IMPROVED TEMPLATE BASED CHORD RECOGNITION USING THE CRP FEATURE

Ken O'Hanlon, Sebastian Ewert, Johan Pauwels, Mark B.Sandler

Centre for Digital Music
Queen Mary University of London, UK

ABSTRACT

The task of chord recognition in music signals is often based upon pattern matching in chromagrams. Many variants of chroma exist and quality of chord recognition is related to the feature employed. Chroma Reduced Pitch (CRP) features are interesting in this context as they were designed to improve timbre invariance for the purpose of query recognition. Their reapplication to chord recognition, however, has not been successful in previous studies. We consider that the default parametrisation of CRP attenuates some tonal information, as well as timbral, and consider alternatives to this default. We also provide a variant of a recently proposed compositional chroma feature, adapted for music pieces, rather than one instrument. Experiments described show improved results compared to existing features.

Index Terms— Chromagram, chord recognition

1. INTRODUCTION

Chroma affords a summary feature of the tonal content of a musical frame in a 12-dimensional pitch class vector [1], and has been applied for a variety of tasks in the field of music information retrieval, such as chord recognition [2] [3], key estimation [4] and query retrieval [5] [6]. A time-chroma representation, referred to as a chromagram, is often formed through summation of time-pitch elements that may themselves be appropriated through addition of time-frequency elements. Many chromagram variants have been proposed which are often differentiated by transforms of the underlying pitch-time representations. Log compression of the pitch representation was proposed in [7], while a group sparse signal representation was employed in [8]. A harmonic product spectrum was employed in [9]. Mauch and Dixon [10] propose using approximate transcription as the pitch representation, while recently joint transcription and chord recognition was performed in [11]. Of particular interest to this paper are the Chroma Reduced Pitch (CRP)[6] which applies a high pass filter to the log-compressed chroma and a compositional approach [12] that remodels a chromagram through decomposition with a dictionary of single note chroma vectors.

This research was funded by EPSRC Program Grant EP/L019981/1 and AHRC Grant AH/L006820/1.

Perhaps the most popular application of chroma is for the purpose of chord recognition, a task that is generally reduced to recognition of major or minor classes of chords. A recent paper by Cho and Bello [2] compared many different chroma variants for this task, using both a learning based approach with Gaussian mixture models, and template based chord recognition. Results were given for a basic framewise classification, and for several filtering approaches of which it was found that HMM filtering performed best, with GMM-based classification performing better than the template-based approaches. It was found that the log compressed chromagram combined with a spectral weighting performed best, and the importance of the chroma feature in chord recognition was emphasised, as it was found that pre- and post-processing steps have little effect on the order of performance relative to the feature employed. The depth and detail of this paper has led to it being used as a reference for recent chroma based chord recognition research e.g. [13]. More recently, the use of deep neural networks has been proposed for the chord recognition problem [14] [13]. While improvements in chord recognition are observed using the DNN approaches, chroma is still a popular, flexible feature, and its semantic interpretability affords simple template matching approaches for chord recognition. However, the lack of a reasonable mechanism for performing no-chord detection with template-based approaches is noted [2].

Considering the importance of chroma, we revisit variants of two previously proposed chroma features, the CRP [6] and the compositional chroma [12] for the task of chord recognition. We consider different variants of the pitch feature reduction step for the CRP, which has previously been reported to perform quite poorly for chord recogntion [2] and introduce regularisation for the compositional approach [12] which has not been applied previously on a realistic dataset. Experimental results show improvement over the baseline log compressed weighted spectrum chroma which is found to perform best in [2].

In the rest of this paper, we first describe a template-based approach to chord recognition, which mostly adheres to the methodology used in [2]. We then describe chroma features, in particular re-introducing the CRP and compositional chroma features, before giving experimental validation. Finally, we conclude with pointers to future work.

2. TEMPLATE-BASED CHORD RECOGNITION

A template-based recognition system based, similar to that employed in [2] is used. Binary chord templates are formed for each chord simply through setting the pitch classes relating to the notes in the chord to one and setting all other dimensions to zero. For the major chord the 1st, 5th and 8th pitch classes are active:

$$\tau_{maj} = [1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0]^T$$

while the minor chord template, $\tau_{\min}$, assigns the 1st, 4rd and 8th pitch classes active. A dictionary, $\mathbf{T} \in \mathbb{R}^{12\times K}$, is constructed in which each column, $\mathbf{t}_k$, is a chord template formed by circular shifting of either τ_{maj} or τ_{maj} with appropriate labelling. Classification is performed by simple multiplication of the templates with the most correlated template selected :

$$\gamma_n = \arg\max_k \ \mathbf{t}_k^T \mathbf{c}_n \tag{1}$$

where $\mathbf{c}_n$ denotes the nth frame of the chromagram $\mathbf{C} \in \mathbb{R}^{12\times N}$ and γ_n is the chord selected at the nth frame.

A probabilistic interpretation of template-based chord recognition is given in [3]. This is effected by considering that the minimum measure of fit between a template and a chroma feature can be expressed as a probability drawn from an underlying distribution. While Gamma and Poisson distributions are also considered in [3], we employ the Gaussian as the underlying distribution as initial experiments indicate little difference between the different distributions in this regard in the context of our approaches. Given ℓ_2 normalised chroma vectors and chord templates a likelihood can be assigned

$$P(\gamma_n = k) = e^{(\mathbf{t}_k^T \mathbf{c}_n - 1)/\sigma^2} \tag{2}$$

where σ^2 is a user tunable parameter, in which case, similar to (1), $\gamma_n = \arg\max_k P(\gamma_n = k)$.

2.1. HMM-based smoothing

HMM-based classification is employed as a final step in the chord recognition system, after normalisation of the likelihoods at each frame. A transition matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$ with homogenous off-diagonal entries is formed, such that $[A]_{k,k} = 1/(1 + \phi(K-1))$, and off-diagonal entries $[A]_{k,j} = \phi/(1 + \phi(K-1))$, where ϕ is a tunable parameter. Similar to [2], the assumption that $\mathbf{A}$ relates the actual transition probabilities is ignored, and the HMM is considered simply a smoother of the quasi-probabilities in an optimisation problem. We note that selection of a good value of ϕ is sensitive to the value of σ^2 employed in the model of fit (2), and also to the particular chroma feature employed. However, we observe that adaptation of the quasi-probabilistic approach with a suitable value of σ^2 selected renders the system less sensitive to the selection of the value of ϕ than the cosine value based approach (1) which is seen to be effective only in a small locality of ϕ when employed in [2].

3. CHROMA FEATURES

As a time-frequency representation, we employ the Constant-Q Transform (CQT) toolbox [15], producing a magnitude CQT, $\mathbf{Y}$, with 200ms windows and 50% overlap. A 36 bins per octave resolution is specified for the CQT from which a pitch feature, $\mathbf{F}$, is derived by placing a Gaussian window over three bins centred on the bin of the expected frequency on the pitch scale as in [2], which we find to be more effective than the 88-pitched filterbank employed in [16]. An additive chromagram, C^A, can be derived by summing the magnitudes of the pitch bins over O octaves:

$$[C]_{p,n} = \sum_{o=1}^{O} [F]_{p+12o,n} \tag{3}$$

Often a power spectrogram, $\mathbf{Y}^{[2]}$, is used in (3), however we find that the use of the magnitude leads to improvements in chord recognition. The log chromagram, C^L, [7] is derived by log compressing the power spectra before addition (3)

$$[F]_{i,n} \longleftarrow \log(1 + \alpha [F]_{i,n}^2 / \max_i [F]_{i,n}^2) \tag{4}$$

where $\alpha = 1000$ is employed in accordance with [2], which we also find to be optimal.

A common technique is to employ a spectral weighting, using a Gaussian window centred on e.g. middle C, or MIDI note 60, as in [10] and [2], where it is considered an overtone removal strategy. Improved chord detection is reported employing the spectral window with various chroma estimation methods [2]. However, we consider this approach should be employed with caution as it may be misleading when the spectral energy is centred significantly lower, or higher than middle C, although it is probably safe to assume that the tonal centre of much e.g. pop music is emphasised by this window.

3.1. Compositional Chroma

We proposed a compositional model of chroma [12] that decomposes an additive chromagram with a dictionary, $\mathbf{D} \in \mathbb{R}^{12\times 12}$, in which each column is a circularly shifted version of the chroma vector of a synthesised note, containing ten harmonic partials in which the amplitude of the nth partial is given by 0.6^n. A Powered Euclidean Distance (PED)

$$\mathcal{C}_P^{(\eta)}(\mathbf{C}|\mathbf{DX}) = \sum_{m,n} ([C]_{m,n}^{\eta} - [DX]_{m,n}^{\eta})^2 \tag{5}$$

was employed as the cost function in NMF-based regression to derive the compositional chroma feature $\mathbf{X} \in \mathbb{R}_+^{12\times N}$. The experiments described in [12] were on a database of MIDI chords played in isolation, and improved performance was found for small values of $\eta \to 0$. A more realistic dataset is employed here, and comparison with the more common β-divergence again showed relative improvement using PED,

which is related to β-divergence through membership of the $\alpha\beta$ divergence family [17], with similar error function scaling e.g. linear scaling is found for $\beta = 2\eta = 1$ [18]. In the current context, we introduce a regulariser to the decomposition, resulting in the cost function $\mathcal{C}_P^{(\eta)}(\mathbf{C}|\mathbf{DX}) + \lambda \sum_n \|\mathbf{x}_n\|_2^2$. Optimisation of the regularised cost function is performed through iterations of a monotonic multiplicative update derived using majorisation minimisation method as in [19] :

$$\mathbf{X} \leftarrow \mathbf{X} \oslash \left[\frac{\mathbf{D}^T[\mathbf{C}^{[\eta]} \oslash [\mathbf{DX}]^{[\eta-1]}]}{\mathbf{D}^T[[\mathbf{DX}]^{[2\eta-1]}] + 2\lambda\mathbf{X}} \right]^{\left[\frac{1}{2-\eta}\right]}. \quad (6)$$

3.2. CRP features

The Chroma Reduced Pitch (CRP) feature was originally proposed in [6], and was designed in order to provide a timbre-invariant feature, which was effected through two separate mechanisms. First, the log compression step (4) is employed, which itself induces timbre-invariance [7]. Second, a high pass filter is applied to the pitch vector, based on the premise that much of the timbral information is present in the lower frequencies, similar to the approach for MFCCs [6].

The high pass filter in CRP is effected through discarding the lower frequency elements of a DCT applied to the pitch vector, before reforming the filtered pitch vector. In particular, a 120 dimensional pitch vector is used, representing MIDI notes $\{1, ..., 120\}$ although only the dimensions $p \in \{21, ..., 108\}$ are populated representing the notes on a piano scale. This design specification of CRP affords cosine vectors that repeat on a octave basis, found in the set of dimensions $\mathcal{O} = \{21, 41, 61, 81, 101, 120\}$ of the DCT dictionary, completing $\{1, ..., 6\}$ cycles in each octave, respectively. More specifically the DCTs in $\hat{\mathcal{O}} = \mathcal{O} \backslash 21$ possess sub-octave cycles whitch emphasise repetitive tonal structure. Although it is moot to explicitly classify a given DCT dimension as tonal, or timbral, in the given context, it may be considered that the dimensions in $\mathcal{O}$ or $\hat{\mathcal{O}}$ carry predominately tonal information. In some cases, these dimensions have a strong relationship with particular tonal intervals, e.g. DCT(61), can be considered to have a strong relationship to the major 3 interval of four semitones [6], as it repeats 3 times per octave.

Experiments described in [6] focus on the application of query recognition, for which timbre-invariance is particularly desirable, and results show that discarding the lower 55 DCT coefficients, referred to as CRP(55), is the optimal strategy for this task. However, CRP(55) performs relatively poorly for template-based and GMM-based chord recognition [2] [12]. although it performs well for this task using nearest-neighbour classification [12], indicating greater timbre-invariance. Other CRP variants proposed in [6] emphasise the DCTs with sub-octave cycles, $\hat{\mathcal{O}}$, rather than employing a high-pass filter. One variant, referred to here as CRP(CS), incorporates sinusoids at the same frequencies as the cosine vectors in $\hat{\mathcal{O}}$ in order to account for phase shifting.

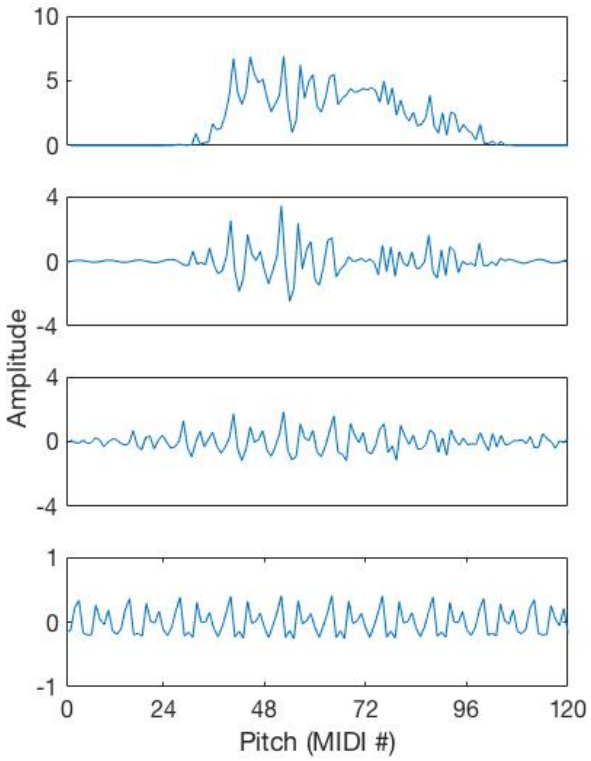

Fig. 1. Pitch features for the same audio frame. From top, Log compressed chroma, CRP(35), CRP(S), CRP(CS).

Another variant, CRP(S), includes the DCT vectors adjacent to $\hat{\mathcal{O}}$ in order to approximate phase-shifted cosine vectors.

A large vocabulary of chord classes were employed in experiments in [12] and it was observed, although unreported, that CRP(55) was extremely inconsistent relative to the different classes of chords examined. Here we consider that this inconsistency may be due to lower frequency octave-cyclic components such as DCT(41) being filtered out in CRP(55). We therefore re-examine CRP(35) [6], which discards the 35 lowest frequency DCT components, and CRP(CS) which uses the set $\mathcal{O}$ of complex sinusiods. We also modify the CRP(S) by using five frequency bins centred at each member of $\hat{\mathcal{O}}$, e.g. DCT(41) is represented by DCTs $\{39, ..., 43\}$. The effect of DCT(21) is considering by testing CRP(15) and variants of CRP(S) / (CS) employing the full set of octave repeating DCTs, $\mathcal{O}$, which we notate as $\mathrm{CRP}^{\mathcal{O}}$(S) and $\mathrm{CRP}^{\mathcal{O}}$(CS). While most chroma features are non-negative, the high-pass filtering in CRP induces negativity [6], as seen in Fig.1. We apply a simple linear transform from the range $[-1,\ 1]$ to $[0,\ 1]$, which we observe to effect a slight improvement in template-based chord recognition when the HMM is used.

A weighted log compressed pitch spectrum and its related CRP vectors are shown in Fig. 1. It is observed here that the CRP(35) retains more of the information from the original log

Chroma	F	HMM	F (W)	HMM (W)
Add	50.5	67.6	56.0	72.7
Log	53.9	69.4	58.1	74.4
Comp.	53.7	70.5	57.8	74.5
CRP (55)	52.5	68.9	56.4	73.2
CRP (15)	54.3	70.6	58.0	74.3
$\mathrm{CRP}^{\mathcal{O}}(\mathrm{S})$	56.4	72.5	58.5	75.0
$\mathrm{CRP}^{\mathcal{O}}(\mathrm{CS})$	54.7	70.3	58.8	74.9
CRP (35)	56.2	72.2	60.0	76.0
CRP (S)	**57.9**	**73.8**	60.2	**76.2**
CRP (CS)	56.4	71.7	**60.5**	76.1

Table 1. Overlap score (%) for various chroma features. F denotes framewise classification, HMM denotes hidden Markov model, (W) denotes spectral weighting applied.

spectrum, with less energy in the zero activity pitch frames $\{1, ..., 20, 109, ...120\}$ than the other variants. Furthermore a plateau is seen around MIDI note 70 in the log compressed pitch vector that is preserved in CRP(35) and not in the other CRP variants. The pitch vector for CRP(CS), on the bottom, produces a repeating pattern across its full width, which is to be expected as it consists only of a few harmonic complex sinusoids. Meanwhile CRP(S) strikes a middle ground between the higher reconstruction of the CRP(35) pitch vector and the emphasis on repetition of (CS) pitch vector. As chroma ultimately tries to capture harmonic repetition, the lesser importance of reconstruction in the CRP(S) /(CS) variants may be considered useful, with the potential to e.g. filter out some non-harmonic elements.

4. EXPERIMENTS

Chord recognition experiments comparing the various chroma methods were performed on the popular Beatles dataset, which was annotated by Harte [20]. For the compositional approach the parameters were set, $\eta = 0.05$ and $\lambda = 0.5$. Classification was performed on the typical maj/min basis with all minor variants noted as *minor* and all other chords denoted as *major*. Frames with no chord annotations are ignored, similar to [21]. Comparison is made between an unweighted spectrum and a weighted spectrum as this was seen to create a difference in performance in [2]. Performance is related in terms of the overlap score, which relates the percentage of correctly selected frames. Results are recorded for unfiltered and HMM filtered approaches. Several values of the HMM regularisation parameter ϕ were employed and the optimal results are given for each chroma feature.

The results are shown in Table 1, where it is seen that the compositional method performs similar to the log compressed chroma, and improves over CRP(55) and the additive chroma. The CRP(15) and $\mathrm{CRP}^{\mathcal{O}}$ variants improve over all these methods, while the CRP(35) and associated CRP(S) and CRP(CS) perform the best. Spectral weighting is seen to have a positive effect on results for all features. For the weighed spectra associated CRPs perform very similarly, however an interesting effect is seen when the spectra are not weighted in which case the CRP(S) variants improve over their associated CRPs by around $\sim 2\%$. It is also noteworthy that in this case the CRP(S) performs almost as well as the spectral weighted log compressed chroma feature, previously considered optimal [2], while not requiring the spectral weighting. This may be particularly advantageous when the tonal profile of the music is different to that which the window suggests. It would seem from these results that inclusion of DCT(41) is required in order to perform template based chord recogntion well, in which case CRP(55) is not apt. However, the case of DCT(21) may be less clear cut; while the overlap rate for CRP(15) and $\mathrm{CRP}^{\mathcal{O}}$ is less than for the CRP(35) / (S) / (CS), the difference in performance is not very large, and further investigation on more complex chords may be warranted.

Chroma	HMM (W)
Log	79.2
Comp.	78.8
CRP (55)	79.0
CRP (S)	79.2
CRP (35)	78.8
CRP (35S)	79.6

Table 2. Results for GMM-based classification of different chroma features in chord recognition experiments.

Finally, GMM-based classifiers were trained for some of the various chroma. Training / test partition and cross validation were not performed as it is solely our intention to compare the different features. For each class of chords, training was performed using the ground truth with all chroma vectors aligned to the root chord to simulate a larger training set [5]. Five Gaussians were learned for each chord label, as deemed sufficient in [2], and weighted spectra and HMM filtering were employed. The results are shown in Table 2, where it is seen that CRP variants perform similar to the compositional method and the log compressed chroma. The performance gap between the template and GMM-based approaches is closer when the CRP features are employed.

5. CONCLUSIONS

We have presented improved template-based chord recognition using variants of the CRP feature, and introduced a regularised variant of compositional chroma which performed similar to the previously considered optimal chroma feature. Possible avenues for future work include employing prior information to enhance performance, and development of a no-chord detection system for template-based chord recognition.

6. REFERENCES

[1] T. Fujishima, "Realtime chord recogntion of musical sound: A system using common lisp music," in *Proceedings of the International Computer Music Conference*, 1999, pp. 464–467.

[2] T. Cho and J. P. Bello, "On the relative importance of individual components of chord recognition systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 477–492, Feb 2014.

[3] L. Oudre, Y. Grenier, and C. Fevotte, "Chord recognition by fitting rescaled chroma vectors to chord templates," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2222–2233, Sept 2011.

[4] J. Pauwels and J.-P. Martens, "Combining musicological knowledge about chords and keys in a simultaneous chord and local key estimation system," *Journal of New Music Research*, vol. 43, no. 3, pp. 318–330, 2014.

[5] D. P. W. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 2007, vol. 4, pp. IV–1429–IV–1432.

[6] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 649–662, 2010.

[7] M. Müller, S. Ewert, and S. Kreuzer, "Making chroma features more robust to timbre changes," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009, pp. 1869–1872.

[8] T. Kronvall, M. Juhlin, S. I. Adalbjornsson, and A. Jakobsson, "Sparse chroma estimation for harmonic audio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 2015.

[9] K. Lee, "Automatic chord recognition from audio using enhanced pitch class profile," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2006.

[10] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, 2010.

[11] Y. Ojima, E. Nakamura, K. Itoyama, and K. Yoshii, "A hierarchical Bayesian model of chords, pitches, and spectrograms for multipitch analysis," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2016.

[12] K. O'Hanlon and M. Sandler, "A compositional approach to chroma estimation," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2016.

[13] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[14] S. Sigtia, N. Boulanger-Lewandowski, and S. Dixon, "Audio chord recognition with a hybrid neural network," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2015.

[15] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *Proceedings of the 7th Sound and Music Computing Conference (SMC)*, 2010.

[16] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 215–220.

[17] A. Cichocki, S. Cruces, and S. Amari, "Generalized alpha-beta divergences and their application to robust non-negative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, January 2011.

[18] K. O'Hanlon and M. B. Sandler, "An iterative hard thresholding approach to ℓ_0 sparse Hellinger nmf," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4737–4741.

[19] V. Y. F. Tan and C. Fevotte, "Automatic relevance determination in nonnegative matrix factorization with the beta/-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, 2013.

[20] C. Harte and M. Sandler, "Automatic chord identification using a quantised chromagram," in *Proceedings of the Audio Engineering Society Convention*, 2005.

[21] N. Jiang, P. Grosche, V. Konz, and M. Müller, "Analyzing chroma feature types for automated chord recognition," in *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*, Jul 2011.