Musical Key Extraction from Audio Using Profile Training

Steven van de Par, Martin McKinney, André Redert
Philips Research Laboratories Eindhoven
High Tech Campus 36
5656 AE Eindhoven, the Netherlands
{Steven.van.de.Par,Martin.McKinney,Andre.Redert}@philips.com

Abstract

A new method is presented for extracting the musical key from raw audio data. The method is based on the extraction of chromagrams using a new approach for tonal component selection taking into account auditory masking. The extracted chromagrams were used to train three key profiles for major and three key profiles for minor keys. The three trained key profiles differ in their temporal weighting of information across the duration of the song. One profile is based on uniform weighting while the other two apply emphasis on the beginning and ending of the song, respectively. The actual key extraction is based on comparing the key profiles with three average chromagrams that were extracted from a particular piece of music using the same temporal weighting functions as used for the key profile training. A correct key classification of 98% was achieved using non-overlapping test and training sets drawn from a larger set of 237 CD recordings of classical piano sonatas.

Keywords: Key Extraction, Chromagram, Audio, Music.

1. Introduction

Musical key extraction is important for a number of reasons; the major/minor distinction has been linked to the emotional connotation of music [1], the key can be used for further automatic music analyses, and also DJing and mixing requires key.

Automatic key extraction from raw audio data poses two challenges. First, tonal components have to be extracted from the raw audio data to obtain some kind of harmonic representation of the actual notes that are being played. Second, based on the harmonic representation, an interpretation has to be performed that leads to a key classification. Systems that work with symbolic data, e.g. MIDI, are able to skip the first step. Here we address the more difficult problem of also extracting a harmonic representation from raw audio data such as done in a number of recent studies [2, 3, 4].

The second step has been addressed in various ways, e.g. by mapping the harmonic representation to a so-called spiral array which is a topology that allows a better interpretation of harmonic relatedness of notes [4], or by extracting chromagrams (e.g. [2]) and comparing these to key profiles based on perceptual data [5].

The method presented here is an extension of the work of Pauws [2] and is based on chromagram extraction from raw audio data and comparison of the chromagrams to a set of key profiles that are based on training data.

2. Method

In the training phase of the algorithm, chromagrams are extracted on a segment-by-segment basis from the input audio (1024 sample segments with a sampling frequency of 16 kHz). Each chromagram consists of 12 numbers that each represent the prevalence of a particular note value in a segment. Chromagrams are averaged across time with three different weighting functions. Because the key is known in the training phase, all chromagrams can be normalized to C-major or c-minor keys and averaged across all training pieces such that two sets of key profiles result (for major and minor key) each consisting of three key profiles with different temporal emphasis.

2.1. Chromagram extraction

As a first step for the chromagram calculation, tonal (sinusoidal) components need to be extracted. The initial tonal component selection is based on a peak-picking method in the DFT (discrete Fourier transform) domain that selects local maxima. In order to refine the frequency resolution, the method of Desainte-Catherine and Marchand [6] is applied where the spectrum of the windowed *differentiated* signal is divided by the spectrum of the windowed *original* signal. Using the peak positions in the original spectrum, the divided spectra allow for a very accurate estimate of frequency.

A low complexity masking model is applied which calculates the energy within three consecutive auditory filter bands (cf. [7]) of the signal spectrum using a spectral weighting function yielding a so-called excitation value. When a tonal component within that band has an energy below the excitation value by a certain margin, it is assumed to be masked, and that component is discarded. In this way only lower-order harmonics will generally be used for the chromagram calculation. This leads to improvements in performance of the algorithm.

For the selected components, note values are calculated assuming that A has a frequency of 440 Hz. Amplitudes of each component are transformed to the dB domain to obtain more equal weighting of components with different amplitudes. Based on these components a histogram is made for each note value which constitutes the chromagram.

Because pieces of music typically start and end in the tonic, we weight these sections more in the key calculation. The temporal weighting functions used for this purpose are exponentially damped and ramped functions, respectively. A damping of 90% is reached after 15 seconds from the beginning or ending of the piece.

2.2. Key extraction

Once the key profiles have been trained according to the method described above, a key can be extracted for a particular piece of music. Three chromagrams are extracted using the three temporal weighting functions.

Each of these extracted chromagrams is correlated to their respective key profiles, yielding three correlation values that relate to the beginning ($c_b(k, s)$), ending ($c_e(k, s)$), and totality of the song ($c_t(k, s)$). Here k denotes the key index ($k = 1$ for C, $k = 2$ for C-sharp, etc.) and s denotes the key denominator (major/minor).

The correlation values are combined into a final correlation value according to $c_f(k, s) = \alpha c_t(k, s) + c_b(k, s) + c_e(k, s)$, where α adjusts the relative contribution of c_t versus c_b and c_e. The values of k and s that result in the maximum value of $c_f(k, s)$ indicate the key that was detected by the algorithm.

3. Evaluation

The key extraction algorithm was evaluated using a database of 237 classical piano pieces including pieces of Bach, Shostakovich, Brahms, and Chopin identical to the database used by Pauws [2].

This database was randomly split into a disjunct training and test set, where the training set covered 98% of the total database. This procedure was repeated 5000 times to obtain an accurate estimate of the average performance of the key extraction algorithm on this database.

In Table I, key classification results are shown for a number of conditions. The standard and best condition (first column) was with $\alpha = 2$, masking module on, and dB conversion of the tonal components before accumulation into a chromagram. The other columns show several variations of the best algorithm where the algorithm was retrained for the particular variation. When $\alpha = \infty$ it means that only the chromagram from the whole the song is used.

Table 1. Correct key classifications for various modifications of parameters and settings. The first column shows the best condition with $\alpha = 2$, including dB transformation on tonal components and masking. The next columns show different values of α, and the last two columns show the results with $\alpha = 2$ but without using a dB transformation and without using the masking model. Standard errors of the mean percentages in this table are all less or equal to 0.2.

$\alpha = 2$	$\alpha = 0$	$\alpha = 1$	$\alpha = 4$	$\alpha = \infty$	No dB	No Mask
98.1	95.6	96.8	97.6	83.7	91.8	88.7

4. Conclusions

The method for extracting musical key that was presented here achieves 98% correct key classification on a set of classical piano pieces. This method is based on different key profiles for beginning, ending, and totality of the song. When instead of different profiles only one key profile is used, only 75.1% (no profile training) to 83.7% (with profile training) performance is achieved (see [2] and Table I with $\alpha = \infty$). An advantage of this method is that it is based on the training of key profiles from raw audio reference data which allows for a good match between training and test data. In addition, it is beneficial to have a compressive (dB) transformation of amplitudes of tonal components to get a more equal contribution of components of different levels and to use a masking model to remove non-salient components.

5. Acknowledgments

We thank Armin Kohlrausch and Janto Skowronek for their helpful comments.

References

[1] M.P. Kastner and R.G. Crowder, “Perception of the major/minor distinction: IV. Emotional connotations in young children.” Music Perception, 1990, Vol. 8, No.2, pp 189-202.

[2] S. Pauws, “Musical key extraction from audio,” in *ISMIR 2004 Fifth Int. Conf. on Music Inf. Retr. Proc.*, 2004, paper 142.

[3] W. Chai and B. Vercoe, “Detection of key change in classical piano music,” in *ISMIR 2005 Sixth Int. Conf. on Music Inf. Retr. Proc.*, 2005, pp 468-473.

[4] C-H. Chuan and E. Chew, “Fuzzy analysis in pitch class determination for polyphonic audio key finding,” in *ISMIR 2005 Sixth Int. Conf. on Music Inf. Retr. Proc.*, 2005, pp 468-473.

[5] C.L. Krumhansl, *Congnitive Foundations of Music Pitch,*, Oxford University Press, New York, 1990.

[6] M. Desainte-Catherine and S. Marchand, “High-precision Fourier analysis of sounds using signal derivatives” J. Audio Eng. Soc.,vol. 48, no. 7/8, July/Aug. 2000, pp. 654-667

[7] B.R. Glasberg and B.C.J. Moore, “Derivation of auditory filter shapes from notched-noise data,” Hearing Research, 1990, Vol. 47, pp. 103-138