Audio Features

Time domain physical audio features can be classified into the following categories: zero crossing-based features, amplitude-based features, power-based features and rhythm-based features.

The following paragraphs describe the most commonly used time domain features belonging tothese categories.

1.1.1. Zero-Crossing Rate-Based Physical Features

This kind of physical features are based on the analysis of the sing-rate change of the analyzed audio input, which is a simple yet effective parameterization used in several machine hearing applications.

- **Zero-crossing rate (ZCR):** it is defined as the number of times the audio signal waveform crosses the zero amplitude level during a one second interval, which provides a rough estimatorof the dominant frequency component of the signal (Kedem [39]). Features based on this criterion have been applied to speech/music discrimination, music classification (Li *et al*. [40],Bergstra *et al*. [41], Morchen *et al*. [42], Tzanetakis and Cook [28], Wang *et al*. [9])), singing voice detection in music and environmental sound recognition (see the works by Mitrovic´ *et al*. [17] andPeltonen *et al*. [18]), musical instrument classification (Benetos *et al*. [10]), voice activity detection in noisy conditions (Ghaemmaghami *et al*. [43]) or for audio-based surveillance systems (as inRabaoui *et al*. [24]).
- **Linear prediction zero-crossing ratio (LP-ZCR):** this feature is defined as the ratio between the ZCR of the original audio and the ZCR of the prediction error obtained from a linear predictionfilter (see El-Maleh *et al*. [44]). Its use is intended for discriminating between signals that showdifferent degree of correlation (e.g., between voiced and unvoiced speech).

1.1.2. Amplitude-Based Features

Amplitude-based features are based on a very simple analysis of the temporal envelope of the signal. The following paragraphs describe the most commonly used amplitude-based temporal features, including the one from the Moving Picture Experts Group (MPEG) [45], (previously reviewed by Mitrovic´ *et al*. [17]), and a feature extraction approach typically used to characterize voice pathologies which has recently found application in music analysis.

- **Amplitude descriptor (AD):** it allows for distinguishing sounds with different signal envelopes, being applied, for instance, for the discrimination of animal sounds (Mitrovic´ *et al*. [46]). It is based on collecting the energy, duration, and variation of duration of signal segments based ontheir high and low amplitude by means of an adaptive threshold (a level-crossing computation).
- **MPEG-7 audio waveform (AW):** this feature is computed from a downsampled waveform envelope, and it is defined as the maximum and minimum values of a function of a non-overlapping analysis time window [45]. AW has been used as a feature in environmental sound recognition, like in the works of Muhammad and Alghathbar [47], or by Valero and Alías [48].
- **Shimmer:** it computes the cycle-to-cycle variations of the waveform amplitude. This feature has been generally applied to study pathological voices (Klingholz [49], Kreiman and Gerratt [50], Farrús *et al*. [51]). However, it has also been applied to discriminate vocal and non-vocal regions from audio in songs (as in Murthy and Koolagudi [52]), characterize growl and screaming singing styles (Kato and Ito [53]), prototype, classify and create musical sounds (Jenssen [54]) or to improve speaker recognition and verification (Farrús *et al*. [51]) to name a few.

1.1.3. Power-Based Features

The following paragraphs describe the most relevant and classic temporal audio features basedon signal power.

- **Short-time energy:** using a frame-based procedure, short-time energy (STE) can be defined as the average energy per signal frame (which is in fact the MPEG-7 audio power descriptor [45]). Nevertheless, there exist also other STE definitions in the literature that compute power in the spectral domain (e.g., see Chu *et al*. [55]). STE can be used to detect the transition from unvoiced to voices speech and *vice versa* (Zhang and Kuo [56]). This feature has also been used in applications like musical onset detection (Smith *et al*. [57]), speech recognition (Liang andFan [58]), environmental sound recognition (Peltonen *et al*. [18], Muhammad and Alghathbar [47], Valero and Alías [48]) and audio-based surveillance systems (Rabaoui *et al*. [24]).
- **Volume:** according to the work by Liu *et al*. [59], volume is defined as the Root-Mean Square (RMS) of the waveform magnitude within a frame. It has been used for speech segmentation applications, e.g., see Jiang *et al*. [60].
- **MPEG-7 temporal centroid:** it represents the time instant containing the signal largest average energy, and it is computed as the temporal mean over the signal envelope (and measuredin seconds) [45]. The temporal centroid has been used as an audio feature in the field of environmental sound recognition, like in the works by Muhammad and Alghathbar [47], and Valero and Alías [48]).
- **MPEG-7 log attack time:** it characterizes the attack of a given sound (e.g., for musical sounds, instruments can generate either smooth or sudden transitions) and it is computedas the logarithm of the elapsed time from the beginning of a sound signal to its first local maximum [45]. Besides being applied to musical onset detection (Smith *et al*. [57]), log attack time (LAT) has been used for environmental sound recognition (see Muhammad and Alghathbar [47], and Valero and Alías [48]).

1.1.4. Rhythm-Based Physical Features

Rhythm represents an relevant aspect of music and speech, but it can also be significant in environmental and human activity related sounds (e.g., the sound of a train, finger tapping, *etc.*), sinceit characterizes structural organization of sonic events (changes in energy, pitch, timbre, *etc.*) along the time axis. Since the review by Mitrovic´ *et al*. [17], there have been little significant contributions to the derivation of rhythm-based features. Thus, the following paragraphs describe the most relevant and classic rhythm-based features found in the literature.

- **Pulse metric:** this is a measure that uses long-time band-passed autocorrelation to determinehow rhythmic a sound is in a 5-second window (as defined by Scheirer and Slaney [61]). Its computation is based on finding the peaks of the output envelopes in six frequency bands and its further comparison, giving a high value when all subbands present a regular pattern. This feature has been used for speech/music discrimination.
- **Pulse clarity:** it is a high-level musical dimension that conveys how easily in a given musical piece, or a particular moment during that piece, listeners can perceive the underlying rhythmicor metrical pulsation (as defined in the work by Lartillot *et al*. [62]). In that work, the authors describe several descriptors to compute pulse clarity based on approaches such as the analysisof the periodicity of the onset curve via autocorrelation, resonance functions, or entropy. Thisfeature has been employed to discover correlations with qualitative measures describing overall properties of the music used in psychology studies in the work by Friberg *et al*. [63].
- **Band periodicity:** this is a measure of the strength of rhythmic or repetitive structures in audio signals (see Lu *et al*. [64]). Band periodicity is defined within a frequency band, and it is obtained as the mean value along all the signal frames of the maximum peak of the subband autocorrelation function.

- **Beat spectrum/spectrogram:** it is a two-dimensional parametrization based on time variations and lag time, thus providing an interpretable representation that reflects temporal changes of tempo (see the work by Foote [22,65]). Beat spectrum shows relevant peaks at rhythm periods that match the rhythmic

properties of the signal. Beat spectrum can be used for discriminating between music (or between parts within an entire music signal) with different tempo patterns.

- **Cyclic beat spectrum:** or CBS for short, this is a representation of the tempo of a music signalthat groups multiples of the fundamental period of the signal together in a single tempo class (Kurth *et al*. [66]). Thus, CBS gives a more compact representation of the fundamental beat period of a song. This feature has been employed in the field of audio retrieval.
- **Beat tracker:** this a feature is derived following an algorithmic approach based on signal subband decomposition and the application of a comb filter analysis in each subband (see Scheirer [67]). Beat tracker mimics at large extent the human ability to track rhythmic beats in music and allows obtaining not only tempo but also compute beat timing positions.
- **Beat histogram:** it provides a more general tempo perspective and summarizes the beat tempos present in a music signal (Tzanetakis and Cook [28]). In this case, Wavelet transform (see Section 4.3 for further details) is used to decompose the signal in octaves for performing subsequent accumulation of the most salient periodicities in each subband to generate the so-called beat histogram. This feature has been used for music genre classification [28].

1.2. Frequency Domain Physical Features

Audio features on the frequency domain constitute the largest set of audio features reported inthe literature (Mitrovic´ *et al*. [17]). They are usually obtained from the Short-Time Fourier Transform (STFT) transform or derived from an autoregression analysis. In general terms, physical frequency domain features describe physical properties of the signal frequency content. Moreover, this type offeatures can be further decomposed as follows:

- Autoregression-based
- STFT-based
- Brightness-related
- Tonality-related
- Chroma-related
- Spectrum shape-related

The following paragraphs describe these subcategories of physical frequency-based features.

1.2.1. Autoregression-Based Frequency Features

Autoregression-based features are derived from linear prediction analysis of signals, which usually captures typical spectral predominances (e.g., formants) of speech signals.

The most commonly employed physical frequency features based on signal autoregression aredescribed below.

- **Linear prediction coefficients:** or LPC for short, this feature represents an all-pole filter that captures the spectral envelope (SE) of a speech signal (formants or spectral resonances that appear in the vocal tract), and have been extensively used for speech coding and recognition applications. LPC have been applied also in audio segmentation and general purpose audio retrieval, like in the works by Khan *et al*. [68,69].
- **Line spectral frequencies:** also referred to as Line Spectral Pairs (LSP) in the literature, Line Spectral Frequencies (LSF) are a robust representation of LPC parameters for quantization and interpolation purposes. They can be computed as the roots phases of the palindromic and the antipalindromic polynomials that constitute the LPC polynomial representation, which in turns

represent the vocal tract when the glottis is closed and open, respectively (see Itakura [70]). Due to its intrinsic robustness they have been widely applied in a diverse set of

classification problems like speaker segmentation (Sarkar and Sreenivas [71]), instrument recognition and in speech/music discrimination (Fu [13]).

- **Code excited linear prediction features:** or CELP for short, this feature was introduced by Schroeder and Atal [72] and has become one of the most important influences in nowadays speech coding standards. This feature comprises spectral features like LSP but also two codebook coefficients related to signal's pitch and prediction residual signal. CELP features have been also applied in the environmental sound recognition framework, like in the work by Tsau *et al*. [73].

1.2.2. STFT-Based Frequency Features

This kind of audio features are generally derived from the signal spectrogram obtained from STFT computation. While some of the features belonging to this category are computed from the analysis of the spectrogram envelope (e.g., subband energy ratio, spectral flux, spectral slope, spectral peaks or MPEG-7 spectral envelope, normalized spectral envelope, and stereo panning spectrum feature), others are obtained from the STFT phase (like group delay functions and/or modified group delay functions).

The following list summarizes the most widely employed STFT-based features.

- **Subband energy ratio:** it is usually defined as a measure of the normalized signal energy along a predefined set of frequency subbands. In a broad sense, it coarsely describes the signal energy distribution of the spectrum (Mitrovic´ *et al*. [17]). There are different approximations as regards the number and characteristics of analyzed subbands (e.g., Mel scale, ad-hoc subbands, *etc.*). Ithas been used for audio segmentation and music analysis applications (see Jiang *et al*. [60], or Srinivasan *et al*. [74]) and environmental sound recognition (Peltonen *et al*. [18]).
- **Spectral flux:** or SF for short, this feature is defined as the 2-norm of the frame-to-frame spectral amplitude difference vector (see Scheirer and Slaney [61]), and it describes sudden changes inthe frequency energy distribution of sounds, which can be applied for detection of musical noteonsets or, more generally speaking, detection of significant changes in the spectral distribution.It measures how quickly the power spectrum changes and it can be used to determine the timbre of an audio signal. This feature has been used for speech/music discrimination (like in Jiang *et al*. [60], or in Khan *et al*. [68,69]), musical instrument classification (Benetos *et al*. [10]),music genre classification (Li *et al*. [40], Lu *et al*. [12], Tzanetakis and Cook [28], Wang *et al*. [9])and environmental sound recognition (see Peltonen *et al*. [18]).
- **Spectral peaks:** this feature was defined by Wang [8] as constellation maps that show the most relevant energy bin components in the time-frequency signal representation. Hence, spectral peaks is an attribute that shows high robustness to possible signal distortions (low signal-to-noise ratio (SNR)–see Klingholz [49], equalization, coders, *etc.*) being suitable for robust recognition applications. This feature has been used for automatic music retrieval (e.g., the well-known Shazam search engine by Wang [8]), but also for robust speech recognition (see Farahani *et al*. [75]).
- **MPEG-7 spectrum envelope and normalized spectrum envelope:** the audio spectrum envelope(ASE) is a log-frequency power spectrum that can be used to generate a reduced spectrogram ofthe original audio signal, as described by Kim *et al*. [76]. It is obtained by summing the energy of the original power spectrum within a series of frequency bands. Each decibel-scale spectral vector is normalized with the RMS energy envelope, thus yielding a normalized log-power version of the ASE called normalized audio spectrum envelope (NASE) (Kim *et al*. [76]). ASE feature has been used in audio event classification [76], music genre classification (Lee *et al*. [77]) andenvironmental sound recognition (see Muhammad and Alghathbar [47], or Valero and Alías [48]).
- **Stereo panning spectrum feature:** or SPSF for short, this feature provides a time-frequency representation that is intended to represent the left/right stereo panning of a stereo audio

signal (Tzanetakis *et al*. [78]). Therefore, this feature is conceived with the aim of capturing relevant information of music signals, and more specifically, information that reflects typical

postproduction in professional recordings. The additional information obtained through SPSF can be used for enhancing music classification and retrieval system accuracies (Tzanetakis *et al*. [79]).

- **Group delay function:** also known as GDF, it is defined as the negative derivative of the unwrapped phase of the signal Fourier transform (see Yegnanarayana and Murthy [80]) and reveals information about temporal localization of events (*i.e.*, signal peaks). This feature has been used for determining the instants of significant excitation in speech signals (like in Smits and Yegnanarayana [81], or Rao *et al*. [82]) and in beat identification in music performances (Sethares *et al*. [83]).
- **Modified group delay function:** or MGDF for short, it is defined as a smoother version of the GDF, reducing its intrinsic spiky nature by introducing a cepstral smoothing process prior to GDF computation. It has been used in speaker identification (Hegde *et al*. [84]), but also in speech analysis, speech segmentation, speech recognition and language identification frameworks (Murthy and Yegnanarayana [85]).

1.2.3. Brightness-Related Physical Frequency Features

Brightness is an attribute that is closely related to the balance of signal energy in terms of high and low frequencies (a sound is said to be bright when it has more high than low frequency content).

The most relevant brightness-related physical features found in the literature are the following:

- **Spectral centroid:** or SC for short, this feature describes the center of gravity of the spectral energy. It can be defined as the first moment (frequency position of the mean value) of the signal frame magnitude spectrum as in the works by Li *et al*. [40], or by Tzanetakis and Cook [28], or obtained from the power spectrum of the entire signal in MPEG-7. SC reveals the predominant frequency of the signal. In the MPEG-7 standard definition [45], the audio spectrum centroid (ASC) is defined by computing SC over the power spectrum obtained from an octave-frequency scale analysis and roughly describes the sharpness of a sound. SC has been applied in musical onset detection (Smith *et al*. [57]), music classification (Bergstra *et al*. [41], Li *et al*. [40], Lu *et al*. [12], Morchen *et al*. [42], Wang *et al*. [9]), environmental sound recognition (like in Peltonen *et al*. [18], Muhammad and Alghathbar [47], Valero and Alías [48]) and, more recently, to music mood classification (Ren *et al*. [86]).
- **Spectral center:** this feature is defined as the median frequency of the signal spectrum, where both lower and higher energies are balanced. Therefore, is a measure close to spectral centroid. It has been shown to be useful for automatic rhythm tracking in musical signals (see Sethares *et al*. [83]).

1.2.4. Tonality-Related Physical Frequency Features

The fundamental frequency is defined as the lowest frequency of an harmonic stationary audio signal, which in turn can be qualified as tonal sound. In music, tonality is a system that organizes the notes of a musical scale according to musical criteria. Moreover, tonality is related to the notion of harmonicity, which describes the structure of sounds that are mainly constituted by a series of harmonically related frequencies (*i.e.*, a fundamental frequency and its multiples), which are typical characteristics of (tonal) musical instruments sounds and voiced speech.

The following paragraphs describe the most widely employed tonality-related features that do not incorporate specific auditory models for their computation.

- **Fundamental frequency:** it is also denoted as F0. The MPEG-7 standard defines audio fundamental frequency feature as the first peak of the local normalized spectro-temporal autocorrelation function [45]. There are several methods in the literature to compute F0,e.g., autocorrelation-based methods, spectral-based methods, cepstral-based methods, and

combinations (Hess [87]). This feature has been used in applications like musical onset detection (Smith *et al*. [57]), musical genre classification (Tzanetakis and Cook [28]), audio retrieval (Wold *et al*. [88]) and environmental sound recognition (Muhammad and Alghathbar [47], Valero and Alías [48]). In the literature F0 is sometimes denoted as *pitch* as it may represent a rough estimate of the perceived tonality of the signal (e.g., pitch histogram and pitch profile).

- **Pitch histogram:** instead of using a very specific and local descriptor like fundamental frequency, the pitch histogram describes more compactly the pitch content of a signal. Pitch histogramhas been used for musical genre classification by Tzanetakis and Cook [28], as it gives a general perspective of the aggregated notes (frequencies) present in a musical signal along a certain period.
- **Pitch profile:** this feature is a more precise representation of musical pitch, as it takes into account both pitch mistuning effects produced in real instruments and also pitch representation of percussive sounds. It has been shown that use of pitch profile feature outperforms conventional chroma-based features in musical key detection, like in Zhu and Kankanhalli [89].
- **Harmonicity:** this feature is useful for distinguishing between tonal or harmonic (e.g., birds, flute, *etc.*) and noise-like sounds (e.g., dog bark, snare drum, *etc.*). Most traditional harmonicity features either use an impulse train (like in Ishizuka *et al*. [90]) to search for the set of peaks in multiples of F0, or uses the autocorrelation-inspired functions to find the self-repetition of the signal in the time- or frequency-domain (as in Kristjansson *et al*. [91]). Spectral local harmonicity is proposed in the work by Khao [92], a method that uses only the sub-regions of the spectrum that still retain a sufficient harmonic structure. In the MPEG-7 standard, two harmonicity measures are proposed. Harmonic ratio (HR) is a measure of the proportion of harmonic components in the power spectrum. The Upper limit of harmonicity (ULH) is an estimation of the frequency beyond which the spectrum no longer has any harmonic structure. Harmonicity has been used also in the field of environmental sound recognition (Muhammad and Alghathbar [47], Valero and Alías [48]). Some other harmonicity-based features for music genre and instrument family classification have been defined, like harmonic concentration, harmonic energy entropy or harmonic derivative (see Srinivasan and Kankanhalli [93]).
- **Inharmonicity:** this feature measures the extent to which the partials of a sound are separated with respect to its ideal position in a harmonic context (whose frequencies are integersof a fundamental frequency). Some approaches take into account only partial frequencies (like Agostini *et al*. [94,95]), while others also consider partial energies and bandwidths(see Cai *et al*. [96]).
- **Harmonic-to-Noise Ratio:** Harmonic-to-noise Ratio (HNR) is computed as the relation between the energy of the harmonic part and the energy of the rest of the signal in decibels (dB) (Boersma [97]). Although HNR has been generally applied to analyze pathological voices (like in Klingholz [49], or in Lee *et al*. [98]), it has also been applied in some music-related applications such as the characterization of growl and screaming singing styles, as in Kato and Ito [53].
- **MPEG-7 spectral timbral descriptors:** the MPEG-7 standard defines some features that are closely related to the harmonic structure of sounds, and are appropriate for discrimination of musical sounds: MPEG-7 harmonic spectral centroid (HSC) (the amplitude-weighted average of the harmonic frequencies, closely related to brightness and sharpness), MPEG-7 harmonic spectraldeviation (HSD) (amplitude deviation of the harmonic peaks from their neighboring harmonicpeaks, being minimum if all the harmonic partials have the same amplitude), MPEG-7 harmonicspectral spread (HSS) (the power-weighted root-mean-square deviation of the harmonic peaks from the HSC, related to harmonic bandwidths), and MPEG-7 harmonic spectral variation (HSV) (correlation of harmonic peak amplitudes in two adjacent frames, representing the harmonic variability over time). MPEG-7 spectral timbral descriptors have been employed for environmental sound recognition (Muhammad and Alghathbar [47],Valero and Alías [48]).

- **Jitter:** computes the cycle-to-cycle variations of the fundamental frequency (Klingholz [49]), that is,the average absolute difference between consecutive periods of speech (Farrús *et al*. [51]). Besides typically being applied to analyze pathological voices (like in Klingholz [49], or in Kreiman and Gerratt [50]), it has also been applied to prototyping, classification and creation of musical sounds (Jensen [54]), improve

speaker recognition (Farrús *et al.* [51]), characterize growl and screaming singing styles (Kato and Ito [53]) or discriminate vocal and non-vocal regions from audio songs (Murthy and Koolagudi [52]), among others.

1.2.5. Chroma-Related Physical Frequency Features

Chroma is related to perception of pitch, in the sense that it is a complement of the tone height. In a musical context, two notes that are separated one or more octaves have the same chroma (e.g., C4 and C7 notes), and produce a similar effect on the human auditory perception.

The following paragraphs describe chroma-related frequency features, which are basically computed from direct physical approaches:

- **Chromagram:** also known as chroma-based feature, chromagram is a spectrum-based energy representation that takes into account the 12 pitch classes within an octave (corresponding to pitch classes in musical theory) (Shepard [99]), and it can be computed from a logarithmic STFT (Bartsch and Wakefield [100]). Then, it constitutes a very compact representation suited for musical and harmonic signals representation following a perceptual approach.
- **Chroma energy distribution normalized statistics:** or CENS for short, this feature was conceived for music similarity matching and has shown to be robust to tempo and timbre variations (Müller *et al.* [101]). Therefore, it can be used for identifying similarities between different interpretations of a given music piece.

1.2.6. Spectrum Shape-Related Physical Frequency Features

Another relevant set of frequency features are the ones that try to describe the shape of the spectrum of the audio signal. The following paragraphs describe the most widely employed, and some of the newest contributions in this area.

- **Bandwidth:** usually defined as the second-order statistic of the signal spectrum, it helps to discriminate tonal sounds (with low bandwidths) from noise-like sounds (with high bandwidths) (see Peeters [34]). However, it is difficult to distinguish between complex tonal sounds (e.g., music, instruments, *etc.*) from complex noise-like sounds using only this feature. It can be defined over the power spectrum or in its logarithmic version (see Liu *et al.* [59], or Srinivasan and Kankanhalli [93]) and it can be computed over the whole spectrum or within different subbands (like in Ramalingam and Krishnan [102]). MPEG-7 defines audio spectrum spread (ASS) as the standard deviation of the signal spectrum, which constitutes the second moment while (being the ASC the first one). Spectral bandwidth has been used for music classification (Bergstra *et al.* [41], Lu *et al.* [12], Morchen *et al.* [42], Tzanetakis and Cook [28]), and environmental sound recognition (Peltonen *et al.* [18], Muhammad and Alghathbar [47], Valero and Alías [48]).
- **Spectral dispersion:** this is a measure closely related to spectral bandwidth. The only difference is that it takes into account the spectral center (median) instead of the spectral centroid (mean) (see Sethares *et al.* [83]).
- **Spectral rolloff point:** defined as the 95th percentile of the power spectral distribution (see Scheirer and Slaney [61]), spectral rolloff point can be regarded as a measure of the skewness of the spectral shape. It can be used, for example, for distinguishing between voiced from unvoiced speech sounds. It has been used in music genre classification (like in Li and Ogihara [103], Bergstra *et al.* [41], Li *et al.* [40], Lu *et al.* [12], Morchen *et al.* [42], Tzanetakis and Cook [28], Wang *et al.* [9]), speech/music discrimination (Scheirer and Slaney [61]), musical instrument classification (Benetos *et al.* [10]), environmental sound recognition

(Peltonen *et al*. [18]), audio-based surveillance systems (Rabaoui *et al*. [24]) and music mood classification (Ren *et al*. [86]).

- **Spectral flatness:** this is a measure of uniformity in the frequency distribution of the power spectrum, and it can be computed as the ratio between the geometric and the arithmetic mean of a subband (see Ramalingam and Krishnan [102]) (equivalent to the MPEG-7 audio spectrum flatness (ASF) descriptor [45]). This feature allows distinguishing between noise-like sounds (high value of spectral flatness) and more tonal sounds (low value). This feature has been used in audio fingerprinting (see Lancini *et al*. [104]), musical onset detection (Smith *et al*. [57]), music classification (Allamanche *et al*. [105], Cheng *et al*. [106], Tzanetakis and Cook [28]) and environmental sound recognition (Muhammad and Alghathbar [47], Valero and Alías [48]).
- **Spectral crest factor:** in contrast to spectral flatness measure, spectral crest factor measures how peaked the power spectrum is, and it is also useful for differentiation of noise-like (lower spectral crest factor) and tonal sounds (higher spectral crest factor). It can be computed as the ratio between the maximum and the mean of the power spectrum within a subband, and has been used for audio fingerprinting (see Lancini *et al*. [104], Li and Ogihara [103]) and music classification (Allamanche *et al*. [105], Cheng *et al*. [106]).
- **Subband spectral flux:** or SSF for short, this feature is inversely proportional to spectral flatness, being more relevant in subbands with non-uniform frequency content. In fact, SSF measures the proportion of dominant partials in different subbands, and it can be measured accumulating the differences between adjacent frequencies in a subband. It has been used for improving the representation and recognition of environmental sounds (Cai *et al*. [96]) and music mood classification (Ren *et al*. [86]).
- **Entropy:** this is another measure that describes spectrum uniformity (or flatness), and it can be computed following different approaches (Shannon entropy, or its generalization named Renyi entropy) and also in different subbands (see Ramalingam and Krishnan [102]). It has been used for automatic speech recognition, computing the Shannon entropy in different equal size subbands, like in Misra *et al*. [107].
- **Octave-based Spectral Contrast:** also referred to as OSC, it is defined as the difference between peaks (that generally corresponds to harmonic content in music) and valleys (where non-harmonic or noise components are more dominant) measured in subbands by octave-scale filters and using a neighborhood criteria in its computation (Jiang *et al*. [108]). To represent the whole music piece, mean and standard deviation of the spectral contrast and spectral peak of all frames are used as the spectral contrast features. OSC features have been used for music classification (Lee *et al*. [77], Lu *et al*. [12], Yang *et al*. [109]) and music mood classification, as in Ren *et al*. [86].
- **Spectral slope:** this is a measure of the spectral slant by means of a simple linear regression (Morchen *et al*. [42]), and it has been used for classification purposes in speech analysis applications (Shukla *et al*. [110]) and speaker identification problems (Murthy *et al*. [111]).
- **Spectral skewness and kurtosis:** spectral skewness, which is computed as the 3rd order moment of the spectral distribution, is a measure that characterizes the asymmetry of this distribution around its mean value. On the other hand, spectral kurtosis describes the flatness of the spectral distribution around its mean, and its computed as the 4th order moment (see Peeters *et al*. [34]). Both parameters have been applied for music genre classification (Baniya *et al*. [112]) and music mood classification