# AUTOMATIC MUSICAL KEY ESTIMATION WITH ADAPTIVE MODE BIAS

*Gilberto Bernardes*[1], *Matthew E. P. Davies*[1] *and Carlos Guedes*[1,2]

[1]INESC TEC, Sound and Music Computing Group, Porto, Portugal
[2]New York University Abu-Dhabi, Abu Dhabi, United Arab Emirates

## ABSTRACT

In this paper we present the INESC Key Detection (IKD) system which incorporates a novel method for dynamically biasing key mode estimation using the spatial displacement of beat-synchronous Tonal Interval Vectors (TIVs). We evaluate the performance of the IKD system at finding the global key on three annotated audio datasets and using three key-defining profiles. Results demonstrate the effectiveness of the mode bias in favoring either the major or minor mode, thus allowing users to fine tune this variable to improve correct key estimates on style-specific music datasets or to balance predictions across key modes on unknown input sources.

***Index Terms***— Audio key estimation, tonal pitch representation, music signal processing, music information retrieval.

## 1. INTRODUCTION

Key or tonality is a prominent concept in Western music. It is defined by a pitch class (*tonic*) and a mode (*major* or *minor*), whose combination establishes a system of relations between pitches in both the vertical and horizontal dimensions of musical structure [1]. Key estimation from musical audio has been extensively researched within the music information retrieval community [2, 3, 4], as it provides important annotations for enhanced navigation and retrieval in large music collections as well as contributing to music-creative tasks such as harmonic DJ mixing [5].

However, the majority of existing key estimation systems rely on the same fundamental principle, that of using key profiles expressing pitch class distributions to which a similar representation obtained from analyzing a musical piece or excerpt is compared to estimate the most probable key. Research on key finding devotes great effort to the creation and evaluation of different key profiles proposed in the literature [6, 7, 2]. Yet, these findings must be understood in the context of the datasets on which they are evaluated, as it has been shown that different key profiles explicitly favor either major or minor key modes [8]. Given that the most widely used datasets in the evaluation of audio key estimation systems have pronounced divergences in key mode distribution (e.g., a strong bias towards the major mode in the Beatles collection [9] and the minor mode in the GiantSteps [3] dataset), we believe this has led to an intrinsic bias in current key detection systems which are adapted to convey better estimations in either minor or major modes, but not both. In turn, this limits the generality of a particular method in finding the key for unknown musical inputs.

In light of these findings, we introduce a strategy to explore key mode estimates of the IKD system [10, 11], a key detection method based on the Tonal Interval Space [11], without the need to hand-tune key profiles. The geometric properties of Tonal Interval Space allow us to easily adapt key mode estimation by introducing spatial displacements to the input – a non-trivial task in commonly used metric spaces used in related literature. This not only enables users to bias the systems towards major or minor correct modes estimates, which has been shown to be an important feature for style-specific key detection [4], but can also balance the correct number of estimates across modes for enhanced results on unknown musical inputs. We demonstrate the efficacy of our approach by explicitly manipulating its accuracy on existing annotated datasets comprised of excerpts in predominantly major and minor modes.

The remainder of this paper is structured as follows. Section 2 provides an overview of the Tonal Interval Space, as well as distance metrics computed in the space relevant to the IKD system. Section 3 starts by presenting the architecture of our system, followed by a detailed description of each of the component modules, with particular emphasis on the novelty of our approach, i.e. the use of mode bias in the key detection method. Sections 4 and 5 present an objective evaluation of the IKD system and, finally, in Section 6 we draw conclusions and state areas for future work.

## 2. OVERVIEW OF THE TONAL INTERVAL SPACE

The system reported in this paper is based on the Tonal Interval Space [12], an extended tonal pitch space in the context of the *Tonnetz* [13]. The most salient pitch levels of tonal Western music – pitches, chords and keys – can be represented as unique locations in the space as Tonal Intervals Vectors (TIVs) from music encoded as both symbolic or audio data. A predominant feature of the Tonal Interval Space is the ability to compute theoretical and perceptual aspects of Western tonal music, such as indicators of multi-level tonal pitch relatedness and consonance, as distances.

In this paper, we focus on audio signal representations in the Tonal Interval Space, due to its relevance in the key estimation system under discussion. To represent an audio signal in the Tonal Interval Space, we first aggregate the energy of each pitch class in a 12-dimensional chroma vector, $c(n)$, and compute a 12-dimensional Tonal Interval Vector, $T(k)$ as its $L_1$ normalized Discrete Fourier Transform (DFT), such that:

$$T(k) = w(k) \sum_{n=0}^{N-1} \bar{c}(n) e^{\frac{-j2\pi kn}{N}}, \; k \in \mathbb{Z} \qquad (1)$$

where $N = 12$ is the dimension of the chroma vector and $w(k) = \{2, 11, 17, 16, 19, 7\}$ are weights derived from empirical conso-

Project TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020, financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF). GB is also supported by FCT, the Portuguese Foundation for Science and Technology, under the post-doctoral grant SFRH/BPD/109457/2015.

nance ratings of dyads used to adjust the contribution of each dimension $k$ of the space (or interpreted musical interval), making it a perceptually relevant space in comparison to its non-weighted version [12]. We set $k$ to $1 \leq k \leq 6$ for $T(k)$ since the remaining coefficients are symmetric. $T(k)$ uses $\bar{c}(n)$ which is $c(n)$ normalized by the DC component $T(0) = \sum_{n=0}^{N-1} c(n)$ to allow the representation and comparison of different hierarchical levels of tonal pitch [12].

The resulting spatial location of tonal pitch in the Tonal Interval Space ensures that configurations understood as perceptually related within the Western tonal music context correspond to small Euclidean distances [12]. Relevant here are the distances of the 24 major and minor TIV keys, which are sparsely and (per mode) equidistantly represented in the space. Neighboring key TIVs adhere to theoretical and perceptual relations (e.g., in the neighborhood of each key TIV, we find its dominant, subdominant, and relative keys) [12]. Furthermore, the set of diatonic pitch classes and chords of each key are at smaller distances than non-diatonic pitch configurations, allowing us to infer the key TIV from a collection of pitch and chord configurations. A final property of the space relevant to our study is the constant vector norm of transposition invariant configurations. This property indicates that, for example, all major keys are at the same distance from the center (the same applies to all minor keys). Additionally, due to the difference of intervallic relations between major and minor keys, a consistent vector norm difference exists between these two sets of configurations, thus the ideal binarised TIVs (containing only the notes of each scale) for harmonic minor keys are closer to the centre of the Tonal Interval Space than for major keys.

## 3. AUDIO KEY ESTIMATION METHOD

Fig. 1 shows the architecture of the IKD system. The first module is responsible for performing a beat segmentation on a musical audio input, whose onset times are then used to compute beat-synchronous TIVs. Given that harmonic changes typically occur on beats [14], we adopt beat segments as the temporal resolution for representing the harmonic content, in order to maximize the efficiency of the system while minimizing the likelihood of two chords being temporally merged. The second module introduces a spatial displacement to the input beat-synchronous TIVs to bias or balance the inference of key mode based on the vector norm difference between major and minor keys in the Tonal Interval Space. Finally, the third module computes the distance between the displaced input TIVs from 12 major and 12 minor TIV key-defining profiles and finds the most probable key as that with the smallest distance.

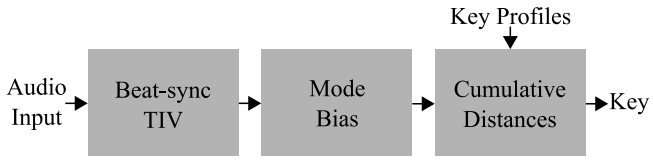


**Fig. 1**. Architecture of the IKD system.

### 3.1. Beat-synchronous TIV

Given an audio signal (with sampling frequency 44.1kHz), we first extract chroma vectors using the NNLS chroma [15] plugin within Sonic Annotator [16] with default parameters, including both tuning correction and spectral whitening. Each chroma vector is calculated over a 46 ms frame. Next, we extract beat locations from the same input audio signal, using the QM-VAMP bar and beat tracking [17] also within Sonic Annotator. To compute the beat-synchronous chroma vectors we then take the median value per chroma bin across all frames within each beat, $b$. Finally, we apply Eq. 1 to compute the beat-synchronous TIVs.

### 3.2. Mode Bias

The principal novelty of our key estimation method in comparison to related template-based key estimation methods is the introduction of a key mode bias, $\alpha$, which exploits the vector norm difference between major and minor key in the Tonal Interval Space. This variable adjusts the location of input beat-synchronous TIVs to favor key estimates in one of the major or minor modes. This can be better understood in the 2-dimensional illustration of the key level in the Tonal Interval Space shown in Fig. 2. When $\alpha < 1$ we "pull" input vectors towards the center of the space (i.e., decrease their norm), thus favoring minor keys estimates. On the other hand, when $\alpha > 1$ we "push" them towards the edge of the space (i.e. increasing their norm), thus favoring major keys estimates.

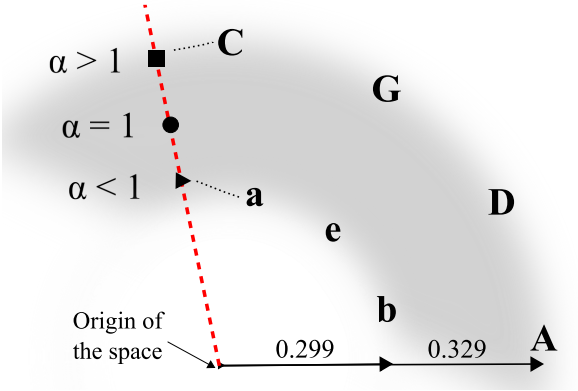


**Fig. 2**. Illustrative example of the key level in the Tonal Interval Space mapped into 2 dimensions. Upper and lower case letters represent major and minor keys, respectively, along with their corresponding vector norm. By altering the norm of an input TIV (represented as a circle) using different values of $\alpha$, we show the impact of the mode bias on key estimates, which oscillates between key modes, notably the relative C major (for values of $\alpha > 1$, represented as a square) and A minor key (for values of $\alpha < 1$, represented as a triangle).

### 3.3. Key TIV Profiles

Fig. 3 shows three key-defining profiles, $p$, adopted in this study, which expose the pitch class distribution of the C major and C minor keys. Their selection was based on their different nature: the knowledge-based profiles by Temperley's ($T^t$) [6], the corpus-driven profiles by Aarden ($T^a$) [7] and Shat'ath ($T^s$) [2], from folk and electronic dance music (EDM) corpora, respectively. These profiles are considered here as chroma vectors, $c(n)$, which we convert to key TIVs using Eq. 1. The key TIVs of remaining keys are computed by rotating the C major and C minor key TIVs, $T(k)$, by

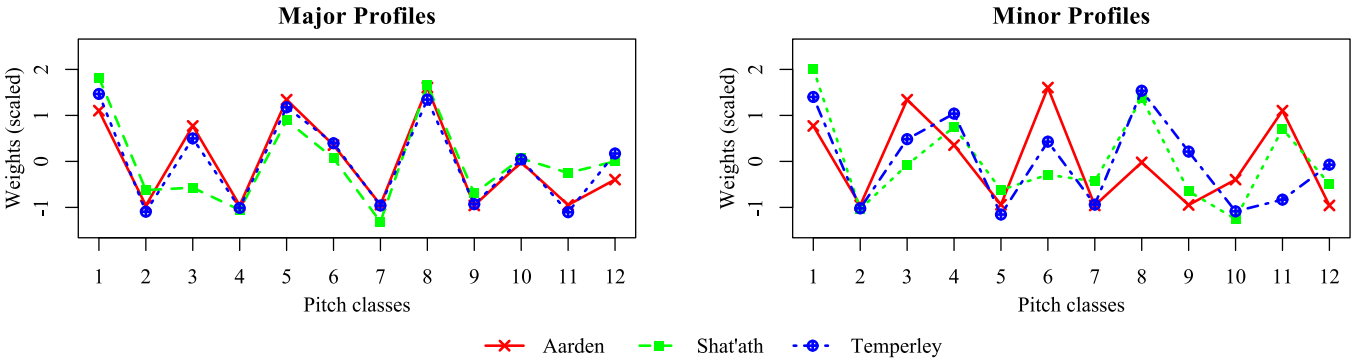


**Fig. 3**. The major and minor key profiles for the four set of profiles used: Aarden ($T^a$) [7], Shat'ath ($T^s$) [2], and Temperley ($T^t$) [6] (for enhanced visualization the profiles were normalized to zero mean and unit variance).

$\varphi(r) = (-2\pi kr)/N$ radians, where $r = [0, 11]$ semitones. Further details on the transposition of pitch configurations in the Tonal Interval Space by means of TIV rotation can be found in [10].

### 3.4. Key Estimates as Minimal Cumulative Distances

Based on the assumption that a key-indicating element is the use of its diatonic pitch set and chords, we define a method for estimating the global key of a musical example, $R_{\min}$, in the Tonal Interval Space by finding the minimum in a function which accumulates over time the distances of the total number of query beat-synchronous TIV, $T_b(k)$, from the 12 major and 12 minor key TIVs, such that:

$$R_{\min} = \text{argmin}_r \sum_{b=1}^{B} \sqrt{\sum_{k=1}^{6} \left| T_b(k) \cdot \alpha - T_r^{p^*}(k) \right|^2} \qquad (2)$$

where $T_r^{p*}$ are 24 major and minor key TIVs, derived from the collection of three different key profiles, $p$. When $r \leqslant 11$, we adopt the major profile and when $r \geqslant 12$, the minor profile. To limit the influence of silent or noisy (inharmonic) beats, $b$, we only retain those for which $T(0) > 0.1$, where $B$ is the total number of retained beat-synchronous TIVs. By default, the mode bias $\alpha = 1$ (i.e. no spatial displacement to the input beat-synchronous TIVs is introduced) and can be adjusted to favor one of the two major and minor modes as detailed in Section 3.2. The system output is a number, $R_{\min}$, ranging between 0-11 for major keys and 12-23 for minor keys, where 0 corresponds to C major, 1 to C# major, and so on through to 23 being B minor.

## 4. EVALUATION

We undertake an objective assessment of the IKD system in estimating the global key from musical audio, focusing on the implications of the mode biasing strategy on three different datasets and for three key-defining templates. By adopting different values of $\alpha$ (both greater than and less than 1) in Eq. 2, we aim to show that: i) our mode bias can improves performance on either major or minor modes by increasing and decreasing $\alpha$, respectively, ii) overall results on correct key estimates can be improved by adopting a balanced $\alpha$ value, and iii) key-defining profiles have a tendency to privilege one of the major or minor modes.

We use three audio datasets with key annotations made by experts in our evaluation. When combined, this collection provides a total of 879 musical examples, which include heterogeneous genre and timbre qualities. The first datatset consists of the initial 30 seconds of 96 classical musical examples evenly distributed across modes and tonics (4 musical examples per key) used in the MIREX Audio Key Estimation task [18]. The second dataset includes the first 30 seconds of 179 Beatles' songs [9], with 89.4% examples in the major mode. The third dataset is the GiantSteps collection [3], which consists of the initial 2 minutes of 604 EDM examples across 23 sub-genres, with 84.8% of the data in the minor mode. As discussed in the introduction, the use of datasets with even mode distribution is an important design decision in the evaluation of systems for the key estimation on unknown input. While the MIREX training set fulfills this criterion, the two remaining datasets favor different modes, which we use as a strategy to understand the behavior of our mode bias algorithm. To this end, we expect to improve the baseline results (i.e. when $\alpha = 1$) on the Beatles and GiantSteps datasets by increasing and decreasing the $\alpha$, respectively.

## 5. RESULTS

Fig. 4 shows the performance of our IKD system on the three datasets under evaluation, for which we provide a score for $\alpha = [0.05, 20]$ and across each profile, $p$, as well. To allow a fair comparison with previous studies, we use the MIREX evaluation procedure [19], which is widely applied in key estimation studies, where correct and neighboring keys estimates are weighted and averaged into a final score according to the following point assignment: correct $(1)$, dominant/subdominant $(.5)$, relative $(.3)$, parallel $(.2)$, and others $(0)$.

The most immediate observation we can draw from our results in Fig. 4 is the effectiveness of the mode bias in regulating the tendency of mode prediction, confirming the expected tendencies on the evolution of the correct estimates in the Beatles and GiantSteps datasets shown in Fig. 4 (b) and (c). As the vast majority of musical examples in the Beatles dataset are in major mode, the ascending accuracy curve shows the expected improvements for the three key-defining templates when $\alpha$ increases. Equally, the GiantSteps datatset reinforces the mode bias effectiveness by showing the contrary tendency, i.e. smaller values of $\alpha$ result in better predictions. On the other hand, the results for the evenly distributed MIREX training set generate a less asymmetric curve for the same range of $\alpha$ values. The inflection point for each key profile curve on the MIREX training set results shown in Fig. 4 (a) can be considered the optimal value of $\alpha$, which provides the best, and most balanced key mode, results for this dataset.

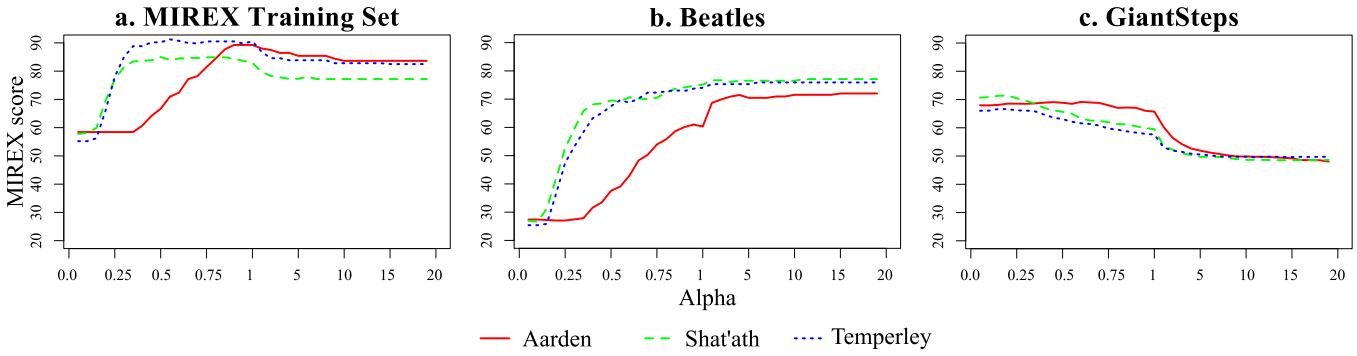


**Fig. 4**. Performance of the IKD system for the three datasets under evaluation: a. MIREX training set, b. Beatles, and c. GiantSteps. Each dataset was evaluated using three key profiles (Aarden ($T^a$) [7], Shat'ath ($T^s$) [2], and Temperley ($T^t$) [6]), on a range of values for the modes bias $\alpha = [0.05, 20]$.

| | NS [20] | SH [2] | FGJH [4] | Rekordbox [21] | IKD |
|---|---|---|---|---|---|
| Beatles | 72.4 | 59.3 | **76.0** | 56.37 | 65.9 |
| GiantSteps | 52.9 | 59.3 | 74.6 | **79.6** | 68.5 |
| Combined | 57.4 | 59.9 | **69.3** | 74.3 | 67.9 |

**Table 1**. Comparison of different key estimation software on the Beatles and Giantstep datasets. The best score for each dataset is shown in bold.

Overall, adopting Temperley [6] profile, $T^t$ when $\alpha = 0.55$ gives the best performance for the MIREX training set (91.3% score) and Shat'ath [2] profile, $T^s$ when $\alpha = 0.35$ provides the best results for the combined set of the Beatles and GiantSteps datasets (67.9% score). Combining the two latter datasets indicates which $\alpha$ value provides the best results for a system to which unknown audio content input is presented. Moreover, in cases where a known tendency for one of the modes exists (such as the minor mode in EDM music), the system can achieve much better performance (77.1% for the Beatles dataset and 71.3% for the GiantSteps, when $\alpha$ equals 10 and 0.15, respectively). The best performing $\alpha$ values for all datasets indicate that the Shat'ath [2] and Temperley [6] profiles without the mode bias (i.e., when $\alpha = 1$) favor correct major mode estimates, where as the Aarden [7] profiles show the opposite behavior.

In Table 1 we present the performance of different systems for audio key estimation on the Beatles and GiantSteps datatsets reported in [4], to which we include the scores for our IKD system (using Shat'ath profiles and $\alpha = 0.35$) and for the Rekordbox [21] software on the Beatles dataset. While our system outperforms Noland and Sandler [20] and Shat'ath [2] systems, it provides worse results than Rekordbox [21] and Falardo et al.'s systems [4]. We believe that the poorer performance of our algorithm in relation to the two latter systems is due to the high optimization of their algorithms for EDM. Not only is the Tonal Interval Space designed to provide 'general' inferences for Western tonal music, without being fitted to any particular style, the NNLS chroma representation used in the IKD also aims at finding perfectly tuned harmonic pitch templates, which may not be the case in most EDM and pop/rock music, namely when using synthesizers. While the performance of Faraldo et al.'s system [4] is most accurate across these two datasets, an initial version of our IKD system outperformed it on the (closed) dataset of 1252 classical music examples from MIREX 2016 Audio Key Detection task [22].

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an enhanced key mode estimation on the IKD system using a mode bias strategy, which introduces a spatial displacement to input TIVs. We demonstrated that the mode bias not only allows users to favor correct key estimates on one of the two major or minor modes – relevant for key detection on style-specific datasets – but also provides a strategy to balance correct key mode predictions when in the presence of unknown input sources. The method sheds some light on breaking the computational key estimation problem into two parts: one for mode estimation and the other for tonic estimation. To this end, we envisage our method could particularly benefit end-users who may be readily be able to distinguish between major and minor modes, but are unable to infer the root, as well as provide insight into enhanced discrimination between the two modes at the various hierarchical pitch levels.

The major contribution of this paper and the strength of the IKD system in relation to related research is its flexibility in correctly estimating the key from audio inputs with dynamic control over mode prediction – a feature that, to the best of our knowledge, has never been considered before in key detection system other than adjusting key-defining profiles in an *ad hoc* manner to fit to existing datasets. Finally, an important consideration resulting from this study is the relevance and influence of evenly distributed datasets in the evaluation design of key detection systems in order to avoid unsound conclusions resulting from the tendency of particular key-defining templates to favor one of the two major or minor modes.

Stylistic instantiations of the current IKD system are planned for future work towards improving the chroma representation used to accommodate musical sounds with non-harmonic timbral qualities [23], such as those typically featured in EDM, as well as understanding the distribution of modes across different musical genres to tune the system for its optimal performance in style-specific datasets. A style-specific key estimation system that profits from the IKD mode bias algorithm, must know in advance the bias towards one of the major or minor modes to provide better key estimates. To this end, we plan on using timbral features to automatically select $\alpha$ from the audio signal .

## 7. REFERENCES


[1] A. J. Milne, *A Computational Model of Cognition of Tonallity*, Ph.D. thesis, The Open University, 2013.

[2] I. Shat'ath, "Estimation of key in digital music recordings," M.S. thesis, Birkbeck College, University of London, 2011.

[3] P. Knees, Á. Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, and M. Le Goff, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 364–370.

[4] Á. Faraldo, E. Gómez, S. Jordà, and P. Herrera, "Key estimation in electronic dance music," in *Proceedings of European Conference on Information Retrieval*. 2016, pp. 335–347, Springer International Publishing.

[5] R. Gebhardt, M. Davies, and B. Seeber, "Psychoacoustic approaches for harmonic music mixing," *Applied Sciences*, vol. 6, no. 5, 2016.

[6] D. Temperley, "What's key for key? the Krumhansl-Schmuckler key-finding algorithm reconsidered," *Music Perception: An Interdisciplinary Journal*, vol. 17, no. 1, pp. 65–100, 1999.

[7] B. Aarden, *Dynamic melodic expectancy*, Ph.D. thesis, Ohio State University, 2003.

[8] J. Albrecht and D. Shanahan, "The use of large corpora to train a new type of key-finding algorithm: An improved treatment of minor mode," *Music Perception: An Interdisciplinary Journal*, vol. 31, no. 1, pp. 59–67, 2016.

[9] C. Harte, *Towards Automatic Extraction of Harmony Information from Music Signal*, Ph.D. thesis, Queen Mary University of London, 2010.

[10] G. Bernardes, D. Cocharro, C. Guedes, and M. Davies, "Harmony generation driven by a perceptually motivated tonal interval space," *ACM Computers in Entertainment*, vol. 14, no. 2, 2016.

[11] G. Bernardes and M. Davies, "Audio key finding in the tonal interval space," in *Submission to the Music Information Retrieval Evaluation eXchange (MIREX) Audio Key Detection task*, 2016.

[12] G. Bernardes, D. Cocharro, M. Caetano, C. Guedes, and M. Davies, "A multi-level tonal interval space for modelling pitch relatedness and musical consonance," *Journal of New Music Research*, vol. 45, no. 4, pp. 281–294, 2016, http://dx.doi.org/10.1080/09298215.2016.1182192.

[13] L. Euler, *Tentamen novae theoriae musicae*, Broude, New York/St. Petersburg, 1968/1739.

[14] H. Papadopoulos and G. Tzanetakis, "Models for music analysis from a markov logic networks perspective," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 19–34, 2017.

[15] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proceedings of the 11th International Society Music Information Retrieval Conference*, 2010, pp. 135–140.

[16] C. Cannam, M. Jewell, C. Rhodes, M. Sandler, and M. d'Inverno, "Linked data and you: Bringing music research software into the semantic web," *Journal of New Music Research*, vol. 39, no. 4, pp. 313–325, 2010.

[17] M. Davies, M. Plumbley, and D. Eck, "Towards a musical beat emphasis function," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2009, pp. 61–64.

[18] "MIREX Audio Key Finding (Training) Dataset," http://www.music-ir.org/mirex/wiki/2005:Audio_and_Symbolic_Key_Finding, [Accessed: 2016-09-12].

[19] J. S. Downie, A Ehmann, M. F., Bay, and M. C. Jones, "The music information retrieval evaluation exchange: Some observations and insights," in *Advances in music information retrieval)*. 2010, pp. 93–115, Springer Berlin Heidelberg.

[20] K. Noland and M. B. Sandler, "Key estimation using a hidden markov model," in *Proceedings of The International Society for Music Information Retrieval Conference*, 2006, pp. 121–126.

[21] Pioneer DJ, "Rekordbox," https://rekordbox.com, [Accessed: 2016-09-12].

[22] "MIREX Audio Key Finding Results," http://nema.lis.illinois.edu/nema_out/mirex2016/results/akd/mrx_05/summary.html, [Accessed: 2016-09-12].

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