LOCALIZED KEY FINDING FROM AUDIO USING NON-NEGATIVE MATRIX FACTORIZATION FOR SEGMENTATION

Özgür İzmirli

Center for Arts and Technology
Computer Science
Connecticut College

ABSTRACT

A model for localized key finding from audio is proposed. Besides being able to estimate the key in which a piece starts, the model can also identify points of modulation and label multiple sections with their key names throughout a single piece. The front-end employs an adaptive tuning stage prior to spectral analysis and calculation of chroma features. The segmentation stage uses groups of contiguous chroma vectors as input and identifies sections that are candidates for unique local keys in relation to their neighboring key centers. Non-negative matrix factorization with additional sparsity constraints and additive updates is used for segmentation. The use of segmentation is demonstrated for single and multiple key estimation problems. A correlational model of key finding is applied to the candidate segments to estimate the local keys. Evaluation is given on three different data sets and a range of analysis parameters.

1. INTRODUCTION

Music being bought on digital media, aired through radio broadcasts, streamed or downloaded from Internet sites is almost exclusively in audio format and generally not accompanied by metadata that would be useful for music information retrieval (MIR). On the other hand, many retrieval tasks require the acquisition, playback, browsing or content analysis to be in audio format. This emphasizes the importance of audio content analysis tools that operate at the front-end and become the eyes and ears of higher level and general MIR tools. Many categories in the MIREX competitions aim at extracting structural information from audio. In this regard, key finding is one of these areas that adds considerably to the structural knowledge that can be extracted from a musical piece. Being able to reliably detect the key of a tonal piece (in the context of Western music) remains an important step in content analysis for MIR. Tonal music constitutes a significant portion of the music consumed today. Hence, models of key finding are applicable to a wide range and large portion of available music. In the same vein, localized key finding is essential for other methods in MIR research to work reliably.

Many audio key finding models exist in the literature. Most of these deal with identifying the main key in a musical piece. Although this is an important task, it does not provide useful information for structural analysis. That is, by knowing the main key of a piece we cannot infer any additional information regarding the time evolution of its harmonic structure. On the other hand, modulation through one or multiple keys is very common in classical music and is utilized in popular music quite often.

From the listener's viewpoint a musical fragment in a single key implies a most stable pitch, the tonic, and a musical scale associated with that key. Throughout this fragment, if the music has a well-formed tonal structure, a change in key center will not be sensed. Secondary functions and tonicizations are heard as short deviations from the well-grounded key in which they appear - although the boundary between modulation and tonicization is not clear cut. A modulation unambiguously instigates a shift in the key center.

Structure discovery aims at providing high-level representations of music. It deals with problems such as similarity, repetition and thumbnailing. Segmentation is used for identifying points of structural change and it can be based on a multitude of features. In this paper, we investigate segmentation from a tonality perspective. The presented method aims to identify points of modulation, the names of the key centers and their corresponding modes without attempting to perform transcription or chord recognition. It also performs this in an unsupervised manner.

In order to infer key from audio input, the music needs to be observed for a certain duration to ensure all necessary elements have been encountered. In other words, one might develop an initial estimate of the key after hearing the first chord of a piece. However, at this point there are multiple competing estimates and one cannot arrive at a reliable decision until subsequent musical events have been heard. Every new musical event works in the direction of weakening some estimates and disambiguating and strengthening others.

The optimal duration of key locality depends on musical context. The model presented here works on this premise and aims to group and segment an appropriate duration of music that belongs to and characterizes a key. This is done with non-negative matrix factorization with chroma features as input. This approach flies in the face of sliding window key center tracking techniques which need the window duration to be fixed and predetermined.

2. RELATED WORK

Symbolic and audio key finding differ in their methods and accuracy. On the symbolic end Chew [5] and Temperley [23] have addressed the problem of modulation. Shmulevic and Yli-Harja [21] employ sliding windows to find local key estimates. Although most researchers working on key finding allude to modulation detection, very little systematic research on performance of these algorithms has been reported. On the audio end Purwins et al. [19] used a fuzzy distance measure between constant-Q profiles and reference constant-Q sets to track key centers. Operation of the method is demonstrated on a single piece of piano music. Chai and Vercoe [4] used a Hidden Markov Model to detect key changes from audio. They used 10 classical pieces to test their method. Gómez [7] uses a specialized form of PCP feature and a sliding window method to track tonality. Izmirli [13] used symbolic representations to perform efficient comparisons of tonal evolution between different renditions of the same piece, in turn proposing a measure of similarity between entire pieces. Harte et al. [10] proposed a harmonic change detection function to detect transitions between tonal regions which were defined by chords in their case. Chord segmentation and recognition are akin to key finding in that common methods are employed for solutions to these problems. For example, among the many models, Sheh & Ellis [20] used a Hidden Markov Model to perform segmentation of chords and chord recognition on Beatles songs.

Segmentation has been an active research topic in the field of MIR. We refer to recent work that relates to local key finding and tonality. Chai [3] proposed models for analysis of musical form and recurrent structure as well as harmony analysis. Ong [17] studied audio-based music structural analysis and used tonal features in measuring similarity of cover songs.

Non-negative matrix factorization (NMF) was initially proposed by Lee and Seung [16] for part-based learning of images. Smaragdis and Brown [22] demonstrated the application of NMF to polyphonic music transcription. Abdallah and Plumbley [1] used a similar method for transcription and demonstrated its performance on piano music. Cont [6] used NMF to learn spectral note templates off line and then used NMF with sparsity constraints to perform real time note recognition. The reader is referred to Izmirli [14] for work in the field of audio key finding.

3. TUNING FRONT END

A tuning front end is used to adjust the frequency reference of the system to each input file. Factors such as transcoding effects and intentional tuning preferences may result in different tunings for each piece. For example, Peeters [18] and Harte and Sandler [9] have proposed methods for tuning adjustment. In order to find the reference tuning of the input audio file, our method analyzes the first 10-15 seconds of the music and compares it to synthetically generated spectral key templates. A detailed description of spectral templates from audio samples is given in [12, 14] and a summary for line spectra is provided below. The frequency that maximizes the integral of the product between the templates and the spectrum serves as the tuning frequency estimate of the input audio:

$$(c,k) = \arg\max_{c,k} \left(\int_{f_{\min}}^{f_{\max}} Y(f) \sum_{i=0}^{R-1} X_{i,c}(f)\, F_{i,k}\; df \right) \quad (1)$$

Y(f) represents the mean of the short-time amplitude spectra over the first 10-15 seconds of the piece. $X_{i,c}(f)$ represents the line spectra of note *i* (a Dirac comb with decaying weights) with its fundamental frequency calculated with respect to the reference frequency *c*. For example, c=442 Hz. would mean the fundamental frequency of note A4 is at that frequency and all other notes are determined according to equally tempered intervals. Each *X* is constructed using 20 harmonics with amplitudes decaying at 12 dB per octave. The limits on the integration are chosen to be in the range 55 - 1250 Hz. R is the total number of notes used for the synthetic spectral templates, typically spanning 5 octaves. Eq. 1 can be directly implemented with a high resolution FFT with zero padding applied to the input signal and a compatible discrete representation for the line spectra.

Profiles are incorporated into the calculation of spectral templates to approximate the distribution of pitch classes in the spectrum. In Eq. 1 $F_{i,k}$ are the composite profile weights rotated k steps for note i within each mode.

$$F_{i,k} = \begin{cases} P_M(\,(i-k+12) \bmod 12) & \mathit{if} \quad 0 \le k \le 11 \\ P_m(\,(i-k+24) \bmod 12) & \mathit{if} \quad 12 \le k \le 23 \end{cases} \quad (2)$$

The composite profile P is given by the elementwise product of the diatonic (D) and Temperley (T) profiles: $P_e(k)=D_e(k)T_e(k)$. The index e is either M for major or m for the minor mode.

Not surprisingly, the second index (k) over which the product is maximized in Eq. 1 gives an estimate of the key of the initial section of the piece. However, at this stage no segmentation has been done nor has any attempt been made to capture the most relevant parts of the music for key estimation. Therefore, this estimate is treated as a by product of the optimization. The results of this estimation are given in the evaluation section.

4. SEGMENTATION

4.1. Non-negative Matrix Factorization

Non-negative matrix factorization aims to decompose a matrix V with n rows and m columns into a product of two matrices W and H. An internal dimension p is chosen such that W has n rows and p columns and H has p rows and m columns. As implied in its name the main constraint that separates this decomposition from other similar ones is its non-negativity constraint on all three matrices. The usefulness of this method originates from its summarization property that when p is chosen to be smaller than n, the columns of V are summarized in columns of W. Hence, W can now be interpreted as a compressed collection of basis vectors that could be used to reconstruct an approximation to the original input in V. Due to the smaller internal dimension p the reconstruction WH will not be exact. Thus, a distance measure between WH and V is used as the cost function to be minimized during the factorization.

NMF which was originally proposed for decomposition of images has also been applied to the problem of polyphonic transcription as mentioned above. This method is suitable for transcription because the method tries to find the sparse additive constituents, i.e. notes, of the observed polyphonic frequency spectrum. It also does not allow for negative contributions of components that would lead to reconstruction through cancellation. In this case, the basis functions in W represent approximations to note spectra and the corresponding weights (in H) can be viewed as the mixing matrix. Parallel to this, the problem of segmentation based on tonal features, the topic of this paper, aims to reveal additive contributions of tonal elements in the analyzed piece.

In the original formulation of NMF by Lee and Seung [16] multiplicative updates were used for the factorization. Later, Hoyer [11] proposed a formulation that used additive updates with sparsity constraints that could be imposed independently on W and H. In this context, sparsity is a measure that quantifies the distribution of energy in a vector. By definition, if the total energy is in a single component the sparsity measure is equal to 1. Similarly, if the energy is spread equally among the components then the measure is equal to 0.

4.2. Segmentation

In this work, NMF is used for segmentation. The columns of V are composed of grouped chroma vectors obtained from the entire length of the musical piece. A group is found by taking the mean of consecutive chroma vectors. The calculation of chroma vectors comprises the following steps: the audio is downsampled to 11025 Hz. The spectrum is calculated with a Hann windowed 2048-point FFT. The 12-element chroma vector is obtained from the spectrum in the range 50 Hz. to 2000 Hz. The tuning frequency, *c*, found in Section 3 is used as the frequency reference while calculating the chroma representation. The details regarding the calculation of the chroma vectors can be found in [12, 14]. The grouped chroma vectors are found by averaging the chroma vectors over a span of *s* seconds. The value of the parameter *s* is on the order of 5-15 seconds. Groups are heavily overlapped. The factorization is performed using the Euclidean norm as the cost function. Finally, the maxima in columns of H are found and all existing segments are identified with each segment defined as a consecutive sequence of maxima with the same row index.

The sparsity constraint is only imposed on the H matrix. Forcing the columns of H to be sparse affects the resultant structure of W. As a result, the input matrix is factored such that the basis vectors learned in W represent the best approximation to the specific clusters of chroma vectors where each cluster approximates a chroma pattern particular to one or a group of keys. This makes the factorization function similar to vector quantization because the columns of W will mimic global representations rather than capturing local features or sparse basis functions as in the case of transcription. The work reported here is part of ongoing research in optimal representations for tonality and key finding. In this context, NMF is preferred over other clustering methods to maintain the flexibility of global vs. local representations regarding chroma and as a means to explore the possibility of lower dimensional representations for tonality.

The clustering behavior may need some more clarification. For example, if there is a single modulation in an input piece then it should suffice for H to have two rows. In order to minimize the cost function, the performed factorization will result in a summarization of the two keys in columns of W. An example of a factorization for a pop piece that contains three keys is shown in Figure 1.a. The horizontal axis is time and each column represents a grouped chroma vector. Figure 1.b. shows the W matrix with internal dimension p=2 and chroma group window of approximately 7 seconds. Figure 1.c. shows the H matrix factored with a sparseness value of 0.3. The ground truth is given at the bottom of the plot. This is an example where the number of keys was underestimated. It can be seen in part c that the first two key regions would be segmented such that they map to the same basis vector. This demonstrates an undesirable situation where a new segment boundary is missed. Nevertheless, the detection of the segment boundary would not be a problem if two closely related keys were interleaved by a distant key. A remedy to this situation would be to increase the internal dimension p to attain less summarization. Figure 2 shows W and H for p=3 using the same song. In this case, the three key regions are clearly detectable. The method gives satisfactory results for a simple case like this, however, in general, it is not realistic to assume that the number of modulations would be known a priori. Therefore, one approach might be to consistently overestimate the number of keys in the input. The down side of this is that there will be more jumps between smaller size segments. This idea is considered in the evaluation using the different data sets.

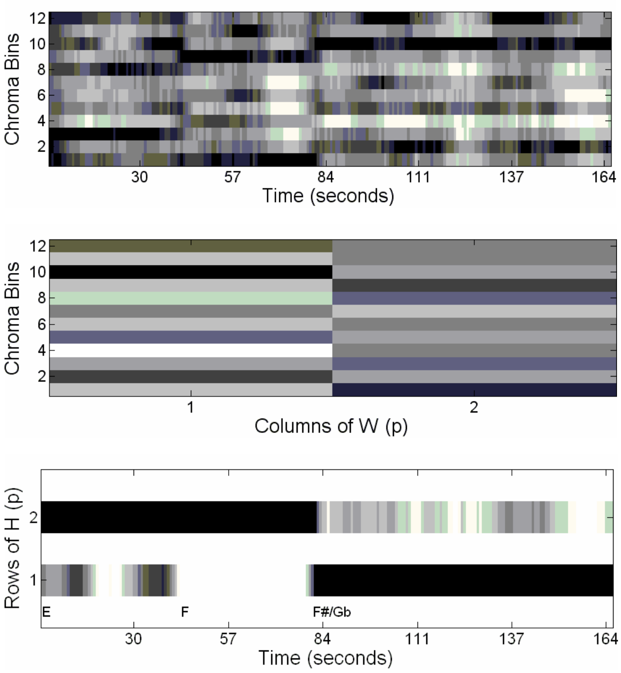

Figure 1. (a) Top plot. The input matrix V for Shania Twain's Come on Over. The summary chroma vectors are the columns of this matrix. (b) Middle plot. The W matrix (p=2). (c) Bottom plot. The H matrix (p=2).

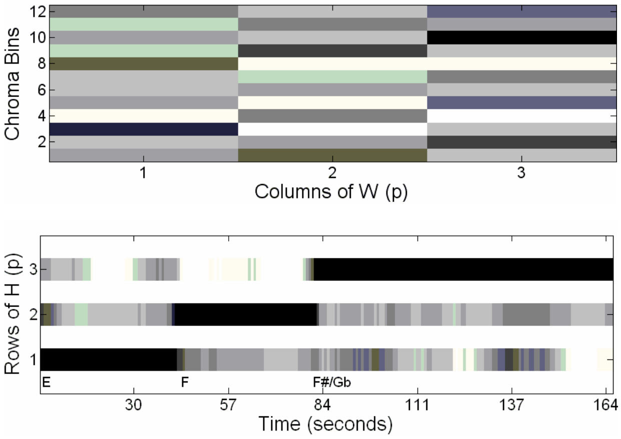

Figure 2. (a) Top plot. W with p=3. (b) Bottom plot. H with p=3.

5. LOCALIZED KEY ESTIMATION

Localized key finding is important for structural segmentation methods in MIR research. For example when using a method of chromagram matching to detect verse or chorus repetitions [2] it is desirable to detect the repetitions regardless of any modulation. Goto [8] addresses this problem by rotating the chroma in all possible keys to account for the possibility of modulation in repetitions of the music. Reliable localized key finding would be helpful in converting all single-key regions to a reference key enabling the existing similarity algorithms to be employed.

5.1. Single Key Estimation

Key finding is generally understood to be the estimation of the main key of a piece. Some models only look at the beginning of a piece. This was also part of the specification in the MIREX 2005 audio key finding competition in which our model ranked first, but only with a slight margin ahead of the other competing algorithms. Some other models look at different parts of entire pieces: beginning, middle and end. Given the performance of the model (model I) in [14] we maintain that analyzing approximately the first half minute of the piece suffices to produce a reliable key estimate of that section. The reason for this, in the case of common practice classical music, is that the main key of a piece is almost always introduced at the beginning of that piece. Furthermore, the key name is spelled out in the name of the piece conveniently saving the researcher some annotation time.

Although this approach works fairly well, the question of optimal segment length for reliable estimation still remains open. A short segment may put too much focus on a particular chord and an excessively long segment may extend into a section where the piece modulates into another key. In [14] this issue was circumvented by using progressive overlapping windows, all pivoting off the starting point of the piece. For each window the key was estimated and an associated confidence was calculated. The final estimate was determined by selecting the key with the overall highest confidence.

The approach presented here for single key estimation uses the segmentation step described in Section 4 to find the length of one segment at the beginning of the piece, on which a key estimation algorithm is run, and consequently render a key estimate. The correlational model in [12] is used to estimate the key.

5.2. Multiple Key Estimation

A key finding algorithm is applied to the entire span of every segment determined by the segmentation method discussed in the previous section. The key is estimated assuming that the optimal key locality has been correctly delineated by the segmentation step. Figure 3.a. shows an example of key estimation on segments found in H for p=3 using a fragment of classical music. The dark letters at the bottom of the figure (above the time axis) are the ground truth and the light letters indicate the key estimates at the beginnings of the respective segments. The audio is taken from one of the data sets used in the evaluation. Figure 3.b. shows segmented key estimation on another pop piece. Note that the key estimates are correct but not continuous in the first section. This is fortunately not a problem for the frame based evaluation explained in the next section. If continuous segments were required a simple algorithm for stitching neighboring segments with the same key could be implemented.

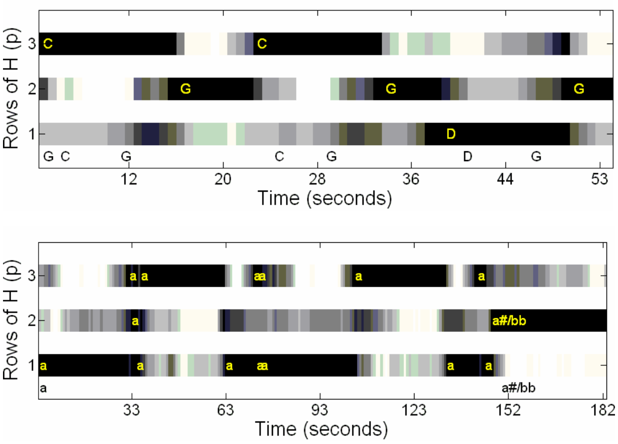

Figure 3. (a) Top plot. Segmented key estimation shown on H for audio from a Bach Choral. (b) Bottom plot. Segmented key estimation for Abba's Money Money Money.

6. EVALUATION

Several types of evaluation were carried out to test the performance of the model. There were three data sets. The first set was a collection of 17 complete pop songs that contained at least one modulation. All pieces were carefully annotated with all keys and modulation points. The second set was the initial fragments of 152 classical music pieces from the Naxos set (www.naxos.com). The ground truth was obtained from the names of the pieces. The third set consisted of short fragments of classical music with modulations. The music was taken from the Tonal Harmony textbook by Kostka and Payne [15]. This data set (K&P) also had 17 short fragments. The recordings on the accompanying CDs were used. The ground truth was obtained from the accompanying instructor's manual.

In all evaluations a raw measure of accuracy was accompanied by a composite score. In order to partially reward closely related key estimates the following (MIREX) fractional allocations were used while calculating the composite score: correct key, 1 point; perfect fifth, 0.5; relative major/minor, 0.3; parallel major/minor, 0.2 points. In the following, the composite score follows the raw figure in parentheses.

Initially, we report on the key finding accuracy of the front end tuning stage. As this stage generates a single estimate from the beginning of each piece, that value was compared to the first key in the annotation. The results were as follows: pop set 58.8 % (74.1%), Naxos set 51.3% (62.9%) and K&P set 76.5% (79.4%). Note however, that this method is not intended for key finding.

Three different types of estimates were calculated for each data set with a chroma group window size of 7.4 seconds. Note that a much smaller window size will focus the chromagram on individual chords and much larger window will degenerate the model to a sliding window implementation at the frame level – but even then, it will be useful at the global level and for visualization. Several window durations have been tested and this length has been determined to be a good compromise. It should also be noted that using longer windows will cause the segmentation boundaries to blur, however, simply picking the maximum element in H will suffice to identify the modulation point. Method 'I' is an unweighted correlation estimate with elements of the chromagram raised to the power of 0.5. All frames within a segment are averaged and correlated to the 24 chroma templates. The index of the template with the highest correlation is the key estimate of that segment. The accuracy is calculated for all available frames in the input piece. Method I is a frame based multiple key estimation measure for the pop and K&P sets and a frame based single key estimation measure for the Naxos set as only the main key data is available as ground truth. It is the frame accuracy of the beginning section that ends on the earliest segment boundary between 10 and 30 seconds of each piece. Method 'II' denotes a single key estimation accuracy measure. It uses a confidence weighted estimate, as explained in Section 5.1, only for the first segment. Method 'III' denotes confidence weighted estimates for all segments in the piece. The accuracy for the three methods and three values of p are given in Table 1.

	p=2 (%)	**p=3 (%)**	**p=4 (%)**
Pop I	79.6 (83.9)	82.4 (87.0)	76.6 (83.5)
Pop II	64.7 (72.6)	70.6 (78.2)	58.8 (70.6)
Pop III	71.8 (76.0)	73.5 (79.2)	67.8 (75.6)
Naxos I	75.1 (80.9)	78.8 (83.5)	72.8 (78.5)
Naxos II	80.9 (85.8)	78.9 (84.1)	78.3 (83.5)
Naxos III	77.1 (83.2)	74.2 (80.2)	74.0 (79.4)
K&P I	69.7 (77.2)	71.5 (77.4)	72.5 (78.2)
K&P II	94.1 (97.1)	94.1 (97.1)	82.4 (85.3)
K&P III	67.6 (74.5)	64.2 (70.9)	68.0 (73.7)

Table 1. Evaluation results for the three different data sets.

These preliminary results are encouraging. It can be seen that the accuracy figures are relatively stable over values of p. This shows that the method is not too sensitive to the internal dimension parameter p, and overestimating the number of modulations does not drastically degrade the performance. For comparison, the results of two evaluations are given: the frame accuracies with unweighted key estimates using ground truth segmentation are 88.3% (92.8%) for the pop set and 76.9% (84.1%) for the K&P set. The unsegmented frame accuracies are pop set 66.49% (75.6%); Naxos set 70.6% (77.1%); K&P set 71.3% (76.1%). This shows that segmentation has improved the frame accuracy for the pop and Naxos sets but not for the K&P set. This is mainly due to the short audio length in the K&P examples and particularly due to insufficient time span in the last modulated key in these recordings. The very high accuracy of model II on this set also supports this point. On the Naxos set, the unsegmented evaluation is done on the segmentation used in method I to make the number of frames equal. The actual difference is probably greater. The accuracy of the modulation points depend on the group window duration and the nature of the modulation. Overall, the proposed method is able to identify modulations and estimate all local key labels in a given piece as seen by the evaluation.

7. CONCLUSIONS

A modulation detection and local key labeling model with a preprocessing stage for tuning adjustment and non-negative matrix factorization for segmentation has been proposed. The model identifies segments that are candidates for unique local keys in relation to the neighboring key centers. A correlational key finding model is run on every segment in order to label each one with a key center. Encouraging results are obtained on three different data sets and it has been shown that the model does not necessarily have to be tuned to the number of keys for the piece of interest, although a slight drop in performance is experienced as a penalty for overestimating the number of keys.

8. REFERENCES

[1] Abdallah S. A. and Plumbley M. D. “Polyphonic Transcription by Non-negative Sparse Coding of Power Spectra.” *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[2] Bartsch, M. and Wakefield, G. H. “To Catch a Chorus: Using Chroma-Based Representations For Audio Thumbnailing,” *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2001.

[3] Chai, W., “Automated Analysis of Musical Structure,” *Ph.D. Dissertation*, MIT, 2005.

[4] Chai, W. and Vercoe, B. “Detection of Key Change in Classical Piano Music,” *Proceedings of the International Conference on Music Information Retrieval*, London, UK, 2005.

[5] Chew, E. “The Spiral Array: An Algorithm for Determining Key Boundaries,” *Proceedings of the Second International Conference*, ICMAI, Edinburgh, Scotland, UK, 2002.

[6] Cont, A. “Realtime Multiple Pitch Observation using Sparse Non-negative Constraints,” *International Symposium on Music Information Retrieval*, Victoria, Canada, 2006.

[7] Gómez, E. “Tonal Description of Music Audio Signals,” *Ph.D. Dissertation*, Pompeu Fabra University, Barcelona, 2006.

[8] Goto, M. “Music Scene Description,” in Anssi Klapuri and Manuel Davy, eds., *Signal Processing Methods for Music Transcription*, pp.327-359, Springer, 2006.

[9] Harte, C. and Sandler, M. “Automatic Chord Identification using a Quantised Chromagram,” *AES 118th Convention*, Barcelona, Spain, 2005.

[10] Harte, C., Sandler M., and Gasser, M. “Detecting Harmonic Change in Musical Audio,” *Proceedings of AMCMM’06*, Santa Barbara, California, USA, 2006.

[11] Hoyer, P. O., “Non-negative Matrix Factorization with Sparseness Constraints,” *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[12] İzmirli, Ö. “Template Based Key Finding from Audio,” *Proceedings of the International Computer Music Conference*, Barcelona, Spain, 2005.

[13] İzmirli, Ö. “Tonal Similarity from Audio Using a Template Based Attractor Model,” *Proceedings of the International Symposium on Music Information Retrieval*, London, UK, 2005.

[14] İzmirli, Ö. “Audio Key Finding Using Low-Dimensional Spaces,” *Proceedings of the International Conference on Music Information Retrieval*, Victoria, Canada, 2006.

[15] Kostka S. and Payne D. *Tonal Harmony* (4th edition), Boston: McGraw Hill, 1999.

[16] Lee, D. D. and Seung, H. S. “Learning the Parts of Objects by Non-negative Matrix Factorization”. *Nature* 401, pp. 788-791, (1999).

[17] Ong, B. “Structural Analysis and Segmentation of Music Signals,” *Ph.D. Dissertation*. Pompeu Fabra University, Barcelona, 2007.

[18] Peeters, G. “Musical Key Estimation of Audio Signal Based on HMM Modeling of Chroma Vectors,” Proceedings of DAFX, McGill, Montreal, Canada, 2006.

[19] Purwins, H., Blankertz, B. and. Obermayer, K. “Constant Q Profiles for Tracking Modulations in Audio Data,” *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001.

[20] Sheh A. and Ellis, D. “Chord Segmentation and Recognition using EM-trained Hidden Markov Models,” *Proceedings of the International Conference on Music Information Retrieval*, Baltimore, Maryland, USA, 2003.

[21] Shmulevich, I. and Yli-Harja, O. “Localized Key Finding: Algorithms and Applications.” *Music Perception, Special Issue in Tonality Induction*, 17(4):531–544, 2000.

[22] Smaragdis, P. and Brown, J. “Non-negative Matrix Factorization for Polyphonic Music Transcription,” *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2003.

[23] Temperley, D. *The Cognition of Basic Musical Structures*, Cambridge, MA: MIT Press, 2001.