A large set of audio features for sound description (similarity and classification) in the CUIDADO project

Geoffroy Peeters

Ircam, Analysis/Synthesis Team, 1 pl. Igor Stravinsky,
75004 Paris, France
peeters@ircam.fr
http://www.ircam.fr/

version: 1.0 (23 avril 2004)

1 Introduction

In this report, we review the set of audio descriptors which has been developed and used in the framework of the CUIDADO I.S.T. project at Ircam.

1.1 Features taxonomy

Many different types of signal features have been proposed for the task of sound description coming from the speech recognition community, previous studies on musical instrument sounds classification (Foote 1997; Scheirer and Slaney 1997; Brown 1998; Martin and Kim 1998; Serra and Bonada 1998; Brown 1999; Wold, Blum et al. 1999; Jensen 2001) (Peeters and Rodet 2002; Peeters 2003; Peeters and Rodet 2003) and results of psycho-acoustical studies (Krimphoff, McAdams et al. 1994; Misdariis, Smith et al. 1998; Peeters, McAdams et al. 2000).

A systematic taxonomy of features is outside the scope of this paper; nevertheless we could distinguish features at least according to four points of view:

1. The *steadiness or dynamicity of the feature*, i.e., the fact that the features represent a value extracted from the signal at a given time, or a parameter from a model of the signal behavior along time (mean, standard deviation, derivative or Markov model of a parameter);
2. *The time extent of the description provided by the features*: some description applies to only part of the object (e.g., description of the attack of the sound) whereas other apply to the whole signal (e.g., loudness of a note);

We can thus distinguish between the time extend validity of the description

- **Global descriptors**: descriptors computed for the whole signal, which meaning is for the whole signal. Example of this are the attack duration of a sound. These descriptors requires to have a previous time localization of the sound events: the signal is either a sound sample or has been segmented into non-overlapping events.
- **Instantaneous descriptors**: descriptors computed for each time frame (a time frame is a short time segment of the signal which duration is around 60msec length). Example of this are the spectral centroïd of a signal which can varies along time.
  A temporal modeling module then process the time vectors of instantaneous descriptors in order to give the final descriptors.

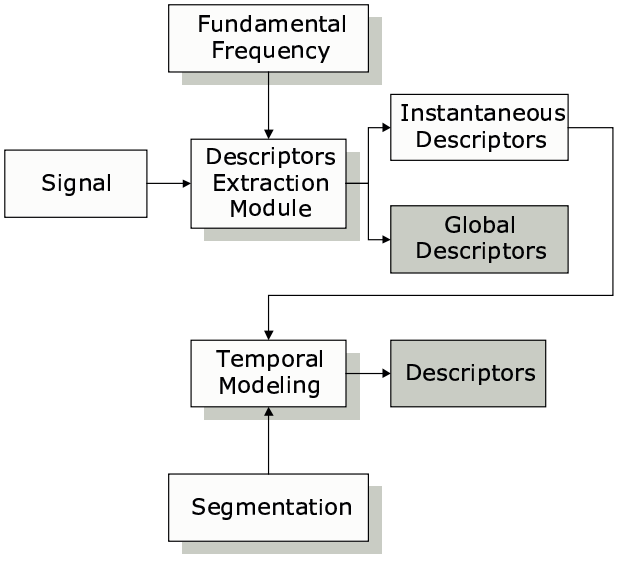


Figure 1 Global and instantaneous descriptor extraction flowchart

3. The *"abstractness" of the feature*, i.e., what the feature represents (e.g., cepstrum and linear prediction are two different representation and extraction techniques for representing spectral envelope, but probably the former one can be considered as more abstract than the latter);
4. The *extraction process of the feature*. According to this point of view, we could further distinguish:
   - Features that are directly computed on the waveform data as, for example, zero-crossing rate (the rate that the waveform changes from positive to negative values);
   - Features that are extracted after performing a transform of the signal (FFT, wavelet . . .) as, for example, spectral centroid (the "gravity center" of the spectrum);
   - Features that relate to a signal model, as for example the sinusoidal model or the source/filter model;
   - Features that try to mimic the output of the ear system (bark or erb bank filter output).

In the CUIDADO project, a large set of features has been implemented, including features related to the

- **Temporal shape:** features (global or instantaneous) computed from the waveform or the signal energy (envelop). Example: attack-time, temporal increase/decrease, effective duration,
- **Temporal feature**: auto-correlation coefficients, zero-crossing rate,
- **Energy features:** features (instantaneous) referring to various energy content of the signal. Example: global energy, harmonic energy, noise energy,
- **Spectral shape features:** features (instantaneous) computed from the Short Time Fourier Transform (STFT) of the signal. Example: centroid, spread, skewness, kurtosis, slope, roll-off frequency, variation), Mel-Frequency Cepstral Coefficients (plus Delta and DeltaDelta coefficients)

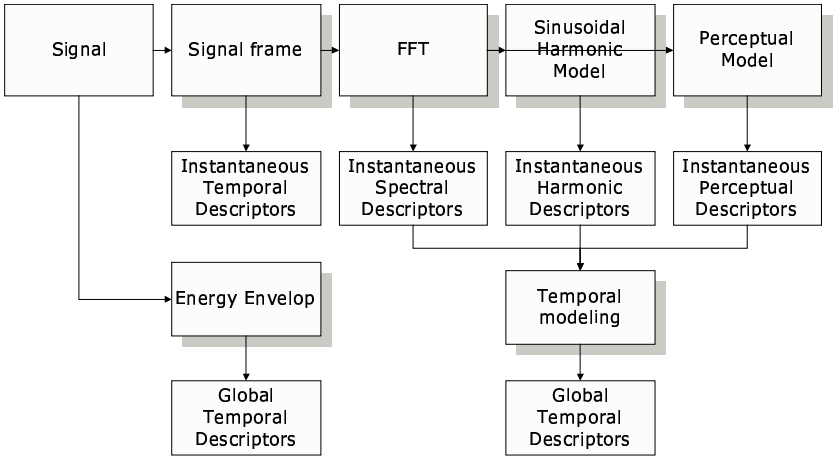


Figure 2 Details of temporal, energy, spectral, harmonic and perceptual descriptors extraction proces

- **Harmonic features:** features (instantaneous) computed from the Sinusoidal Harmonic modeling of the signal. Example: harmonic/noise ratio, odd to even and tristimulus harmonic energy ratio, harmonic deviation,
- **Perceptual features:** features (instantaneous) computed using a model of the human earring process. Example: relative specific loudness, sharpness, spread,
- MPEG-7 Low Level Audio Descriptors (spectral flatness and crest factors (MPEG-7 2002)).

1.2 Organization of the paper

In part 2, we indicate the various pre-processing stages needed for the extraction of the descriptors.
In part 3, 4, 5, 6, 7, 8 and 9, we present by family the various global and instantaneous descriptors.
In part 10, we present the temporal models applied to the instantaneous descriptors which have been used.
In part 11, we sum up in a table the whole list of extracted descriptors and their corresponding xml tag used in the CUIDADO project.

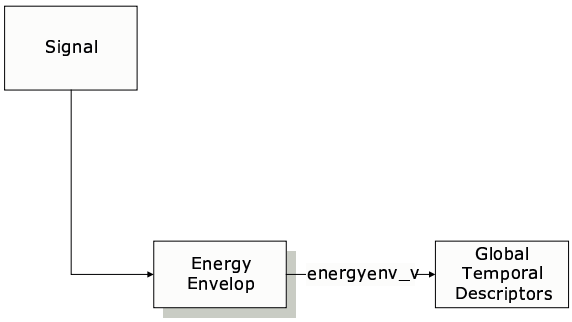
2 Pre-computing

The pre-computing stage of the extraction module provides the adequate signal representations for latter processing of descriptors extraction. It concerns

- the estimation of the energy envelop of the signal,
- the Short Time Fourier Transform (STFT)
- the sinusoidal harmonic modeling of the signal,
- a cascade of processing trying to mimic human earring process.

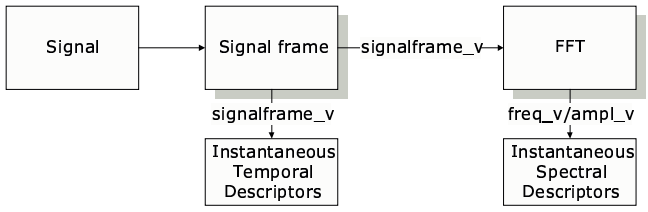
2.1 Energy envelop

Description: the energy envelop is used for the calculation of the global temporal descriptors: log-attack time, temporal centroïd, … It can be computed in several way: low-pass filtering of the analytical signal amplitude, … Although a simple and efficient implementation relies on the computing of the instantaneous *rms* (root mean square) values of the local signal. The window size (L=100msec) has been chosen in order to apply an equivalent low-pass filter with a cut-off frequency of 5 Hz.



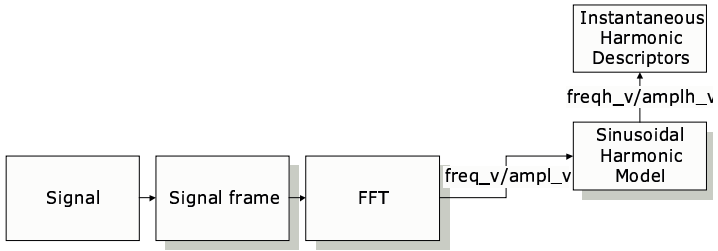
2.2 Short-Time Fourier Transform

A frame-by-frame analysis is performed using a window size of 60 msec and a hop size of 20 msec. This is the double of what is defined in the mpeg-7 scalable series (30 msec and 10 msec). This doubling allows a correct description of harmonic sounds with pitch down to 50 Hz (3 periods of 50 Hz = 60 msec).
The Short-Time Fourier Transform for a given time frame is obtained from the FFT of the corresponding signal frame.



2.3 Sinusoidal Harmonic modeling

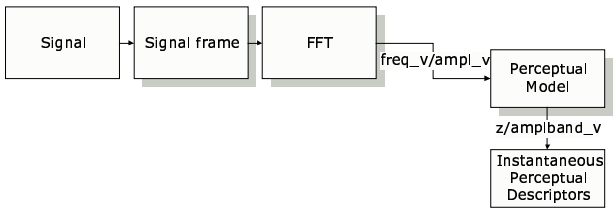
At each time frame, the peaks of the STFT of the windowed signal segment are estimated. Peaks close to multiples of the fundamental frequency at this frame are then choose in order to estimate the sinusoidal harmonic frequency and amplitude (Depalle, Garcia et al. 1993).



2.4 Perceptual model

Before computing the perceptual descriptors, several process are applied to the spectrum:

- mid-ear filtering,
- logarithmic band conversion (Mel or Bark bands).



2.4.1 Mid-ear filtering

In order to simulate the attenuation due to the human middle ear, we applied a filter to the STFT of each signal frame (Moore, Glasberg et al. 1997). The filter frequency response is represented in Figure 3.

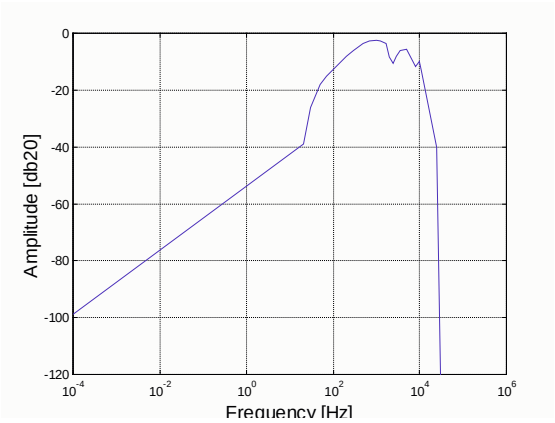


Figure 3 Mid-ear attenuation filter frequency response

2.4.2 Mel scale

Description: Human Auditory system behavior can be modeled by a set of critical band filter. The Mel bands are one of these. It is based on the Mel frequency scale, which is linear at low frequencies (below 1000 Hz) and logarithmic at high frequencies (above 1000 Hz). The Mel scale is specially popular in the Automatic Speech Recognition community where it is used for the calculation of the Mel Frequency Cepstral Coefficient (MFCC) (Rabiner and Juang 1993).

Conversion from Hz scale to Mel scale

- for f<1000Hz, $M = f$
- for f>1000Hz, $M = fc \cdot \left(1 + \log_{10}\left(\frac{f}{fc}\right)\right)$

Where M is the frequency expressed in Mel, f in Hz, and fc=1000Hz.

This is represented in Figure 4.

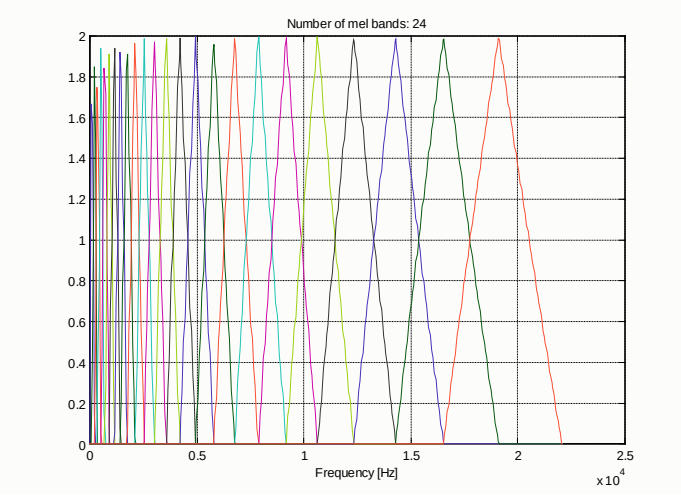


Figure 4 Mel bands

Calculation of the critical band energy

The linear frequency axe is first converted into Mel scale. The Mel scale axe is then divided into 20 equally spaced bands. After weighting by the Mel band window (we've chosen triangle shape windows), the energy of the bins k of the FFT corresponding to each Mel band z ($begin(z)<k<end(z)$) are then summed up to form the contribution to the band z.

$$amplband\_v(z) = \sum_{k=begin(z)}^{end(z)} A_k^2$$

where A_k is the amplitude of the bin k of the FFT

2.4.3 Bark scale

Description: Although we used the Mel bands for the MFCC calculation (because of its popularity in the ASR community), the Bark bands (Zwicker and Terhardt 1980) can model a better approximation of the Human Auditory system. This latter will be used for the calculation of the Loudness, Specific Loudness, Sharpness and Spread.

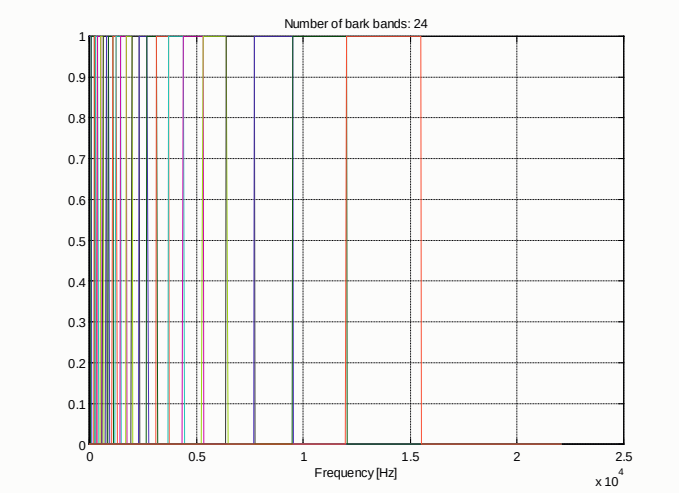


Figure 5 Bark bands

Conversion from Hz scale to Bark scale

$$B = 13 \cdot a\tan\left(\frac{f}{1315.8}\right) + 3.5 \cdot a\tan\left(\frac{f}{7518}\right)$$

Where B is the frequency expressed in Bark, and f in Hz.

Calculation of the critical band energy

The linear frequency axe is converted into Bark scale. The Bark scale axe is then divided into 24 equally spaced bands. The energy of the bins k of the FFT corresponding to each Bark band z are then summed up to form the contribution to the band z.

$$amplband\_v(z) = \sum_{k=begin(z)}^{end(z)} A_k^2$$

where A_k is the amplitude of the bin k of the FFT

2.5 Amplitude and Frequency scale

2.5.1 Amplitude scales

When features are extracted from the signal spectrum, from the harmonic peaks or from filter-banks, various amplitude scales are considered: linear amplitude, amplitude converted to an energy scale and amplitude converted to a log-amplitude scale:

- Linear Amplitude: amplitude
- Energy: amplitude^2
- Log-amplitude: log-amplitude

2.5.2 Frequency scales

When features are extracted from the signal spectrum, from the harmonic peaks or from filter-banks, various frequency scales are considered: linear frequency and frequency converted to a log-frequency scale centered on 1000 Hz.

- Linear frequency
- Log-frequency: Defined with respect to a frequency of 1000 Hz

logfreq = log2(freq /1000);

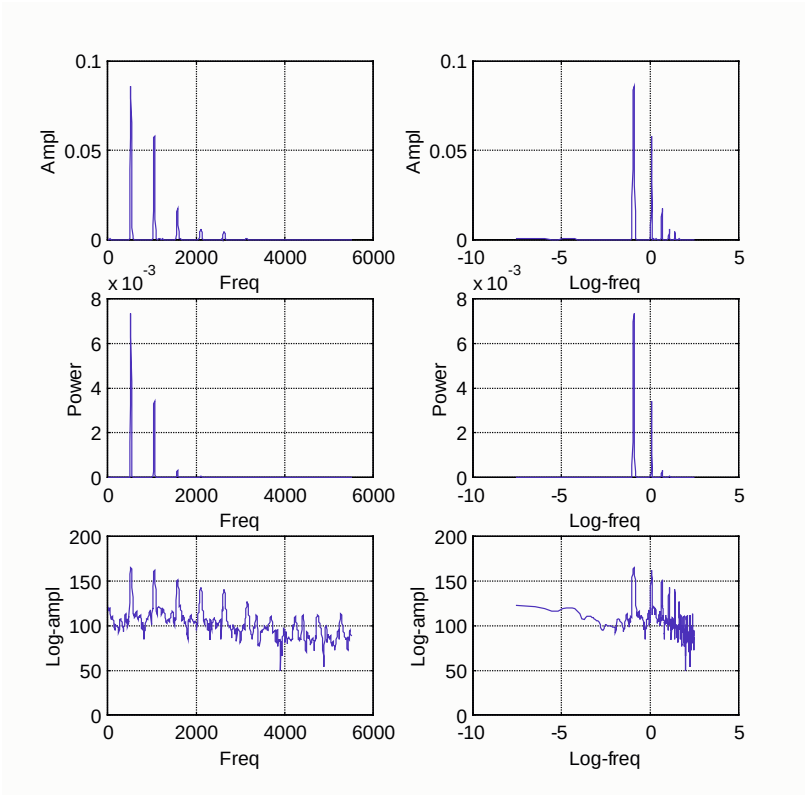


Figure 6 Spectrum in [top-left] lin-frequency/lin-amplitude, [middle-left] lin-frequency/energy, [bottom-left] lin-frequency/log-amplitude, [top-right] log-frequency/lin-amplitude, [middle-right] log-frequency/energy, [bottom-right]] log-frequency/log-amplitude

2.6 Descriptors on Spectrum / Harmonic peaks / Bark bands

Most of the spectral shape descriptors (detailed in part 6.1, namely the centroid, spread, skewness, kurtosis, slope, decrease, roll-off and variation) are also computed on the harmonic peaks and on a Bark band representation.

Some harmonic descriptors (detailed in part 7, namely the deviation, odd to even energy ratio and tristimulus) are also computed on a Bark band representation.

3 Global temporal features

3.1 Envelop characterization

3.1.1 Attack / Decay / Sustain / Release envelop modeling

As it is the case in most synthesizers, it is usual to represent the evolution along time of the energy of a sound sample using an attack, decay, sustain, release (ADSR) envelop (see **Figure 7**). However, this representation is hardly achievable for most real sounds, since a) the decay part is often not clearly present, b), the sustain part is not present if the sound is not sustained (guitar sounds), c) the release part is not present if the sound has been truncated which is the case with some "sampler" sounds.

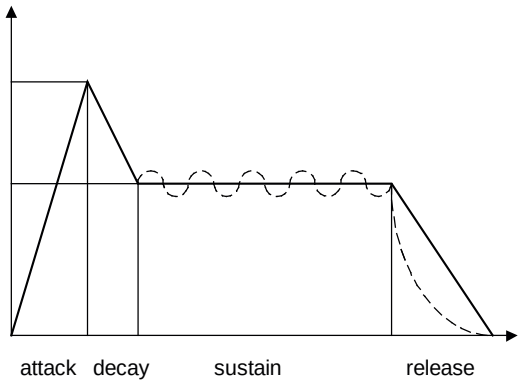


Figure 7 Envelop modeling: ADSR envelop, with a modulation on the sustain part and an exponential release

For this reason, in the following we will deal with a simpler representation: the attack/rest envelop; in this representation the decay part is not estimated and the sustain and release parts are merged. If the sound is a sustained sounds the rest will represent the sustain, if the sound is not sustained the rest will represent the release (see Figure **8**).

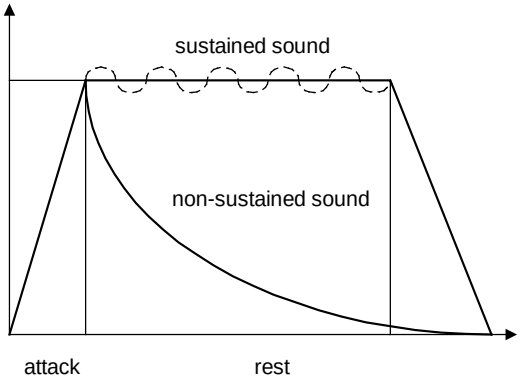


Figure 8 Envelop modeling: AR envelop, with a modulation on the sustain part and an exponential release

3.1.2 Attack part

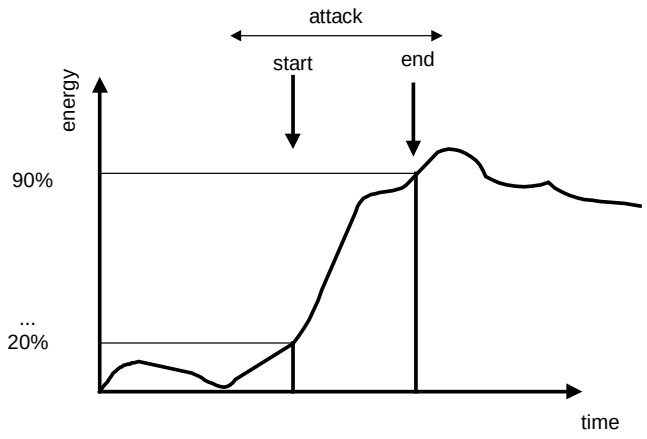
The attack of the sound is described using two parameters:

- The duration of the attack
- The average slope of the energy of the attack: the increase factor

For both parameters, we need to estimate when does the attack actually starts and end which is not an easy task considering that the attack is a fuzzy concept.

3.1.2.1 Estimation of the start and end of the attack

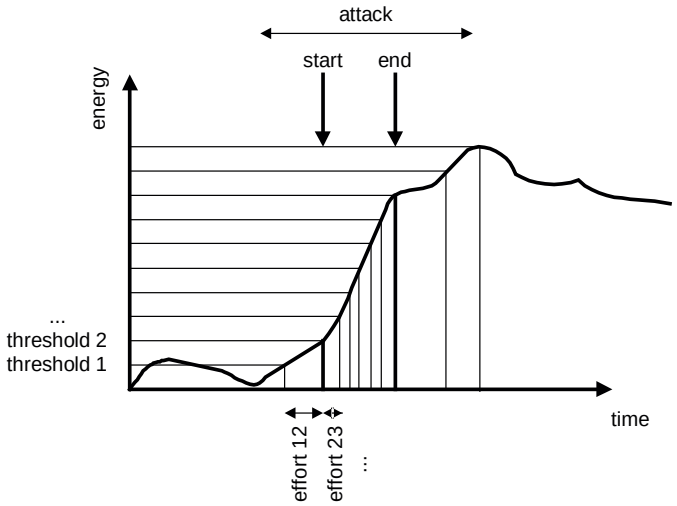
Fixed threshold method: In this method, the start and end of the attack are estimated by applying thresholds on the energy envelop of the signal. In order to take into account the possible presence of noise in the signal the "start-attack" threshold is usually set to 20% (in the following all threshold values are expressed as percentage of the maximum value of the energy of the sound along time). In order to take into account the possibility that the maximum of the envelop does not occur at the end of the attack but possibly latter in the signal (as in a trumpet sound), the "end-attack" threshold is set to 90%. However these threshold are to be set empirically for each sound set.



Adaptative threshold method (weakest effort method): In this method, the value of the "start-attack" and "end-attack" thresholds are not fixed but estimated according to the behavior of the signal during the attack. For a specific threshold value th_i, the time the energy envelop reaches for the first time this threshold is estimated, we not it t_i. For successive values of the threshold th_i, we define an effort as the time the signal goes from one threshold value to the next threshold value: $w_i=t_{i+1}-t_i$. The average effort value, w, is then computed.

We then determine the best threshold for the starting of the attack th_{st} as the first threshold for which the effort w_i goes below $M*\underline{w}$. In a similar way, we determine the best threshold for the ending of the attack th_{end}, as the first threshold for which the effort w_i goes above $M*\underline{w}$. We've used a value of $M=3$.

Finally, the exact start time t_{st} and end time t_{end} of the attack are refined around the time corresponding to th_{st} and th_{end} by taking the local minimum and local maximum respectively.



3.1.2.2 Log-Attack Time (mpeg7:LogAttackTime) DT.g_lat

Description: The log-attack time is the logarithm (decimal base) of the time duration between the time the signal starts to the time it reaches its stable part. It has been proved to be one of the most perceptually important descriptors. It can be estimated taking the logarithm of the time from the start to the end of the attack.

Formulation:

$$\text{lat} = \log_{10}(\text{stop\_attack} - \text{start\_attack})$$

3.1.2.3 Temporal increase (cuidado:TemporalIncrease) DT.g_incr

Description: The increase time is defined as the average temporal slope of the energy during the attack time.

Formulation: We compute the local slopes of the energy corresponding to each effort w_i. We compute the weighted average of the slopes. The weights are chosen in order to emphasizes slope values in the middle of the attack (weights = values of a gaussian function centered around threshold=50% and of std=0.5).

3.1.3 Sustain part

The sustain of the sound is described using two parameters:

- The decrease slope
- The modulation of the energy and the modulation of fundamental frequency

3.1.3.1 Decrease part: Temporal decrease (cuidado:TemporalDecrease) DT.g_decr

Description: The temporal decrease is a measure of the amount of decrease of the signal energy. It allows distinguishing non-sustained (percussive, pizzicato, …) sounds from sustained sounds. Its calculation is based on the following envelop temporal model starting from the maximum of the energy envelop (t_{max}):

$$S(t) = A \cdot \exp(-\alpha(t - t_{\max})) \quad t > t_{\max}$$

α is estimated by linear regression on the logarithm of the energy envelope of the signal.

3.1.3.2 Sustain part: Energy Modulation and Fundamental frequency modulation

(mpeg7:AudioPower ScalableSeriesType element name="Modulation")
(mpeg7:AudioFUndamentalFrequency ScalableSeriesType element name="Modulation")

Description: During the sustained part of a note played on natural musical instruments, tremolo and vibrato are often used for expressiveness. The energy modulation and fundamental frequency modulation computed on the energy and fundamental frequency signal during the sustained part of the sound aims at describing those. Each modulation is represented by a frequency and amplitude factor.

Formulation:
The modulation is estimated only during the sustained part of a sound using a peak detection algorithm in the amplitude spectrum of the instantaneous descriptor.

1. Locate the sustained part of the sound
2. Correct the energy envelope (fundamental frequency) by subtracting its logarithmic (linear) tendency during the sustained part
3. Compute the amplitude spectrum of the corrected envelope
4. Locate the maximum peak within the range [1 Hz, 10 Hz].

3.1.4 Example

In Figure 9 and Figure 10, we illustrate the computation of the attack parameters and sustain/decrease parameters.

On the top of Figure 9 we illustrate the energy envelop of an alto sound along the various time corresponding to the threshold 10% to 100% (vertical green lines). On the middle of Figure 9 we represents the efforts (y-axis) corresponding to the various thresholds (x-axis), the mean value of the efforts (continuous horizontal line) and the M*mean value of the effort (dotted horizontal line). On the bottom of Figure 9 we illustrate the detected start of the attack (vertical green line), end of the attack (vertical red line), start of the sustained part (vertical black line) and approximation of the increase and decrease (red lines).

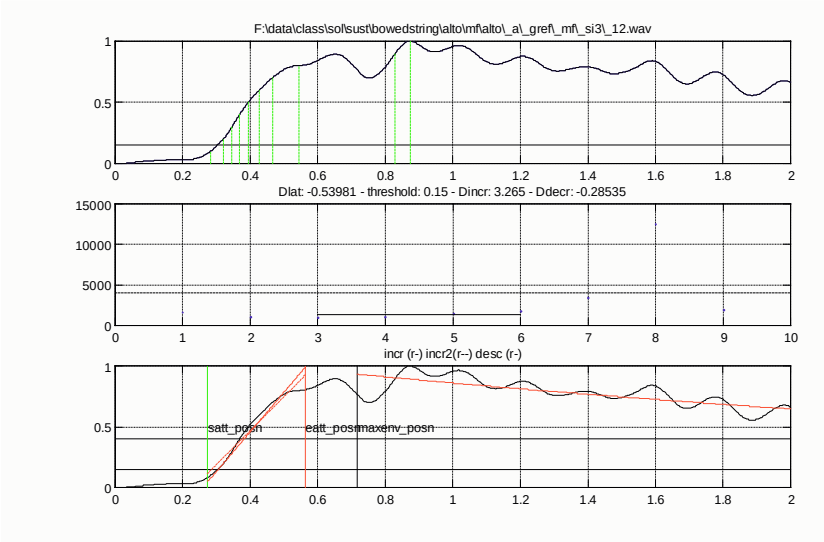


Figure 9 Log-Attack Time and Temporal Increase / Decrease estimation
[Top] Energy Envelop with percent thresholds (vertical lines)
[Middle] Efforts corresponding to the percent thresholds
[Bottom] Resulting attack, increase and decrease segments

On the top of Figure 10, we illustrate the sustain parts on the same sounds (start by green line, end by red line). On the middle of Figure 10, we represent only the sustained part, the decrease approximation (red line) and the estimated modulation (dotted red line). On the bottom of Figure 10, we represents the amplitude spectrum of the corrected temporal energy during the sustained part.

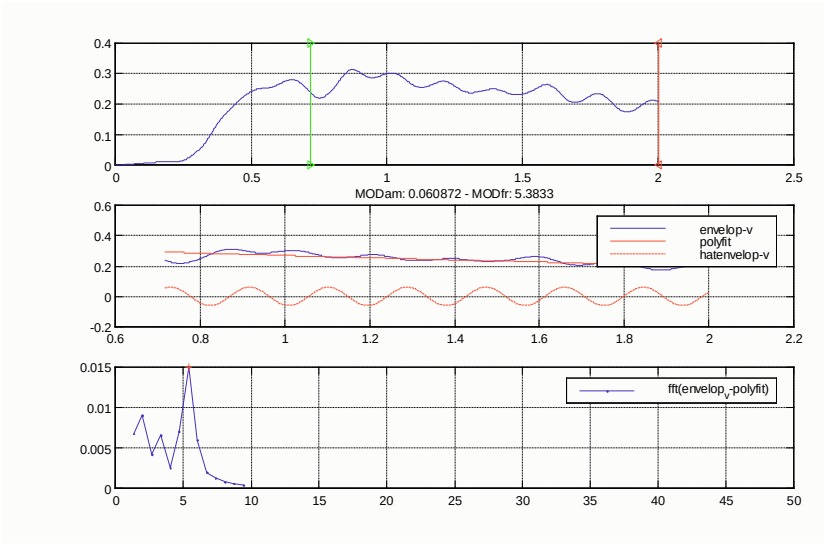


Figure 10 Energy modulation estimation
[Top] Energy Envelop
[Middle] Energy Envelop of the sustained part corrected by linear regression
[Bottom] Amplitude spectrum of the corrected energy envelop

3.2 Others

3.2.1 Temporal centroid (mpeg7:TemporalCentroid) DT.g_tc

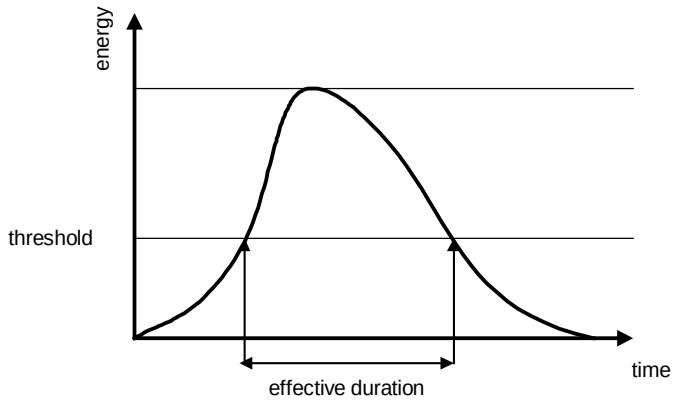
Description: The temporal centroïd is the time averaged over the energy envelop. It allows distinguishing percussive from sustained sounds. It has been proved to be one perceptually important descriptors.

Formulation:

$$tc = \frac{\sum_t e(t) \cdot t}{\sum_t e(t)}$$

3.2.2 Effective Duration (cuidado:TemporalEffectiveDuration) DT.g_ed

Description: The effective duration is a measure of the time the signal is perceptually meaningful. It allows distinguishing percussive sounds from sustained sounds but depends on the recording length. It is approximated by the time the energy envelop is above a given threshold. A threshold of 40% was used.



4 Instantaneous temporal features

4.1 Auto-correlation (cuidado:AudioZcr) DT.i_xcorr_m

Description: The cross-correlation represents the signal spectral distribution but in the time domain (the cross-correlation of a signal is the inverse Fourier Transform of the spectrum energy distribution of the signal). It has been proved to provide a good description for classification (Brown 1998). In order to obtain cross-correlation coefficients independent from the sampling rate of the signal, the signal is first down-sampled at 11025 Hz. From the cross-correlation, we only keep the first 12 coefficients.

Formulation:

$$xcorr(k) = \frac{1}{x(0)^2} \sum_{n=0}^{N-k-1} x(n) \cdot x(n+k)$$

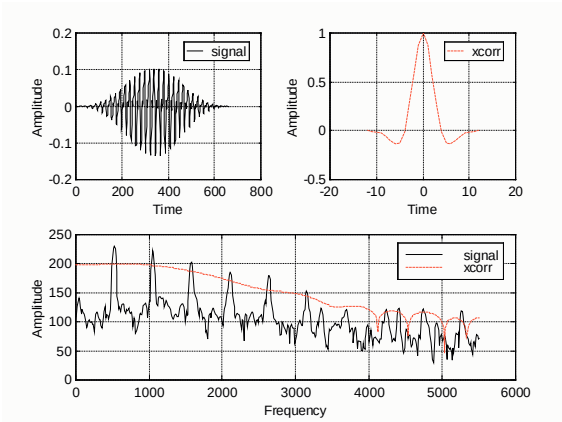


Figure 11 [top-left] signal [top-right] cross-correlation function
[bottom] signal amplitude spectrum and spectrum envelop estimated by cross-correlation (dashed line)

4.2 Zero-crossing rate (cuidado:AudioXcorr) DT.i_zcr_v

Description: The zero-crossing rate is a measure of the number of time the signal value cross the zero axe. Periodic sounds tend to have a small value of it, while noisy sounds tend to have a high value of it. It is computed at each time frame on the signal.

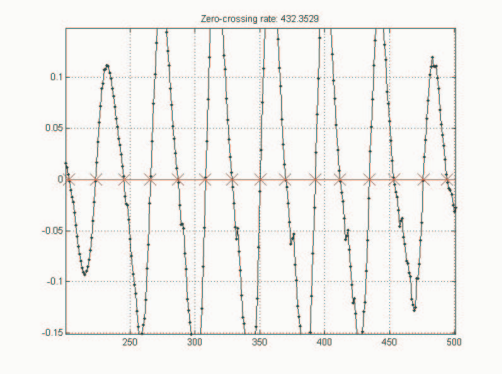


Figure 12 Zero-crossing rate (=432) during voiced speech region

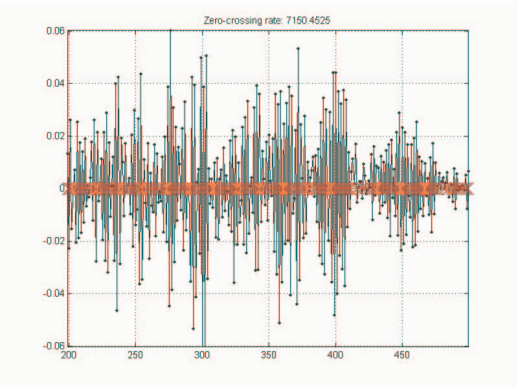


Figure 13 Zero-crossing rate (=7150) during unvoiced speech region

5 Energy features

5.1 Total Energy (mpeg7:AudioPower) DE.i_tot_v

Description: The total energy estimates the signal power at a given time. It is estimated directly form the signal frame around a given time.

5.2 Harmonic Part Energy (cuidado:AudioHarmonicPower) DE.i_harmo_v

Description: The harmonic energy estimates the power of the harmonic part of the signal at a given time. It is estimated from the estimated harmonic amplitude at a given time.

5.3 Noise Part Energy (cuidado:AudioNoisePower) DE.i_noise_v

Description: The noise energy estimates the power of the noise (non-harmonic) part of the signal at a given time. It is estimated from the signal obtained by subtracting the harmonic part from the signal.

6 Spectral features

6.1 Spectral shape description

6.1.1 Spectral centroid (mpeg7:AudioSpectrumCentroid) DS.i_sc_v

The spectral centroïd is the barycenter of the spectrum. It is computed considering the spectrum as a distribution which values are the frequencies and the probabilities to observe these are the normalized amplitude.

$$\mu = \int x \cdot p(x) \quad \delta x$$

where

- x are the observed data: $x = freq\_v(x)$
- p(x) is the probability to observe x: $p(x) = \frac{ampl\_v(x)}{\sum_x ampl\_v(x)}$

6.1.2 Spectral spread (mpeg7:AudioSpectrumSpread) DS.i_ss_v

Following the previous definition, we define the spectral spread as the spread of the spectrum around its mean value, i.e. the variance of the above defined distribution

$$\sigma^2 = \int (x-\mu)^2 \cdot p(x) \quad \delta x$$

6.1.3 Spectral skewness (cuidado:AudioSpectrumSkewness) DS.i_skew_v

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

$$m_3 = \int (x-\mu)^3 \cdot p(x) \quad \delta x$$

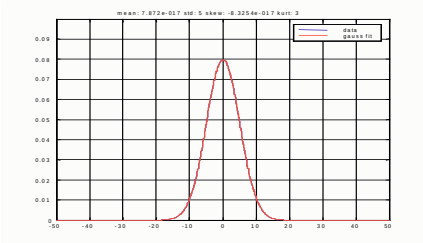
The skewness is then: $\gamma_1 = \frac{m_3}{\sigma^3}$

The skewness SK describes the degree of asymmetry of the distribution.

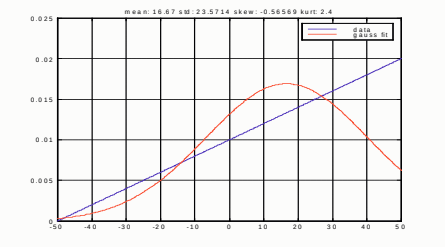
- SK = 0 indicates a symmetric distribution,
- SK < 0 indicates more energy on the right,
- SK >0 indicates more energy on the left.

The following figures represent various values of the spectral skewness depending on the spectral shape (x=frequency, y=amplitude):

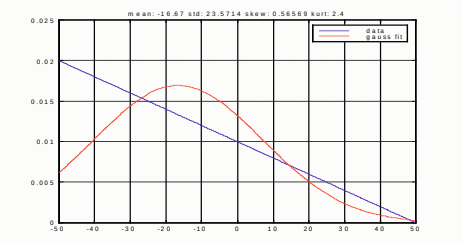
- data (blue line),
- gaussian pdf fitting to the data (red line)



skewness=0



skewness<0



skewness>0

6.1.4 Spectral kurtosis (cuidado:AudioSpectrumKurtosis) DS.i_kurto_v

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

$$m_4 = \int (x-\mu)^4 \cdot p(x) \quad \delta x$$

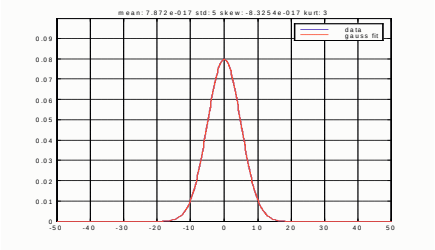
The kurtosis is then: $\gamma_2 = \frac{m_4}{\sigma^4}$

The kurtosis K indicates the peakedness/flatness of the distribution.

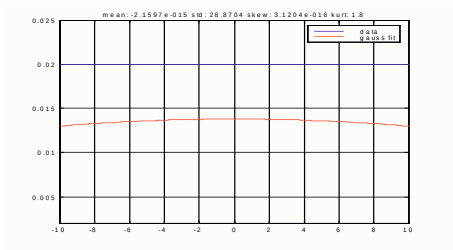
- K = 3 for a normal distribution,
- K < 3 for a flatter distribution,
- K > 3 for a peaker distribution.

The following figures represent various values of the spectral kurtosis depending on spectral shape (x=frequency, y=amplitude):

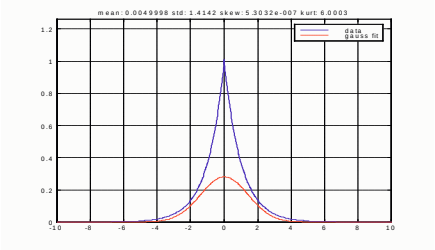
- data (blue line),
- gaussian pdf fitting to the data (red line)



Kurtosis=3



Kurtosis=1.8



Kurtosis=6

6.1.5 Spectral slope (cuidado:AudioSpectrumSlope) DS.i_slope_v

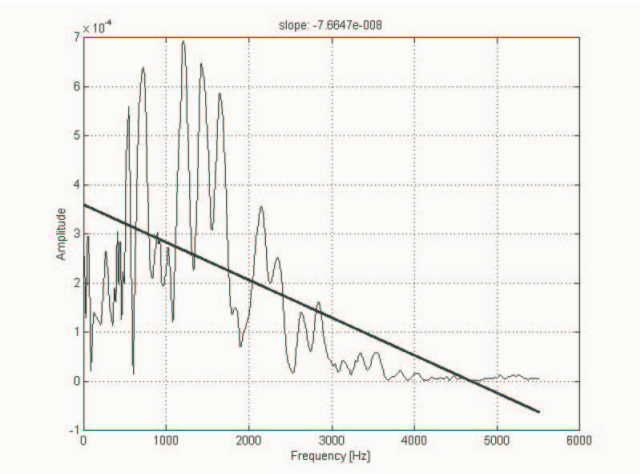
Description: The spectral slope represents the amount of decreasing of the spectral amplitude. It is computed by linear regression of the spectral amplitude.

Formulation:

$$\hat{a}(f) = slope \cdot f + const$$

where

$$slope = \frac{1}{\sum_k a(k)} \frac{N\sum_k f(k)*a(k) - \sum_k f(k) * \sum_k a(k)}{N\sum_k f^2(k) - \left(\sum_k f(k)\right)^2}$$



6.1.6 Spectral decrease (cuidado:AudioSpectrumDecrease) DS.i_decr_v

Description: The spectral decrease also represents the amount of decreasing of the spectral amplitude. This formulation comes from perceptual studies, it is supposed to be more correlated to human perception.

Formulation:

$$decrease = \frac{1}{\sum_{k=2:K} a} \sum_{k=2:K} \frac{a(k) - a(1)}{k-1}$$

6.1.7 Spectral roll-off (cuidado:AudioSpectrumRollOff) DS.i_rolloff_v

Description: The spectral roll-off point is the frequency so that 95% of the signal energy is contained below this frequency. It is correlated somehow to the harmonic/noise cutting frequency.

Formulation:

$$\sum_{0}^{fc} a^2(f) = 0.95 \sum_{0}^{sr/2} a^2(f)$$

where f_c is the spectral roll-off frequency, *sr/2* is the Nyquist frequency.

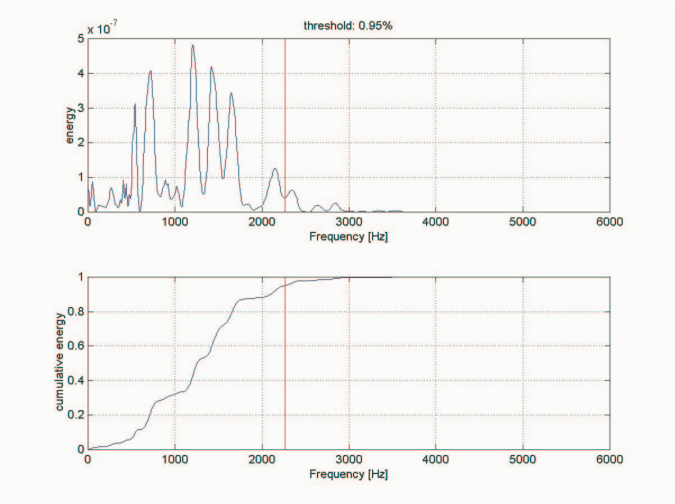


Figure 14 [Top] Energy spectrum along frequency with 95% spectral roll-off frequency (vertical red line) [bottom] cumulative energy along frequency with 95% spectral roll-off frequency (vertical red line)

6.2 Temporal variation of spectrum

6.2.1 Temporal variation of spectrum: spectral variation (cuidado:AudioSpectrumVariation) DS.i_var_v

Description: The spectral variation (also called sometimes in literature spectral flux) represents the amount of variation of the spectrum along time. It is computed from the normalized cross-correlation between two successive amplitude spectra *a(t-1)* and *a(t)*.

Formulation:

$$\text{variation} = 1 - \frac{\sum_k a(t-1,k) \cdot a(t,k)}{\sqrt{\sum_k a(t-1,k)^2} \sqrt{\sum_k a(t,k)^2}}$$

It is close to 0 if the successive spectrum are similar, to 1 of the successive spectrum are highly dissimilar.

6.3 Global spectral shape description

6.3.1 Mel Frequency Cepstral Coefficients (MFCC) (cuidado:AudioMFCC) DP.i_MFCC_m

Description: The MFCC represent represents the shape of the spectrum with very few coefficients. The cepstrum, is the Fourier Transform (or Discrete Cosine Transform DCT) of the logarithm of the spectrum. The Mel-cepstrum is the cepstrum computed on the Mel-bands instead of the Fourier spectrum. The use of mel scale allows better to take better into account the mid-frequencies part of the signal. The MFCC are the coefficients of the Mel cepstrum. The first coefficient being proportional to the energy is not stored, the next 12 coefficients are stored for each frame.

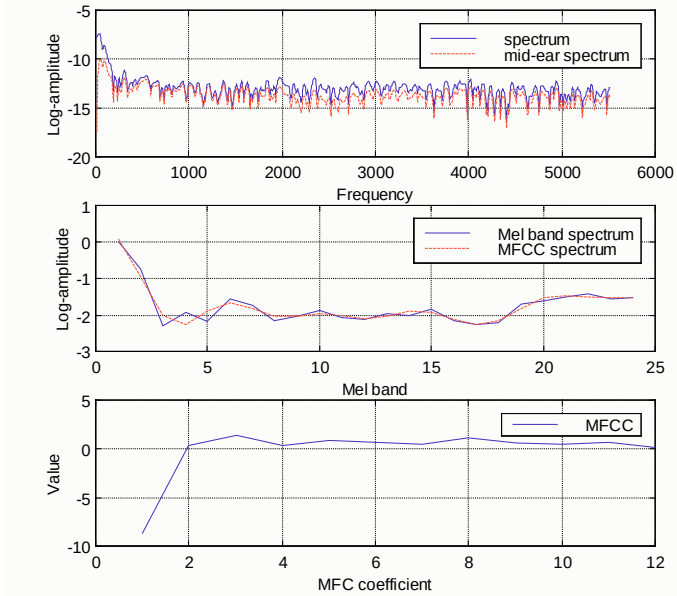
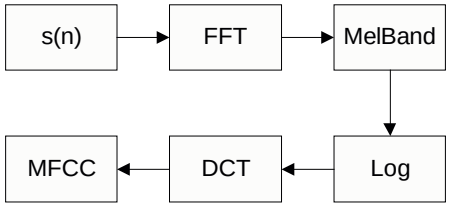


Figure 15 [Top] signal spectrum and mid-ear filtered spectrum (dashed line) [middle] Mel band spectrum and MFCC spectrum (dotted line) [bottom] MFCC coefficients

Formulation:



Delta-MFCC, Delta-Delta-MFCC:
The Delta-MFCC and Delta-Delta MFCC are the first and second order derivative of the MFCC along time

$$DMFCC = \frac{\partial}{\partial t} MFCC(t)$$

$$DDMFCC = \frac{\partial^2}{\partial^2 t} MFCC(t)$$

7 Harmonic features

7.1.1 Fundamental frequency (mpeg7:AudioFundamentalFrequency) DH.i_f0_v

For an harmonic signal, the fundamental frequency is the frequency so that its integer multiple best explain the content of the signal spectrum. The fundamental frequency has been computed using the maximum likelihood algorithm. (Doval 1994) (Doval and Rodet 1993).

7.1.2 Noisiness (mpeg7:AudioHarmonicity) DH.i_noisiness_v

The noisiness is the ratio between the energy of the noise (non-harmonic part) to the total energy. It is close to 1 for purely noise signal and 0 for purely harmonic signal.

$$\text{noisiness} = \frac{ener\_noise}{ener\_tot}$$

7.1.3 Inharmonicity (cuidado:AudioInharmonicity) DH.i_inharmo_v

The inharmonicity represents the divergence of the signal spectral components from a purely harmonic signal. It is computed as an energy weighted divergence of the spectral components from the multiple of the fundamental frequency.

$$inharmo = \frac{2}{f0} \frac{\sum_h |f(h) - h * f0| * a^2(h)}{\sum_h a^2(h)}$$

This coefficient ranges from 0 (purely harmonic signal) to 1 (inharmonic signal). The range is [0,1] since a(h)-h*f0 is at maximum equal to f0.

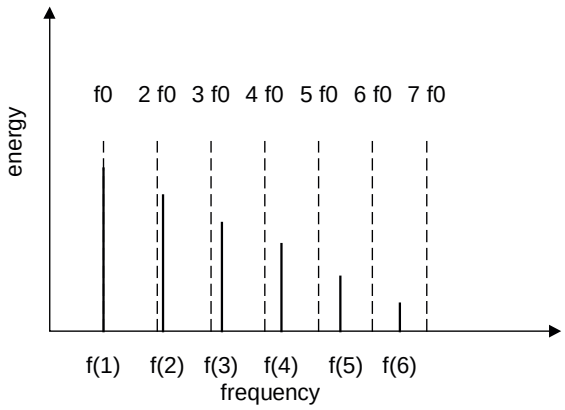


Figure 16 Inharmonicity coefficient computation: harmonic multiple (dotted lines), observed spectral peaks (contiuous lines)

7.1.4 Harmonic Spectral Deviation (mpeg7:HarmonicSpectralDeviation) DH.i_devs_v

The harmonic spectral deviation is the deviation of the amplitude harmonic peaks from a global spectral envelope.

$$HDEV = \frac{1}{H} \sum_h \left(a(h) - SE(h) \right)$$

where H is the total number of considered harmonics, $a(h)$ the amplitude of the h^{th} harmonic, $SE(h)$ the amplitude of the spectral envelope evaluated at the frequency $f(h)$.

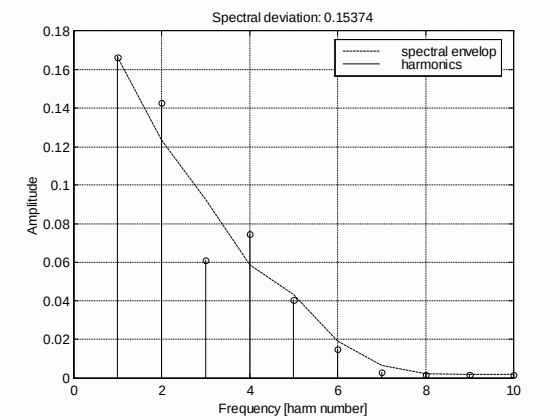


Figure 17 Harmonic of the signal and spectral envelope for the estimation of the spectral deviation for a trumpet sound

7.1.5 Odd to Even Harmonic Energy Ratio (cuidado:HarmonicSpectralOERatio): DH.i_oeratio_v

Description: The odd to even ratio is a measure allowing to distinguish odd harmonic energy predominant sounds (such as clarinet sounds), from equally important harmonic energy sounds (such as the trumpet). It is computed from the ratio between the odd harmonic energy to the even harmonic energy.

Formulation:

$$OER = \frac{\sum_{h=1:2:H} a^2(h)}{\sum_{h=2:2:H} a^2(h)}$$

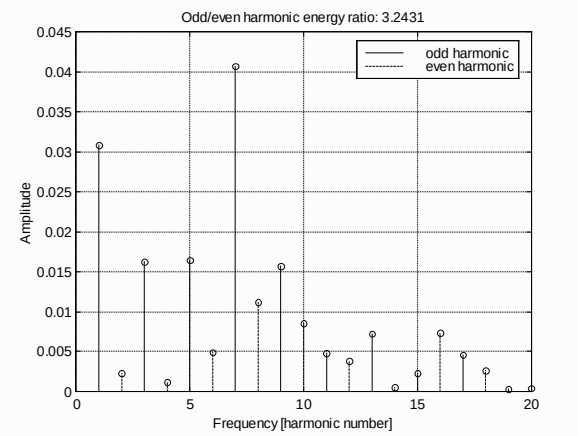


Figure 18 Odd and Even harmonics for a clarinet sounds

7.1.6 Tristimulus (cuidado:HarmonicSpectralTristimulus): DH.i_tri*_v

Description:
The tristimulus values have been introduced in [Pollard *et al.* 1982] as a timbre equivalent to the color attributes in the vision. The tristimulus are three different types of energy ratio allowing a fine description of the first harmonic of the spectrum, which are perceptually more salient.

Formulation:

$$T1 = \frac{a(1)}{\sum_h a(h)}$$

$$T2 = \frac{a(2)+a(3)+a(4)}{\sum_h a(h)}$$

$$T3 = \frac{\sum_{h=5:H} a(h)}{\sum_h a(h)}$$

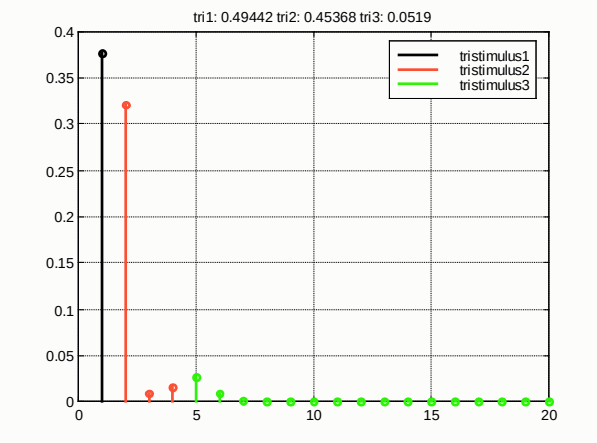


Figure 19 First harmonic (black) second, third and fourth harmonic (red), fifth to the end (black) for the estimation of the tristimulus

8 Perceptual features

8.1 Features

8.1.1 Total Loudness and specific loudness (cuidado:AudioLoudness): DP.i_loud_v

Description : (Moore, Glasberg et al. 1997)
The **specific loudness** is the loudness associated to each Bark band. We note *N'(z)* the loudness in the z^{th} Bark band. The precise expression of the loudness can be found in (Moore, Glasberg et al. 1997). This expression was approximated by neglecting terms of the expression acting only in specific cases (very weak signals) and by expressing it in relative scale (X. Rodet 2001). The specific Loudness can be expressed in a simple form:

$$N'(z) = E(z)^{0.23}$$

The **total loudness** is the sum of individual loudness (Zwicker 1990)

$$N = \sum_{1}^{nb\_band} N'(z)$$

where *N'(z)* is the specific loudness

8.1.2 Relative Specific Loudness (cuidado:AudioRelativeSpecificLoudness): DP.i_specloudnorm_m

We define a relative specific loudness as the specific loudness normalized by the total loudness:

$$Nrel(z) = N'(z) / N$$

The normalized specific loudness is then independent from the total loudness and represents a sort of equalization curve of the sounds.

8.1.3 Sharpness (cuidado:AudioSharpness) DP.i_sharp_v

The sharpness is the perceptual equivalent to the spectral centroïd but computed using the specific loudness of the Bark bands (Zwicker 1977).

$$A = 0.11 \cdot \frac{\sum_{z=1}^{nband} z \cdot g(z) \cdot N'(z)}{N}$$

where *z* is the index of the band and g*(z)* is a function defined by:

$$g(z) = 1 \quad if \quad z < 15$$
$$g(z) = 0.066 \cdot \exp(0.171z) \quad if \quad z \geq 15$$

8.1.4 Spread (cuidado:AudioSpread) DP.i_spread_v

The spread measures the distance from the largest specific loudness value to the total loudness.

$$ET = \left(\frac{N - \max_z N'(z)}{N} \right)^2$$

9 Various features

9.1 Spectral Flatness/Crest measure (mpeg7:AudioSpectrumFlatness) DP.sfm_m

The **Spectral Flatness** is a measure of the noisiness (flat, decorrelation)/ sinusoidality of a spectrum (or a part of it). It is computed by the ratio of the geometric mean to the arithmetic mean of the energy spectrum value.

$$SFM(num\_band) = \frac{\left(\prod_{k \in num\_band} a(k)\right)^{1/K}}{\frac{1}{K}\sum_{k \in num\_band} a(k)}$$

where a(k) is the amplitude in frequency band number k.

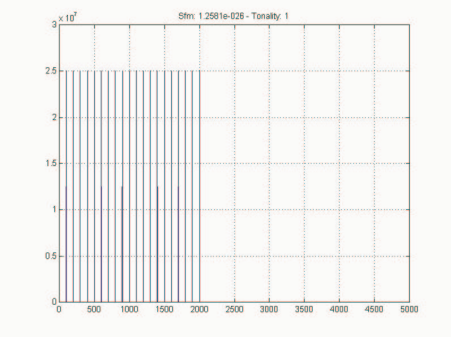
For tonal signals, SFM is close to 0, for noisy signal it is close to 1.

It is compute for several frequency bands. We used the following four frequency bands:

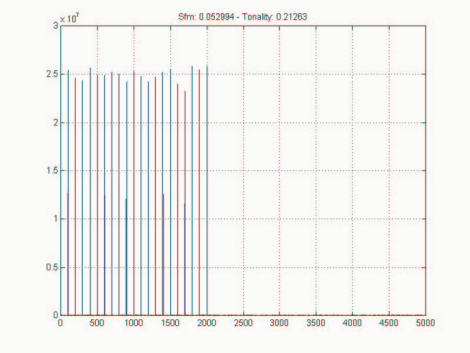
- 250 to 500 Hz
- 500 to 1000 Hz
- 1000 to 2000 Hz
- 2000 to 4000 Hz

Another descriptors related to the flatness of the spectrum is the **Spectral Crest** Factor. It is computed by the ratio of the maximum value within the band to the arithmetic mean of the energy spectrum value.

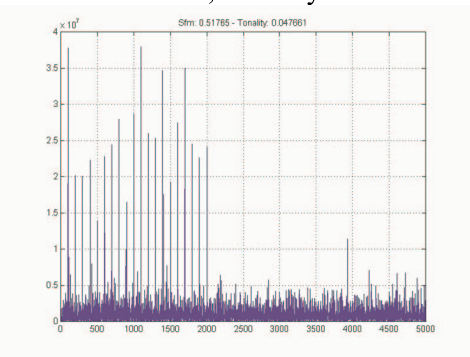
$$SCM(num\_band) = \frac{\max(a(k \in num\_band))}{\frac{1}{K}\sum_{k \in num\_band} a(k)}$$



SFM=0, Tonality=1



SFM=0.05, Tonality=0.21



SFM=0.51, Tonality=0.047

Converting SFM to Tonality measure

The SFM can be converted to the so-called "tonality coefficient" by

$$SFM_{db} = 10 \cdot \log_{10}(SFM)$$

$$Tonality = \min\left(\frac{SFM_{db}}{-60}, 1\right)$$

For tonal signals, Tonality is close to 1, for noisy signal it is close to 0.

Practical computation:

The SFM_{db} can be computed practically using $SFM_{db} = 10 * \frac{1}{N}\sum_{k}\left(\log_{10} a(k) - \log_{10} \mu\right)$

10 Temporal modeling

Each descriptors (except the global descriptors) has been extracted using a frame-by-frame analysis. These instantaneous descriptors can be used for in a real-time context (recognition or description). These descriptors can also be used in order to create a Hidden Markov Model representing the descriptors behavior along time. Another possibility is to model the instantaneous descriptors along time by their statistics (mean values, variance values) during a period of time (a signal segment, an entire sound events, a musical notes).

For a specific time segment, only the part of the signal above a noise threshold is taken into account. The estimation of the mean, variance and derivative values are weighted by the instantaneous loudness of the values of the signal.

10.1.1 Mean

Mean value of the features x weighted by the AudioLoudness w

$$\bar{x} = \frac{\sum_{frame} w(frame)x(frame)}{\sum_{frame} w(frame)}$$

10.1.2 Variance

Variance value of the features x weighted by the AudioLoudness w

$$z = \frac{\sum_{frame} w(frame)\left(x(frame) - \bar{x}\right)^2}{\sum_{frame} w(frame)}$$

10.1.3 Deviation

Derivative value of the features x weighted by the AudioLoudness w

$$deriv = \frac{\sum_{frame} w(frame, frame+1)\left(x(frame+1) - x(frame)\right)}{\sum_{frame} w(frame, frame+1)}$$

10.1.4 Temporal modeling an mpeg-7 audio scalable series

The temporal models considered here extend the existing data reduction techniques of the `mpeg7::scalableseries`. The temporal models are computed using the `weight` element of the `scalableseries` set to the `AudioLoudnessType`. The resulting temporal models are stored in the `mpeg7:scalableseries` with `numOfElements=1` using the following `element name`:

Element Name	Mpeg-7
Mean	yes
Variance	Yes
Derivative	Extension
Modulation	extension

The extraction and storage process of the generic temporal models is illustrated in Figure 20.

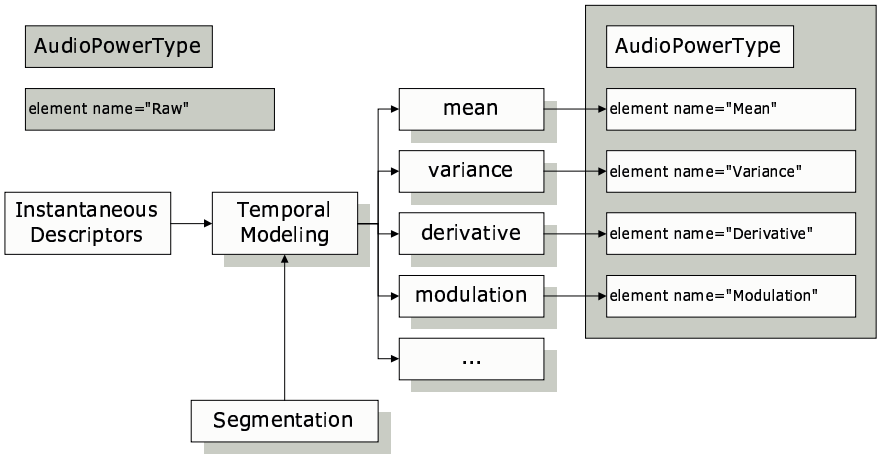


Figure 20 Temporal modeling of instantaneous descriptors: example for the AudioPowerType descriptors

11 List of all descriptors

LLD List	frame based	number of features	acronym	xml tag
Temporal Features				
Global Temporal Features				
Log Attack Time	n	1	DTg_lat	mpeg7:LogAttackTime
Temporal Increase	n	1	DTg_incr	cuidado:TemporalIncrease
Temporal Decrease	n	1	DTg_decr	cuidado:TemporalDecrease
Temporal Centroid	n	1	DTg_tc	mpeg7:TemporalCentroid
Effective Duration	n	1	DTg_ed	cuidado::TemporalEffectiveDuration
Instantaneous Temporal Features				
Signal Auto-correlation function	y	12	DTi_xcorr_m	cuidado:AudioXcorr
Zero-corssing rate	y	1	DTi_zcr	cuidado:AudioZcr
Energy Features				
Total energy	y	1	DEi_tot_v	mpeg7:AudioPower
Total energy Modulation (frequency, amplitude)	n	2	DTg_mod_fr, DTg_mod_am	**ScalableSeriesType element name="Modulation"**
Total harmonic energy	y	1	DEi_harmo_v	cuidado:AudioHarmonicPower
Total noise energy	y	1	DEi_noise_v	cuidado:AudioNoisePower
Spectral Features				
Spectral Shape				
Spectral centroid	y	*6*	*DSi_sc_m*	mpeg7:AudioSpectrumCentroid (mpeg7:SpectralCentroid)
Spectral spread	y	*6*	*DSi_ss_m*	mpeg7:AudioSpectrumSpread
Spectral skewness	y	*6*	*Dsi_skew_m*	cuidado:AudioSpectrumSkewness
Spectral kurtosis	y	*6*	*Dsi_kurto_v*	cuidado:AudioSpectrumKurtosis
Spectral slope	y	*6*	*Dsi_slope_v*	cuidado:AudioSpectrumSlope
Spectral decrease	y	*1*	*Dsi_decs_c*	cuidado:AudioSpectrumDecrease
Spectral rolloff	y	*1*	*Dsi_rolloff_v*	cuidado:AudioSpectrumRollOff
Spectral variation	y	*3*	*Dsi_variation_v*	cuidado:AudioSpectrumVariation
Global spectral shape description				
MFCC	y	12	DPi_mfcc_m	cuidado:AudioMFCC
Delta MFCC	y (post)	12	DPi_Dmfcc_m	
Delta Delta MFCC	y (post)	12	DPi_DDmfcc_m	
Harmonic Features				
Fundamental frequency	y	1	DHi_f0_v	mpeg7:AudioFundamentalFrequency
Fundamental fr. Modulation (frequency, amplitude)	n	2	F0 Mod AM, FR	**ScalableSeriesType element name="Modulation"**
Noisiness	y	1	DHi_noisiness_v	mpeg7:AudioHarmonicity
Inharmonicity	y	1	DHi_inharmo_v	cuidado:AudioInharmonicity
Harmonic Spectral Deviation	*y*	*3*	*DHi_devs_v*	*mpeg7:HarmonicSpectralDeviation*
Odd to Even Harmonic Ratio	*y*	*3*	*Dhi_oeratio_v*	*cuidado:HarmonicSpectralOERatio*
Harmonic Tristimulus	*y*	*9*	*Dhi_tri_v*	*cuidado:HarmonicSpectralTristimulus*
Harmonic Spectral Shape				
HarmonicSpectral centroid	y	6	DHi_sc_m	mpeg7:HarmonicSpectralCentroid
HarmonicSpectral spread	y	6	DHi_ss_m	mpeg7:HarmonicSpectralSpread
HarmonicSpectral skewness	y	6	DHi_skew_m	cuidado:HarmonicSpectralSkewness
HarmonicSpectral kurtosis	y	6	DHi_kurto_v	cuidado:HarmonicSpectralKurtosis
HarmonicSpectral slope	y	6	DHi_slope_v	cuidado:HarmonicSpectralSlope
HarmonicSpectral decrease	y	1	DHi_decs_c	cuidado:HarmonicSpectralDecrease
HarmonicSpectral rolloff	y	1	DHi_rolloff_v	cuidado:HarmonicSpectralRollOff
HarmonicSpectral variation	y	3	DHi_variation_v	mpeg7:HarmonicSpectralVariation
Perceptual Features				
Loudness	y	1	DPi_loud_v	AudioLoudness
RelaitveSpecific Loudness	y	24	DPi_specloud_m	cuidado:AudioRelativeSpecificLoudness
Sharpness	y	1	DPi_sharp_v	cuidado:AudioSharpness
Spread	y	1	DPi_spread_v	cuidado:AudioSpread
Perceptual Spectral Envelope Shape				
Perceptual Spectral centroid	y	6	DPi_sc_m	cuidado:AudioFilterbankCentroid
Perceptual Spectral spread	y	6	DPi_ss_m	cuidado:AudioFilterbankSpread
Perceptual Spectral skewness	y	6	DPi_skew_m	cuidado:AudioFilterbandSkewness
Perceptual Spectral kurtosis	y	6	DPi_kurto_v	cuidado:AudioFilterbankKurtosis
Perceptual Spectral Slope	y	6	DPi_slope_v	cuidado:AudioFilterbankSlope
Perceptual Spectral Decrease	y	1	DPi_decs_c	cuidado:AudioFilterbankDecrease
Perceptual Spectral Rolloff	y	1	DPi_rolloff_v	cuidado:AudioFilterbankRolloff
Perceptual Spectral Variation	y	3	DPi_variation_v	cuidado:AudioFilterbankVariation
Odd to Even Band Ratio	*y*	*3*	*DP_ioeratio_v*	*cuidado:AudioFilterbankOERatio*
Band Spectral Deviation	*y*	*3*	*DPi_devs_v*	*cuidado:AudioFilterbankDeviation*
Band Tristimulus	*y*	*9*	*DPi_tri_v*	*cuidado:AudioFilterbankTristimulus*
Various features				
Spectral flatness	y	4	DPi_sfm_m	mpeg7:AudioSpectrumFlatness
Spectral crest	y	4	DPi_scm_m	cuidado:AudioSpectrumCrest
Total Number of Features		166		

12 Acknowledgement

Many thanks to Patrice Tisserand, Nicolas Misdariis, Patrick Susini, Daniel Preztnizer, Jeremy Marozeau, Olivier Houix, Stephen McAdams.
Part of this work was conducted in the context of the European projects CUIDAD and CUIDADO.

13 References

Brown, J. (1998). Musical Instrument identification using autocorrelation coefficients. Proc. Intern. Symposium on Musical Acoustics.

Brown, J. (1999). "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features." JASA **105**(3): 1933-1941.

Depalle, P., G. Garcia, et al. (1993). Tracking of Partials for Additive Sound Synthesis using Hidden Markov Models. ICASSP, Minneapolis, USA.

Doval, B. (1994). Estimation de la fréquence fondamtenale des signaux sonores. Paris, Université Paris VI.

Doval, B. and X. Rodet (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. ICASSP, Minneapolis.

Foote, J. (1997). Content-based Retrieval of Music and Audio. Multimedia Storage and Archiving Systems II, Proceedings of SPIE.

Jensen, K. (2001). The Timbre model. Mosart, Barcelona, Spain.

Krimphoff, J., S. McAdams, et al. (1994). "Caractérisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique." Journal de physique **4**: 625-628.

Martin, K. and Y. Kim (1998). 2pMU9. Instrument identification: a pattern-recognition approach. 136th Meet. Ac. Soc. of America.

Misdariis, N., B. Smith, et al. (1998). Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres. 135th Meet. Ac. Soc. of America / 16th Int. Cong. on Acoustics, Seattle, Washington, USA.

Moore, Glasberg, et al. (1997). "A Model for the Prediction of Thresholds Loudness and Partial Loudness." J. Audio Eng. Soc. **45**: 224-240.

MPEG-7 (2002). Information Technology - Multimedia Content Description Interface - Part 4: Audio. ISO/IEC JTC 1/SC 29. ISO/IEC FDIS 15938-4:2002.

Peeters, G. (2003). Automatic Classification of Large Musical Instrument Databases using Hierarchical Classifiers with Inertia Ratio Maximization. AES 115th Convention, New York, USA.

Peeters, G., S. McAdams, et al. (2000). Instrument sound description in the context of MPEG-7. ICMC, Berlin, Germany.

Peeters, G. and X. Rodet (2002). Automatically selecting signal descriptors for Sound Classification. ICMC, Goteborg, Sweden.

Peeters, G. and X. Rodet (2003). Hierachical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instrument Database. DAFX, London, UK.

Rabiner, L. and B. Juang (1993). Fundamentals of speech recognition. New-York, Prentice-Hall.

Scheirer, E. and M. Slaney (1997). Construction and evaluation of a robust multifeature speech/music discriminator. ICASSP, Munich, Germany.

Serra, X. and J. Bonada (1998). Sound transformations based on SMS High Level Attributes. DAFX, Barcelona (Spain).

Wold, E., T. Blum, et al. (1999). Classification, search and retrieval of audio. CRC Handbook of Multimedia Computing. B. Furth. Boca Raton, FLA, CRC Press: 207-226.

X. Rodet, P. T. (2001). Projet Ecrins: calcul des descripteurs de bas-niveaux. Paris, Ircam.

Zwicker, E. (1977). "Procedure for calculating loudness of temporally variable sounds." JASA.

Zwicker, E. (1990). Psychoacoustics. Berlin, Springer-Verlag.

Zwicker, E. and E. Terhardt (1980). "Analytical expression for critical-band rate and critical bandwidth as a function of frequency." JASA **68**: 1523-1525.