Low-level features and timbre

Juan Pablo Bello
MPATE-GE 2623 Music Information Retrieval
New York University
Music signal analysis

- **High-level representation**
  - e.g.: style, artist, mood, form

- **Mid-level representation**
  - e.g.: pitch, onsets, beats

- **Low-level representation**
  - e.g.: spectral flux, ACF, cepstrum

 raw input

[... 0.2 0.1 0.05 -0.05 -0.1 ...]
Low-level features

• The raw input data is often too large, noisy and redundant for analysis.

• Feature extraction: input signal is transformed into a new (smaller) space of variables that simplify analysis.

• Features: measurable properties of the observed phenomenon, usually containing information relevant for pattern recognition.

• They result from neighborhood operations on the input signal. If the operation produces a local decision -> feature detection.

• Usually one feature is not enough: combine several features into feature vectors, describing a multi-dimensional space.
Timbre

- Timbre: tonal qualities that define a particular sound/source. It can refer to, e.g., class (e.g. violin or piano), or quality (e.g. bright, rough)

- Oftentimes defined comparatively: attribute that allows us to differentiate sounds of the same pitch, loudness, duration and spatial location (Grey, 75)

- Timbre spaces: empirically measure the perceived (dis)similarity between sounds and project to a low-dimensional space where dimensions are assigned a semantic interpretation (brightness, temporal variation, synchronicity, etc).

- Audio-based: recreate timbre spaces by extracting low-level features with similar interpretations (centroid, spectral flux, attack time, etc). Most of them describe the steady-state spectral envelope.
Temporal features

- The root-mean-square (RMS) level coarsely approximates loudness:

\[
RMS(m) = \sqrt{\frac{1}{N} \sum_{n=-N/2}^{N/2} (x(n + mh))^2 w(n)}
\]

- Zero-crossing rate (ZCR) is a weighted measure of the number of times the signal changes sign in a frame:

\[
ZCR(m) = \frac{1}{2N} \sum_{n=-N/2}^{N/2} |\text{sgn}(x(n + mh)) - \text{sgn}(x(n + mh - 1))|
\]

\[
\text{sgn}(x) = \begin{cases} 
1 & \text{if } x > 0 \\
0 & \text{if } x = 0 \\
-1 & \text{if } x < 0 
\end{cases}
\]
Temporal features

- ZCR is high for noisy (unvoiced) sounds and low for tonal (voiced) sounds.
- For simple periodic signals, it is roughly related to the fundamental frequency.
Spectral features

• The most common is the spectral centroid (SC):

\[ SC(m) = \frac{\sum_k f_k |X(m, k)|}{\sum_k |X(m, k)|} \]

• It is usually associated with the sound’s “brightness”

• Spectral spread (SS) is a measure of the bandwidth of the spectrum:

\[ SS(m) = \frac{\sum_k (f_k - SC(m))^2 |X(m, k)|}{\sum_k |X(m, k)|} \]

• Higher-order moments can be used to characterize the asymmetry and peakedness of the distribution
Spectral features
Spectral flatness

• Spectral flatness is a measure of the noisyness of the magnitude spectrum.

• It is the ratio between the geometric and arithmetic means:

\[
SF(m) = \left( \prod_{k} |X(m, k)| \right)^{1/K} \frac{1}{K} \sum_{k} |X(m, k)|
\]

• Different filterbanks can be used for pre-processing, s.t. k refers to band number and K to total number of bands.

• It is often used as a “tonality” coefficient (in dB)
Spectral flatness
SC and SS define a coarse (unimodal) model of the spectral envelope.
Channel Vocoder

- Decomposes the sound using a bank of bandpass filters + sums magnitude for each bandpass signal

- For a set of $L$-long filters $w$ overlapped by $L-1$ bins:

\[
CV(m) = |X(m, k)| \ast w(k)
\]

\[
CV(m) = \mathbb{R} (\text{IFFT} [\text{FFT}(|X(m, k)|) \times \text{FFT}(w(k))])
\]

- $w(k)$ is normalized to unit sum, zero-padded to the length of $X$, and circularly shifted s.t. its center coincides with the first bin.
Channel Vocoder

- The spectral envelope approximation is coarser/finer depending on L
Remember Cepstrum?

- Treat the log magnitude spectrum as if it were a signal -> take its (I)DFT
- Measures rate of change across frequency bands (Bogert et al., 1963)
- For a real-valued signal it’s defined as:

\[ c_x(l) = \text{real}(\text{IFFT}(\log(|FFT(x)|))) \]

- Followed by low-pass “liftering” in the cepstral domain
Cepstrum

- The real cepstrum can be weighted using a low-pass window of the form:

\[
\begin{align*}
   w_{LP}(l) &= \begin{cases} 
   1 & \text{if } l = 0, L_1 \\
   2 & \text{if } 1 \leq l < L_1 \\
   0 & \text{if } L_1 < l \leq L - 1 
   \end{cases} \\
   c_{LP}(l) &= c_x(l) \times w_{LP}(l) \\
   C_{LP}(k) &= e^{R[\text{FFT}(c_{LP}(l))]} 
\end{align*}
\]

- Such that \( L_1 \leq L/2 \), and \( C_{LP} \) is the spectral envelope.
• The spectral envelope approximation is coarser/finer depending on $L_1$
Cepstrum

Spectrogram (dB)

Cepstrum

Spectral Envelope (dB)
MFCC

- Mel-frequency Cepstral Coefficients (MFCC): variation of the linear cepstrum, widely used in audio analysis.

- Most popular features in speech: due to their ability to compactly represent the audio spectrum

- Introduced to music DSP by Logan (ISMIR, 2000).
MFCC

• The Mel scale is a non-linear perceptual scale of pitches judged to be equidistant:

\[
mel = 1127.01028 \times \log \left(1 + \frac{f}{700}\right)
\]

\[
f = 700 \times \left(e^{\frac{mel}{1127.01028}} - 1\right)
\]

• Approx. linear \( f < 1\text{kHz}; \) logarithmic above that.

• Reference point is at \( f = 1\text{kHz}, \) which corresponds to 1000 Mel: a tone perceived to be half as high is 500 Mel, twice as high is 2000 Mel, etc.
• Filterbank of overlapping windows

• Center frequencies uniformly distributed in mel scale, s.t. the center frequency of one window: starting point of next window and end point of previous window.

• All windows are normalized to unity sum.
MFCC

• An efficient representation of the log-spectrum can be obtained by applying a transform that decorrelates the Mel dB spectrum (see Rabiner and Juang, 93).

• This decorrelation is commonly approximated by means of the Discrete Cosine Transform (DCT)

• DCT: real-valued transform, similar to the DFT. Most of its energy is concentrated on a few low coefficients (effectively compressing the spectrum)

\[
X_{DCT}(k) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x(n) \cos \left( \frac{\pi k}{N} \left( n - \frac{1}{2} \right) \right)
\]
MFCC
MFCC

Spectrogram (dB)

MFCC

Mel-Spectrum envelope (dB)
A reminder

• The feature vector is representing an N-long time segment, and is best mapped in time to the center of the window.

• Zero-padding can be used to map the first vector to $n = 0$, and ensure all the signal is analyzed.
Post-processing

• We can characterize the short-term temporal dynamics of feature coefficients by using delta and acceleration coefficients:

\[
\Delta y = \frac{y(n) - y(n - h)}{h}
\]

\[
\Delta \Delta y = \frac{y(n) - 2y(n - h) + y(n - 2h)}{h^2}
\]

• Normalization is often necessary/beneficial:

\[
\hat{y} = \frac{y - \min(y)}{\max(y - \min(y))}
\]

\[
\hat{y} = \frac{y - \mu_y}{\sigma_y}
\]
Post-processing

• Normalizing features across time avoids bias towards high-range features

• Normalizing feature vectors make them more comparable to each other

• Looses dynamic change information
Summarization

• Global (song/sound) features can be obtained by summarizing frame-level features:

\[
\begin{align*}
&\text{• Resulting on a single } 2 \times P\text{-long feature vector of means and variances.} \\
&\text{• If not independent we measure the covariance:} \\
&\text{\quad } \text{cov} = \sum_{m} (y - \mu_y)(y - \mu_y)^T / M 
\end{align*}
\]
Summarization

- Texture windows can be used to capture local behavior:
Summarization

- Computing simple statistics across time ignores temporal ordering. Same global features for:

  - Original signal
  - Re-shuffled signal
  - Reversed signal
References


References


