They suggest that current ACE systems are starting to overfit to subjective single reference annotations, thereby producing models that fail to accurately represent the variability found in human annotations. Humphrey et al. and Ni et al. both argue in [10, 24] that subjectivity plays a role in the collection of reference annotations, and that subjectivity should be embraced in ACE rather than resolved. The lack of inter-rater agreement between annotators and the observation that current ACE systems are starting to overfit toward a single subjective reference annotation implies that ACE can be improved by taking into account annotator subjectivity.

1.2 Embracing annotator subjectivity in ACE

One approach of embracing annotator subjectivity in ACE is personalization. Instead of providing all users of an ACE system with the same chord labels, chord labels could be tuned for particular use. Personalization could be informed by, for example, a users' chord-label preference, primary instrument, musical proficiency or cultural background.

[4] https://www.ultimate-guitar.com/.

[5] https://www.e-chords.com/.

One way of solving this problem would be to build a model for each subjective annotation or user. This way, for each user, personalized chord labels could be extracted that are tailored to the users' idiosyncrasies. However, this introduces scalability problems: personalization for an arbitrary number of annotators would potentially mean training and testing an arbitrary number of models. For large-scale ACE applications with hundreds of thousands of users, such as *Chordify* and *Riffstation*, this would amount to an unmaintainable problem. A more maintainable and scalable solution would be to create a single model that is capable of taking into account the variability found in multiple heterogeneous reference annotations. With correct modeling of variability and the tuning of model output to users' particular idiosyncrasies, specific chord labels could be provided to particular users.

Many approaches to incorporate user modeling into machine learning methods have been proposed for various applications, for example modeling user historical behavior for scientific paper recommendation [33], integrating user location [35], time [34], or age and gender information for collaborative filtering [12]. With the exception of incorporating user modeling in music recommendation systems (e.g., [26]), we are not aware of similar works in the MIR domain.

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 multiple annotators. After deep learning this shared information, we can create chord labels that are tuned to a particular *annotator vocabulary*, thereby providing an annotator with familiar, and personal chord labels. We test our approach on a 50-song dataset with multiple expert 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 as follows. This paper presents an approach to automatic chord label personalization by taking into account annotator subjectivity. This paper extends [15] by way of a harmony representation that captures more harmonic intervals, use of a larger dataset, and a more complete evaluation that includes other harmony representations. We introduce a mid-level representation that captures harmonic intervals found in chords. The representation introduced in this paper extends the one introduced in [15] by a) taking into account all possible intervals in one octave, and by including the bass note. This way, we can account for all chords (with inversions) that contain intervals within one octave. In contrast to [15], we compare the results of using this representation to other harmony representations that are commonly used in ACE. We show that after integrating these representations from multiple annotators and deep learning, we can provide better chord labels for each individual annotator. To this end, we use a new and open annotator subjectivity dataset that better captures the variance found in popular music than the one used in [15]. Finally, we show that chord label personalization using integrated representations outperforms personalization from a commonly used reference annotation.

2 Chord-label annotator subjectivity

For the goal of automatic chord estimation and personalization, we use two datasets. The *Chordify annotator subjectivity dataset* (CASD): this dataset contains *multiple reference annotations* for fifty songs from four expert annotators. The *Billboard* dataset: this dataset contains *single reference annotations* for the same songs. After briefly discussing the CASD and *Billboard* dataset in Sect. 2.1, we discuss the disagreement between annotators in the CASD in more detail in Sect. 2.2. We investigate the disagreement between annotators in terms of the average pairwise agreement between the annotators using the standard MIREX ACE evaluation measures. This will give us a musically informed idea of the average agreement found in our dataset that will inform our personalization results in further sections. A discussion of the comparison between the CASD and BB will inform us how much the annotators agree with a standard reference annotation that is commonly used in training and testing ACE systems.

2.1 The *Billboard* (BB) and *Chordify annotator subjectivity dataset* (CASD) datasets

The *Billboard* dataset contains *single reference annotations* for songs selected randomly from the Billboard hot 100 chart in the USA between 1958 and 1991 and was introduced by Burgoyne et al. [2]. The transcription process concerned multiple people, but their annotations were integrated into a single annotation. After two annotators finished transcribing a song, a third meta-annotator resolved the differences between the annotators and combined them into a single master transcription. The *Billboard* dataset is a commonly used chord label dataset that is in practice used as the *de facto* ground truth for a large number of studies into harmony and related tasks (e.g., MIREX ACE).

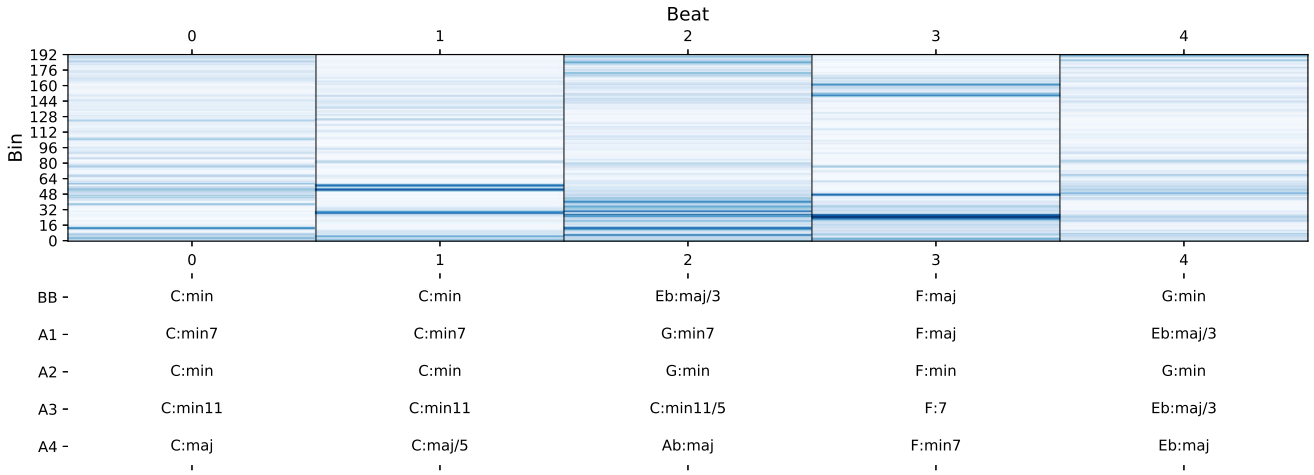

Fig. 1 Example of Constant-Q audio features and corresponding disagreement between the *Billboard* (BB) annotation and annotations from the four annotators A1, A2, A3, A4 from the CASD for a selection of beats. The figure shows that annotators can disagree on root notes, intervals, and inversions. Furthermore, the figure shows that although for beats 0 and 1 the same disagreement is found, their corresponding Constant-Q features are different. See that. In this research, we improve ACE by taking into account this disagreement when automatically estimating chords from audio for the annotators

The CASD[6] dataset was introduced by the authors of this paper and has a different focus: instead of aiming at constructing a chord sequence consensus, it provides *multiple reference annotations*, provided by four expert annotators (two guitarists and two pianists) for fifty songs [16]. All annotators are musical experts: they either studied composition or music performance at the undergraduate or graduate level. All annotators are also successful professional music performers, with between fifteen and twenty years of experience on their primary instrument. The chord labels provided by the annotators are encoded using the chord-label syntax introduced by Harte et al. [9]. This syntax provides a simple and intuitive encoding that is highly structured and unambiguous to parse with computational means. The annotations were created by (a modified version of the) Chordify web service, and an audio recording was provided by a YouTube web player. The YouTube audio was manually checked to be of reasonable quality. Furthermore, by comparing the audio features provided by the *Billboard* dataset with the YouTube audio, we verified that we used a recording that is similar to the one used in the *Billboard* dataset. The musical material was selected from the *Billboard* dataset. As a consequence, the CASD provides *Billboard* dataset identifiers which make it possible to cross-reference with the *Billboard* dataset.

Figure 1 provides an example of the disagreement between the annotations from annotators A1, A2, A3, A4 from the CASD and the annotation from the *Billboard* (BB). The figure shows Constant-Q features calculated from audio and the corresponding chord labels from the BB and the CASD. The figure shows that annotators disagree with each other on the level of root notes (e.g., `G` and `C` in beat 2), intervals (e.g., `C:min` and `C:min11` in beat 0 and 1), and bass notes (e.g., `Eb:maj` and `Eb:maj/3` in beat 4). Furthermore, the figure shows that although the same disagreement can be found for beat 0 and 1, the corresponding Constant-Q features do not look similar. The root cause of this disagreement is not known. The first instrument of annotators (i.e., a bias toward listening to the instrument they are accustomed to listening to), their preferred level of transcription detail, their musical sophistication (e.g., instrument and music theory proficiency), and even their harmonic taste (i.e., simply preferring the sound of a chord over another) could all be reasons why annotators differ in their transcriptions. In any case, we hypothesize that disagreement is an important factor in harmonic transcriptions that, when taken into account, can improve ACE results by providing tailored chord labels.

2.2 Chord-label agreement in the CASD and BB

In [16], we presented a detailed overview of the agreement found in the CASD, as well as the agreement from the annotators from the CASD with the BB [16]. Using the commonly used MIREX evaluation of chord-label overlap between annotations, a relatively low agreement with the annotators was found [25]. Altogether, it was found that the annotators used a particular set of chord labels, or *vocabulary* for their transcriptions. Their vocabularies differ in size and use of particular chord labels. That is, in addition to sharing common chord labels in their transcriptions,

[6] https://github.com/chordify/CASD.

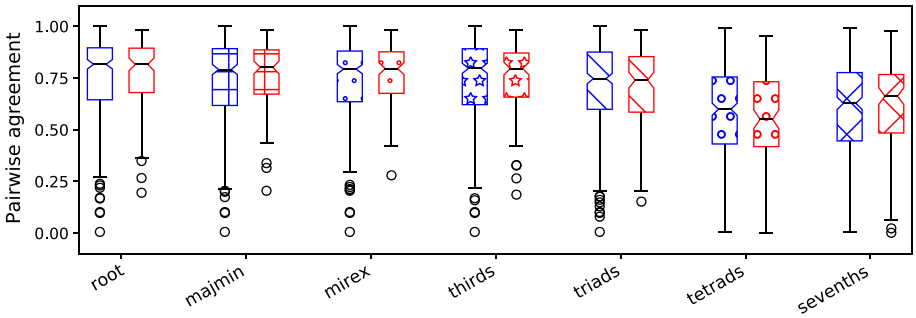

Fig. 2 In red: pairwise agreement among the four annotators from the CASD for all MIREX chord granularity levels (in red). In blue: pairwise agreement between the four annotators with the *Billboard* annotations. The figure shows overall low agreement between the annotators, and lower agreement with increased chord-label granularity. Furthermore, the annotators do on average not agree more with the *Billboard* annotation than with each other. This figure is adapted from Koops et al. [16] (color figure online)

each annotator uses a subset of particular chord labels that are not shared with the other annotators.

The pairwise agreement among all annotators for all fifty songs and all evaluation methods as reported by Koops et al. [16] can be found in Fig. 2 in blue. The figure shows that, overall, there is low agreement between the annotators, and lower agreement with increased chord-label granularity: annotators agree more on the root notes (ROOT) than on complex chord labels (e.g., SEVENTHS). The average agreement of root notes and majmin chords (two common evaluation granularities for ACE tasks) were found to be only 0.76 ($\sigma = 0.19$), and 73 ($\sigma = 0.2$), respectively. These levels of agreement raise questions on the existence of a *subjectivity ceiling* in harmonic transcriptions. It can be argued that if such a ceiling exists, ACE results beyond this ceiling are indicative that an ACE algorithm is tuned to a particular subjective annotation.

We argued in [16] that the relatively low overall chord-label agreement between expert annotators is problematic for the creation of one-size-fits-all chord-label annotations, which are almost universally used in ACE research. One approach to solving the problem of creating chord-label annotations with the broadest appeal is creating a consensus annotation from multiple expert annotations. This was proposed and presented in the *Billboard* dataset. The annotations in this dataset are the result of a meta-annotator creating a consensus from two expert annotations [2]. Assuming that a consensus annotation is on average closer to individual annotations than annotations are to each other, one would expect that the CASD annotators would agree on average more with the *Billboard* annotation than with each other.

In addition to annotator pairwise evaluations, Koops et al. evaluated the annotations from the CASD on the corresponding *Billboard* dataset annotation. Again a relatively low agreement was found. Agreement decreases with an increase in chord-label granularity: annotators agree more on the root notes (ROOT) than on complex chords (e.g., SEVENTHS) of the *Billboard* annotations. It was found that the average agreement of root notes is only 0.77 ($\sigma = 0.16$), with some scores as low as 0.19. Furthermore, a roughly equal chance was found of annotators agreeing more with the *Billboard* than with the other annotators. This last finding indicates that creating an ideal one-size-fits-all "ground truth" chord label annotation is problematic.

The results from analyzing annotator subjectivity in the CASD and BB in [16] indicate that 1) annotator subjectivity results in significantly different annotations and 2) that a single reference annotation for ACE is not expressive enough to provide annotators with tailored chord labels. Therefore, we propose to leverage annotator subjectivity to create an ACE system that can provide chord sequences tailored to specific users.

3 Deep learning harmonic interval subjectivity

We believe that chord-label sequences are inherently subjective. Based on the use-case and transcription purpose, user background, harmonic knowledge, and musical skills, different chord-label sequences can be considered correct for the same piece. In this research, we propose an ACE system that can provide personalized chords which are tailored to the musical preferences of a specific annotator. To accomplish this, we create a harmonic bird's-eye view from different reference annotations, by integrating their chord labels on the level of root, intervals, and bass. More specifically, we introduce a new structured representation that captures the shared harmonic interval profile of multiple chord labels, which we deep learn from audio and use to find the most appropriate chord labels for an annotator. The system is an improved version of the ACE personalization system that was introduced by Koops et al. [15].

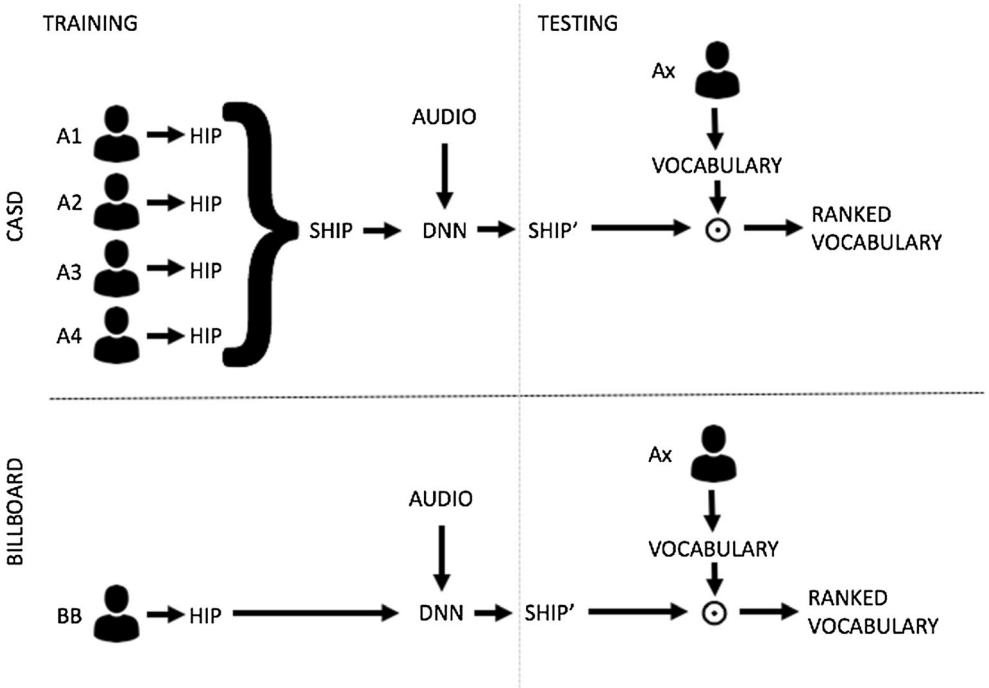

Fig. 3 Pipeline of our proposed CASD chord-label personalization ACE system. We start by creating Harmonic Interval Profiles (HIP) for each chord label for each annotator of the CASD. Then we integrate their HIP into a shared harmonic interval profile (SHIP), capturing the shared perceived harmonic content. We then train a DNN on Constant-Q audio features input to learn SHIP. After training, we rank annotator A_x's chord-label vocabulary using the SHIP. For each annotator individually, their highest ranked chord label is chosen, resulting in possibly different chord labels for each annotator. We compare this system with the same system, but with the DNN trained only on the HIP of the *Billboard* annotation

Figure 3 shows the pipeline of our proposed system, using annotations from multiple annotators from the CASD at the top and the single reference annotation from the BB at the bottom. For the CASD system we convert the annotators' chord labels into harmonic interval profiles (HIP). To capture annotator subjectivity, we integrate the HIP from all annotators into a shared harmonic interval profile (SHIP). SHIP captures the average harmonic content shared between the annotators. After extracting Constant-Q (CQT) features from audio, we train a deep neural network (DNN) to associate a context window of CQT to SHIP features. In the testing phase, we extract new SHIP from unseen CQT. We then use the SHIP to rank the chord labels from an annotator's chord-label *vocabulary*. The top-ranked chord label is chosen for that annotator. We compare this system with another system with a similar pipeline. The only difference is that we train the DNN on only the single BB reference annotation before personalizing chord labels for each annotator. The following sections discuss each of the steps of the system in more detail.

3.1 Structured harmony representation

Several harmony representations have been proposed in computational harmony research, of which *chroma* is most commonly used. Chroma captures the pitch class content of a chord wrapped into a single octave. Korzeniowski et al. [17] showed that this representation can be learned from audio using a deep neural network. This representation is sometimes extended to include a one-hot vector representing the root, or bass note. McFee et al. [21] proposed to learn such an extended chroma representation using deep learning for the goal of large-vocabulary automatic chord estimation. 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 hypothesize that the commonly used chroma representation is not able to capture these properties needed for chord-label personalization. Therefore, we propose to use a new representation that captures a harmonic interval profile (HIP) of chord labels, instead of the common approach of directly learning a chord-label classifier.

3.1.1 Harmonic interval profile (HIP)

The rationale behind the HIP is that most chords can be reduced to the root note and a set of intervals relative to the root, where the amount and combination of intervals determine the chord quality and possible extensions. The

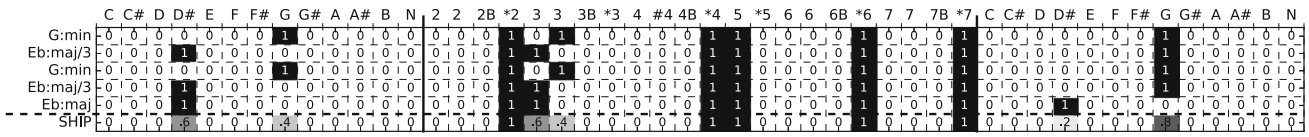

Fig. 4 Example harmonic interval profiles (HIP) of different chord labels and their SHIP. HIP consists of several concatenated vectors. The first one is a one-hot vector representing root notes (12 + 1 'no chord' N bins). The second vector represents the intervals contained in the chord label relative to the root. Each interval, except the fourth and fifth, is represented as present (e.g., 2), present as minor (e.g., ♭2), not present (e.g., ★2) or both present (e.g., 2B). For fourths, 4 refers to the perfect fourth and #4 to the augmented fourth. For fifths, we only include present or not (i.e., 5 and ★5). The last vector represents the bass note (or inversion). Calculating the SHIP results in a concatenation of three probability density functions that describe the distribution of perceived harmonic content among the annotators

HIP captures this intuition by reducing a chord label to its root note, interval profile, and bass note.

Previous research by Koops et al. [15] proposed to use HIP containing a concatenation of three one-hot vectors: roots, thirds sevenths. In this paper, we use a larger HIP, to account for more intervals and inversions. The HIP in this research contains a concatenation of a couple of vectors: roots, intervals, and inversions. With these vectors, we can encode all possible chord labels. 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 4 and denotes if the chord denoted by the chord label contains a major second (2), minor second (♭2), both major and minor (2B) or no second (★2) relative to the root note. The third vector, also of size 4 represents the same, but for thirds (3, ♭3, 3&♭3, ★3). The fourth vector denotes the fourth intervals (♯4, ♭4, ♯4&♭4, ★4). The fifth vector, of size 2, denotes whether the chord contains a fifth (5 or ★5). The sixth vector, size 4, denotes the sixth intervals (♯6, ♭6, ♯6&♭6, ★6). The seventh vector denotes the sevenths intervals (♯7, ♭7, ♯7&♭7, ★7). With these intervals, combined with the root and bass, we can represent and evaluate every chord label in the CASD and *Billboard* using the standard MIREX ACE evaluation measures. The HIP can be extended or reduced to other intervals as well. Figure 4 shows a number of example chord labels and their HIP equivalent.

3.1.2 Shared harmonic interval profile (SHIP)

To capture the harmonic content shared among different HIP, we create a *Shared harmonic interval profile* (SHIP) by computing the columnwise arithmetic mean of multiple HIPs. The last row of Fig. 4 shows an example of the SHIP created from the HIPs above it. By averaging, we create a fuzzy chord representation: the SHIP essentially contains a concatenation of probability density functions for the root, bass, and each stacked intervals. These probability density functions express the average harmonic content shared among the annotators' chord labels. Instead of the classical ACE approach of trying to estimate just a single chord label, we propose to estimate this fuzzy representation from audio.

3.2 Audio feature extraction

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 [28]. CQT have proven to be very successful audio features for ACE and are commonly used in state-of-the-art ACE systems [17]. We calculate these CQT features with a hop length of 4096 samples, a minimum frequency of ≈ 32.7 Hz (C1 note), $24 \times 8 = 192$ bins, with 24 bins per octave, from an audio signal with a sample rate of 44.1 Khz. This way we can capture pitches spanning from low notes to 8 octaves above C1. Two bins per semitone allow for slight tuning variations. In Fig. 1, several examples of CQT can be found.

3.3 Deep learning shared harmonic interval profiles

We use a dense deep neural network to learn SHIP from CQT. Experiments with other deep architectures, such as convolutional neural networks, did not give us significantly better results, yet significantly increased computation time. Based on preliminary experiments, a dense architecture with three leaky ($\alpha = 0.3$) hidden rectifier unit layers of sizes 512, 512, and 512 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 [1, 29]. Therefore, the input to our DNN is a window (i.e., patch) of CQT features from which we learn the SHIP corresponding to the center frame. From experiments we 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 an input layer size of $192 \times 15 = 2880$. The output layer consists of 48 units corresponding with the SHIP features (containing roots, thirds, fifths, sixths, and sevenths) as explained above.

We train the DNN framewise 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 the central audio frame of the 15-frame window). We train the hyper-parameters of the network using mini-batch (size 512) training using the ADAM update rule [13]. Early stopping is applied when validation accuracy does not increase after 10 epochs. During testing, we use the derived SHIP from unseen CQT (which we will refer to as audio-SHIP) to rank the chord labels from an annotators' vocabulary to find the most appropriate chord label for that annotator.

4 Annotator vocabulary-based chord label estimation

The audio-SHIP features are used to rank the chord labels from a given vocabulary. For a chord label **L**, its HIP **h** contains exactly eight ones, corresponding to the root, seconds, thirds, fourths, fifths, sixths sevenths, and bass of the label **L**. From the SHIP **A** of a particular audio frame, we project out eight values for which **h** contains ones (**h**(**A**)). The product of these values is then interpreted as the combined probability $\mathbf{CP}(= \Pi\ \mathbf{h}(\mathbf{A}))$ 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 with **A**.

For the chord label examples in Fig. 4, the products of the nonzero values of the pointwise multiplications ≈ 0.13, 0.29, and 0.07 for `G:min`, `Eb:maj/3`, `Eb:maj`, respectively. If we consider these chord labels to be a vocabulary and normalize the values to sum to unity, we obtain probabilities ≈ 0.27, 0.59, 0.14, respectively. Given multiple annotators providing reference annotations, their chord-label vocabularies, and audio-SHIP, we can now generate annotator specific chord labels.

5 Evaluation

SHIP models multiple (related) chords for a single frame, e.g., the SHIP in Fig. 4 models different closely related chords. For the purpose of personalization, we want to present the annotator with only the chords they understand and prefer. In this research, we assume that the annotators want to be provided chord labels from their own vocabulary. Doing so will produce a high chord label accuracy for each annotator. For example, if an annotator's vocabulary does not contain a `G:maj7` but does contain a `G:maj`, and both are probably from an audio-SHIP, we like to present the latter. In this paper, we evaluate our DNN ACE personalization approach, and the audio-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 from the CASD with training on a commonly used single reference annotation from the *Billboard* dataset. In the first case we train a DNN (DNN_{CASD}) on SHIPs derived from

Table 1 Chord label personalization accuracies for annotators A1, A2, A3 and A4 using the multiple reference annotations from the CASD (DNN_{CASD}), compared to using a single *Billboard* (BB) reference annotation (DNN_{BB})

	DNN_{CASD}						DNN_{BB}					
	A1	A2	A3	A4	$\bar{x}$	BB	A1	A2	A3	A4	$\bar{x}$	BB
ROOT	.78	.79	.77	.65	.75	.77	.74	.74	.74	.61	.71	.85
MAJMIN	.74	.75	.76	.63	.72	.74	.69	.68	.72	.57	.66	.81
MAJMIN_INV	.73	.73	.74	.53	.68	.72	.68	.66	.70	.49	.63	.80
THIRDS	.73	.74	.72	.62	.70	.71	.69	.67	.69	.56	.65	.80
THIRDS_INV	.71	.72	.69	.52	.66	.68	.67	.66	.65	.48	.62	.78
TRIADS	.70	.72	.70	.60	.68	.67	.66	.65	.66	.55	.63	.77
TRIADS_INV	.69	.71	.67	.51	.65	.65	.64	.64	.64	.47	.60	.76
MIREX	.73	.73	.72	.61	.70	.71	.68	.67	.70	.57	.65	.81
TETRADS	.60	.63	.56	.54	.58	.54	.55	.54	.51	.48	.52	.71
TETRADS_INV	.59	.62	.54	.46	.55	.52	.54	.53	.50	.41	.49	.71
SEVENTHS	.64	.66	.63	.57	.62	.62	.58	.56	.57	.50	.55	.75
SEVENTHS_INV	.63	.65	.60	.48	.59	.61	.56	.55	.55	.42	.52	.75
$\bar{x}$	.69	.70	.68	.56	.66	.66	.64	.63	.64	.51	.60	.78

A SHIP with intervals 2,3,4,5,6,7 is used in both cases. Accuracies are consistently higher for all annotators A1, A2, A3, and A4 when taking into account multiple reference annotations (DNN_{CASD}) compared to using just a single reference annotation (DNN_{BB}). Nevertheless, the best performances are found for BB on the right, indicating that DNN_{BB} is tuned to the particular BB annotation and is not ideal for chord-label personalization

Table 2 Chord label personalization accuracies for annotators A1, A2, A3 and A4 using the multiple reference annotations from the CASD using extended chroma proposed in [17, 21]

	DNN_{chroma}					
	A1	A2	A3	A4	$\bar{x}$	BB
ROOT	.50	.48	.48	.46	.48	.49
MAJMIN	.39	.46	.49	.43	.44	.41
MAJMIN_INV	.38	.44	.47	.38	.42	.41
THIRDS	.45	.45	.45	.44	.45	.49
THIRDS_INV	.44	.44	.42	.38	.42	.48
TRIADS	.37	.45	.44	.42	.42	.39
TRIADS_INV	.36	.43	.42	.37	.40	.38
MIREX	.30	.41	.33	.36	.35	.31
TETRADS	.29	.39	.31	.31	.33	.30
TETRADS_INV	.37	.45	.45	.43	.42	.40
SEVENTHS	.32	.43	.39	.38	.38	.32
SEVENTHS_INV	.31	.41	.36	.33	.35	.32
$\bar{x}$	.37	.44	.42	.41	.40	.39

Similar results were found using regular chroma. All results are significantly lower than the results from Table 1

the *Chordify annotator subjectivity dataset* containing fifty popular songs annotated by four expert annotators. In the second case, we train a DNN (DNN_{BB}) on the HIP of the *Billboard* (BB) single reference annotation [19]. For both systems, we derive audio-SHIP and evaluate the systems on every individual annotator. We hypothesize that training a system on SHIP based on multiple reference annotations from CASD captures the annotator subjectivity of these annotations and leads to better chord labels per annotator than training the same system on just a single (BB) reference annotation.

For each song in the dataset, we calculate CQT and SHIP features. We randomly divide our CQT and SHIP dataset framewise into 66% training (163800 frames), 10% evaluation (24818 frames) and 24% testing (59563 frames) sets. To account for the "album effect", we make sure songs are not split across sets [18, 32]. For the testing set, for each annotator, we create chord labels from the deep learned SHIP based on the annotators' vocabulary from the test set. We use the standard MIREX chord label evaluation methods that are introduced in Sect. 2.2 to compare the output of our system with the reference annotation from an annotator [25].

6 Results

Table 1 presents the results of chord label personalization using the DNN_{CASD} on the left and the DNN_{BB} on the right. In the columns A1, A2, A3 and A4, personalization accuracies can be found for each chord label granularity. The $\bar{x}$-column presents the average accuracy per granularity, while the $\bar{x}$-row presents the average accuracy per annotator. In addition to chord label personalization for each annotator, we also treated the BB annotation as an annotator and performed chord label personalization. This way, we can investigate what role a single reference annotation can play in providing personalized chord labels.

The DNN_{CASD} columns of Table 1 for each annotator show average accuracies ranging between .71 and .60. Comparable high-accuracy scores for each annotator show that the system is able to learn an audio-SHIP representation that (i) is meaningful for all annotators (ii) from which chord labels can be accurately personalized for each annotator.

Comparing the annotator columns of DNN_{CASD} and DNN_{BB}, we see that for each annotator DNN_{CASD} models annotator subjectivity better than DNN_{BB}. The average accuracies of DNN_{CASD} are per annotator on average 2.5 to 9.6 percentage points higher than DNN_{BB}, showing that for these annotators, using only BB is not enough to accurately represent the variability found in human annotations. The difference in accuracy between the networks is significant ($p \ll 0.01$). Furthermore, comparing the $\bar{x}$-columns shows that per granularity, DNN_{CASD} outperforms DNN_{BB} on average by 6 percentage points, with up to 7 percentage points for the most complex chord granularity (SEVENTHS). The BB-column shows the results comparable to the average annotator ($\bar{x}$), which shows that even though the DNN_{CASD} system did not take the BB annotation into account when training, it is nevertheless capable of providing personalized chord labels for BB.

This doesn't mean that DNN_{BB} is not trained well. On the contrary, the DNN_{BB} table shows that it provides much better chord labels for the BB than for the four annotators. Comparing the $\bar{x}$ and BB columns for DNN_{BB}, we find that on average, DNN_{BB} provides better chord labels for BB than for $\bar{x}$ with about 19 percentage point difference. For SEVENTHS, this difference is even 25 percentage points. This shows that the system trained solely on the BB is good at providing chord labels only when we evaluate it on that same dataset. It seems that DNN_{BB} has tuned itself to the particular subjective reference annotation found in the *Billboard* dataset. Furthermore, if we assume the existence of a subjectivity ceiling as suggested in Sect. 2.2, we can see that the results from the DNN_{BB} evaluated on the BB lie (far) beyond this ceiling found in the CASD, proving further proof of overfitting to the BB. If we make the same comparison for the DNN_{CASD}, we find accuracies that are very close to the subjectivity ceiling.

To test the contribution of our proposed SHIP representation to our system, we compared our system to one in

Table 3 Average ($\bar{x}$) and standard deviation (σ) pairwise agreement results between all annotators

	ROOT	MAJMIN	MIREX	THIRDS	TRIADS	TETRADS	SEVENTHS
$\bar{x}$	.76	.73	.74	.74	.71	.57	.6
σ	.19	.2	.18	.19	.21	.24	.24

Agreement decreases with increased chord granularity

which SHIP is replaced by common 12-bin chroma and extended chroma as proposed by McFee et al. and Korzeniowski et al. [17, 21]. After integrating them, we deep learn these representations in the same fashion as we learn SHIP, and also perform vocabulary filtering as explained in Sect. 4 to provide each annotator and the BB with chord labels. None of these representations yielded satisfactory results, with low chord-label personalization accuracies ranging between .29 and .50 for all chord label granularities for all annotators and the BB. The results for using extended chroma can be found in Table 2. This table shows that significantly worse results are obtained using extended chroma compared to using SHIP in Table 1. Similar worse results were obtained using common chroma.

Overall, the results from Table 1 show that chord label personalization using the SHIP is improved by taking into account multiple reference annotations, while personalization using a commonly used single reference annotation yields significantly worse results per annotator. Furthermore, the poor personalization results from using the common chroma representation show that, in contrast to chroma, SHIP is capable of capturing the shared harmonic content needed for personalization when the average pairwise agreement between annotators is relatively low (Table 3).

7 Conclusions and discussion

We presented an automatic chord-label estimation and personalization system that takes into account chord labels from multiple reference annotations. Our system uses an interval-based chord label representation that captures the shared subjectivity between annotators and an annotators' specific chord label vocabulary to find appropriate chord labels for that particular annotator. An experiment showed that for each annotator, better chord labels can be provided using this system compared to the conventional ACE approach of using only a single *"ground truth"* reference annotation. Furthermore, we found that chord-label personalization using the common chroma representation yields poor results.

To test the scalability of our system, our experiment needs to be repeated on a larger dataset, with more songs and more annotators. Unfortunately, chord label datasets with multiple reference annotations are scarce, and creating such datasets is costly and time-consuming. Repetition of a similar experiment as presented in this paper on a larger dataset with instrument/proficiency/cultural-specific annotations from different annotators would shed light on whether our system generalizes to providing chord label annotations for such different contexts. Furthermore, finding which audio features correlate with subjectivity and disagreement could provide insight into the audio structures that influence these aspects, and inform future systems to predict them directly from recorded sound.

We showed the relative improvement of using multiple versus a single reference annotation. Other deep architectures and audio features should be explored to investigate if there are techniques that are perhaps even better for learning fuzzy chord representations from audio. Although our ACE accuracies are below the current state of the art, a comparison to state-of-the-art ACE results might not be completely fair. The results from our experiments, as well as the experiments from [24], show that overfitting is a problem in ACE when using a single reference annotation. We believe that the agreement from a subjectivity ceiling should inform future ACE evaluations. From the results presented in this paper, we consider chord label personalization as the next step in the evolution of ACE systems.

Acknowledgements H.V. Koops thanks the organizers of the Systematic Approaches to Deep Learning in Audio workshop and gratefully acknowledges the hospitality and financial support of the Erwin Schrödinger International Institute for Mathematics and Physics, Vienna. H.V. Koops and A. Volk are supported by the Netherlands Organization for Scientific Research, through the NWO-VIDI-Grant 276-35-001 to A. Volk.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest

References

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

2. Burgoyne JA, Wild J, Fujinaga I (2011) An expert ground truth set for audio chord recognition and music analysis. In: Proceedings of the 12th international society for music information retrieval conference, ISMIR, vol 11, pp 633–638
3. Burgoyne JA, Wild J, Fujinaga I (2013) Compositional data analysis of harmonic structures in popular music. In: International conference on mathematics and computation in music. Springer, pp 52–63
4. Chuan CH, Chew E (2007) A hybrid system for automatic generation of style-specific accompaniment. In: Proceedings of the 4th international joint workshop on computational creativity, pp 57–64
5. de Haas WB, Volk A, Wiering F (2013) Structural segmentation of music based on repeated harmonies. In: IEEE international symposium on multimedia (ISM). IEEE, pp 255–258
6. Dong XL, Berti-Equille L, Srivastava D (2009) Integrating conflicting data: the role of source dependence. Proc VLDB Endow 2(1):550–561
7. Flexer A (2014) On inter-rater agreement in audio music similarity. In: Proceedings of the 15th international society for music information retrieval conference, ISMIR, pp 245–250
8. Gauvin HL (2015) The times they were a-changin: a database-driven approach to the evolution of musical syntax in popular music from the 1960s. Empir Musicol Rev 10(3):215–238
9. Harte C, Sandler MB, Abdallah SA, Gómez E (2005) Symbolic representation of musical chords: a proposed syntax for text annotations. In: Proceedings of the 6th international society for music information retrieval conference, ISMIR, vol 5, pp 66–71
10. Humphrey EJ, Bello JP (2015) Four timely insights on automatic chord estimation. In: Proceedings of the 16th international society for music information retrieval conference, ISMIR, pp 673–679
11. Kaliakatsos-Papakostas M, Cambouropoulos E, Kühnberger KU, Kutz O, Smaill A (2014) Concept invention and music: creating novel harmonies via conceptual blending. In: In Proceedings of the 9th conference on interdisciplinary musicology (CIM2014), CIM2014. Citeseer
12. Karatzoglou A, Amatriain X, Baltrunas L, Oliver N (2010) Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In: Proceedings of the fourth ACM conference on Recommender systems. ACM, pp 79–86
13. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations, ICLR
14. Koops HV, de Haas WB, Bountouridis D, Volk A (2016) Integration and quality assessment of heterogeneous chord sequences using data fusion. In: Proceedings of the 17th international society for music information retrieval conference, ISMIR, New York, USA, pp 178–184
15. Koops HV, de Haas WB, Bransen J, Volk A (2017) Chord label personalization through deep learning of integrated harmonic interval-based representations. In: Proceedings of the first workshop on deep learning for music, Anchorage, USA, 18–19 May, 2017, pp 19–25
16. Koops HV, de Haas WB, Burgoyne JA, Bransen J, Volk A (2017) Harmonic subjectivity in popular music. Technical report of UU-CS-2017-018, Department of Information and Computing Sciences, Utrecht University
17. Korzeniowski F, Widmer G (2016) Feature learning for chord recognition: the deep chroma extractor. In: Proceedings of the 17th international society for music information retrieval conference (ISMIR), New York, USA pp 37–43
18. Kruspe A, Lukashevich H, Abeßer J (2011) Artist filtering for non-western music classification. In: Proceedings of the 6th audio mostly conference: a conference on interaction with sound, AM '11. ACM, New York, NY, USA, pp 82–87. https://doi.org/10.1145/2095667.2095679
19. Mauch M, Cannam C, Davies M, Dixon S, Harte C, Kolozali S, Tidhar D, Sandler M (2009) Omras2 metadata project 2009. In: Late-breaking demo session at 10th international society for music information retrieval conference, ISMIR
20. Mauch M, MacCallum RM, Levy M, Leroi AM (2015) The evolution of popular music: USA 1960–2010. R Soc Open Sci 2(5):150081
21. McFee B, Bello J (2017) Structured training for large-vocabulary chord recognition. In: 18th international society for music information retrieval conference, ISMIR
22. McVicar M, Santos-Rodríguez R, Ni Y, De Bie T (2014) Automatic chord estimation from audio: a review of the state of the art. IEEE/ACM Trans Audio Speech Lang Process: TASLP 22(2):556–575
23. Meyer LB (1957) Meaning in music and information theory. J Aesthet Art Crit 15(4):412–424
24. Ni Y, McVicar M, Santos-Rodriguez R, De Bie T (2013) Understanding effects of subjectivity in measuring chord estimation accuracy. IEEE Trans Audio Speech Lang Process 21(12):2607–2615
25. Raffel C, McFee B, Humphrey EJ, Salamon J, Nieto O, Liang D, Ellis DP, Raffel C (2014) Mir_eval: a transparent implementation of common MIR metrics. In: Proceedings of the 15th international society for music information retrieval conference, ISMIR, pp 367–372
26. Rendle S, Schmidt-Thieme L (2010) Pairwise interaction tensor factorization for personalized tag recommendation. In: Proceedings of the third ACM international conference on Web search and data mining. ACM, pp 81–90
27. Schoenberg A (1978) Theory of harmony. University of California Press, Berkeley
28. Schörkhuber C, Klapuri A (2010) Constant-q transform toolbox for music processing. In: Proceedings of the 7th sound and music computing conference, Barcelona, Spain
29. Sigtia S, Boulanger-Lewandowski N, Dixon S (2015) Audio chord recognition with a hybrid recurrent neural network. In: Proceedings of the 16th international society for music information retrieval conference, ISMIR, pp 127–133
30. Van Balen J, Burgoyne JA, Bountouridis D, Müllensiefen D, Veltkamp RC (2015) Corpus analysis tools for computational hook discovery. In: Proceedings of the 16th international society for music information retrieval conference, ISMIR, pp 227–233
31. Van Balen J, Burgoyne JA, Wiering F, Veltkamp RC (2013) An analysis of chorus features in popular song. In: Proceedings of the 14th international society for music information retrieval conference, ISMIR, pp 107–112
32. Vatolkin I, Rudolph G, Weihs C (2015) Evaluation of album effect for feature selection in music genre recognition. In: Proceedings of the 16th international society for music information retrieval conference, ISMIR, Málaga, Spain, pp 169–175
33. Wang Y, Liu J, Dong X, Liu T, Huang Y (2012) Personalized paper recommendation based on user historical behavior. In: Zhou M, Zhou G, Zhao D, Liu Q, Zou L (eds) Natural language processing and Chinese computing. Springer, Berlin, Heidelberg, pp 1–12
34. Xiong L, Chen X, Huang TK, Schneider J, Carbonell JG (2010) Temporal collaborative filtering with bayesian probabilistic tensor factorization. In: Proceedings of the 2010 SIAM international conference on data mining. SIAM, pp 211–222
35. Zheng VW, Cao B, Zheng Y, Xie X, Yang Q (2010) Collaborative filtering meets mobile recommendation: a user-centered approach. In: AAAI, vol 10, pp 236–241