AN ALGORITHM FOR AUDIO KEY FINDING

Özgür İzmirli

Center for Arts and Technology
Connecticut College
270 Mohegan Ave.
New London CT, USA
oizm@conncoll.edu

ABSTRACT

An algorithm for audio key finding that participated in the 2005 Music Information Retrieval Evaluation Exchange (MIREX 2005) is presented. This algorithm takes a sound file that contains polyphonic audio as input and outputs a key estimate for this file. It is designed to operate on short fragments of audio taken from the beginnings of musical works. The algorithm consists of three stages: chroma template calculation, chroma summary calculation and overall key estimation. MIREX evaluation results are given with a breakdown of correctly identified keys and perfect fifth, relative major/minor and parallel major/minor errors. This algorithm produced the highest composite percentage score by a small margin among participating algorithms.

Keywords: MIREX, key finding.

1 INTRODUCTION

This paper outlines an algorithm for key finding from audio. A summary of the algorithm is given in this section. Subsequent sections provide more detail of the algorithm and finally evaluation results are discussed. Other variants and applications of this method are discussed in [1, 2].

The input to the algorithm is a sound file that contains the musical work for which the key is to be estimated. The algorithm analyzes fragments of audio taken from the beginnings of musical works. The input may contain polyphonic input. It is assumed that the pieces input to this algorithm start in the key that is the same as the one designated by the composer. The output consists of a single key estimate that is one of 24 possibilities – 12 for major and 12 for minor. The method has three stages: chroma template calculation using monophonic instrument sounds, chroma summary calculation from the input audio file and overall key estimation using the chroma templates and chroma summary information.

In the first stage, templates are formed using monophonic instrument sounds spanning several octaves. For this, initially, average spectra are calculated for each monophonic note. Next, spectral templates are formed as a weighted sum of the average spectra obtained from individual instrument sounds. Two types of weighting are performed. The first is done according to a pitch distribution profile. In general, the profile can be chosen to be one of Krumhansl's probe tone ratings [3], Temperley's profiles [4], flat diatonic profiles or combinations of these. The second is a weighting that is a function of the contributing note's (MIDI pitch) value. The first weighting is used to model the pitch class distribution and the second weighting is used to account for the registral distribution of notes. The resulting spectral templates are then collapsed into chroma templates. This process comprises a many-to-one frequency mapping for each chroma in order to form a 12-element chroma template. As a result 24 chroma templates are formed. These templates act as prototypes of chroma vectors for major and minor keys.

In the second stage, spectra are calculated from the input audio file and then mapped to chroma vectors. A summary chroma vector is obtained by averaging the chroma vectors in a window of fixed length. Overlapping windows of different lengths are used to obtain a range of localities. All windows start from the beginning of the piece and therefore longer windows contain information in the shorter windows as well as the new information in the later parts of their span. The lengths of the windows start from a single frame and progressively increase up to a maximum time into the piece.

The key is estimated in the third stage using the chroma templates and the summary chroma vectors. For each window, correlation coefficients are calculated between the summary chroma vector and all chroma templates. The template index with the highest correlation is regarded as the estimate of the key for that window. In order to find the most prevalent key estimate for a piece, the confidence of the estimate for each window is also found. Next, the total confidence over all windows is calculated for each plausible key. The key with the maximum total confidence is reported as the overall key estimate.

2 TEMPLATE CALCULATION

Templates act as prototypes to which information obtained from the audio input is compared. The purpose of constructing templates is to have an ideal set of reference chroma patterns for all possible keys. This section outlines the calculation of templates and the following section describes how the input audio is processed in order to perform the key estimation.

2.1 Instrument Sounds

In this algorithm, templates are obtained from recordings of real instruments, but, they could equivalently be obtained from synthetically generated harmonic spectra. A collection of sound files is used, with each file containing a single note and appropriately labelled to reflect its note content. For this algorithm, piano sounds from the University of Iowa Musical Instrument Samples online database have been used. The sounds were converted to mono from stereo and downsampled to a sampling rate of 11025 Hz. The frequency analysis is carried out using

50% overlapping 4096-point FFTs with Hann windows. The analysis frequency range for this algorithm is fixed at 55 Hz on the low end and 2000 Hz on the high end. In general, depending on the spectral content of the input a wider frequency range may be used.

2.2 Pitch Distribution Profiles

Pitch distribution profiles may be used to represent tonal hierarchies in music. Krumhansl [3] suggested that tonal hierarchies for Western tonal music could be represented by probe tone profiles. Her method of key finding is based on the assumption that a pattern matching mechanism between the tonal hierarchies and the distribution of pitches in a musical piece models the way listeners arrive at a sense of key. Many key finding models rely on this assumption and several extensions have been proposed. In one such extension, beside other additions, Temperley (2001) has proposed a pitch distribution profile. We utilize this profile in combination with a diatonic profile as this combination results in the best performance of the current model. The diatonic profile can be viewed as a flat profile which responds to presence or absence of pitches but is not sensitive to the relative importance of pitches. In operation alone, it resembles earlier approaches to key finding in which pattern matching approaches had been used. Table 1 shows the composite profile used in this model.

Profiles are incorporated into the calculation of templates to approximate the distribution of pitches in the spectrum and the resulting chroma representation. The base profile for a reference key (A in this case) has 12 elements, represents weights of individual chroma values and is used to model pitch distribution for that key. Given that this distribution is invariant under transposition, the profiles for all other keys are obtained by rotating this base profile.

Table 1. Diatonic, Temperley's and composite pitch distribution profiles. D_M: major, D_m: harmonic minor, T_M: Temperley major, T_m: Temperley minor, P_M: composite major and P_m: composite minor.

Chroma	D_M	D_m	T_M	T_m	P_M	P_m
0	1	1	5.0	5.0	5.0	5.0
1	0	0	2.0	2.0	0.0	0.0
2	1	1	3.5	3.5	3.5	3.5
3	0	1	2.0	4.5	0.0	4.5
4	1	0	4.5	2.0	4.5	0.0
5	1	1	4.0	4.0	4.0	4.0
6	0	0	2.0	2.0	0.0	0.0
7	1	1	4.5	4.5	4.5	4.5
8	0	1	2.0	3.5	0.0	3.5
9	1	0	3.5	2.0	3.5	0.0
10	0	0	1.5	1.5	0.0	0.0
11	1	1	4.0	4.0	4.0	4.0

2.3 Chroma Templates

The average spectrum of an individual monophonic sound with index i, X_i, is computed by averaging the spectra, obtained from windows that have significant energy, over the duration of the sound and then scaling the average spectrum by its mean value. Here, i=0 refers to the note A in the lowest octave, i=1 refers to Bb a semitone higher etc. R is the total number of notes within the instrument's pitch range used in the calculation of the templates. For this algorithm R is chosen to be 51. The lowest note is A1 and the highest is B5.

Using the average spectra obtained for each individual note, templates are calculated by weighted sums. A template for a certain scale type and chroma value is the sum of X_i weighted by the profile element for the corresponding chroma and by the second weighting that is a function of the note index (i). A template is calculated for each scale type and chroma pair resulting in a total of 24 templates as given in equation (1). The first 12 are major, starting from reference chroma 'A', and last 12 are minor.

$$C_n = \begin{cases} \Psi\left[\sum_{i=0}^{R-1} X_i\, f(i)\, P_M((i-n+12) \bmod 12)\right] \\ \qquad if \quad 0 \le n \le 11 \\ \Psi\left[\sum_{i=0}^{R-1} X_i\, f(i)\, P_m((i-n+24) \bmod 12)\right] \\ \qquad if \quad 12 \le n \le 23 \end{cases} \tag{1}$$

$P_e(k)$ is the profile weight as given in Table 1, where e denotes the scale type (M:major or m:minor) and k denotes the chroma. In this work, the profile is given by the elementwise product of the diatonic and Temperley profiles: $P_e(k)=D_e(k)T_e(k)$. f(i) is the secondary weighting function that accounts for registral distribution of notes. Here, it is chosen to be a simple decreasing function: $f(i)= 1-0.14i^{0.5}$. Ψ is a function that maps the spectrum into chroma bins. The mapping is performed by dividing the analysis frequency range into 1/12th octave regions with respect to the reference A=440 Hz. Each chroma element in the template is found by a summation of the magnitudes of the FFT bins over all regions that have the same chroma value.

3 SUMMARY CALCULATION

Once the profiles are calculated they become part of the model and are used for determining the key estimates for all audio input. i.e. one set of templates is used for all audio files in a dataset. The second stage of the method involves calculation of chroma summary vectors.

Initially, a chroma vector is calculated for each FFT frame from the audio input with the same analysis parameters used for calculating the templates. Next, the actual starting point of the music is found by comparing the signal energy to a threshold. This frame is made the pivot point for the remainder of the analysis. A summary chroma vector is defined to be the average of individual chroma vectors within a window of given length. All windows start from the pivot frame. The first window has a single frame and window length is progressively

increased in succeeding windows until the maximum analysis duration is reached. The maximum length of the audio to be analyzed is chosen to be approximately 30 seconds. This results in a sequence of summary chroma vectors where each summary vector corresponds to a window of specific length.

4 ESTIMATION OF KEY

The key estimate for an input sound file is determined from the individual key estimates corresponding to the various size windows and their associated confidence values. These two entities are determined as follows: For each window a key estimate is produced by computing correlation coefficients between the summary chroma vector and the 24 precalculated chroma templates and then picking the one with the maximum value. The confidence for an individual key estimate is given by the difference between the highest and second highest correlation value divided by the highest value. At this point each window has an associated key estimate and a confidence value. Finally, the total confidence for each plausible key is found by summing confidence values over all windows. A key is plausible if it has appeared at least once as an individual estimate in one of the windows. The key with the maximum total confidence is selected as the key estimate.

5 MIREX EVALUATION

MIREX 2005 provided the opportunity for empirical evaluation and comparison of algorithms in many areas related to music information retrieval. Algorithms participating in MIREX 2005 were submitted directly to the MIREX committee and the evaluations were run without intervention of the participants. The results of the MIREX 2005 evaluations were recently reported for all participating algorithms in audio key finding [5]. Beside other contests that took place during the exchange, a closely related category was symbolic key finding. The MIREX evaluation framework for audio key finding and symbolic key finding used the same dataset that consisted of 1252 pieces. The symbolic key finding algorithms used data directly from MIDI note files whereas sound files were synthesized from the same set of MIDI files for use in audio key finding. This enabled, for the first time, a performance comparison of symbolic key finding and audio key finding methods. It should be stressed, however, that because the audio material in the dataset was synthesized the results of the audio key finding cannot be generalized to actual audio recordings containing the same pieces.

Prior to evaluation, a test set of 96 pieces were made available to the participants for testing and calibrating their algorithms. The performance evaluation criteria were established before the actual evaluation started. According to these the performance of an algorithm was determined by the percentage of correctly identified keys as well as closely related keys. In order not to severely penalize closely related key estimates the following fractional allocations were used: correct key 1 point, perfect fifth 0.5, relative major/minor 0.3 and parallel major/minor 0.2 points. This was determined by the proposers at an early stage of the audio key finding contest proposal.

The audio dataset was reported to have two versions. Different synthesizers were used to generate the different versions - Winamp and Timidity. A percentage score was calculated for each version of the dataset taking into account the fractional allocations mentioned above. The composite percentage score was the average performance of the algorithms on the two datasets.

The algorithm explained in this paper performed as follows: Using the Winamp database, 1086 pieces were estimated correctly. Furthermore, an additional 36 estimates were perfect fifths of the correct key, 38 were relative major/minors and 17 were parallel major/minors. 75 of the estimated keys were considered unrelated. The percentage score for this database was 89.4 percent. Using the Timidity database, the algorithm found the correct key for 1089 pieces. For this database, an additional 42 estimates were perfect fifths of the correct key, 31 were relative major/minors and 18 were parallel major/minors. 72 of the estimated keys were considered unrelated. The percentage score for this database was 89.7 percent. The resulting composite percentage score was 89.55 percent.

This algorithm performed slightly better than the other algorithms in this evaluation exchange for the given dataset. The performances of the 7 participating algorithms ranged from 79.1 percent to 89.55 percent in their composite percentage scores.

6 CONCLUSION

In this paper, an algorithm for key finding from audio has been described and the MIREX 2005 evaluation results for this algorithm have been presented. According to the MIREX results the composite percentage score of this algorithm was 89.55 percent. The MIREX evaluation framework will hopefully continue to bring topics in music information retrieval research into focus and motivate research in this area.

ACKNOWLEDGEMENTS

Many thanks to the IMIRSEL group and the MIREX 2005 committee for organizing and performing the evaluation. Thanks also go to Arpi Mardirossian, Ching-Hua Chuan and Elaine Chew for proposing the audio key finding category for MIREX 2005 and organizing the submissions.

REFERENCES

[1] İzmirli, Ö. (2005) "Tonal Similarity from Audio Using a Template Based Attractor Model," Proceedings of the International Symposium on Music Information Retrieval (ISMIR2005), London, UK.

[2] İzmirli, Ö. (2005) "Template Based Key Finding From Audio," Proceedings of the International Computer Music Conference (ICMC2005), Barcelona, Spain.

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

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

[5] http://www.music-ir.org/evaluation/mirex-results/