MUSICAL KEY EXTRACTION FROM AUDIO

Steffen Pauws
Philips Research Laboratories Eindhoven
Prof. Holstlaan 4
5656 AA Eindhoven, the Netherlands
steffen.pauws@philips.com

ABSTRACT

The realisation and evaluation of a musical key extraction algorithm that works directly on raw audio data is presented. Its implementation is based on models of human auditory perception and music cognition. It is straightforward and has minimal computing requirements. First, it computes a chromagram from non-overlapping 100 msecs time frames of audio; a chromagram represents the likelihood of the chroma occurrences in the audio. This chromagram is correlated with Krumhansl's key profiles that represent the perceived stability of each chroma within the context of a particular musical key. The key profile that has maximum correlation with the computed chromagram is taken as the most likely key. An evaluation with 237 CD recordings of classical piano sonatas indicated a classification accuracy of 75.1%. By considering the exact, relative, dominant, sub-dominant and parallel keys as similar keys, the accuracy is even 94.1%.

1. INTRODUCTION

Besides tempo, genre and music mood, musical key is an important attribute for Western (tonal) music though only musically well-trained people can identify the key in a piece of music easily [3]. Knowing the key of a piece of music is relevant for further music analysis or for music applications such as mood induction; the mode of the key is deemed to provide a specific emotional connotation [6].

1.1. Related work

The extraction of key from music audio is not new, but not often reported in literature (see, for instance, Leman's algorithm [8] for an exception in which human tone center recognition is modeled).

Many algorithms that are found in literature work on symbolic data only (e.g., MIDI or notated music) by eliminating keys if the pitches are not contained in the key scales [9], by looking for key-establishing harmonic aspects [4] or key-establishing aspects at accent locations [1], by using the tonal hierarchy [7] extended with the role of subsidiary pitches and sensory memory [5], by searching for keys in the scalar and chordal domain in parallel [12], by harmonic analysis [11], by median filtering using an inter-key distance [10], or by computing a inter-key distance using a geometric topology of tonality (i.e., the spiral array) [2].

2. METHOD

Our approach to key extraction starts by computing a chromagram over six octaves from A0 (27.5 Hz) to A6 (1760 Hz) from the raw audio data. This chromagram representation is used as input to the maximum-key profile correlation (MKC) algorithm to get the musical key.

2.1. Chromagram computation

The chromagram is defined as the restructuring of a spectral representation in which the frequencies are mapped onto a limited set of 12 chroma values in a many-to-one fashion. This is done by assigning frequencies to the 'bin' that represents the ideal chroma value of the equally tempered scale for that frequency. The 'bins' correspond to the twelve chromas in an octave.

To this end, the spectrum $S(f)$ is modelled as a combination of the spectral content of the perceptual pitch and the musical background, denoted as background level $\eta(f)$ at frequency f,

$$S(f) = \sum_{n=1}^{N} h^{n-1} W(nf) A(n) (\eta(f) + \delta(np_i - f)) \quad (1)$$

where the spectral pitch content is modeled as a scaled impulse train reflecting the interpretation of pitch as a harmonic series; it contains high energy bursts at integral multiples np_i, for harmonic index n. Further, N denotes the number of harmonics, $h(\leq 1)$ denotes the factor controlling peak contribution to the pitch percept, $W(\cdot)$ is an arc-tangent function representing the transfer function of the auditory sensitivity filter, and $A(n)$ is the gain for harmonic n.

The computation of Equation 1 is done by adding harmonically compressed amplitude FFT-based spectrum

representations, for which the following properties are implemented.

1. Spectral content above 5 kHz is cut off by downsampling the signal. It is assumed that harmonics in the higher frequency regions do not contribute significantly to the pitches in the lower frequency regions.

2. Only a limited number of harmonically compressed spectra are added. We use $N = 15$.

3. Spectral components (i.e., the peaks $A(n) = S(np_i)$) are enhanced to cancel out spurious peaks that do not contribute to pitches.

4. Spectral components at higher frequencies contribute less to pitch than spectral components at lower frequencies. We use $h = 0.83$.

5. The frequency abscissa is transformed to a logarithmic one by means of interpolation, since human pitch perception follows logarithmic laws.

6. A weighting function is used to model the human auditory sensitivity; the perceived loudness of a pitch depends on its frequency. We use an arctangent function.

From an algorithmic point of view, the input signal is partitioned in non-overlapping time frames of 100 milliseconds. If the signal is in stereo format, a mono version is created by averaging both channels first.

Since further processing considers only the musical pitches from A0 (27.5 Hz) to A6 (1760.0 Hz), the harmonic compression is done over 6 octaves from 25 Hz until 5 kHz, also to capture some harmonics of the higher pitch frequencies. So, spectral content at frequencies greater than 5 kHz are not taken into account. A low-pass filtering of 10 kHz by FIR approximation and a decimation process bandlimits and downsamples the signal. This down-sampling decreases dramatically the computing time necessities without affecting results seriously. The 'remaining' samples in a frame are multiplied by a Hamming window, zero-padded, and the amplitude spectrum is calculated from a 1024-point FFT. This spectrum consists of 512 points spaced 4.88 Hz on a linear frequency scale. Next, a procedure is applied aiming at enhancing the peaks without seriously affecting frequencies or their magnitudes. Only values at and around the spectral peaks are taking into account by setting all values at points that are more than two FFT points (9.77 Hz) separated from a relative maximum, equal to 0. The resulting spectrum is then smoothed using a Hanning filter.

Since a linear resolution of 4.88 Hz is far too limited for the lower pitch regions (the pitch frequency difference between C2 and C♯2 is 3.89 Hz), the values of the spectrum on a logarithmic frequency scale are calculated for 171 ($\lceil 1024/6 \rceil$) equidistant points per octave by cubic-spline interpolation. The interpolated spectrum is multiplied by a raised arc-tangent function, mimicing the sensitivity of the human auditory system for frequencies below 1250 Hz. The result is shifted along the logarithmic frequency scale, multiplied by a decreasing factor h and added for all harmonics to be resolved ($N = 15$) resulting in the harmonically compressed spectrum defined over at least six octaves.

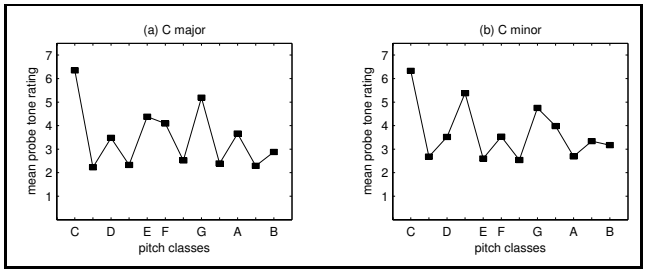

Figure 1. Mean probe tone rating (or key profiles) in the context of the key C major (a) and the key C minor (b). Adopted from [7]

The chromagram for each frame is computed by locating the spectral regions in the harmonically compressed spectrum that correspond with each chroma in A-440 equal temperament. For the pitch class C, this comes down to the six spectral regions centered around the pitch frequencies for C1 (32.7 Hz), C2 (65.4 Hz), C3 (130.8 Hz), C4 (261.6 Hz), C5 (523.3 Hz) and C6 (1046.5 Hz). The width of each spectral region is a half semitone around this center. The amplitudes in all four spectral regions are added and normalized to form one chroma region. Adding and normalizing the chromagrams over all frames results in a chromagram for the complete music sample.

2.2. Maximum key-profile correlation

The maximum key-profile correlation is an algorithm for finding the most prominent key in a music sample [7]. Originally, the algorithm was devised for symbolic encodings of music (e.g., MIDI). Here, it is used as a back-end to a signal processing step that works on raw audio data.

The MKC algorithm is based on key profiles that represent the perceived stability of each chroma within the context of a particular musical key. Krumhansl and Kessler [7] derived the key profile by a probe tone rating task. In this task, subjects were asked to rate, on a scale of 1 to 7, the suitability of various concluding pitches after they had listened to a preceding musical sample that established a particular key. The mean ratings represent the key profiles and show clear differences in the perceived stability of the chromas: highest ratings are given to the tonic, and the other two pitches of the triad, followed by the rest of pitches of the scale to be concluded by the non-scale pitches (see Figure 1).

Key profiles only depend on the relationship between a pitch and a tonal center and not on absolute pitches. Consequently, profiles for different major or minor keys are all transpositions of each other.

The MKC algorithm is based on the assumption that

the most stable chromas occur most often in a music sample. It computes the correlation (i.e., Pearson's product moment correlation) between the distribution of chroma occurrences in the musical sample and all 24 key profiles. Recall the chromagram takes the role of this distribution of chroma occurences given as a vector with 12 elements. The key profile that provides the maximum correlation is taken as the most probable key of the musical sample. The correlation value can be used as the salience of the perceived key or the degree of tonal structure of the music sample.

The strong point of the MKC algorithm is that it uses a differential weighting of all scale pitches and non-scale pitches. This means that the tonic, the perfect fifth and the third and all other pitches vary in importance in establishing a key at listeners.

3. EVALUATION

The evaluation of the algorithm consisted of a performance assessment in finding the correct key from a set of 237 performances of classical piano sonatas on CD. The correct key was defined as the main key for which the musical composition was originally composed. Recall that music composers use various key modulating techniques to build up tension and relaxation in the music. However, many compositions start and end with the same key; these pieces are called *monotonal*.

The following CDs were used. All 237 recordings of the CDs were used in the experiment.

Rosalyn Tureck
J.S. Bach, The Well-tempered Clavier Books I & II, 48 Preludes and Fugues
Recording New York 12/1952 - 5/1953
Remastering 1999, Deutsche Grammophon, 463 305-2

Jeno Jando
J.S. Bach, The Well-tempered Clavier Book I, 24 Preludes and Fugues
Naxos Classical, 8.55379697, 1995

Vladimir Askenazy
D. Shostakovich, 24 Preludes & Fugues, op.87,
Decca, 466 066-2, 1999.

Glenn Gould
J. Brahms, The Glenn Gould Edition,
Sony Classical, Sony Music Entertainment, 01-052651-10, 1993.

Evgeny Kissin
F.F. Chopin, 24 Preludes Op. 28, Sonate no. 2, Marche funebre / Polonaise op.53
Sony Classical, Sony Music Entertainment.

The original key of the compositions was compared with the extracted key from the CD-PCM data of the piano performances in fragments of 2.5, 5.0, 10.0, 15.0, 20.0, 25.0 and 30.0 seconds at the start, at the middle and at the end of the performances. Lastly, the complete performance was analysed. Note that a simple time measure of 4 beats at a tempo of 100 beats per minute takes 2.4 seconds. Also, note that the end of the performances (as found on the CDs) is often played by slowing down and sustaining the closing chord until it 'dies out'.

length (secs.)	start	middle	end
2.5	53.6% (127)	21.9% (52)	38.8% (92)
5.0	59.1% (140)	25.6%(61)	44.3% (105)
10.0	61.2% (145)	29.5% (70)	58.2% (138)
15.0	66.2% (157)	30.0% (71)	58.7% (139)
20.0	67.5% (160)	32.9% (78)	64.1% (152)
25.0	71.7% (170)	34.2% (81)	64.6% (153)
30.0	72.2% (171)	37.1% (88)	66.7% (158)

Table 1. *Classification accuracy for finding the exact main key in 237 piano sonatas for variable-length fragments from 2.5 to 30 seconds at various locations.*

exact	**75.1% (178)**
Rel.	6.8% (16)
V	1.3% (3)
IV	6.3% (15)
Par.	4.6% (11)
total	94.1%(223)

Table 2. *Classification accuracy for finding the exact main key in 237 complete piano sonatas. Confusions with the relative, dominant (V), sub-dominant (IV) and the parallel key are given. In the last row, the classification accuracy is shown if we consider all these related keys as correct keys.*

In Tables 1 and 2, the results are shown in terms of percentage correct. If we consider the algorithm as reliable, it is evident that most of the classical compositions can be termed as monotonal. As shown in Table 1, analysing the start and the end of a performance for at least a 25 second fragment provides a sensible judgement of the main key (i.e., 65-72% correct). The last five seconds of a classical piano performance provides a unreliable judgement of the main key, as too little data on harmony are present due to the last 'dying out' chord of most performances. Also, the middle of a performance should not be used to extract the main key since the composition might already have gone through various key changes. As shown in Table 2, to obtain the best possible judgement of the main key of a performance with an accuracy of 75.1%, the complete performance needs to be analysed.

The algorithm makes mistakes by confusing the exact main key with its relative, dominant, sub-dominant or parallel key. As shown in Table 2, this is considerable (about 4-7%) for the relative, sub-dominant and parallel keys. The cause of these key confusions needs to be sought in the way in which the piano sonatas are composed. However, since these keys are all 'friendly' to each other, they can all be considered as similar in particular music applications. Then, the classification accuracy amounts to 94.1%.

4. CONCLUSION

The present key extraction algorithm starts by computing the chromagram from raw audio data of a musical fragment. To this end, it extracts the likelihood of all possible pitches in the range from A0 (27.5 Hz) to A6 (1760 Hz) by computing harmonically compressed spectra in non-overlapping time frames of 100 msecs. The likelihood of all pitches are collected in a single octave and averaged for the complete musical fragment to arrive at a chromagram. The algorithm needs only minimum amount of computing necessities; it runs in parallel while the system is playing out music allowing online tracking of the harmonic progression in the music.

This chromagram is used in a correlative comparison with the key profiles of all 24 Western musical keys. These key profiles express what chromas are most important (i.e., most stable) in a given key on a rating scale from 0 to 7. The key which profile demonstrates the highest correlation with the provided chromagram is taken as the key of the musical fragment under study.

The algorithm identifies correctly the exact main key in 75.1% of the cases by analysing the complete CD recording of piano sonatas. If we assume exact, relative, dominant, sub-dominant and parallel keys as similar, it achieves a 94.1% accuracy. We have no data on recordings with other instrumentation or from other musical idioms.

The algorithmic performance *seems* to comply with human performance. Note that musically trained people can identify the correct key in 75% of the cases, tough after listening only to the first measure [3]. However, we do not know to what extent the algorithm and humans make similar mistakes. This is concern for further research.

Concluding, the following weak points of the current algorithm need attention:

- The raw audio data are taken as is, whereas a preprocessing stage might reveal fragments in a musical performance that contain key-relevant information and fragments that do not. An check on harmonisity and transients, for instance, may clearly discern fragments with harmonic instruments carrying prime information on musical key from noisy, percussive instruments.
- Music perceptive and cognitive factors that establish a musical key at a human listener can be further integrated into the algorithm. Temporal, rhythmic and musical harmonic factors of pitches are not modelled, whereas it is known, for instance, that the temporal order of pitches and the position of pitches in a metrical organisation (e.g., the first beat, strong accents) influence both the perception of a tonal center (i.e., the tonic of the key).
- Music theoretical and compositional constructs are not modelled in the algorithm. Composers use various key modulation techniques in which they signify how strong a new key will be established (i.e., Schenkerian analysis). For instance, a cadence in root position and more than three chords from the diatonic scale establish a strong (new) key in theoretical sense.

5. REFERENCES

[1] Chafe, C., Mont-Reynaud, B., Rush, L., Toward an intelligent editor of digital audio: Recognition of musical constructs. *Computer Music Journal, 6*, 30-41, 1982.

[2] Chew, E., An Algorithm for Determining Key Boundaries. *In Proceedings of the 2nd Intl Conference on Music and Artificial Intelligence* , 2002.

[3] Cohen, A.J., Tonality and perception: Musical scales prompted by excerpts from Das Wohl-temperierte Clavier of J.S.Bach. *Paper presented at the Second Workshop on Physical and Neuropsychological Foundations of Music, Ossiach, Austria*, 1977.

[4] Holtzman, S.R., A program for key determination, *Interface, 6*, 29-56, 1977.

[5] Huron, D., Parncutt, R., An improved model of tonality perception incorporating pitch salience and echoic memory, *Psychomusicology, 12*, 154-171, 1993.

[6] Kastner, M.P., Crowder, R.G., Perception of the Major/Minor Distinction: OIV. Emotional connations in young children, *Music Perception, 8, 2*, 189-202.

[7] Krumhansl, C.L., *Cognitive Foundations of Musical Pitch*, Oxford Psychological Series, no. 17, Oxford University Press, New York, 1990.

[8] Leman, M., Schema-based tone center recognition of musical signals, *Journal of New Music Research, 23*, 169-204, 1994.

[9] Longuet-Higgins, H.C., Steedman, M.J., On interpreting Bach, *Machine Intelligence, 6*, 221-241, 1971.

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

[11] Temperly, D., An algorithm for harmonic analysis. *Music Perception, 15, 1*, 31-68, 1997.

[12] Vos, P.G., Van Geenen, E.W. A parallel processing key-finding model. *Music Perception, 14, 2*, 185-224, 1996.