TRACKING METRICAL STRUCTURE CHANGES WITH SPARSE-NMF

Elio Quinton, Ken O'Hanlon, Simon Dixon, Mark Sandler

Centre for Digital Music
Queen Mary University of London

ABSTRACT

The estimation of rhythmic properties such as tempo, beat positions or metrical structure are central aspects of Music Information Retrieval (MIR) research. Meter inference algorithms are typically designed to track metrical structure in presence of mild deviations of the feature estimates over time in order to account for performance imprecisions, expressive timing or musical effects such as accelerando. Abrupt changes of metrical structure over time are comparatively rarely addressed. In this paper, we present an unsupervised approach to detect metrical structure changes. Formulating the problem as a metrical structure based segmentation retrieval task, we present a variant of sparse NMF and compare it to existing methods. For evaluation, we introduce a new dataset of music recordings containing metric modulations with the corresponding annotations.

Index Terms— MIR, rhythm, meter, segmentation, NMF

1. INTRODUCTION

Automatic extraction of the metrical structure of music from audio recordings is a complex and challenging task. Meter inference may be approached by tracking several (usually two or three) metrical cycles of different length such as the beat, downbeat and tatum, typically using latent state space models [1, 2, 3] or by analysing the metrical structure in the frequency domain where peaks in the beat spectra relate to metrical level pulse rates [4, 5]. Meter inference algorithms are typically designed to track metrical structure in presence of mild deviations of the feature estimates over time in order to account for expressive timing or performance imprecisions [1, 2]. As a consequence, tracking musical effects such as accelerando is made possible. Abrupt changes of metrical structure over time are comparatively rarely addressed and even less so as a task in itself. See for instance the bar pointer model, which allows the tracking of abrupt meter changes but requires supervised learning of the metrical structures [6].

In this paper we propose an unsupervised method for the detection of metric modulations within a music piece. We formulate the task as a metrical structure based segmentation retrieval problem [7]. We restrict this study to abrupt modulations from one section of stable metrical structure to another section with a different but still stable metrical structure. Segment boundaries represent the time points at which the metric modulation happens. The number, length and metrical structure of each segment is a priori unknown.

Multiplying the Fast Fourier Transform (FFT) and Autocorrelation Function (ACF) based beat spectra, as initially suggested by Peeters [8] was shown to be effective to filter out harmonics in the spectrum so that its peaks more closely relate to the hierarchy of metrical level pulse rates of the music [9]. So far, such beat spectra have only been used as a summary feature (i.e. averaged over time). In this paper we extend this approach to the use of a *metergram* in which every frame is the product of FFT and ACF beat spectra. Changes in metrical structure over time therefore result in apparent structure in the metergram, which we seek to recover.

Segments of consistent metrical structure are expected to correspond to homogeneous regions in the metergram so that segmentation may be retrieved by detection of *homogeneity* [10]. NMF has previously been used for homogeneity-based segmentation [11, 12, 13, 14]. Here we present a variant of sparse-NMF which we compare to existing NMF methods as well as a popular *novelty*-based approach [15] to perform segmentation. As an additional contribution, we introduce a new dataset made of a corpus of music recordings containing metric modulations and the corresponding metrical structure annotations.

We detail the computation of the metergram in section 2 and present our segmentation method as well as baselines in section 3. In section 4 we introduce a new dataset and discuss the algorithms performance. We conclude in section 5.

2. METERGRAM

We first compute a spectrogram of the audio signal using a 11.6ms Hann window with 50% overlap. An onset detection function is derived using the *superflux* method with the parameter values recommended by the authors [16]. We then compute two rhythmograms $\mathbf{R}_F$ and $\mathbf{R}_A$, based on the FFT and ACF of the windowed onset detection function respectively, using 12s Hann windows with 0.24s overlap. The ACF rhythmogram is mapped to the frequency domain, as

This work is supported by the EPSRC award 1325200 and the AHRC Grant AH/L006820/1

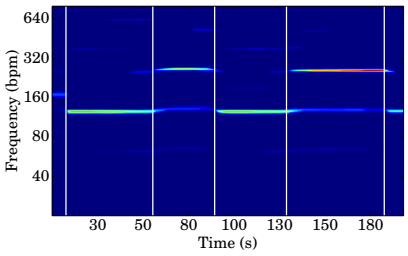


Fig. 1: Metergram with annotated segment boundaries overlaid for the track "Geno (Tribute to Dexys Midnight Runners)" by Union of Sound.

proposed in [8], so that the metergram may be computed as the element-wise product (denoted by $\odot$) of the two rhythmograms $\mathbf{R}(m,n) = \mathbf{R}_F(m,n) \odot \mathbf{R}_A(m,n)$. The bins of the metergram are then re-assigned to a logarithmic frequency scale with $\omega_m = \omega_0 \times 2^{\left(\frac{m}{b}\right)}$ where $\omega_0 = 20$ BPM and $b = 100$ bins/octave.

Moreover, the energy distribution of interest in a metergram resembles what would be called a *harmonic* structure in an audio spectrogram, each horizontal line typically corresponding to a metrical level pulse rate [9]; whereas broad band energy distributions/noise are not informative about the metrical structure. As a result, we apply median filtering on the metergram with a 15s window, in order to only keep the "harmonic" part [17]. An example of such a metergram overlaid with the annotated metrical structure change boundaries is given in Figure 1. It is apparent in this example that the metric modulations correlate with changes in energy distribution in the metergram and this is what we seek to capture.

3. SEGMENTATION

3.1. Self-similarity novelty-based segmentation

Metric modulations such as those apparent in Figure 1 may be characterised by a change of energy distribution in the metergram frames so that the segmentation may be performed by detection of *novelty* over time. Foote introduced a novelty-based method that has since become a standard for automatic structural segmentation and which we apply to the Self-Similarity Matrix (SSM) of the metergram [15]. We assume that sections of consistent metrical structure will be at least around 10s long. We found that varying the Gaussian tapered checkerboard kernel size by a few seconds did not have a significant impact and set it to 15s. The segmentation has been computed for a range of novelty curve peak-picking threshold values. We present the results obtained with the threshold value resulting in the highest performance.

3.2. NMF for segmentation by frame clustering

Non-negative Matrix Factorisation (NMF) seeks to factorise a non-negative matrix $\mathbf{V} \in \mathbb{R}_{\geq 0}^{M \times N}$ such that $\mathbf{V} \approx \mathbf{WH}$ where $\mathbf{W} \in \mathbb{R}_{\geq 0}^{M \times K}$ contains a *template* vector in each column, with *activations* in the corresopnding row of $\mathbf{H} \in \mathbb{R}_{\geq 0}^{K \times N}$ and the decomposition rank K is set in advance. NMF is typically performed through gradient-based optimisation [18]. The generalised β-divergence

$$D_\beta(\mathbf{s}|\mathbf{z}) = \frac{1}{\beta(\beta-1)} \sum_i s_i^\beta + (\beta-1) z_i^\beta - \beta(s_i z_i^{\beta-1}) \quad (1)$$

includes Euclidean distance ($\beta = 2$), Kullback-Leibler (KL) and Itakuro-Saito (IS) divergences as limit cases as $\beta \to \{1, 0\}$, respectively. Penalty terms γ can be used to encourage certain behaviours in NMF, such as sparse activation [19]. The cost function to minimise is then $D_\beta(\mathbf{V}|\mathbf{WH}) + \lambda\gamma$ where λ controls the impact of the penalty. Multiplicative updates for penalised β-NMF are given by

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left[\frac{\mathbf{W}^T \left(\mathbf{V} \odot (\mathbf{WH})^{(\beta-2)}\right)}{\mathbf{W}^T (\mathbf{WH})^{(\beta-1)} + \lambda \Psi_\mathbf{H}} \right]^{\varphi(\beta)} \quad (2)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left[\frac{\left(\mathbf{V} \odot (\mathbf{WH})^{(\beta-2)}\right) \mathbf{H}^T}{(\mathbf{WH})^{(\beta-1)} \mathbf{H}^T + \lambda \Psi_\mathbf{W}} \right]^{\varphi(\beta)} \quad (3)$$

where $\Psi_\mathbf{H}$ and $\Psi_\mathbf{W}$ typically describe the gradient of the penalty and $\varphi(\beta)$ is a parameter that ensures monotonicity.

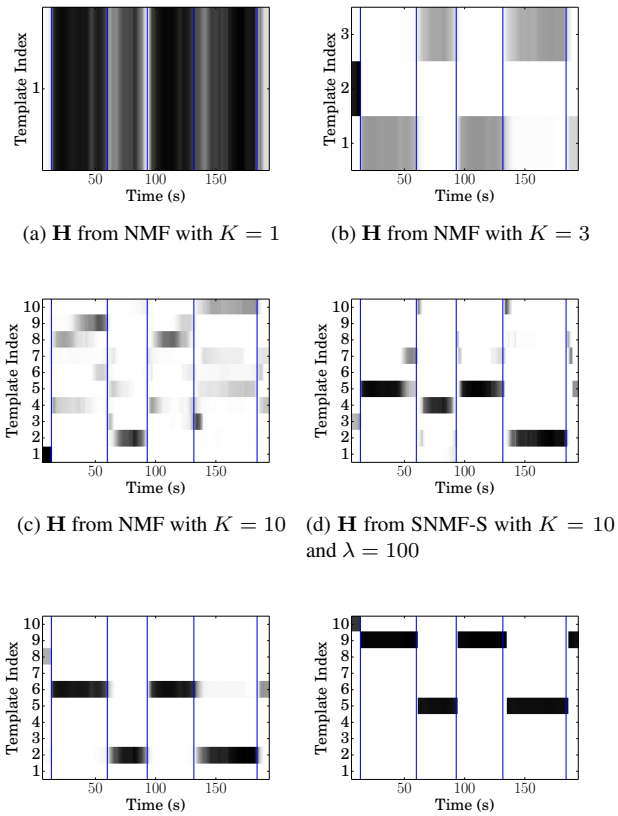
Applying NMF to the metergram ($\mathbf{V} = \mathbf{R}$), the templates should correspond to the beat spectrum of each section and the activations revealing their temporal extent, i.e. the segmentation. From a metrical structure point of view, the track with metergram shown in Figure 1 has three different parts: a short introduction and two main alternating parts. Figure 2 (a), (b) and (c) depicts the corresponding activation matrix $\mathbf{H}$ obtained with NMF ($\beta = 1$) for $K = \{1, 3, 10\}$ respectively, illustrating that segmentation is performed best when the rank equals the number of parts in the track. However, in our scenario this number is not known a priori.

3.2.1. Heuristic automatic rank determination baseline

When the chosen rank is too small, the factorisation can not be accurate (cf. Figure 2), implying a large reconstruction error. This error is expected to decrease when the rank increases, becoming reasonably small when K is equal to the number of different segments in the track, with small decreases in error for further rank augmentation. On this premise, we devise a baseline automatic rank estimation method, notated NMF-K_e. For each track, an NMF decomposition and the reconstruction error is computed for a range of ranks, i.e. $K \in \{1, ..., 10\}$. The effective rank K_e is selected so that

$$K_e \triangleq K : D_1(\mathbf{V}, \mathbf{WH})_K \geq \epsilon \text{ and } D_1(\mathbf{V}, \mathbf{WH})_{K+1} < \epsilon \quad (4)$$

with $\epsilon = 2.10^{-4}$. The factorisation of rank K_e is then used to retrieve segmentation.



(a) $\mathbf{H}$ from NMF with $K = 1$ (b) $\mathbf{H}$ from NMF with $K = 3$

(c) $\mathbf{H}$ from NMF with $K = 10$ (d) $\mathbf{H}$ from SNMF-S with $K = 10$ and $\lambda = 100$

(e) $\mathbf{H}$ From ARD, with $K = 10$,$\beta = 1$, $\phi = 0.1$ and $a = 500$ (f) $\mathbf{H}$ from sparse L_β-S-β-NMF with $K = 10$, $\beta = 0.5$ and $\lambda = 5$

Fig. 2: Activation matrices for the track "Geno (Tribute to Dexy's Midnight Runners)" by Union of Sound for a range of NMF variants with segment boundary annotations overlaid as vertical lines

As NMF is employed as a clustering tool here, we also compare to the popular K-means clustering method. A data representation similar to NMF is formed, where the columns of $\mathbf{W}$ are the cluster centroids, and each entry of $\mathbf{H}$ is defined as $h_{k,n} = \delta_{kc_n}$ where c_n is the index of the cluster to which the n^{th} frame is assigned. The effective rank is also determined using (4). Let us notate this method K-means-K_e.

3.3. Sparse-NMF / Automatic Relevance Determination

We now introduce a sparse NMF algorithm for β-divergence which we label L_β-S-β-NMF, inspired by the approach in [20], employing the penalty $\gamma = \frac{1}{\beta}\sum_{n=1}^{N}\|\mathbf{y}_n\|_\beta^\beta$ where $y_{k,n} = h_{k,n}\|\mathbf{w}_k\|_2$ and $h_{k,n}$ is the activation coefficient of the k^{th} component at n^{th} time frame. Multiplicative updates for L_β-S-β-NMF are given by substituting $\Psi_{\mathbf{W}} = \mathbf{W} \odot \text{repmat}\left(\sum_n \sqrt{h_{k,n}^T}, M, 1\right)$, $\Psi_{\mathbf{H}} = 1/\sqrt{\mathbf{H}}$ and $\varphi(\beta) = 1/(3-\beta)$ into (3) and (4). Stronger sparsity is enforced by setting $\beta < 1$, and we employ $\beta = \frac{1}{2}$ in the experiments in this paper.

For comparison then, we employ the method proposed by Tan and Févotte to perform Automatic Relevance Determination (ARD) by aggressively and jointly sparsifying $\mathbf{H}$ and $\mathbf{W}$ row-wise and column-wise respectively so that unnecessary components are de-activated while iteratively optimising the factorisation w.r.t. the β-divergence in [21]. We refer to the original publication for details of the algorithm, but we note that multiplicative update rules for $\mathbf{H}$ and $\mathbf{W}$ proposed by the authors consist of the standard β-NMF updates with the addition of a penalty term whose influence is controlled by a parameter ϕ. It was noted in [22] that ARD bears some similarities with reweighted L_1-minimisation [23] which suggests that it enforces a sparsity constraint stronger than L_1 penalty.

We compare to a sparse KL factorisation using the more typical L_1 norm of $\mathbf{H}$ [19], which we denote as SNMF-S. For all approaches a rank purposefully too large is selected, hoping that the sparse penalisation will select few components [11]. In particular, we set K =10 in our experiments as we assume unlikely that a music piece contains more than 10 different metrical structures.

3.4. Hidden Markov Model for final segmentation

Transitional and simultaneous activations may arise in $\mathbf{H}$ (cf Figure 2). We employ a Hidden Markov Model (HMM) to filter out these artefacts and produce the final segmentation by decoding the state sequence using the Viterbi algorithm. The number of hidden states is set equal to the number of NMF components K, each state being associated with the *true* activation of a component. We define the probability of emitting the component index k from the k^{th} hidden state ψ_k as:

$$\pi(k|\psi_k) = \frac{\exp\left(-A_{k,n}\right)}{\sum_{k=1}^{K}\exp\left(-A_{k,n}\right)} \tag{5}$$

where $A_{k,n} = (h_{k,n} - \mu)^2/2\sigma^2$, $\mu = \max_{k,n}(h_{k,n})$, and $\sigma = \mu$. $\pi(k|\psi_k)$ is therefore large for large activation coefficients and vice versa. The transition probabilities are defined in two classes: remain in the same state $P(\psi_i|\psi_i)$ and transition from state ψ_i to state ψ_j notated $P(\psi_j|\psi_i)$.

$$\forall(i,j) \in \{1,...,K\}, \begin{cases} P(\psi_j|\psi_i) = P_d & i = j \\ P(\psi_j|\psi_i) = \frac{1-P_d}{K-1} & i \neq j \end{cases} \tag{6}$$

We set $P_d = 0.9$ in our experiments, as it has empirically been found to work well.

4. EVALUATION

4.1. Dataset

In order to evaluate the performance of the proposed system, we introduce a new dataset of music pieces containing metric modulations[1]. The music corpus was composed thanks

[1] http://isophonics.net/content/metric-modulations-dataset

to a crowdsourcing campaign. No musical style constraints were enforced so that a variety of genres are represented, although submissions predominantly consisted of western popular, rock and progressive rock music. From all the suggestions, only the pieces featuring exclusively abrupt modulations from a section of stable metrical structure to a section of different, yet stable structure were kept. Each track was manually annotated with beat, downbeat and segment boundary positions as well as with all metrical level pulse rates present in each segment. The segment boundaries are located at the first downbeat of the new section. The dataset contains 67 tracks featuring 2 to 16 segments each, that being 378 segments in total and an average of 5.6 segments per track. However, due to repetition, the number of unique metrical structures per track rarely exceeds 4 and is always largely lower than 10. A large variety of metric modulation types are represented, including changes of tempo, meter and combinations thereof.

4.2. Evaluation metrics

We use a range of standard metrics to evaluate the quality of the segmentation produced by the various methods under test. The pairwise precision, recall and f-measure, notated ppr, prr and pfm respectively are calculated as $ppr = \frac{|P_e \cap P_a|}{|P_e|}$, $prr = \frac{|P_e \cap P_a|}{|P_a|}$ and $pfm = 2\frac{ppr \cdot prr}{ppr + prr}$ where P_e is the set of similarly-labelled pairs of frames estimated by the machine and P_a is the set of similarly-labelled pairs of frames annotated in the human-generated reference annotations.

We also compute the over- and under-segmentation scores, S_o and S_u as proposed in [24] as well as the corresponding simile F-measure metric: $S_f = 2\frac{S_o S_u}{S_o + S_u}$.

Boundary hit rate whereby segment boundaries estimated by the algorithm are regarded as correct if they are within a tolerance window from an annotated boundary is also computed. We report the hit rate F-measure with threshold of 3s and 8s, notated Fm_3 and Fm_8 respectively.

4.3. Results

The evaluation results obtained for all methods considered in this paper are given in Table 1. All methods have at least one adjustable parameter. For conciseness, we only present here the results obtained with the parameter configuration that produces the highest pfm. It is to be noted that configurations that produce the highest pfm also produce the highest hit rate F-measure and S_f in the vast majority of the cases. In other words, the performance of the methods tends to peak in the same area of the parameter space for all F-measure metrics.

Considering the pairwise frame clustering metrics, it is interesting to note that the ARD method leans towards high recall whereas other NMF-based methods lean towards higher precision, with the exception of the L_β-S-β-NMF which exhibits a very balanced performance. ARD with $\beta = 1$ produces the best pfm performance, with L_β-S-β-NMF closely following. The examination of under- and over-segmentation scores reveals that all methods tend to over-segment more than they under-segment and L_β-S-β-NMF produces the highest S_f score.

Table 1: Segmentation performance for all methods considered in this paper. For each method, we present the results obtained with the parameter configuration leading to the best pfm. The highest score for each metric is in bold characters.

Methods	ppr	prr	pfm	S_o	S_u	S_f	Fm_3	Fm_8
ARD $\beta = 0$	0.59	**0.92**	0.66	0.59	0.58	0.58	0.22	0.38
ARD $\beta = 1$	0.70	0.91	**0.75**	0.69	0.70	0.70	**0.42**	**0.53**
ARD $\beta = 2$	0.61	0.84	0.66	0.61	0.63	0.62	0.28	0.36
L_β-S-β-NMF	0.77	0.78	0.73	**0.72**	0.84	**0.75**	0.41	0.52
SNMF-S	0.84	0.50	0.57	0.58	0.85	0.69	0.31	0.43
NMF-K_e	0.80	0.67	0.67	0.66	0.81	0.73	0.39	**0.53**
Kmeans-K_e	**0.89**	0.47	0.57	0.61	**0.90**	0.73	0.37	0.45
SSM Foote	0.66	0.81	0.68	0.68	0.68	0.68	0.07	0.42

For every NMF+HMM method, raising the hit rate threshold from 3s to 8s improves the F-measure score by about 0.1 points, which suggests that the precise localisation of the boundaries is a significantly challenging problem on which to improve in future work. The effect is even more pronounced in the case of novelty-based segmentation, which suggests that the NMF+HMM strategy leads to more precise boundary locations estimates than peak-picking a Foote novelty curve.

Overall it appears that ARD with $\beta = 1$ and L_β-S-β-NMF share the highest scores on all F-measure metrics (i.e. pfm, S_f,Fm_3 and Fm_8), often exhibiting close scores. These are the only two methods to consistently equal or outperform SSM Foote and automatic rank determination baselines and may therefore be considered as the two best performing methods. They are also the two methods enforcing the strongest sparsity constraints in the NMF decomposition — cf. Figure 2 for illustration. In addition, SNMF-S performs best when the weight of its sparsity constraint, which is comparatively weaker, is extremely large ($\lambda = 100$). This suggests in more general terms that very strong sparsity constraints are beneficial for the quality of the segmentation produced.

5. CONCLUSION

We have presented an unsupervised method to segment musical recordings with respect to metrical structure and therefore took a step towards the automatic tracking of metric modulations. In addition, we presented a new dataset to evaluate such systems. The results show that homogeneity-based NMF-powered methods outperform the standard novelty-based approach and that very strong sparsity constraints are instrumental in achieving such a result. Directions for future work consist in improving the segment boundary location precision and extended evaluation of the metrical structure templates learnt.

6. REFERENCES

[1] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[2] F. Krebs, A. Holzapfel, A. T. Cemgil, and G. Widmer, "Inferring metrical structure in music using particle filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 817–827, 2015.

[3] A. Holzapfel, F. Krebs, and A. Srinivasamurthy, "Tracking the "Odd": Meter Inference in a Culturally Diverse Music Corpus.," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2014, pp. 425–430.

[4] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Audio Engineering Society Convention 114*, 2003.

[5] M. Robine, P. Hanna, and M. Lagrange, "Meter Class Profiles for Music Similarity and Retrieval.," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2009, pp. 639–644.

[6] N. Whiteley, A. T. Cemgil, and S. J. Godsill, "Bayesian Modelling of Temporal Structure in Musical Audio.," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2006, pp. 29–34.

[7] B. Thoshkahna, M. Müller, V. Kulkarni, and N. Jiang, "Novel Audio Features for Capturing Tempo Salience in Music Recordings," in *IEEE International Conference on Acoustics Speech and Signal Processing*. 2015.

[8] G. Peeters, "Time variable tempo detection and beat marking," in *Proceedings of the ICMC*, 2005.

[9] E. Quinton, C. Harte, and M. Sandler, "Extraction of Metrical Structure from Music Recordings," in *Int. Conference on Digital Audio Effects (DAFx)*, 2015.

[10] J. Paulus, M. Müller, and A. Klapuri, "State of the Art Report: Audio-Based Music Structure Analysis.," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2010, pp. 625–636.

[11] R. J. Weiss and J. P. Bello, "Identifying Repeated Patterns in Music Using Sparse Convolutive Non-Negative Matrix Factorization," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2010.

[12] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 236–240.

[13] F. Kaiser and T. Sikora, "Music Structure Discovery in Popular Music using Non-negative Matrix Factorization.," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2010, pp. 429–434.

[14] P. Seetharaman and B. Pardo, "Simultaneous Separation and Segmentation in Layered Music," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2016.

[15] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *IEEE International Conference on Multimedia and Expo (ICME)*. 2000, vol. 1, pp. 452–455.

[16] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Int. Conf. on Digital Audio Effects (DAFx)*, 2013.

[17] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Int. Conf. on Digital Audio Effects (DAFx)*, 2010.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[19] J. Eggert and E. Korner, "Sparse coding and NMF," in *Proceedings of the International Joint Conference on Neural Networks*. 2004, vol. 4, pp. 2529–2533, IEEE.

[20] 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, 2016.

[21] V. YF Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the β-Divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592–1605, 2013.

[22] V. YF Tan and C. Févotte, "Supplementary material to "Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence"," 2012, https://www.ece.nus.edu.sg/stfpage/vtan/supp_mat.pdf.

[23] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[24] H. M. Lukashevich, "Towards Quantitative Measures of Evaluating Song Segmentation.," in *International Society for Music Information Retrieval (ISMIR) Conference*, 2008, pp. 375–380.