COMPOSITIONAL CHROMA ESTIMATION USING POWERED EUCLIDEAN DISTANCE

Ken O'Hanlon and Mark B. Sandler

Centre for Digital Music
Queen Mary University of London

ABSTRACT

Chroma features are a popular tool in musical signal processing and information retrieval tasks, providing a compact representation of the tonal content of a piece of music. A variety of approaches to chroma estimation have been proposed, most of which rely on the summation of related frequency partials. However, frequency partials may be incorrectly assigned due to the log/linear relationship of frequency and pitch. Variations of chroma employing overtone suppression strategies are found in the literature. We propose a compositional model of chroma, which considers a coarse modelling of the effects of overtones in the expected chroma vectors of single notes. Synthetic chord recognition experiments indicate the usefulness of the proposed approach.

***Index Terms*—** Compositional model, chromagram, powered Euclidean distances, non-negative

1. INTRODUCTION

Compositional models [1] are used in many audio signal processing tasks, such as source separation [2] [3] and automatic transcription [4] [5], and typically consider inverse problems which can be solved through the approximation

$$\mathbf{y} \approx \mathbf{A}\mathbf{x} \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^M$ is a given signal or vector, $\mathbf{A} \in \mathbb{R}^{M \times N}$ is a dictionary matrix, and $\mathbf{x} \in \mathbb{R}^N$ is an unknown vector. A matrix variant of (1), $\mathbf{Y} \approx \mathbf{AX}$, is often used for processing audio spectrogram in a non-negative framework, with $\mathbf{Y}, \mathbf{A}, \mathbf{X} \geq 0$ in an entrywise sense, using methods based upon those of Non-negative Matrix Factorisation (NMF) [6]. Many variants of NMF incorporate different priors in the decomposition [2] [7] [4], while a variety of different objective functions have been adapted to multiplicative update NMF algorithms [6] [8] [3] [9].

Chroma, or pitch class profile [10], is a compact representation of audio indicating the activity of semi-tone separated pitch classes, without reference to the pitch height, or octave, in which the class is active. Chroma-time representations, or chromagrams, are popular for music information retrieval tasks, such as chord recognition [11], key estimation [12], and audio similarity search [13]. Typically the salience of each chroma is defined by summation of windowed elements of a spectrum that are assigned to a given pitch class. Fourier transforms may be used; however the pitch ambiguity of low frequency elements led to the use of high-resolution instantaneous frequency [14], or logarithmic frequency scale filterbanks such as the Constant Q-Transform (CQT) [15] and a filterbank with 88 bandpass filters centred on the fundamental frequencies of the keys of grand piano [16].

A problem when performing chroma estimation is that overtones of a fundamental frequency may be assigned to an incorrect chroma bin. For example, the second overtone of a fundamental frequency is applied to the chroma bin of the perfect fifth, as the first overtone of the perfect fifth and the second overtone of the fundamental overlap. Different approaches to countering this problem have been proposed in the literature. A weighted spectrogram using a Gaussian window centred on middle C is used to temper the effects of overtones, assuming that most fundamental frequencies are local to this selected centre [17]. The NNLS chroma of Mauch and Dixon [15] performs an approximate transcription of a CQT transform using a dictionary of harmonic templates, then sums the pitch-time activations to estimate the chromagram. More recently group sparsity has been used to assign harmonic partials to pitch classes [18]. An alternative approach [19] builds chord chroma templates that include the expected contributions of overtones with various distance measures used to compare such templates with calculated chroma vectors.

Perhaps surprisingly, given their popularity in audio processing in general, a compositional approach to chroma has not been previously proposed. In this paper, we propose such an approach using a dictionary of single note chroma templates. We also propose using powered Euclidean distance cost functions, that generalise the Hellinger distance, to perform the decompositions. These proposed approaches, and some matching methods, are described in the next section. Some previous chroma methods are then outlined, before they are compared experimentally to the proposed approach. Results indicate the usefulness of the compositional method. Finally, we conclude with pointers to future work.

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

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

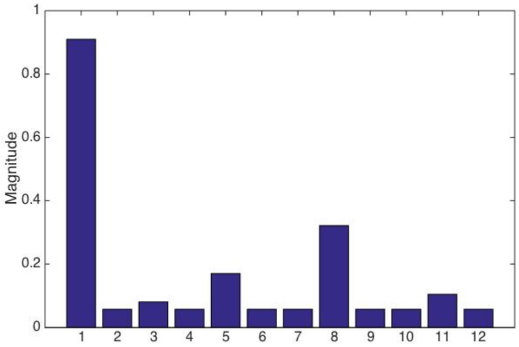

Fig. 1. Chroma template for single note

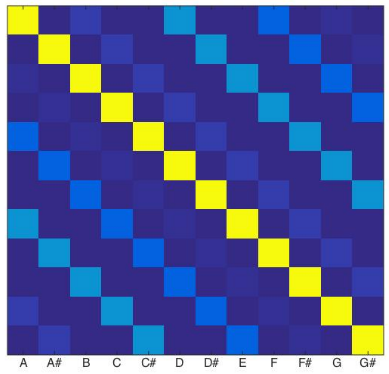

Fig. 2. Chroma dictionary for compositional approach

2. PROPOSED APPROACH

The 88-pitched filterbank, proposed in [16] is used. A recent result [3] states that additivity of audio spectra is better approximated in magnitude spectra than power spectra. Hence, the ℓ_2 norm of the filterbank outputs over a time section is employed as the pitch-time activation, rather than the ℓ_2^2 employed in [16]. From such a pitch vector, or matrix, $\mathbf{P}$ a chroma feature can be calculated

$$\mathbf{y}_m = \sum_o \mathbf{p}_{\hat{m}+12\times o} \tag{2}$$

where o is the octave number, $m \in \{1, .., 12\}$ represent pitch classes $\{A, ..., G\sharp\}$, and $\hat{m} = m + 20$ is the MIDI number of a given note.

2.1. Compositional Chroma Model

In order to employ a compositional model for chroma a single note chroma vector is simulated. This is performed by calculating a harmonic note template, with magnitudes assigned

$$m_k = 0.7^k \tag{3}$$

where m_k is the magnitude of the kth harmonic partial of the fundamental frequency f_0. The magnitudes of the first ten partials are assigned to the nearest chroma bin, with a floor noise added. Fig. 1, shows such a single note chroma template, which could be considered to represent a major chord. A dictionary, $\mathbf{A}$ is formed using shifted versions of the single note chroma template, as displayed in Fig. 2.

2.2. Powered Euclidean Distance

Different measures of fit may be employed to perform the approximation (1). In processing of audio spectrograms, the β-divergence [20] is typically employed. However, we propose to use Powered Euclidean Distance (PED) as an alternative to the β-divergence [20]. Both PED and the β-divergence are special cases of the generalised $\alpha\beta$ divergence [8], a recently proposed flexible family of measures of fit:

$$\mathcal{C}^{(\alpha,\beta)}(\mathbf{y}|\mathbf{z}) = -\frac{1}{\alpha\beta}\sum_m y_m^\alpha z_m^\beta - \frac{\alpha}{\alpha+\beta} y_m^{\alpha+\beta} - \frac{\beta}{\alpha+\beta} z_m^{\alpha+\beta} \tag{4}$$

which also contains the α-divergence [21] as a special case. The β-divergence is given by (4) with $\alpha = 1$.

The powered Euclidean distances

$$\mathcal{C}_{PE}^{\eta}(\mathbf{y}|\mathbf{z}) = \sum_m (y_m^\eta - z_m^\eta)^2 \tag{5}$$

are $\alpha\beta$-divergences with $\alpha = \beta = \eta$, including Euclidean ($\eta = 1$) and Hellinger distances ($\eta = 0.5$), which we previously employed in the context of ℓ_0 sparse NMF in [22]. We note that a scaling parameter used in (4) is omitted from (5).

Empirical results given in [8], [23] indicate that a diagonal structure underlies the $\alpha\beta$ divergence, with some equivalence between divergences of similar summed parameters $\alpha + \beta$, parallel to the α-divergence. However, PED is the only family of $\alpha\beta$-divergences that are symmetric, and indeed PEDs are metrics [8]. Furthermore, unlike β-divergence, and $\alpha\beta$-divergence in general the PED does not consist of ratios of $\mathbf{y}$ and $\mathbf{z}$ as can be seen by comparing (4) and (5). We believe that this makes the PED a more robust measure when the model is less tuned to the data. As the chroma dictionary, displayed in Fig. 2, is quite general, we consider that the PED may be preferable as a measure-of-fit rather than the popular β-divergence for chroma decompositions.

A multiplicative NMF update for the PE distance can be simply derived from the $\alpha\beta$ NMF updates, and is given by

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^T[\mathbf{Y}^{[\eta]} \otimes \mathbf{Z}^{[\eta-1]}]}{\mathbf{A}^T[\mathbf{Z}^{[2\eta-1]}]} \tag{6}$$

where $\mathbf{Z} = \mathbf{AX}$ and $\mathbf{Y}^{[\nu]}$ denotes elementwise exponentiation of $\mathbf{Y}$.

Chord	Active	Chord	Active
maj	1,5,8	min	1,4,8
aug	1,5,9	dim	1,4,7
sus	1,6,8	dom7	1,5,8,11
maj7	1,5,8,12	min7	1,4,8,11
minmaj7	1,4,8,12	min6	1,4,8,10
dim7	1,4,7,10	dim maj7	1,4,7,12
maj+7	1,5,9,12	+7	1,5,9,11

Table 1. List of chords and their active notes on a chromatic scale. The root of the chord is denoted 1.

2.3. Matching methods

A popular approach to chord recognition is to match given chroma vectors with either chord templates or other datapoints, and we propose different matching methods for these two cases.

2.3.1. Template matching for mixed cardinality

Many previous template-based chord recognition implementations have focussed on classifying three note chords [17]. In this case, a binary template can be used for matching with expected active chroma bins set to one and inactive bins to zero, e.g. for a major chord

$$t_k = [1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0]. \tag{7}$$

This approach simply sums the activations in the expected chroma bins through multiplication with the chroma vector, and the selected chord is that which displays the largest sum. As we consider comparison of three-note and four-note chords it is necessary to form a template model that affords comparison of chords of these different cardinalities. To this end, a binary vector such as above is centred simply through subtraction of its mean, e.g. the vector in (7) transforms to

$$t_k^c = \left[\frac{3}{4}, \frac{-1}{4}, \frac{-1}{4}, \frac{-1}{4}, \frac{3}{4}, \frac{-1}{4}, \frac{-1}{4}, \frac{3}{4}, \frac{-1}{4}, \frac{-1}{4}, \frac{-1}{4}, \frac{-1}{4}\right] \tag{8}$$

leading to template vectors that are more discriminative between chords of different note cardinalities.

2.3.2. Minimum Hellinger Distance

For nearest neighbour experiments we propose to use the minimum Hellinger distance (MHD), which we previously proposed in [22], in a matching pursuit type algorithm:

$$\hat{k} = \arg\max_k \frac{\mathbf{d}_k^{[0.5]T}\mathbf{y}^{[0.5]}}{\mathbf{d}_k^T\mathbf{1}} \tag{9}$$

where $\mathbf{d}_k$ is the kth column of a dictionary of chord-labelled chroma vectors of unit sum and $\hat{k}$ denotes the selected chroma vector. The MHD can be derived from the cosine distance, $\frac{\mathbf{d}_k^T\mathbf{y}}{\|\mathbf{d}_k\|_2\|\mathbf{y}_k\|_2}$, between the square root vectors, $\mathbf{d}_k^{[0.5]}$, $\mathbf{y}^{[0.5]}$, or from considering the value of the Hellinger distance at the point where the gradient of the Hellinger distance between two vectors is zero, similar to the scaling approach used for β-divergences in [19]. Again the symmetry of the Hellinger / PED is noted.

3. BASELINE CHROMA METHODS

It has previously been stated that most chroma methods can be classed as either overtone suppressing or timbre-invariant [17]. We consider some well-known methods of each of these classes for comparison with the proposed approach. All methods use the 88 pitch filterbank [16], to allow fair comparison with the compositional approach.

For timbre invariance we compare a chromagram derived by summing from a log compressed spectrum where

$$[P]_{log} = \log_{10}(1 + 100 * [P]). \tag{10}$$

This approach is referred to as LogC in this paper. Another timbre invariant approach, the CRP feature proposed in [16] is compared. The CRP feature is derived by forming the log pitch spectrum (10), to which a DCT is applied. Only the higher frequency elements of the DCT are kept in the recomposition of the log spectrum, which is then summed, in typical fashion, to form the chroma vector.

For overtone suppression, Gaussian filtering, seen to be effective in [17], is used. In this case the pitch vector is filtered by a Gaussian window centred on C4, described by

$$[GW]_q = e^{-\frac{(60-q)^2}{450}} \tag{11}$$

where q is the MIDI note number, with $q = 60$ referring to C4. After filtering, the chroma vector is formed using typical pitch class addition. This approach is referred to as GW in the rest of this paper.

A variant of the Chord Template (CT) matching approach of [19] is also compared. In order to have the closest comparison to the proposed approach, the dictionary of chord templates is formed by multiplying the dictionary of single note chroma vectors, displayed in Fig. 2 by a dictionary of binary chord vectors such as in (7).

4. EXPERIMENTS

A synthetic dataset was formed using chords played on several MIDI instruments, including an electric piano, a grand organ, a harp and a string ensemble. A list of the fourteen different chord types used, and their active notes is outlined in Table 1. Each of these chords was formed for all root notes, leading to 151 different chord classes, as the *aug* and *dim7* have redundant expressions in terms of chroma. Each chord was reproduced with a range of root notes from MIDI 33 to MIDI 92, i.e. 5 octaves.

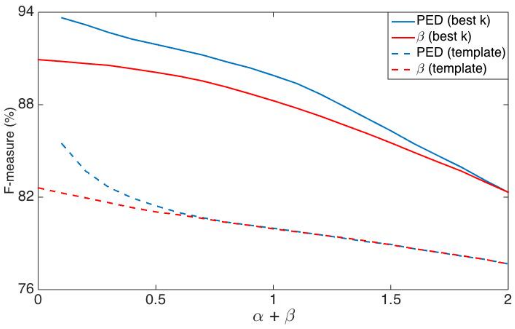

Fig. 3. Comparison of PED and β-divergence for chord recognition, using both the best k-peaks and the template matching approaches

.

4.1. Template matching

The proposed approach was compared to the other chroma methods, CRP, LogC, WG, and CT, in experiments for chord recognition using template-based matching. Several metrics were compared for template based matching, the cosine distance, the cosine distance using centred templates, such as in (8), and the Hellinger distance. An alternative best-k classification was also compared, in which the largest k coefficients in each chroma vector were taken, where k is the number of notes in the known chord displayed by the vector. This can be considered equivalent to template matching with known note cardinality. For the proposed compositional approach, both the PED and β-divergence were compared for a range of values in $\alpha + \beta = [0, ..., 2]$ in steps of 0.1. However, for PED, when $\alpha + \beta = 0$, the update (6) is not valid, and this value was omitted.

In all cases the correct detections, tp, and incorrect detections, fp are labelled, from which the accuracy measure

$$Acc = \frac{tp}{tp + fp} \times 100\%$$

is derived. True positives were noted only when both the chord type and the root were correctly identified.

The results for the comparison of the two generalised divergences is shown in Fig. 3. In both cases, the performance increases monotonically as $\alpha + \beta \to 0$, while the PED is seen to improve upon the β-divergence by around 3% for the optimal values. A comparison of the optimal PED compositional approach with the other approaches is given in Table 2, where a large improvement is seen for the proposed method, both for the best-k and template matching approaches. For all methods using binary templates, the centred cosine distance is seen to improve upon other measures, while a significant improvement is seen when Hellinger distance is employed with the chord templates (CT) [19].

Alg.	Best-K	Cosine	Centred	Hellinger
PED(0.05)	**93.6**	**79.3**	**85.5**	**85.0**
LogC	84.6	71.8	71.7	69.2
CRP	75.3	52.5	58.5	-
GW	83.4	69.0	73.0	69.7
CT	-	70.8	71.7	76.8

Table 2. Comparison of compositional approach with PED to other approaches for best-k, and template matching with cosine distance for binary templates (6), and centred templates, and Hellinger distance for binary templates.

4.2. Nearest Neighbour labelling

For this set of experiments, classification was performed on a nearest neighbour basis, comparing chords of a given instrument with those of other individual instruments. Two alternative approaches are used for the datasets, from which to find a neighbour. In the first approach, for each individual instrument a dataset is formed from the union of all datapoints for all other instruments. In the second case, a mean vector for each root-transposed chord is calculated and shifted, as in Fig. 2. A dictionary is then formed by concatenation of the chord-specific subdictionaries.

All datapoints from the original signals are then compared to the dictionary and classified according to the labelling of the nearest dictionary element. Both the cosine and minimum Hellinger distances were used as measures of similarity. Again, accuracy was defined relative to the numbers of correct and incorrect detections.

The results are given in Table 3, where it is seen that the proposed PED based compositional method performs best overall. For the datapoint search, the performance for CRP is similar to that of the PED, although this method does not perform so well when matching is performed with the mean vectors. Again the Hellinger distance improves over the cosine distance, while it was found that centring did not improve performance relative to the cosine distance. When the datapoint dictionaries were used the difference between all algorithms was less than 3%; however when the dictionary of mean vectors was employed, the relative performance was more variable.

Alg.	NN(C)	NN(H)	M(C)	M(H)
PED(0.05)	82.0	**85.5**	77.1	**84.7**
LogC	80.8	84.1	**77.8**	79.6
CRP	**85.4**	-	52.8	-
GW	79.4	82.6	70.5	76.7

Table 3. Results for the nearest neighbour experiments. NN denotes datapoint search, M denotes search over the dictionary of mean vectors. Bracketed letters denote cosine (C) and minimum Hellinger (H) distances.

5. CONCLUSIONS

We have described a compositional model of chroma, employing a dictionary of a shifted synthetic single note chroma vector. The use of the powered Euclidean distance was considered for these decompositions and was seen to improve on the well known β-divergence. While the initial results are promising, further work will seek to employ the proposed method in more realistic scenarios, while the incorporation of prior information, such as temporal continuity will be introduced.

6. REFERENCES

[1] T. Virtanen, J.F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *Signal Processing Magazine, IEEE*, vol. 32, no. 2, pp. 125–144, March 2015.

[2] T. Virtanen, "Monaural sound source separation by non-negative matrix factorisation with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[3] A. Liutkus, D. Fitzgerald, and R. Badeau, "Cauchy nonnegative matrix factorization," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, 2015, pp. 1–5.

[4] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, March 2010.

[5] K. O'Hanlon, H. Nagano, N. Keriven, and M. D. Plumbley, "Non-negative group sparsity with subspace note modelling for polyphonic transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 530–542, March 2016.

[6] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS 14)*, Denver, 2000, pp. 556–562.

[7] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Barcelona, 2004, pp. 318–325.

[8] 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.

[9] K. Yoshii, K. Itoyama, and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 51–55.

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

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

[12] Johan Pauwels and Jean-Pierre 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.

[13] C. Rhodes, T. Crawford, M. Casey, and M. d'Inverno, "Investigating music collections at different scales with audiodb," *Journal of New Musical Research*, vol. 39, pp. 337–348, 2010 2010.

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

[15] 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.

[16] Meinard Müller and Sebastian 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.

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

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

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

[20] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He, "Extended smart algorithms for non-negative matrix factorization," *Lecture notes in Artificial Intelligence, 8th International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, vol. 4029, pp. 548–562, 2006.

[21] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," in *6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, Florida, 2006, pp. 32–39.

[22] K. O'Hanlon, M. D. Plumbley, and M. Sandler, "Non-negative matrix factorisation incorporating greedy Hellinger sparse coding applied to polyphonic music transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, Brisbane, 2015.

[23] E. Yilmaz, J.F. Gemmeke, and H. Van hamme, "Noise-robust speech recognition with exemplar-based sparse representations using alpha-beta divergence," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 5502–5506.