Signal Processing 117 (2015) 198–207

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Musical key extraction using diffusion maps

Ofir Lindenbaum [a,*], Arie Yeredor [a], Israel Cohen [b]

[a] Tel Aviv University, Tel Aviv 69975, Israel
[b] Technion-Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form
30 April 2015
Accepted 10 May 2015
Available online 19 May 2015

Keywords:
Key extraction
Dimensionality reduction
Diffusion maps

ABSTRACT

We propose a method for automatic musical key extraction using a two-stage spectral dimensionality reduction (two consecutive mappings). First we build a data set representing the 24 Western musical keys, and then we use a nonlinear dimensionality reduction method, in order to understand the true manifold on which the musical keys lie. The order of the keys along the manifold is perfectly correlated with a cognitive model for the key space. We exploit this manifold in order to extract the musical key from a musical piece. Furthermore we propose three classifiers using the extracted manifold. The Classifiers work in two stages, by first estimating the mode and then by estimating the key within the estimated mode. Finally we examine our method on The Beatles data set and demonstrate its improved performance compared to various existing methods.

1. Introduction

Most Western musical pieces were written using the 24 possible diatonic musical keys. Musical keys describe the relation among pitches in the piece. This musical feature, which is retrievable by most musicians, is difficult to extract automatically using a computer. In this study we develop and examine an algorithm for automatic key extraction from raw data. Automatic key extraction has driven much recent research due to a large number of applications involved, for example, content based search [1], playlist generation, mosaicing, automatic accompaniment and disc jockey work. Recent studies focus on the task of extracting a musical key from raw data without the use of symbolic transcription; these studies have achieved reasonable results but cannot compete with a gifted musician.

To imitate the musician's intelligence, it is helpful to understand how an amateur musician might perform the task (at least according to the authors' own experience). By listening to a musical piece, the musician could create a tonal description (pitches or notes) of the song. Then, possibly by playing a musical instrument, the musician is able to determine which key is most appropriate for the given piece (a more proficient musician could improve this procedure). Most studies try to imitate this two-stage task by first extracting a tonal representation of the musical piece using a histogram of pitches [2] (12 semitones of the chromatic scale) represented in the chroma domain, or using a pitch class profile [3]. In the second stage the representation is classified into one of the 24 possible Western keys.

For the classification task Krumhansl and Kessler [4] have performed a cognitive probe-tone experiment and derived 24 typical key chroma profiles. These profiles aim to describe the significance of each pitch to the musical key; they represent the typical distribution of notes in a musical key. The profiles were modified by Temperley [5] and Gomez [6], who improved the classification results when analyzing Western music. The classification is performed by computing the correlation values to all 24

* Corresponding author. Tel.: +972 528783883.
E-mail address: ofirlin@gmail.com (O. Lindenbaum).

http://dx.doi.org/10.1016/j.sigpro.2015.05.006
0165-1684/

profiles and estimating the most prominent key. Recent algorithms such as [7] build a training set and derive a statistical average profile to represent each of the 24 keys. Both the statistical and the cognitive approaches for creating a typical profile can be described geometrically by a partition of the chroma space into 24 unrelated zones. However, the computation of the partition differs between the methods. In our study we will use a non-linear dimensionality reduction algorithm for the chroma space partition.

Many studies have been done in the field of cognitive music to model the structure of the tonal space [8]. One of the oldest models of this kind, which was developed by Leonard Euler in 1739 [9], is called Tonnetz (Tonal network), ordering all the harmonical similarities between notes in a 2-dimensional diagram. Other examples of these models are the Circle-of-Fifth, a double helix [10] and a torus [4]. In [11, 8] dimensionality reduction methods were used to visualize and confirm the existence of low dimensional structures in the tonal space. Chuan and Chew [12] used a Spiral Array space and proposed the Center of Effect model. According to this model each key is represented by a point on a spiral (more precisely a helix) and the classification is done by measuring the geodesic distance between points on the spiral. This approach can be viewed as creating an artificial manifold, then projecting a new data point to the manifold, and finally classifying the data using a partition of the manifold. Recent works by Peeters [13] use a Hidden Markov Model for the classification task.

The first hurdle is to extract the exact pitches from the musical piece. Due to the physical property of most musical instruments, playing a single note generates a fundamental frequency and all of its harmonics. This creates a problem when trying to extract pitches from a polyphonic musical piece because the harmonics of all of the instruments and human voices are mixed together. This problem can be solved by identifying the fundamental pitch and removing the harmonics from the spectral representation. These techniques have been used by Pauws [14]. Such models take into account the perceptual pitch and the musical background simultaneously. Chuan and Chew [12] proposed using a fuzzy analysis system, and Cremer and Derboven [15] proposed an overtone removal process. An alternative solution, implemented by Gomez [6], extends the Pitch Class Profile to Harmonic Pitch Class Profile by considering a theoretical amplitude contribution of the first four harmonics of each pitch within the three main triads in a given key. Genussov and Cohen [16] proposed approaching the problem using sparse representation methods. Various recent studies used Diffusion Maps (DM) [17,18], a non-linear dimensionality reduction method to extract and analyze unknown parameters from physical systems. These include among others: speaker identification [19], audio-visual recognition [20], classification of skeletal fibers [21]. In [8] tonal atonal classification was performed after applying DM to chroma representations of audio signal.

In this study, we extend this technique and address the challenging related task of key extraction. In the first part, we demonstrate the use of a dense DM for classification tasks of time varying signals. We show the improvement resulting from a dense time-domain blocks-processing, and we propose a novel approach for tuning the width-parameter of the DM kernel. The resulting dense diffusion mapping elucidates the low dimensional structure on which the keys lie (thereby corroborating the results of [11,8,22,23]). In the second part we use a two-stage mapping (one for the mode the other for the key) to propose three new classifiers of musical keys. Finally, we use the Beatles 179 songs data set as a test set and demonstrate the advantages of the proposed method compared to recent state-of-the-art algorithms.

The structure of the paper is as follows: Section 2 describes the methods and algorithms used and proposed in this work. In Section 3, we describe how we build and analyze the 24 keys training set. Experimental results are presented and analyzed in Section 4, followed by conclusions in Section 5.

2. Methods and algorithms

2.1. Tonal description

The first step of key extraction from raw audio data is extracting some tonal description of the musical piece. It is difficult to create an accurate transcription of the piece. However, we are not interested in a time representation of the piece, but rather in finding a description of the spectral energy corresponding to the pitches throughout the piece. We use a 12-D feature vector called Pitch Class Profile (PCP) to represent the tonal properties of the musical piece [2]. The PCP is derived from the chromatic scale. This scale is a 12 note musical scale, spaced with equal distances on a logarithmic scale starting at a basic note. It is a frequency domain vector showing the distribution of energy along the pitch classes [6] of a given musical piece. The frequencies are mapped onto a limited set of 12 chroma values (i.e., all octaves are wrapped into one). A common method for computing a PCP is the constant Q transform (CQT) [24] (used by [25] to track modulations in audio), a discrete spectral analysis of logarithmically spaced bins (similar to DFT). The L-bins CQT coefficients of a signal $\boldsymbol{s}[n]$ are computed as follows. The frequency range is first determined by selecting its lowest frequency $f_{\min}$ and its highest frequency $f_{\max}$. Then, denoting the desired number of bins per octave as β, the frequency center of the ℓ-th frequency bin set to $f_\ell = f_{\min} \cdot 2^{\ell/\beta}$, so that the total number of bins is $L = \beta \cdot \log_2(f_{\max}/f_{\min})$. The constant frequency-to-binwidth ratio is determined as $Q = (2^{1/\beta}-1)^{-1}$. The CQT coefficients are then given by

$$s_{\mathrm{cq}}[\ell] = \frac{1}{N_\ell} \sum_{n=o}^{N_\ell - 1} w_\ell[n] \cdot s[n] \cdot e^{-j2\pi n Q/N_\ell}, \quad \ell = 0,1,2,\ldots,L-1, \qquad (1)$$

where $w_\ell[n]$ is a window-function of n for extracting the ℓ-th CQT coefficient, and N_ℓ is the length of that window. The minimum required length of $w_\ell[n]$ is given (in samples) by $N_\ell = \lceil Q\frac{f_s}{f_\ell} \rceil$, where f_s is the sampling frequency.

Using $\boldsymbol{s}_{\mathrm{cq}} = [s_{\mathrm{cq}}[1], \ldots, s_{\mathrm{cq}}[72]]$, we compute the PCP vector $\boldsymbol{c}_s$ of $\boldsymbol{s}[n]$ by summing all corresponding bins from different octaves into a 12-D vector $\boldsymbol{c}_s$ whose bth element