technische universität
dortmund

Definitions of Audio Features for Music Content Description

Wolfgang Theimer
Igor Vatolkin
Antti Eronen

Algorithm Engineering Report
TR08-2-001
Feb 2008
ISSN 1864-4503

Faculty of Computer Science
Algorithm Engineering (LS 11)
44221 Dortmund / Germany
http://ls11-www.cs.uni-dortmund.de/

Definitions of Audio Features for Music Content Description

Wolfgang Theimer[1], *Igor Vatolkin*[2], *Antti Eronen*[3]

[1]Nokia Research Center Bochum, Germany, [2]Technische Universität Dortmund, Germany,
[3]Nokia Research Center Tampere, Finland

Wolfgang.Theimer@ieee.org, Igor.Vatolkin@cs.uni-dortmund.de,
Antti.Eronen@nokia.com

CONTENTS

Contents 2
1. Introduction 4
2. Definition of Mathematical Terms and Variables 4
3. Mathematical Definition of Music Features 6
3.1 Tempo & Rhythm Features 6
3.1.1 Relative periodicity amplitude peaks 6
3.1.2 Beat times 6
3.1.3 Downbeat times 6
3.1.4 Duration of music piece 7
3.1.5 Fluctuation patterns 7
3.1.6 Median beat period 7
3.1.7 Median tatum period 7
3.1.8 Median tempo 8
3.1.9 Meter (duple / triple) 8
3.1.10 First periodicity peaks 8
3.1.11 Ratio of second and first periodicity peak 8
3.1.12 Sum of correlated components 9
3.1.13 Sum of periodicity amplitudes 9
3.1.14 Tatum times 9
3.1.15 Centroid of periodicity 9
3.2 Timbre & Energy Features 9
3.2.1 Angle average in phase domain 9
3.2.2 Distance average in phase domain 10
3.2.3 Normalized energy of harmonic components 10
3.2.4 Harmonic spectral deviation 11
3.2.5 Inharmonicity 12
3.2.6 Linear prediction coefficients 12
3.2.7 Loudness 12
3.2.8 Low energy 13
3.2.9 MFCC coefficients 13
3.2.10 Noisiness 13
3.2.11 Odd-to-even harmonic energy ratio 14
3.2.12 Perceived sharpness 14
3.2.13 Perceived spread of sound 15
3.2.14 Relative perceived loudness 15
3.2.15 Root mean square (RMS) 15
3.2.16 Spectral bandwidth 15
3.2.17 Spectral centroid 16
3.2.18 Spectral crest factor 16
3.2.19 Spectral discrepancy 16
3.2.20 Spectral flatness measure 17
3.2.21 Spectral flux 17
3.2.22 Spectral kurtosis 17
3.2.23 Spectral extent 18
3.2.24 Spectral skewness 18
3.2.25 Spectral slope 18
3.2.26 Sub-band energy ratio 19
3.2.27 Tristimulus 19
3.2.28 Variance & distance average between extremal spectral values 19
3.2.29 Variance & distance average between zerocrossings of the time-domain signal 20

3.2.30 Y-axis intercept 20
3.2.31 Zero-crossing rate 20
3.3 Harmony & Melody Features 20
3.3.1 Amplitude of maximum in the chromagram 20
3.3.2 Amplitudes of five main peaks 21
3.3.3 Chroma vector 21
3.3.4 Tone with maximum strength in the chromagram 21
3.3.5 Fundamental frequency 21
3.3.6 Fundamental tone of musical key 23
3.3.7 Musical mode of musical key 23
3.3.8 Pitch interval between two maximum peaks of chromagram 23
3.3.9 Positions of the main peaks 23
3.3.10 Width of the main peaks 23
3.4 Structural Features 24
3.4.1 Representative section 24
3.4.2 Sections 25
4. Derived Features 25
5. References 26

1. INTRODUCTION

This document is a result of a collaboration project between Nokia Research in Bochum and Tampere, the Institute of Music and Music Science and the Chair of Algorithm Engineering at Dortmund University. A comprehensive set of music features for content-based recommendation of music has been developed during the collaboration project. In order to make the features comparable and to verify the correct calculation, a mathematical definition, either as a closed-form formula or as an algorithm, must be agreed. In this technical report definitions of the music features plus references to the original sources are provided. The features can be categorized into four different groups:

1. tempo & rhythm
2. timbre & energy
3. harmony & melody
4. structural features.

2. DEFINITION OF MATHEMATICAL TERMS AND VARIABLES

- Closed interval $[a,b]$: All values including the boundaries a and b are taken into account.
- Open interval (a,b): All values excluding the boundaries a and b are taken into account. Also semi-open intervals exist: $(a,b]$, $[a,b)$.
- The infimum operator $\lfloor x \rfloor$ returns the biggest integer number equal or smaller to the argument x.
- Length of time window (number of samples): N
  Different time intervals need to be considered. For timbre information windows of 10 – 20 ms duration provide a quasi-stationary spectrum. For beat analysis the window must comprise several seconds to be able to observe multiple beats within one window.
- Total number of samples in a piece of music: N_{total}
- Total number of windows: L_{total}
- Discrete time signal: $x(n), \quad n \in [0, N_{total} - 1]$
- A signal estimate is marked with the ^ sign, e.g. an estimated signal in the time domain by $\hat{x}(n)$.
- Total number of frequency values: K
- Window function: $w_N(n), \quad w_N(n) = 0$ for $n < 0 \vee n \geq N$
- Discrete spectral signal: $X(k), \quad k \in [0, K-1]$

- Discrete Fourier transform and its inverse:

  $$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, \quad W_N = e^{-j\frac{2\pi}{N}}, K = N$$

  $$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) W_N^{-kn}$$

  The first $\frac{N}{2}$ values represent positive frequencies, the second $\frac{N}{2}$ values describe the spectral content for negative frequencies. For a real-valued signal (like we have) we only need to analyse the positive frequencies since the negative ones are redundant.

- Spectrum amplitude (in a given time window): $A(k) = |X(k)|$

- Spectrum amplitude in time window $l \in [1, L_{total}]$: $A^{(l)}(k)$

- Sampling frequency, measured in Hz: f_s

- Frequency $f \in \left[0, \frac{f_s}{2}\right)$ in Hz and its relationship to a frequency index $k \in \left[0, \frac{N}{2}\right)$:

  $$f(k) = \frac{k\ f_s}{N}$$

- Fundamental frequency, measured in Hz: f_0
  Used in conjunction with harmonic features evaluating (harmonic) multiples of a fundamental frequency.

- Total number of harmonics: I

- Amplitudes of harmonics: A_{h1} is the amplitude at the fundamental frequency, A_{hi} is the amplitude of the *i*th. partial.

- Total number of spectral peaks considered: J

- Spectral peaks with indices k_i ,and amplitudes A_i : k_1 is the location of the peak with the highest amplitude A_1, k_2 is location of the second highest peak with amplitude A_2, … .

- Autocorrelation function: $$r_{xx}(i) = \sum_{n=0}^{N-i-1} x(n)x(n+i)$$

- Estimated spectrum (e.g. as linear approximation): $\hat{A}(k)$.

3. MATHEMATICAL DEFINITION OF MUSIC FEATURES

3.1 Tempo & Rhythm Features

3.1.1 Relative periodicity amplitude peaks

- **Meaning:** The normalized amplitudes of first two peaks of the signal autocorrelation function with argument > 0, so that the sum of all periodicity values sums up to 1.

- **Definition:**
  Simple solution
  The relative peaks are defined as the maxima of the autocorrelation function $r_1 = \frac{r_{xx}(n_1)}{\sum_{i=0}^{N-1} r_{xx}(i)}$ and $r_2 = r_{xx}(n_2)$, where n_1 and n_2 are the displacements of the second and third highest peaks relative to absolute maximum at $n = 0$.

- **Reference:** /1/ modified by using autocorrelation function directly, /10/

3.1.2 Beat times

- **Meaning:** The beat is the basic metric level in music. It corresponds to the rate at which most people would tap their foot on the floor while listening to music. Beat times correspond to the points in time when the foot would hit the floor.

- **Definition:** The beat times can be produced e.g. using the method described in /10/ (no closed form solution exists). The main steps in the method are as follows. First, the signal is resampled to a fixed sample rate, to support arbitrary input sample rates. Second, an accent filter bank transforms the acoustic signal of music into a form that is suitable for beat and tatum analysis. In this stage, subband accent signals are generated, which constitute an estimate of the perceived accentuation on each subband. The accent filter bank stage significantly reduces the amount of data. Then, the accent signals are accumulated into four-second frames. Periodicity estimation looks for repeating accents on each subband. The subband periodicities are then combined, and summary periodicity is computed. Next, the most likely beat and tatum periods are estimated from each periodicity frame. This uses a probabilistic formulation of primitive musicological knowledge, including the relation, the prior distribution, and the temporal continuity of beats and tatums. Finally, the beat phase is found and beat and tatum times are positioned. The accent signal is filtered with a pair of comb filters, which adapt to different beat period estimates.

  The output is a sequence of floating point values, units given in seconds.

- **Reference:** /10/

3.1.3 Downbeat times

- **Meaning:** The downbeat is the first beat of a measure in music. For example, each measure in music with a 4/4 time signature consists of four beats, and the first one in each sequence of four beats is a downbeat.

- **Definition:** A sequence of floating point values, units given in seconds. The sequence indicates a subset of the song beats which are downbeats. Thus, the number of

downbeats is always less than the number of beats, and the downbeats coincide with beat times.

- **Reference:** /10/

3.1.4 Duration of music piece

- **Meaning:** Time duration of the complete piece of music in seconds
- **Definition:** $T_{total} = \frac{1}{f_s} N_{total}$
- **Reference:** --

3.1.5 Fluctuation patterns

- **Meaning:** Fluctuation patterns describe the amplitude modulation of loudness at different frequency bands /14/. Depending on frequency, loudness modulation has different effects on the perception. The perceived fluctuation strength is the strongest around 4Hz and gradually decreases up to a modulation frequency of 15Hz. At 15Hz the sensation of roughness starts to increase.
- **Definition:** The calculation of fluctuation patterns utilizes the energies from the bands of the mel-frequency filterbank. The steps in the processing are the following (/14/): 1) Process the mel-frequency spectrogram in short frames (e.g. three seconds) 2) Combine the mel-frequency bands into a smaller number of bands (e.g. from 36 down to 12). 3) For each frequency band use an FFT to compute the strength of the amplitude modulation at frequencies in the range of 0-10Hz. 4) Apply a weighting to the modulation frequencies using a model of perceived fluctuation strength. 5) In addition, /14/ suggests applying some additional filters to smooth the patterns. The resulting fluctuation pattern is a matrix with columns corresponding to frequency bands and rows corresponding to modulation frequencies. Further processing may involve vectorizing and averaging over the whole song to get a single vector descriptor for the song.
- **Reference:** Description of fluctuation pattern analysis (/14/), MFCC calculation 3.2.9

3.1.6 Median beat period

- **Meaning:** A representative beat interval calculated over the song.
- **Definition:** Median value of the sequence of beat times.
- **Reference:** /10/

3.1.7 Median tatum period

- **Meaning:** A representative tatum interval calculated over the song.
- **Definition:** Median value of the sequence of tatum times.
- **Reference:** /10/

3.1.8 Median tempo

- **Meaning:** Averaged tempo in beats per minute (BPM) over the song.
- **Definition:** $60 / \tau_B$, where τ_B is the median beat period in seconds.
- **Reference:** --

3.1.9 Meter (duple / triple)

- **Meaning:** After Goyon in /12/: “Two categories of musical meter are generally distinguished: duple and triple. This notion is contained in the numerator of the time signature: if the numerator is a multiple of two, then the meter is duple, if not a multiple of two but of three, the meter is triple. For instance, 2/4 and 4/4 signatures are duple, 3/4 and 9/8 are triple.” In addition, there are time signatures that do not fit the usual duple/triple categories, but their estimation is left for future work.
- **Definition:** Value of two for duple meter, three for triple meter. Goyon presents one method for performing the duple/triple decision /12/, but we have not implemented or tested it yet. A more detailed definition will be added when the analysis has been implemented and tested.
- **Reference:** /12/.

3.1.10 First periodicity peaks

- **Meaning:** The arguments of the periodicity peaks (displacements) describe the periods of repeating signal patterns like the tatum or beat. They are measured e.g. as beats / minute (BPM).
- **Definition:**
  The numbers n_1 and n_2 are the displacements of the second and third highest maxima of the autocorrelation function $r_{xx}(i)$ relative to the absolute maximum at $n = 0$.

  The autocorrelation function can replaced with a weighted summary periodicity function produced by the method described in /10/. More precisely, the weighted summary periodicity function can be taken as the summary periodicity vector $S[k]$ weighted with the beat prior. As the autocorrelation function, this function represents the strength of different periodicities in the signal, but is calculated in a more elaborate manner taking into account some aspects from musicological knowledge (the prior likelihood of different beat periods). For accurate beat estimation this kind of weighting and periodicity estimation is essential, and may thus give more robust periodicity features as well.
- **Reference:** /1/

3.1.11 Ratio of second and first periodicity peak

- **Meaning:** The ratio of the second and first periodicity peaks describes the dominance of either faster or slower beat structures.
- **Definition:** $\frac{r_2}{r_1}$

- **Reference:** /1/

3.1.12 Sum of correlated components

- **Meaning:** The (normalized) sum of correlated components provides a measure of the total amount of correlation within one window. White noise e.g. would have a value of $S = 0$ since there is no correlation for displacements $i > 0$.

- **Definition:** $S = \sum_{i=1}^{N-1} \frac{|r_{xx}(i)|}{|r_{xx}(0)|}$

- **Reference:** signal processing definition

3.1.13 Sum of periodicity amplitudes

- **Meaning:** This measures the periodicity power in the signal /12/.

- **Definition:** Sum of the values of the weighted summary periodicity vector (see 3.1.10).

- **Reference:** The weighted summary periodicity vector is calculated with the method presented in /10/.

3.1.14 Tatum times

- **Meaning:** The tatum is the shortest durational pulse that is more often than incidentally encountered in music. The term was coined by J. A. Bilmes in an MIT Master's thesis "Timing is of essence" published in 1993, and named after the jazz musician Art Tatum /11/.

- **Definition:** A sequence of floating point values in seconds. Every beat time coincides with a tatum time.

- **Reference:** The tatum times are produced with the method described in /10/.

3.1.15 Centroid of periodicity

- **Meaning:** The centroid of the periodicity vector is defined as the tempo for which half of the periodicity energy is contained in lower tempi.

- **Definition:** Centroid (or balancing point) of the weighted summary periodicity vector. The value is given in BPM.

- **Reference:** The feature has been described in /12/.

3.2 Timbre & Energy Features

3.2.1 Angle average in phase domain

- **Meaning:** The distribution of angles in the phase domain is a summary indicator of the signal dynamics.

- **Definition:**
  Vectors with m dimensions in the phase domain are created by taking a subsampled

sequence of m signal samples (subsampling factor d) starting at position i:
$\mathbf{p}_i = \left(x(i), x(i+d), x(i+2d), \ldots x(i+(m-1)d)\right)^T$

The number of dimensions is typically set to 2, but is not limited to this value.
An angle α_i in the phase domain is defined as the angle subtended between three consecutive phase domain points $\mathbf{p}_i$:

$\alpha_i = \angle(\mathbf{p}_{i-1}, \mathbf{p}_i, \mathbf{p}_{i+1})$, $\cos\alpha_i = \frac{\mathbf{s}_{iv}{}^T \mathbf{s}_{in}}{\| \mathbf{s}_{iv} \| \| \mathbf{s}_{in} \|}$ with $\mathbf{s}_{iv} = \mathbf{p}_{i-1} - \mathbf{p}_i$, $\mathbf{s}_{in} = \mathbf{p}_{i+1} - \mathbf{p}_i$.

The angle average is defined as $\mu_\alpha = \frac{1}{N-2-(m-1)d} \sum_{i=1}^{N-2-(m-1)d} |\alpha_i|$.

- **Reference:** /2/, /5/

3.2.2 Distance average in phase domain

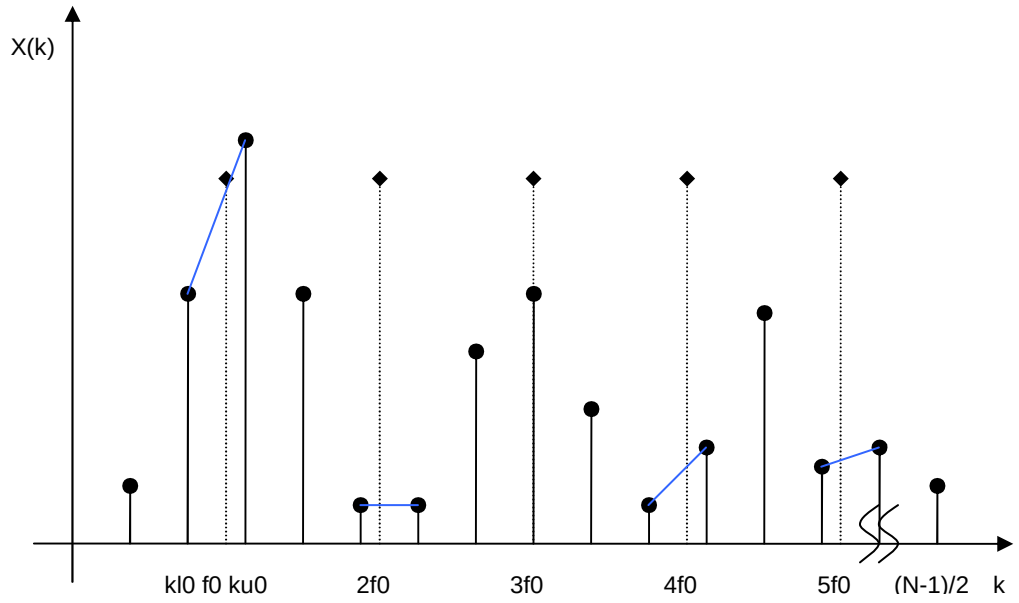
- **Meaning:** Another indicator for signal changes are the distances between adjacent points in the phase domain.
- **Definition:** The distribution of phase domain points can be described by the mean values for the distances between subsequent points $\mathbf{p}_i$:

  Mean value for distances: $\underline{\mu}_s = \frac{1}{N-2-(m-1)d} \sum_{i=1}^{N-1-(m-1)d} \left\| \mathbf{s}_{iv} \right\|$

- **Reference:** /2/, /5/

3.2.3 Normalized energy of harmonic components

- **Meaning:** The harmonic components are the amplitudes at the fundamental frequency of the signal and its integer multiples. The total harmonic energy characterizes the strength of a stationary sound without noise. It decreases during transitions or if the sound is noise-like. The harmonic energy is computed by linear interpolation of spectral values at the positions if_0, i.e. at multiples of the fundamental frequency. if_0 is not necessarily an integer number.
- **Definition:**



- The harmonic energy can be computed as a summation across the spectral energy at multiples of the fundamental frequency if_0: $E_{harmonic} = \sum_{i=1}^{I} \left(\tilde{X}(if_0) \right)^2$

  We need to compute the spectral values $\tilde{X}$ at multiples of the fundamental frequency by a linear interpolation

  $\tilde{X}(if_0) = X(k_{li}) + \left(\frac{if_0 K}{f_s} - k_{li} \right) \left(X(k_{ui}) - X(k_{li}) \right)$ between the adjacent sampled values

  with $k_{li} = \left\lfloor \frac{if_0 K}{f_s} \right\rfloor$, $k_{ui} = \left\lfloor \frac{if_0 K}{f_s} \right\rfloor + 1 = k_{li} + 1$ and the summation boundary index

  $I = \left\lfloor \frac{f_S}{2 f_0} \right\rfloor$.

  The normalized energy of harmonic components is the ratio of energy in the harmonics and the total spectral energy:

  $$S_{harmonic} = \frac{\sum_{i=1}^{I} \left(\tilde{X}(if_0) \right)^2}{\sum_{k=0}^{K/2-1} A^2(k)}$$

- **Reference:** own definition

3.2.4 Harmonic spectral deviation

- **Meaning:** The harmonic spectral deviation is the deviation of the amplitude at harmonic peaks from the global spectral envelope (averaged spectrum).

- **Definition:** $S_{dev} = \frac{1}{I} \sum_{i=1}^{I} \left(A_i - S(k_i) \right)$, where $S(k)$ is the global (averaged) spectrum

- **Reference:** /3/

3.2.5 Inharmonicity

- **Meaning:** The inharmonicity represents the divergence of the signal spectral components form a purely harmonic signal.
- **Definition:** $S_{inharm} = \frac{2}{f_0} \frac{\sum_{i=1}^{I} |f(k_i) - i\ f_0|\ A^2(k_i)}{\sum_{i=1}^{I} A^2(k_i)}$
- **Reference:** /3/

3.2.6 Linear prediction coefficients

- **Meaning:** The linear predicition coefficients $a(i)$ are the coefficients of an all-pole predictor of the signal $x(n)$ and can be used to estimate the value at point *n* as a linear combination of past values: $\hat{x}(n) = -\sum_{i=1}^{p} a(i)\, x(n-i)$.
- **Definition:** The LPC coefficients can be computed via the Levinson-Durbin recursion:

  1. For $n = 0$ set $E_0 = R(0),\ a_0(0) = 1$

  2. For step $n = 1...p$ do

  o $k_n = \frac{-1}{E_{n-1}} \sum_{i=0}^{n-1} a_{n-1}(i)\, R(n-i)$

  o $a_n(n) = k_n$

  o For $i = 1$ to $n-1$ do
  $a_n(i) = a_{n-1}(i) + k_n\, a_{n-1}(n-i)$

  o $E_n = E_{n-1}(1 - k_n^{\ 2})$

  3. The final result at iteration step p is $a(i) = a_p(i)$ for $i = 1...p$.

- **Reference:** /7/

3.2.7 Loudness

- **Meaning:** Loudness means the subjective judgment of the intensity of a sound. It can be derived from an auditory spectrogram by adding the amplitudes across all frequency bands. An example implementation of an auditory spectrogram is presented in /13/.
- **Definition:** A rough approximation of the loudness is obtained as the zeroth MFCC coefficient. This represents the sum across the amplitudes of the mel-frequency filterbank. The main difference to a “proper” auditory spectrogram is that masking is not taken into account in the mel-filterbank.

- **Reference:** /13/.

3.2.8 Low energy

- **Meaning:** Percentage of analysis windows where the RMS of the amplitudes are below the average RMS value across a texture window (consisting of a larger number of analysis windows). The analysis window has a length of around 10 – 20 ms with almost stable spectral characteristics. The texture window consists of several analysis windows with different short-time spectra that form a sound pattern, the “music texture”.

- **Definition:**
  Let N_a be the number of analysis windows in a texture window and $x_{rms}(n)$ be the RMS value in analysis window n (see later this section for the definition of RMS).
  The low energy rate r_{low} is defined as
  $r_{low} = \frac{1}{N_a}\sum_{i=0}^{N_a-1} u(\mu_{rms} - x_{rms}(n))$ where $\mu_{rms} = \frac{1}{N_a}\sum_{i=0}^{N_a-1} x_{rms}(n)$, where
  $u(x) = \begin{cases} 1 & for\ x > 0 \\ 0 & for\ x \le 0 \end{cases}$.
  In jAudio the preceeding 100 windows are taken into account as texture window ($N_a = 100$).

- **Reference:** /6/

3.2.9 MFCC coefficients

- **Meaning:** The cepstrum is the Fourier transform (or DCT) of the logarithm of the amplitude spectrum. The Mel frequency cepstral coefficients (MFCC) is the cepstrum that is computed on the Mel bands instead of the Fourier spectrum. The first coefficient, which is proportional to the signal energy is not stored, but the next 12 coefficients are stored.

- **Definition:** $x(n) \xrightarrow{FFT} A(k) \xrightarrow{Mel\ band} M(k) \xrightarrow{\log} \log M(k) \xrightarrow{DCT} MFCC(n)$
  The conversion from linear frequency values f to Mel values m is given by a logarithmic function: $m = 2595\ \log_{10}(1 + f/700)$:

  The spectral values on the linear frequency scale are integrated in triangular windows which are uniformly spaced on the Mel scale (i.e. they are logarithmically spaced on the linear frequency scale):
  $M(k') = \sum_{k=0}^{K/2-1} A(k) w_{k'}(k)$, where $w_{k'}(k)$ are triangular windows with increasing width for higher *k*.

- **Reference:** /3/

3.2.10 Noisiness

- **Meaning:** The noisiness of a signal is the ratio of the non-harmonic energy and the total energy.

- **Definition:** $S_{noi\sin ess} = \frac{\sum_{k=0}^{\frac{K}{2}-1} A(k)^2 - E_{harmonic}}{\sum_{k=0}^{\frac{K}{2}-1} A(k)^2}$

- **Reference:** /3/

3.2.11 Odd-to-even harmonic energy ratio

- **Meaning:** The odd to even harmonic energy ratio allows to distinguish odd harmonic energy predominant sounds (e.g. clarinet sounds) from sounds that contain harmonics in a more balanced way. It serves to differentiate between instruments. E.g. at the low frequency end of a clarinet playing range the odd harmonics are much stronger than the even partials.

- **Definition:** $S_{OER} = \frac{\sum_{i=1}^{\frac{I+1}{2}} A^2(k_{2i-1})}{\sum_{i=1}^{\frac{I}{2}} A^2(k_{2i})}$

- **Reference:** /3/

3.2.12 Perceived sharpness

- **Meaning:** The sharpness is the perceptual equivalent to the spectral centroid, but computed using the specific loudness of the Bark bands.

- **Definition:** The Bark frequency scale ranges from z=1 to 24 and corresponds to the first 24 critical bands of hearing. The subsequent band edges $f_{Bark}(i), \quad i \in [0,24]$ measured in Hz are 0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500. The conversion from Hz to Bark can be described by the following equation:
  $Bark = 13\arctan\left(\frac{f}{1315.8}\right) + 3.5\arctan\left(\frac{f}{7518}\right).$
  The perceived sharpness is defined as $S_{sharp} = 0.11\frac{\sum_{z=1}^{Z} z\, g(z)\, N'(z)}{N'_{total}}$, where z is the index of the Bark band and $g(z)$ is defined as
  $g(z) = \begin{cases} 1 & if\ z < 15 \\ 0.066\exp(0.171z) & if\ z \geq 15 \end{cases}$. The total loudness N'_{total} is defined as $N'_{total} = \sum_{z=1}^{Z} N'(z)$ with the individual loudness values $N'(z) = E(z)^{0.23}$ based on the

energy in the Bark bands:

$$E(z) = \sum_{k=k_{Bark}(z-1)}^{k_{Bark}(z)-1} A^2(k) \text{ and } k_{Bark}(z) = \left\lfloor \frac{K}{f_s} f_{Bark}(z) \right\rfloor.$$

- **Reference:** /3/

3.2.13 Perceived spread of sound

- **Meaning:** The spread of sound measures the distance from the largest specific loudness value to the total loudness.
- **Definition:** $S_{spread} = \left(\frac{N'_{total} - \max_z N'(z)}{N'_{total}} \right)^2$
- **Reference:** /3/

3.2.14 Relative perceived loudness

- **Meaning:** The relative perceived loudness is the specific loudness divided by the total loudness.
- **Definition:** $S_{relloud} = \frac{N'(z)}{N'_{total}}$
- **Reference:** /3/

3.2.15 Root mean square (RMS)

- **Meaning:** The root mean square provides a normalized measure of the signal energy in a time window.
- **Definition:** $x_{rms} = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} x^2(n)}$
- **Reference:** --

3.2.16 Spectral bandwidth

- **Meaning:** The spectral bandwidth is defined as the extent of the power transfer function around the the center frequency.
- **Definition:** $B^2_{rms} = \frac{\sum_{k=0}^{\frac{K}{2}-1} (k - S_{centroid})^2 A^2(k)}{\sum_{k=0}^{\frac{K}{2}-1} A^2(k)}$
- **Reference:** /8/

3.2.17 Spectral centroid

- **Meaning:** The spectral centroid is the center of gravity of the amplitude spectrum of the signal.
- **Definition:** $S_{centroid} = \frac{\sum_{k=0}^{\frac{K}{2}-1} k\ A(k)}{\sum_{k=0}^{\frac{K}{2}-1} A(k)}$
- **Reference:** /3/

3.2.18 Spectral crest factor

- **Meaning:** The spectral crest factor is a measure of the spectral flatness in a frequency band. It is a value between one (all spectral amplitudes are equal, example: white noise) and infinity (no spectral flatness at all).
- **Definition:** $S_{crest} = \frac{\max_{k\in[k_{il},k_{iu}]} A(k)}{\frac{1}{k_{iu}-k_{il}+1}\sum_{k=k_{il}}^{k_{iu}} A(k)}$, where the crest factor is computed in different frequency bands:
  - 250 – 500 Hz (k_{1l} - k_{1u})
  - 500 – 1000 Hz (k_{2l} - k_{2u})
  - 1000 – 2000 Hz (k_{3l} - k_{3u})
  - 2000 – 4000 Hz (k_{4l} - k_{4u}).
- **Reference:** /3/

3.2.19 Spectral discrepancy

- **Meaning:** The spectral discrepancy is the normalized sum of all spectral amplitude deviations from the linear regression line $\hat{A}(k)$.
- **Definition:** $S_{disc} = \frac{2}{K}\sum_{k=0}^{\frac{K}{2}-1}\left|A(k)-\hat{A}(k)\right|$
- **Reference:** /2/, /5/

3.2.20 Spectral flatness measure

- **Meaning:** The spectral flatness measure describes the noisiness versus sinusoidality of signal. If there is a sinusoidal structure, the spectral flatness is zero. If the signal consists mainly of noise, the spectral flatness measure is near one.

- **Definition:** $S_{flatness} = \frac{\left(\prod_{k=k_{il}}^{k_{iu}} A(k)\right)^{\frac{1}{k_{iu}-k_{il}+1}}}{\frac{1}{k_{iu}-k_{il}+1}\sum_{k=k_{il}}^{k_{iu}} A(k)}$, where flatness measure is computed for the following frequency bands:
  - 250 – 500 Hz (k_{1l} - k_{1u})
  - 500 – 1000 Hz (k_{2l} - k_{2u})
  - 1000 – 2000 Hz (k_{3l} - k_{3u})
  - 2000 – 4000 Hz (k_{4l} - k_{4u}).

- **Reference:** /3/ (used for audio fingerprinting)

3.2.21 Spectral flux

- **Meaning:** The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions.

- **Definition:** $S_{flux}(l) = \frac{2}{K}\sum_{k=0}^{K/2-1}\left(A^{(l)}(k) - A^{(l-1)}(k)\right)^2$

- **Reference:** /6/

3.2.22 Spectral kurtosis

- **Meaning:** The kurtosis gives a measure of flatness of a distribution around its mean value. It is computed from the 4th order moment.

- **Definition:** $S_{kurtosis} = \frac{m_4}{\sigma^4}$ with

  $m_4 = \frac{\sum_{k=0}^{\frac{K}{2}-1}(k - S_{centroid})^4 A(k)}{\sum_{k=0}^{\frac{K}{2}-1} A(k)}$ and $\sigma^2 = \frac{\sum_{k=0}^{\frac{K}{2}-1}(k - S_{centroid})^2 A(k)}{\sum_{k=0}^{\frac{K}{2}-1} A(k)}$.

- **Reference:** /3/

3.2.23 Spectral extent

- **Meaning:** The spectral extent is the frequency k_r so that a defined percentage p_{extent} of the signal energy is contained below this frequency. In Yale $p_{extent} = 80\%$.

- **Definition:** $\sum_{k=0}^{k_r} A^2(k) = p_{extent} \sum_{k=0}^{\frac{K}{2}-1} A^2(k)$

- **Reference:** /3/. Please note that this term is called roll-off frequency in /3/.

3.2.24 Spectral skewness

- **Meaning:** The skewness gives a measure of asymmetry of a distribution around its mean value. It is computed from the 3rd order moment.

- **Definition:** $S_{skew} = \frac{m_3}{\sigma^3}$ with

$$m_3 = \frac{\sum_{k=0}^{\frac{K}{2}-1} (k - S_{centroid})^3 A(k)}{\sum_{k=0}^{\frac{K}{2}-1} A(k)} \text{ and } \sigma^2 = \frac{\sum_{k=0}^{\frac{K}{2}-1} (k - S_{centroid})^2 A(k)}{\sum_{k=0}^{\frac{K}{2}-1} A(k)}.$$

- **Reference:** /3/

3.2.25 Spectral slope

- **Meaning:** The spectral slope represents the amount of spectral decrease as a function of frequency. It assumes that the amplitude spectrum follows a linear model: $\hat{A}(k) = m\,k + b$. The slope m is computed by linear regression.
  The spectral slope has been used e.g. in speaker recognition. It can also be derived from the output of the MFCC spectrum.

- **Definition:** $$m = \frac{\frac{K}{2}\sum_{k=0}^{\frac{K}{2}-1} k\,A(k) - \sum_{k=0}^{\frac{K}{2}-1} k \sum_{k=0}^{\frac{K}{2}-1} A(k)}{\frac{K}{2}\sum_{k=0}^{\frac{K}{2}-1} k^2 - \left(\sum_{k=0}^{\frac{K}{2}-1} k\right)^2}$$

  Reference: /2/, /5/

3.2.26 Sub-band energy ratio

- **Meaning:** The sub-band energy ratios are the fractions of the spectral energy in the frequency intervals between $\left[0,\frac{f_s}{16}\right),\left[\frac{f_s}{16},\frac{f_s}{8}\right),\left[\frac{f_s}{8},\frac{f_s}{4}\right),\left[\frac{f_s}{4},\frac{f_s}{2}\right)$ and can be translated to index ranges $\left[0,\frac{K}{16}\right),\left[\frac{K}{16},\frac{K}{8}\right),\left[\frac{K}{8},\frac{K}{4}\right),\left[\frac{K}{4},\frac{K}{2}\right)$
- **Definition:** $S_1=\frac{\sum_{k=0}^{\frac{K}{16}-1}A^2(k)}{\sum_{k=0}^{\frac{K}{2}-1}A^2(k)}$, $S_2=\frac{\sum_{k=\frac{K}{16}}^{\frac{K}{8}-1}A^2(k)}{\sum_{k=0}^{\frac{K}{2}-1}A^2(k)}$, $S_3=\frac{\sum_{k=\frac{K}{8}}^{\frac{K}{4}-1}A^2(k)}{\sum_{k=0}^{\frac{K}{2}-1}A^2(k)}$, $S_4=\frac{\sum_{k=\frac{K}{4}}^{\frac{K}{2}-1}A^2(k)}{\sum_{k=0}^{\frac{K}{2}-1}A^2(k)}$
- **Reference:** /8/

3.2.27 Tristimulus

- **Meaning:** The tristimulus values S_{T1}, S_{T2} and S_{T3} have been introduced as the timbre equivalent for the color attributes in vision. The tristimuli are three types of energy ratios allowing a description of the first harmonics of a signal. In /9/ the classification of different instruments based on S_{T2} and S_{T3} is sketched. Note that the sum of all tristimulus values is one, thus one value is redundant.
- **Definition:** $S_{T1}=\frac{A_{h1}}{\sum_{i=1}^{I}A_{hi}}$, $S_{T2}=\frac{A_{h2}+A_{h3}+A_{h4}}{\sum_{i=1}^{I}A_{hi}}$ and $S_{T3}=\frac{\sum_{i=5}^{I}A_{hi}}{\sum_{i=1}^{I}A_{hi}}$.
- **Reference:** /3/, /9/

3.2.28 Variance & distance average between extremal spectral values

- **Meaning:** The maximum spectral values represent the strongest tones in the music. The variance estimates the uniformity of the music: The variability is proportional to the variance.
- **Definition:** $\mu_k=\frac{1}{J-1}\sum_{j=2}^{J}(k_j-k_{j-1})$, $\sigma^2{}_k=\frac{1}{J-2}\sum_{j=2}^{J}\left[(k_j-k_{j-1})-\mu_k\right]^2$

  where the *J* highest spectral peaks are used with k_j and k_{j-1} being the frequency indices of neighboring peaks (i.e. the maxima are not sorted according to their magnitude).
- **Reference:** /2/, /5/

3.2.29 Variance & distance average between zerocrossings of the time-domain signal

- **Meaning:** The distance between zero values of the signal is a measure of the frequency content. The variance estimates the spectral variation.
- **Definition:**
  Let $y(n) = \begin{cases} 1 & if\ \mathrm{sgn}(x(n)) \neq \mathrm{sgn}(x(n-1)) \\ 0 & otherwise \end{cases}$ and the distances between peaks of $y(n)$ be $n_i - n_{i-1}$. Mean and variance between peaks of the new function are defined as
  $\mu_{zc} = \frac{1}{P-1}\sum_{i=2}^{P}(n_i - n_{i-1})$, $\sigma^2{}_{zc} = \frac{1}{P-2}\sum_{i=2}^{P}[(n_i - n_{i-1}) - \mu_{zc}]^2$ where P is the number of zerocrossings of the original signal $x(n)$.
- **Reference:** /2/, /5/

3.2.30 Y-axis intercept

- **Meaning:** The y-axis intercept is the interpolated spectral strength for frequency zero. It assumes that the amplitude spectrum follows a linear model: $\hat{A}(k) = m\,k + b$. The slope m is computed by linear regression (see above).
- **Definition:** $b = \frac{2}{K}\sum_{k=1}^{\frac{K}{2}-1} A(k) - \frac{2m}{N}\sum_{k=1}^{\frac{K}{2}-1} k$
- **Reference:** /2/, /5/

3.2.31 Zero-crossing rate

- **Meaning:** The zero-crossing rate is a measure for the high frequency content of a signal. It is strongly correlated with the spectral centroid (which can be computed using the MFCC spectrum).
- **Definition:** $r_{zero-crossing} = \frac{1}{2(N-1)}\sum_{n=0}^{N-2}\left|\mathrm{sgn}\,x(n+1) - \mathrm{sgn}\,x(n)\right|$
- **Reference:** --

3.3 Harmony & Melody Features

3.3.1 Amplitude of maximum in the chromagram

- **Meaning:** The maximum amplitude of the chromagram describes the strength of a tone on various octave levels.
- **Definition:** $A_{\max chroma} = \max_{\tilde{k}} A_{chroma}(\tilde{k})$

- **Reference:** /4/

3.3.2 Amplitudes of five main peaks

- **Meaning:** The amplitude of the five main peaks of the spectrum describes the strongest tones in the music.
- **Definition:** The definition can be found in the introductory section.
- **Reference:** --

3.3.3 Chroma vector

- **Meaning:** The chroma vector summarizes all spectral components belonging to the same pitch class. All frequencies are folded to one octave and the spectral amplitudes at the same half or quarter tone are summed together.
- **Definition:**

$$A_{chroma}\left(\tilde{k}\right) = \sum_{k:\ p(k)=\tilde{k}} A(k) \text{ with } p(k) = \left\lfloor 24\log_2\left(\frac{2k}{N}\frac{f_s}{f_1}\right)\right\rfloor \bmod 24 \text{ and } f_1 = 440\ Hz$$

  This is the definition for quarter tones (24 bins per octave). For the half-tone chroma vector replace 24 by 12.

- **Reference:** /4/

3.3.4 Tone with maximum strength in the chromagram

- **Meaning:** The frequency of the maximum of the chromagram shows the strongest tone.
- **Definition:** $\tilde{k}_{\max} = \arg\max_{\tilde{k}} A_{chroma}\left(\tilde{k}\right)$
- **Reference:** /4/

3.3.5 Fundamental frequency

- **Meaning:** The fundamental frequency is the frequency where the spectral contributions of multiples of this frequency contribute mostly to the total signal.
- **Definition:**
  1. Perform Hamming windowing of signal window and create FFT:

$$Y(k) = FFT(x(n)\cdot w_N(n)), \quad w_N(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

  2. Determine the positions k_j of spectral maxima A_j. Please note that the amplitudes are not sorted according to magnitude, i.e. $k_j > k_{j-1}\ \forall j$. Set all adjacent spectral values to the left to zero where derivative is positive and set all neighbouring spectral values to the right to zero where the derivative is negative. The search continues until more than M partials are found or until the amplitude of a partial is lower than a fraction of the strongest partial.
  3. Interpolate peak frequencies and amplitudes to improve resolution by the following formula for all peaks:

$$cor_j = \frac{0.5\left(\ln(A(k_j - 1) - \ln(A(k_j + 1))\right)}{\ln(A(k_j - 1)) - 2\ln(A(k_j)) + \ln(A(k_j + 1))}$$

Frequency: $f_j^* = \frac{f_s(k_j + cor_j)}{N}$

Amplitude: $A_j^* = \exp\left(\ln(A(k_j)) - 0.25\, cor_j \left(\ln(A(k_j - 1)) - \ln(A(k_j + 1))\right)\right)$

4. Perform dynamical thresholding by setting all modified amplitudes $A^*{}_j$ around k_j to zero that are less than $0.9 \max_{k \in [k_j - \Delta k, k_j + \Delta k]} \left(A^*(k)\right)$

5. Compute sequence of frequency differences

$\mathbf{f}_d = f_1^* - 0, f_2^* - f_1^*, f_3^* - f_2^*, \ldots$.

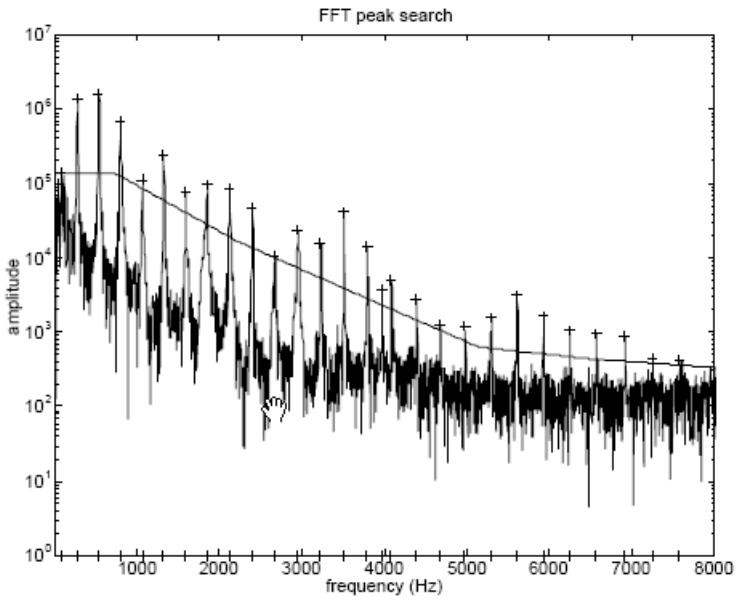
The first fundamental frequency estimate is the mean value: $fund = mean(\mathbf{f}_d)$

6. Take into consideration inharmonicity of sound, i.e. that the frequency difference of the higher partials can be very different from the fundamental frequency.
This is done by computing the difference of frequency differences
$\mathbf{f}_{dd} = f_{d2} - f_{d1}, f_{d3} - f_{d2}, \ldots$, e.g. $f_{dd2} = f_{d2} - f_{d1}$ and subtracting an estimate for the harmonic deviation from the actual estimate:

$f'_{di} = f_{di} - \frac{1}{L}\sum_{l=1}^{L} f_{dd(i-l)}$, where L is in the order of a few overtones (partials).

6. The fundamental frequency is the mean of the harmonic deviation-corrected frequency estimates:

$fund = mean(\mathbf{f}'_d)$



- **Reference:** /9/

3.3.6 Fundamental tone of musical key

- **Meaning:** Musical key describes the harmonics of a given music sequence or a complete music piece.
- **Definition:** An ordered group of tones defines musical key and describes a musical sequence or a complete musical piece. The central role is played by the fundamental tone, or tonic, which is strongly represented and establishes the relationship structure between used tones. In the occident music a fundamental tone is one of the twelve octave tones, e.g. C (of C Major).
- **Reference:** --

3.3.7 Musical mode of musical key

- **Meaning:** Musical key describes the harmonics of a given music sequence or a complete music piece.
- **Definition:** The distances between tones in the tone set of a musical key (measured in half-tones) influence the feeling and perceived impression of music. A concrete constellation of the tones corresponds to a certain musical mode, e.g. Major (of C Major) or Phrygian (of C Phrygian).
- **Reference:** --

3.3.8 Pitch interval between two maximum peaks of chromagram

- **Meaning:** The pitch interval describes the octave relationship between the strongest tones in the piece of music.
- **Definition:** The pitch interval is the frequency difference (measured in quarter tones) of the two highest peaks of the chromagram.
- **Reference:** /4/

3.3.9 Positions of the main peaks

- **Meaning:** The position k_i of the five main peaks = highest maxima in the spectrum provide the frequencies of the strongest tones.
- **Definition:** The definition can be found in the introductory section.
- **Reference:** --

3.3.10 Width of the main peaks

- **Meaning:**
- **Definition:** The algorithm determines the “peak borders” left (at position k_{li}) and right (at position k_{ri}) from each peak (at position k_i) and sets the width of each peak to $k_{ri} - k_{li}$. Starting from the highest amplitude (the first peak), the iterative divide-and-conquer technique finds the most prominent peaks to the left and to the right.

For the calculation of left peak border, the values $A(k_i), A(k_{i-1}),..., A(k_l)$ are consecutively analyzed. k_l is here the left border of the currently analyzed frequency interval. This interval is set to $\left[0, \frac{K}{2}\right]$ at the beginning of divide-and-conquer for the first peak. After the calculation of the peak width, the values from the interval $[k_{li}; k_{ri}]$ are taken away from the current interval. The divide-and-conquer proceeds with the new two remaining intervals and finds the further peaks and so on.

The left border is achieved if the first two of the following three constraints are satisfied at least s times ("sloppy values parameter" $s \in [1; \infty)$) or if the third constraint is satisfied:

Peak width where amplitude goes down to a certain percentage.

$$\begin{cases} A(k_{li}) < \frac{2}{K} \sum_{k=0}^{\frac{K}{2}-1} A(k) \\ A(k_{li}) > \tau \cdot A(k_{li+1}) \\ k_{li} = k_l \end{cases}$$

The first constraint is fulfilled if a current value is below the average amplitude. The meaning of the second constraint is to discard minor noise around the peak ("tolerance of variance" $\tau = 1.1$ on default). The last constraint is fulfilled if the left boarder of the examined frequency interval is achieved.

The right border is calculated similarly by analyzing the amplitudes to the right of peak position.

- **Reference:** /2/, /5/

3.4 Structural Features

3.4.1 Representative section

- **Meaning:** Representative section is a part of the music file which contains a representative part of the whole file. It may be e.g. a single instance of the chorus section.

- **Definition:** A representative section can be obtained with the method presented in /15/. The method utilizes a self-similarity matrix representation that is obtained by summing two separate distance matrices calculated using the mel-frequency cepstral coefficient and pitch chroma features. This is followed by detection of the off-diagonal segments of small distance from the distance matrix. From the detected segments, an initial chorus section is selected using a scoring mechanism utilizing several heuristics, and subjected to further processing. This further processing involves using image processing filters in a neighborhood of the distance matrix surrounding the initial chorus section.

The representative section is defined by its start and end time in seconds.

- **Reference:** /15/

3.4.2 Sections

- **Meaning:** Western popular music consists of distinguishable sections such as intro, verse, chorus, bridge outro. The goal of this descriptor is to present the different sections in the music file.

- **Reference:** See /16/ for one approach to music segmentation.

4. DERIVED FEATURES

A lot of features can be derived from the single-window oriented features to describe the relationship of features across time windows.

There are several ways to exploit the feature evolution over time in music:

- Compute the first derivative or higher derivates for features as a function of the window index.

- Compute the histogram of feature values across the complete piece of music. The sequence order of the values is lost in this case.

- Model the distribution of feature values by parameterized functions. A typical example is the Gaussian mixture model consisting of a summation of bell-shaped probability functions.

- Perform arithmetic operations on one or several windows of the basic features.

5. REFERENCES

/1/ D. McEnnis, C. McKay, I. Fujinaga and P. Depalle, “jAudio: A Feature Extraction Library”, in Proc. of the 6th International Conference on Music Information Retrieval (ISMIR), pp. 153 – 160, 2006

/2/ I. Mierswa and K. Morik, “Automatic Feature Extraction for Classifying Audio Data”, Machine Learning Journal, vol 58, pp. 127 – 149, 2005

/3/ G. Peeters, “A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project”, IRCAM, France, 2004

/4/ W. Chai, “Semantic Segmentation and Summarization of Music”, IEEE Signal Processing Magazine, March 2006

/5/ I. Mierswa, “Automatisierte Merkmalsextraktion aus Audiodaten”, Master Thesis, University of Dortmund, 2004

/6/ G. Tzanetakis and P. Cook, “Musical Genre Classification of Audio Signals”, IEEE Transactions on Speech and Audio Processing Vol 10, No. 5, pp, 293 – 302, 2002

/7/ T. Parsons, “Voice and Speech Processing”, McGraw-Hill, 1987

/8/ Z. Liu, J. Huang, Y. Wang and T. Chen, “Audio Feature Extraction & Analysis for Scene Classification”, in IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing (MMDSP97), 1997

/9/ K. Jensen, “Timbre Models of Musical Sounds”, PhD Thesis, Datalogisk Institut, Copenhagen University, Denmark, 1999

/10/ J. Seppänen, A. Eronen, J. Hiipakka, “Joint Beat & Tatum Tracking from Music Signals”, in Proc. of the 6th International Conference on Music Information Retrieval (ISMIR), p. 23 – 28, 2006

/11/ J. Seppänen, "Computational Models of Musical Meter Recognition", M.Sc. thesis, Tampere University of Technology, Tampere, Finland, 2001.

/12/ F. Goyon, “A Computational Approach to Rhythm Description – Audio Features for the Computation of Rhythm Periodicity Functions and Their Use in Tempo Induction and Music Content Processing”, PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2005.

/13/ T. Jehan, “Creating Music by Listening”, PhD Thesis, Massachusetts Institute of Technology, 2005.

/14/ E. Pampalk, “Computational Models for Music Similarity and their Application in Music Information Retrieval”, PhD Thesis, Technische Universität Wien, 2006.

/15/ A. Eronen, "Chorus Detection with Combined Use of MFCC and Chroma Features and Image Processing Filters", in Proceedings of 10th International Conference on Digital Audio Effects (DAFX-07), pp. 229-236, Bordeaux, France, 2007.

/16/ J. Paulus, A. Klapuri, "Music Structure Analysis by Finding Repeated Parts", in Proc. of the 1st Audio and Music Computing for Multimedia Workshop (AMCMM2006), Santa Barbara, California, USA, pp. 59-68, 2006