An introduction to audio features

Carmine-Emanuele Cella

Conservatorio di Padova

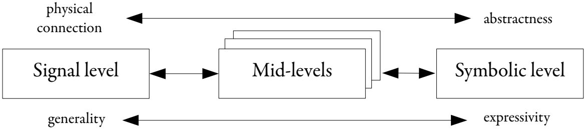
23 march 2015

1. Introduction

2. Features

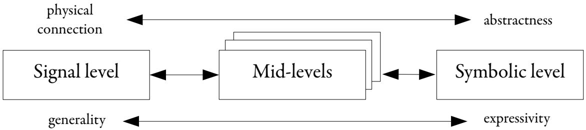
3. Classification

Different representations



- Music, in its final stage of *performance*, can be described in many ways (time-varying signal, symbolic system, etc.).

Different representations



- Music, in its final stage of *performance*, can be described in many ways (time-varying signal, symbolic system, etc.).
- Each approach selects a particular degree of abstraction: the *signal level*, the *symbolic level*, a fixed mixture of both (*mid-levels*).

Basic signal models (1)

The decomposition of a signal $x[n]$ into expansion functions is a linear combination of the form:

$$\overset{n}{\vec{x}} = \sum_{k=1}^{K} \alpha_k \, \overset{n}{\vec{g}}_k \ .$$

The coefficients α_k are derived from the analysis stage, while the functions $g_k[n]$ can be determined by the analysis stage or fixed beforehand.

Basic signal models (2)

An example of signal decomposition is given by the Discrete Fourier transform (DFT):

$$\vec{X}_k = \sum_{i=0}^{n-1} x_i \cdot e^{-j\cdot\frac{2\cdot\pi}{n}\cdot k}.$$

Time-frequency decompositions of signals of N samples are also possibile, such as the Short-time Fourier transform (STFT), taking n samples at a time and hopping by t samples:

$$\overset{N}{\vec{X}}_{\overset{n}{\vec{k}}} = \sum_{i=0}^{N/t} \overset{n}{\vec{h}} \cdot \overset{n}{\vec{x}}_{i\cdot t} \cdot e^{-j\cdot\frac{2\cdot\pi}{n}\cdot\overset{n}{\vec{k}}}$$

where $\overset{n}{\vec{h}}$ is a window of length n-samples.

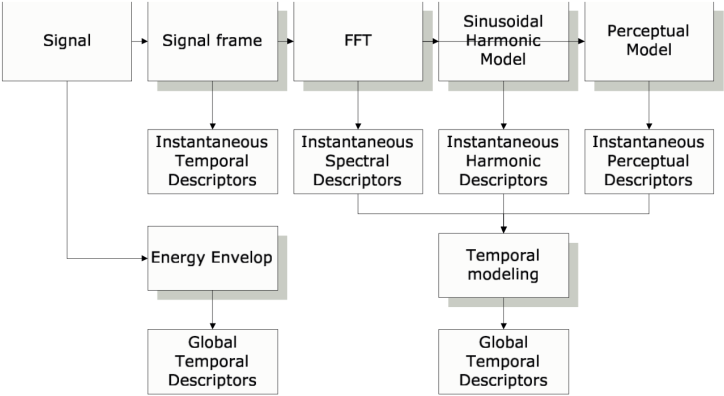
Definition of low-level features

- Low-level features: numerical values describing the contents of a signal according to different kinds of inspection: temporal, spectral, perceptual, etc.

Definition of low-level features

- Low-level features: numerical values describing the contents of a signal according to different kinds of inspection: temporal, spectral, perceptual, etc.
- They are computed on a small time scale which is usually between 40 ms and 80 ms; different kinds of temporal modeling (like mean and variance computation) can then be applied.

Families of descriptors


Signal
Signal frame
FFT
Sinusoidal Harmonic Model
Perceptual Model
Instantaneous Temporal Descriptors
Instantaneous Spectral Descriptors
Instantaneous Harmonic Descriptors
Instantaneous Perceptual Descriptors
Energy Envelop
Temporal modeling
Global Temporal Descriptors
Global Temporal Descriptors

Spectral features (1)

- Spectral centroid (brightness) and spectral spread (bandwidth).

$$\mu = \int x \cdot p(x) dx.$$

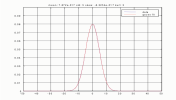
$$\sigma^2 = \int (x - \mu)^2 \cdot p(x) dx.$$

Here x are the observed data (i.e. the frequencies of the spectrum) while $p(x)$ are the probabilites to observe x (i.e. the amplitudes of the spectrum).

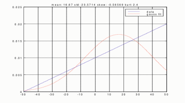
Spectral features (2)

- Spectral skewness (asymmetry).

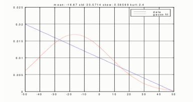
$$\gamma^3 = \frac{\int (x-\mu)^3 \cdot p(x)dx}{\sigma^3}.$$



skewness=0



skewness<0

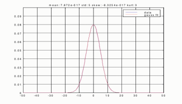


skewness>0

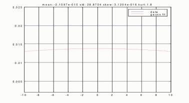
Spectral features (3)

- Spectral kurtosis (flatness).

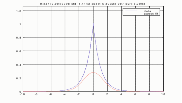
$$\gamma^4 = \frac{\int (x-\mu)^4 \cdot p(x)dx}{\sigma^4}.$$



Kurtosis=3



Kurtosis=1.8



Kurtosis=6

Other features (1)

- There are many other spectral features, such as: spectral flux, high-frequency content, spectral crest, spectral rolloff, spectral decrease, spectral irregularity, etc.

Other features (1)

- There are many other spectral features, such as: spectral flux, high-frequency content, spectral crest, spectral rolloff, spectral decrease, spectral irregularity, etc.
- Other families are computed, principally, by applying the same formulas on modeled data, such as harmonic or perceptual modeling.

Other features (1)

- There are many other spectral features, such as: spectral flux, high-frequency content, spectral crest, spectral rolloff, spectral decrease, spectral irregularity, etc.
- Other families are computed, principally, by applying the same formulas on modeled data, such as harmonic or perceptual modeling.
- In this way it is possible to compute, for example: harmonic centroid or harmonic spread, perceptual centroid or perceptual spread and so on.

Other features (1)

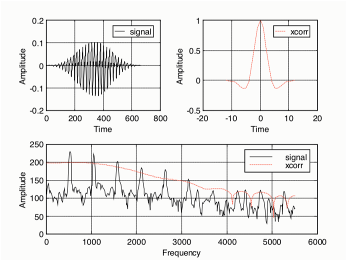
- There are many other spectral features, such as: spectral flux, high-frequency content, spectral crest, spectral rolloff, spectral decrease, spectral irregularity, etc.
- Other families are computed, principally, by applying the same formulas on modeled data, such as harmonic or perceptual modeling.
- In this way it is possible to compute, for example: harmonic centroid or harmonic spread, perceptual centroid or perceptual spread and so on.
- There exist also *domain-specific* descriptors (usually middle-level) such as MFCC or inharmonicity.

Other features (2): autocorrelation

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:

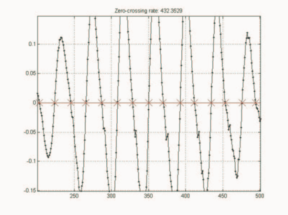
$$xcorr(k) = \frac{1}{x(0)^2} \sum_{n=0}^{N-k-1} x(n) \cdot x(n+k)$$



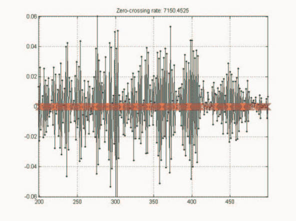
[top-left] signal [top-right] cross-correlation function
[bottom] signal amplitude spectrum and spectrum envelop estimated by cross-correlation (dashed line)

Other features (3): zero-crossing rate

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.



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



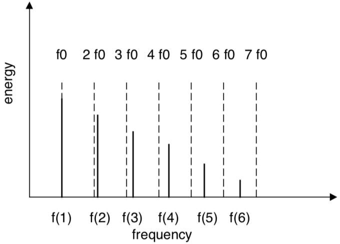
Zero-crossing rate (=7150) during unvoiced speech region

Other features (4): Inharmonicity

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.

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

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.

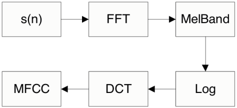


Inharmonicity coefficient computation: harmonic multiple (dotted lines), observed spectral peaks (contiuous lines)

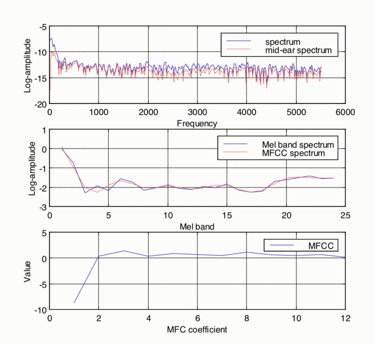
Other features (5): MFCC [Mel-frequency cepstral coeff.]

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.

Formulation:

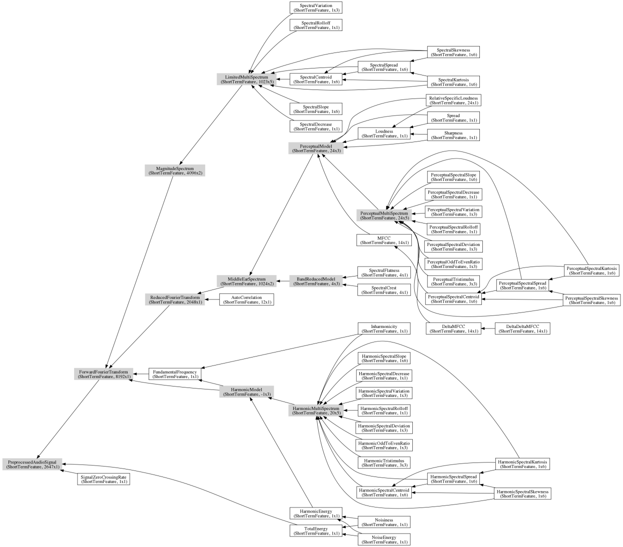


s(n)
FFT
MelBand
Log
DCT
MFCC

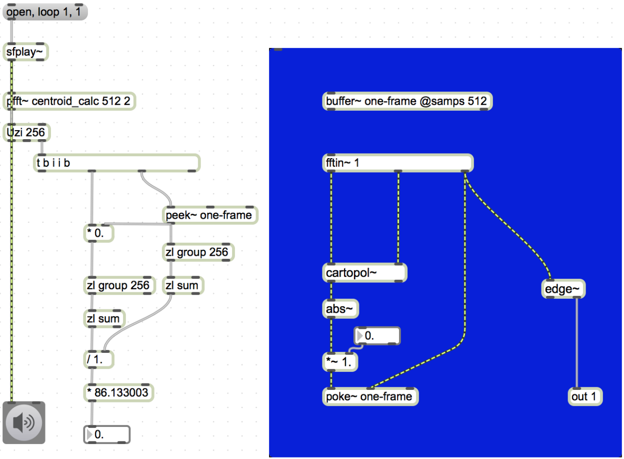


[Top] signal spectrum and mid-ear filtered spectrum (dashed line) [middle] Mel band spectrum and MFCC spectrum (dotted line) [bottom] MFCC coefficients

Families in IrcamDescriptor (2009)


SpectralVariation (ShortTermFeature, 1x3)
SpectralRolloff (ShortTermFeature, 1x1)
SpectralSkewness (ShortTermFeature, 1x6)
SpectralSpread (ShortTermFeature, 1x6)
LimitedMultiSpectrum (ShortTermFeature, 1023x5)
SpectralCentroid (ShortTermFeature, 1x6)
SpectralKurtosis (ShortTermFeature, 1x6)
RelativeSpecificLoudness (ShortTermFeature, 24x1)
SpectralSlope (ShortTermFeature, 1x6)
Spread (ShortTermFeature, 1x1)
SpectralDecrease (ShortTermFeature, 1x1)
Loudness (ShortTermFeature, 1x1)
Sharpness (ShortTermFeature, 1x1)
PerceptualModel (ShortTermFeature, 24x3)
MagnitudeSpectrum (ShortTermFeature, 4096x2)
PerceptualSpectralSlope (ShortTermFeature, 1x6)
PerceptualSpectralDecrease (ShortTermFeature, 1x1)
PerceptualMultiSpectrum (ShortTermFeature, 24x5)
PerceptualSpectralVariation (ShortTermFeature, 1x3)
PerceptualSpectralRolloff (ShortTermFeature, 1x1)
MFCC (ShortTermFeature, 14x1)
PerceptualSpectralDeviation (ShortTermFeature, 1x3)
PerceptualOddToEvenRatio (ShortTermFeature, 1x3)
PerceptualSpectralKurtosis (ShortTermFeature, 1x6)
SpectralFlatness (ShortTermFeature, 4x1)
MiddleEarSpectrum (ShortTermFeature, 1024x2)
BandReducedModel (ShortTermFeature, 4x3)
PerceptualTristimulus (ShortTermFeature, 3x3)
PerceptualSpectralSpread (ShortTermFeature, 1x6)
SpectralCrest (ShortTermFeature, 4x1)
ReducedFourierTransform (ShortTermFeature, 2048x1)
AutoCorrelation (ShortTermFeature, 12x1)
PerceptualSpectralCentroid (ShortTermFeature, 1x6)
PerceptualSpectralSkewness (ShortTermFeature, 1x6)
Inharmonicity (ShortTermFeature, 1x1)
DeltaMFCC (ShortTermFeature, 14x1)
DeltaDeltaMFCC (ShortTermFeature, 14x1)
HarmonicSpectralSlope (ShortTermFeature, 1x6)
ForwardFourierTransform (ShortTermFeature, 8192x1)
FundamentalFrequency (ShortTermFeature, 1x1)
HarmonicSpectralDecrease (ShortTermFeature, 1x1)
HarmonicModel (ShortTermFeature, -1x3)
HarmonicSpectralVariation (ShortTermFeature, 1x3)
HarmonicMultiSpectrum (ShortTermFeature, 20x5)
HarmonicSpectralRolloff (ShortTermFeature, 1x1)
HarmonicSpectralDeviation (ShortTermFeature, 1x3)
HarmonicOddToEvenRatio (ShortTermFeature, 1x3)
HarmonicTristimulus (ShortTermFeature, 3x3)
PreprocessedAudioSignal (ShortTermFeature, 2647x1)
HarmonicSpectralKurtosis (ShortTermFeature, 1x6)
SignalZeroCrossingRate (ShortTermFeature, 1x1)
HarmonicSpectralSpread (ShortTermFeature, 1x6)
HarmonicSpectralCentroid (ShortTermFeature, 1x6)
HarmonicSpectralSkewness (ShortTermFeature, 1x6)
HarmonicEnergy (ShortTermFeature, 1x1)
Noisiness (ShortTermFeature, 1x1)
TotalEnergy (ShortTermFeature, 1x1)
NoiseEnergy (ShortTermFeature, 1x1)

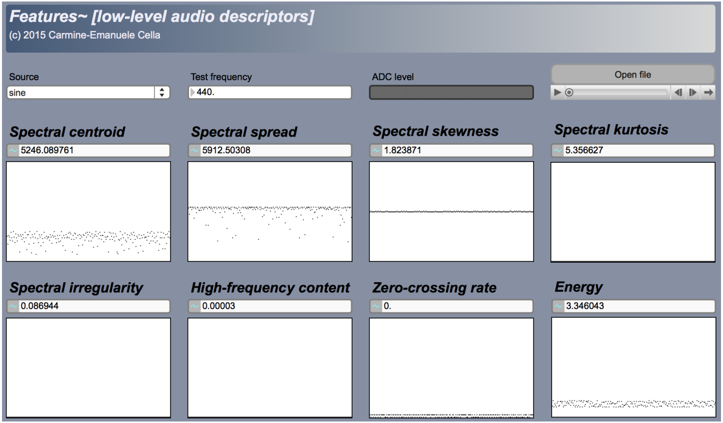
A minimal implementation in Max/MSP



```
open, loop 1, 1
sfplay~
pfft~ centroid_calc 512 2
Uzi 256
t b i i b
peek~ one-frame
* 0.
zl group 256
zl group 256
zl sum
zl sum
/ 1.
* 86.133003
0.
```

```
buffer~ one-frame @samps 512
fftin~ 1
cartopol~
abs~
0.
*~ 1.
poke~ one-frame
edge~
out 1
```

A more complete set in features∼ (2015)



Features~ [low-level audio descriptors]

(c) 2015 Carmine-Emanuele Cella

Source: sine

Test frequency: 440.

ADC level

Spectral centroid: 5246.089761

Spectral spread: 5912.50308

Spectral skewness: 1.823871

Spectral kurtosis: 5.356627

Spectral irregularity: 0.086944

High-frequency content: 0.00003

Zero-crossing rate: 0.

Energy: 3.346043

Features and classification (1)

- Audio indexing: sound classification method based on the projection of low-level features over a set of sounds in a multi-dimensional space (*feature space*).

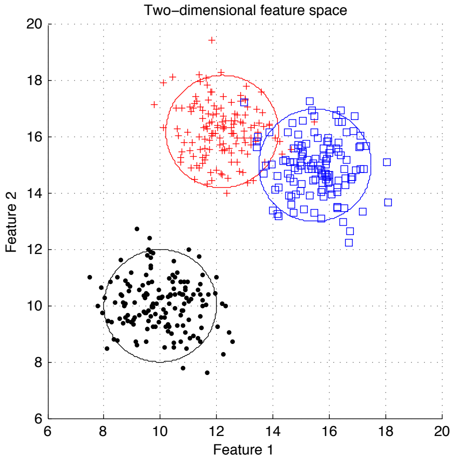
Features and classification (1)

- Audio indexing: sound classification method based on the projection of low-level features over a set of sounds in a multi-dimensional space (*feature space*).
- By analysing the space with some combined geometrical and statistical techniques (like K-means, Gaussian Mixture Models, Principal Component Analysis, etc.) it is possible to find the clusters of sounds present in the space.

Features and classification (1)

- Audio indexing: sound classification method based on the projection of low-level features over a set of sounds in a multi-dimensional space (*feature space*).
- By analysing the space with some combined geometrical and statistical techniques (like K-means, Gaussian Mixture Models, Principal Component Analysis, etc.) it is possible to find the clusters of sounds present in the space.
- With specific techniques, such as the BIC measure or the gap-statistic, it is possible to *evaluate* the computed clustering.

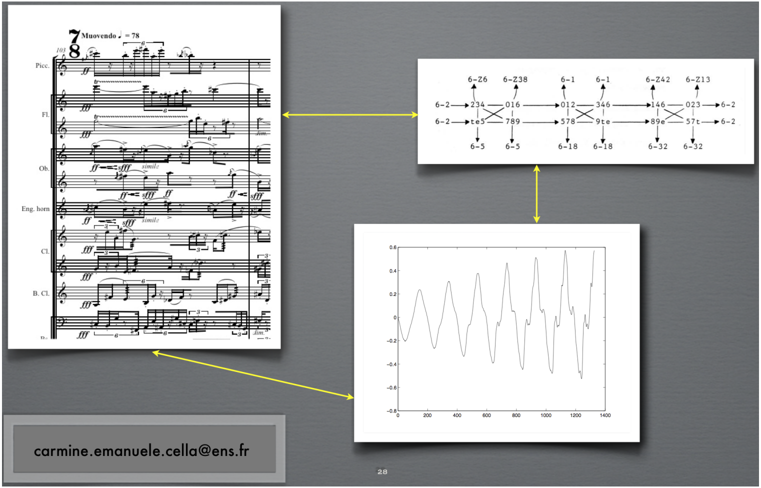
Features and classification (2)



Some references...

- G. Peeters, **A large set of audio features for sound description (similarity and classification) in the CUIDADO project**, IRCAM internal report, 2004.
- J.J. Burred, C. E. Cella, et al., **Using the SDIF sound description interchange format for audio features**, ISMIR Philadelphia, 2007.
- C. E. Cella, **On symbolic representations of music**, PhD dissertation, 2011, University of Bologna/IRCAM.

Any questions?



carmine.emanuele.cella@ens.fr

28