TEMPLATE BASED KEY FINDING FROM AUDIO

Özgür İzmirli
Center for Arts and Technology
Connecticut College, New London
Connecticut 06320, USA

ABSTRACT

A model for template based key finding from audio is presented and two methods that implement this model are compared. Templates are computed from a weighted combination of spectra obtained from sound recordings of monophonic musical notes. Individual weights of notes contributing to the templates are determined by profiles representing tonal hierarchies in Western music. Key determination is based on the correlations between spectral summary information obtained from audio input and the precomputed templates. The first method that implements the template based model utilizes a pure spectral representation and the second uses a chroma based representation. An overall evaluation of the model and comparison between the two methods are shown using a test audio collection. Performance results are presented for different profiles and a variety of analysis durations. Results are encouraging and show that template based models are viable for key finding from audio.

1. INTRODUCTION

Tonality is the system by which listeners organize pitches around a most stable pitch while listening to music. Key finding, as it relates to Western tonal music, is the problem of estimating the key of a piece in terms of the most stable pitch, called the tonic, and the mode of the musical scale used. Applications of key finding in music understanding reach such diverse fields as automated music analysis, music information retrieval, music perception and machine learning.

Many models that address the problem of key finding have been reported in the literature and these can be classified into two main groups. The first is the group of models that use symbolic event-based input, such as MIDI data, and do not directly deal with the sonic properties of the input. All processing is performed on unambiguous and complete symbolic input. The second group of models directly use audio data as input. This group can also be further divided into two approaches, the first approach aims at performing transcription to convert audio data into symbolic form before applying a symbolic algorithm. Given that the state-of-the-art transcription accuracy is not satisfactory for this purpose, the algorithms in this category have to deal with incomplete, missing or probabilistic data. The second approach bypasses the transcription phase and directly operates on audio data. The proposed model in this paper belongs to the latter category and is a template based correlational model. It also adopts a structural approach to key finding in which note order and local timing have minimal effect on key estimation.

2. RELATED WORK

In the first group, many models have been proposed. A few examples demonstrating the variety of approaches are included here. Barthélemy and Bonardi [1] first extract figured bass information using chord templates and go on to finding key regions by applying an "island growing" approach. Krumhansl's algorithm [11] also known as the Krumhansl-Schmuckler algorithm, calculates correlations between a vector of pitch-class durations obtained from a musical fragment, and probe tone ratings obtained from human subjects [10]. Chew [3] uses a geometrical representation called the Spiral Array in which pitches are represented by points on a spiral and determines key with a Center of Effect Generator method. Temperley [17] uses a variation of Krumhansl's algorithm and finds keys in his preference rule-based analysis system. Shmulevich and Yli-Harja [16] extend Krumhansl's algorithm to operate in the form of sliding windows and apply median based filters to the output for smoothing. More detailed reviews can be found in [3][11] and [17].

In the second group there has been relatively less research. Leman [12] proposed a cognitive model for determination of tonal context and consequently tone centers based on an ear model front-end. The approach takes audio input and addresses issues such as consonance, fusion and self-organization in modeling tonal context. İzmirli [9] reported on a model that had a pitch-class note recognition front-end followed by a stage that consisted of leaky integrators to model recency effects and decay. In this model, as musical events were encountered, leaky integrators were charged according to respective strengths of pitch events. Huron and Parncutt [8] used a psychoacoustic model of pitch perception that employed echoic memory and pitch salience to model key perception.

Chroma based representations have been used in key finding [6][13], discovering similarity and repetition in audio recordings [2][7], and chord segmentation, recognition and alignment in audio [15]. A chroma based representation is a compact form of spectral representation obtained by a many-to-one mapping from the short-time spectrum of audio. Fujishima [5] originally proposed the Pitch Class Profile (PCP) for use in chord recognition. His algorithm is based on training the system with synthesized chords and

determining the nearest chord to the calculated PCP from the input.

Purwins, Blankertz and Obermayer [14] have proposed a model which calculates a constant-Q (CQ) transform, collapses the spectrum into CQ profiles, and calculates distances using a fuzzy distance measure between the profiles and reference CQ sets. They discuss their model's use in tonal center and modulation tracking. Gómez and Herrera [6] present a comparison of cognition-inspired models based on Krumhansl's method and feature-based machine learning methods for key finding from polyphonic audio. One of the features they use is the Harmonic Pitch Class Profile which is a specialized version of PCP that uses the peaks in the spectrum. Pauws' model [13] uses an auditory perception inspired front-end to compute a chromagram which is then used to compute the correlations with the Krumhansl and Kessler profiles [10]. Chuan and Chew's model [4] estimates pitch strength using peaks in the spectrum which are then used by the Spiral Array model to estimate key. Zhu, Kankanhalli and Gao [18] first find the tuning frequency of the input, perform partial tracking, apply consonance filtering, obtain a pitch profile, and determine the scale root and key separately. They report that their system performs better finding scale roots than scale roots and keys simultaneously.

3. MOTIVATION

As mentioned above, chroma based representations such as the PCP, chromagram or CQ profiles have been shown to be useful in key finding. In these approaches a chroma based representation is obtained by a many-to-one mapping from a spectrum into a 12-element vector that associates each element with a chroma. This is done by partitioning the spectrum into semitone-size regions, summing or averaging the energy within all semitone regions that belong to the same chroma over the entire analysis frequency range, and assigning the result to the related chroma element in the chroma vector (see [5] or [15] for the formula). It should be kept in mind that this mapping results in loss of specificity of the original note distribution and spectral characteristics of the instruments due to the inherent ‘aliasing’ during the mapping. It is well-known that this kind of representation is only a crude approximation to the actual chroma information and hence it is not possible to directly compare it to profile data such as the probe tone ratings [10]. Therefore, in order to perform comparative operations, templates need to be formed that are similar in their representation to the chroma vectors obtained from the musical input.

Although the key finding methods differ in the ways they process the spectrum before obtaining the chroma vector they all calculate correlations between this vector and some profile representing the ideal chroma distribution. The approach described in this paper uses templates obtained from real instrument sounds. The templates are formed from weighted spectra of these sounds organized in scale patterns. As the computation method of templates and the computation method of spectral and chroma representations from audio are the same, they hold comparable information. The motivation is to use templates as focal points in tonal space in order to classify incoming information and ultimately determine the key. A second aim of this work is to compare the pure spectral and chroma based representations in key finding from audio.

4. KEY ESTIMATION

4.1. Pitch Distribution Profiles

Krumhansl [11] suggested that tonal hierarchies are acquired through experience and that a pattern matching mechanism between the tonal hierarchies and the distribution of pitches in a musical piece might model the way listeners arrive at a sense of key. She suggested that tonal hierarchies for Western tonal music could be represented by the probe tone profiles found experimentally in an earlier study [10]. In the model presented here, profiles are incorporated into the calculation of templates to approximate the distribution of pitches in the spectrum. A base profile represents weights of individual chroma values and is used to model pitch distribution within a key. The profiles for all 12 keys are obtained by rotating this base profile, as, this distribution is invariant under transposition. Three different profiles are used in this study to approximate the pitch distributions. These are Krumhansl's probe tone ratings [11], Temperley's profiles [17] and a flat diatonic profile. The base profiles for all three are given in Table 1.

Chroma	TM	Tm	KM	Km	DM	Dm
0	5.0	5.0	6.35	6.33	1	1
1	2.0	2.0	2.23	2.68	0	0
2	3.5	3.5	3.48	3.52	1	1
3	2.0	4.5	2.33	5.38	0	1
4	4.5	2.0	4.38	2.6	1	0
5	4.0	4.0	4.09	3.53	1	1
6	2.0	2.0	2.52	2.54	0	0
7	4.5	4.5	5.19	4.75	1	1
8	2.0	3.5	2.39	3.98	0	1
9	3.5	2.0	3.66	2.69	1	0
10	1.5	1.5	2.29	3.34	0	0
11	4.0	4.0	2.88	3.17	1	1

Table 1. Profiles used in this study. T: Temperley, K: Krumhansl, D: diatonic, M: major, m: minor.

4.2. Spectral Templates

Recordings from monophonic instrument sounds are used in the calculation of templates. The sounds are sampled at 5512.5 Hz. The analysis is carried out using 50% overlapping 2048-point FFTs with a Hanning window. Analysis frequency range is taken to be from 50Hz to 2000 Hz. The spectrum of an individual

monophonic sound with index i, X_i, is computed by averaging windows that have significant energy over the duration of each sound and then scaling the average spectrum by its mean value.

A total of 24 templates are formed of which each template corresponds to a tonic and mode pair. The following indexing convention is used to represent the tonic-mode pairs for the templates: n=0:A (A major), n=1:Bb, n=2:B, n=3:C,... n=11:Ab, n=12:a (A minor), n=13:bb, n=14:b,...n=23:g#. The calculation of template n is performed by summing the spectra of sounds weighted by one of the profiles given in Section 4.1:

$$T_n = \begin{cases} \sum_{i=0}^{N-1} X_i \circ P_M((i-n+12) \bmod 12) & 0 \le n \le 11 \\ \sum_{i=0}^{N-1} X_i \circ P_m((i-n+24) \bmod 12) & 12 \le n \le 23 \end{cases} \quad (1)$$

X_i denotes the spectrum of note i where i=0 refers to note A in the lowest octave, i=1 refers to Bb a semitone higher and i=12 refers to A an octave higher etc. N is the number of sounds used which usually spans several octaves. The operation $\circ$ denotes multiplication of the spectrum vector X_i by the scalar weight $P_e(k)$, where e denotes the mode (M:major or m:minor) and k denotes the chroma index in the profiles given in Table 1. Note that T_n refers to a single template of choice and also that the individual bin indices have been omitted in notating T_n and X_i.

4.3. Chroma Templates

Chroma templates, C_n, are calculated from the spectral templates T_n. The mapping from the spectral templates to the chroma vectors is performed by dividing the analysis frequency range into 1/12th octave regions with respect to the reference A=440 Hz. such that the reference frequency is at the logarithmic center of the region associated with chroma A. Each chroma element in the template is then found by a summation of the magnitudes of the FFT bins over all regions with the same chroma value.

4.4. Summary Vectors

Spectral and chroma summary vectors are calculated from the polyphonic audio input using the same parameters used for calculating the templates. The spectral summary vector, U, is a summation of all windows in the analysis duration (e.g. 5 secs.) of the input. At this stage additional filtering is performed to leave out parts of the spectrum that do not contain pitch bearing data. The spectral flatness measure (SFM) is calculated for half octave subbands over the analysis frequency range. SFM is a measure of how peaky the spectrum is within the subband in question. The SFM is given by the ratio of the geometric mean to the arithmetic mean of the bins of the magnitude of the spectrum in the subband. Values closer to 1 indicate a flat spectrum and values closer to 0 indicate the existence of sharp peaks. Before being added to the summary vectors all half octave frequency regions of the spectrum are tested for spectral flatness and those regions that do not contain significant peak information are overwritten with values of 0. A threshold of 0.6 is used. Chroma summary vectors, D, are then obtained from spectral summary vectors, U, by applying the usual mapping.

4.5. Estimation

Spectral and chroma summary vectors are calculated for the beginning fragments of each piece. It is assumed that the musical works input to this model unambiguously establish the key indicated in their title in the first few seconds of the piece. The final step is the estimation of the key. This is done by computing correlation coefficients, $r_{Tn,U}$, between the spectral summary vector and all 24 spectral templates for a given profile. The index n that corresponds to the maximum correlation coefficient is reported as the key of the piece for the spectral method. Similarly, for the chroma method another correlation coefficient is calculated, $r_{Cn,D}$, and the winning index is reported as the key estimate of the piece. An evaluation of the model is given in the following section.

5. RESULTS

The spectral and chroma methods were tested on 85 pieces that were chosen randomly from a music streaming service (www.naxos.com.) The pieces were mainly chosen from the common practice period and an effort was made to obtain an even balance of keys in the test set. The distribution of keys is given in Table 2. Works from the following composers were used; the number of works taken from each composer is given in parentheses: J.S. Bach (2), Beethoven, (5), Brahms (1), Chopin (13), Clementi (5), Corelli (1), Dvorak (1), Handel (5), Haydn (2), Hofmann (8), Kraus (5), Mozart (18), Pachelbel (1), Scarlatti (7), Schubert (3), Scriabin (1), Telemann (1), Tchaikovsky (2), Vivaldi (4). The templates were obtained using piano sounds from the University of Iowa Musical Instrument Samples collection (http://theremin.music.uiowa.edu/MIS.html.)

Key	Frequency	Key	Frequency
A	3	a	4
Bb	5	bb	3
B	3	b	4
C	3	c	4
Db	3	c#	3
D	5	d	4
Eb	4	d#/eb	3
E	3	e	4
F	3	f	3
F#/Gb	4	f#	3
G	4	g	4
Ab	3	g#	3

Table 2. Distribution of keys in the audio file collection.

Both methods, denoted as spectral and chroma in Table 3, were tested for various analysis durations taken from the beginning of each piece as well as for different profiles. The table summarizes the performance of the model. Temperley, Krumhansl and Diatonic refer to the profiles given in Table 1. The combined profile in the first two rows of the table is obtained by multiplying Temperley's profile with the Diatonic profile. The highest recognition figure is 86% and is achieved with the Chroma-Temperley-Diatonic and the Chroma-Temperley combinations. The spectral method performs better with the Temperley-Diatonic profile than with the Temperley profile alone. Krumhansl's profile produces acceptable results but offers lower recognition than Temperley's and the combination profiles. The poor performance of the flat diatonic profile with respect to the other profiles shows that a non-uniform weighting is necessary.

Method-Profile	5 s.	7.5 s.	10 s.	15 s.
Spectral-Temperley-Diatonic	78%	81%	80%	77%
Chroma-Temperley-Diatonic	71%	86%	86%	82%
Spectral-Temperley	69%	80%	78%	75%
Chroma-Temperley	81%	85%	86%	84%
Spectral-Krumhansl	68%	71%	74%	67%
Chroma-Krumhansl	75%	78%	80%	77%
Spectral-Diatonic	57%	58%	55%	54%
Chroma-Diatonic	57%	59%	58%	58%

Table 3. Model performance according to method, profile and duration from the beginnings of the pieces.

6. CONCLUSIONS

The results from the evaluation of the template based approach to key finding are promising. The performance has been reported for four different profiles and four separate analysis durations ranging from 5 seconds to 15 seconds to show the dependency of the accuracy on the type of profile and analysis duration. Results indicate that the model reaches its peak recognition rate of 86% with an analysis duration between 7.5 and 10 seconds. Temperley's profile combined with a flat diatonic profile gives the best result. Results also show that both spectral and chroma methods could be utilized for key finding.

As future work, determination of templates from a training set will be considered. It is expected that forming the templates based on information of actual average spectral distributions that could be obtained from the training data will help improve performance. Another possible track of research entails the exploration of using a larger selection of templates, possibly with different scales, and profiles. The model also has direct applications in modulation tracking. Modulation points can be identified by implementing the method described in this paper as a sliding window model.

7. REFERENCES

[1] Barthélemy, J. and Bonardi, A. "Figured Bass and Tonality Recognition," *Proceedings of the International Conference on Music Information Retrieval*, Bloomington, Indiana, USA, 2001.

[2] Bartsch, M. A. and Wakefield, G. H. "To Catch a Chorus: Using Chroma-based Representations for Audio", *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2001.

[3] Chew, E. "Towards a Mathematical Model of Tonality", *Doctoral Dissertation*, Department of Operations Research, Massachusetts Institute of Technology, Cambridge, MA, USA, 2000.

[4] Chuan, C. and Chew, E. "Polyphonic Audio Key Finding Using the Spiral Array CEG Algorithm", *Proceedings of the International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 2005*.

[5] Fujishima, T. "Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music", *Proceedings of the International Computer Music Conference*, Beijing, China, 464–467, 1999.

[6] Gómez, E. and Herrera, P. "Estimating the Tonality of Polyphonic Audio Files: Cognitive versus Machine Learning Modelling Strategies", *Proceedings of the Fifth International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[7] Hu, N., Dannenberg, R.B. and Tzanetakis, G., "Polyphonic Audio Matching and Alignment for Music Retrieval," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2003.

[8] Huron, D. and Parncutt, R. "An Improved Model of Tonality Perception Incorporating Pitch Salience and Echoic Memory", Psychomusicology, 12(2), 154-171, 1993.

[9] İzmirli, Ö. and Bilgen, S. "A Model for Tonal Context Time Course Calculation from Acoustical Input", *Journal of New Music Research,* Vol.25, No. 3, 276-288, 1996.

[10] Krumhansl, C. and Kessler, E. "Tracing the Dynamic Changes in Perceived Tonal Organization in a Spatial Representation of Musical Keys", *Psychological Review*, 89, 334–368, 1982.

[11] Krumhansl, C. *Cognitive Foundations of Musical Pitch*. Oxford University Press, New York, 1990.

[12] Leman, M. "Tonal Context by Pattern Integration Over Time", *In D. Baggi (Ed.), Readings in Computer-Generated Music*, Los Altimos, CA: IEEE Computer Society Press. pp. 117-137, 1992.

[13] Pauws, S. "Musical Key Extraction from Audio", *Proceedings of the Fifth International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.

[14] 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.

[15] Sheh, A. and Ellis, D. P. W. "Chord Segmentation and Recognition using EM-Trained Hidden Markov Models", *Proceedings of the International Conference on Music Information Retrieval*, Baltimore, Maryland, USA, 2003.

[16] Shmulevich, I. and Yli-Harja, O. "Localized Key-Finding: Algorithms and Applications", *Music Perception*, 17, 4, p. 531-544, 2000.

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

[18] Zhu, Y., Kankanhalli, M. S. and Gao, S. "Music Key Detection for Musical Audio", *Proceedings of the 11th International Multimedia Modelling Conference*, Melbourne, Australia, 2005.