Melody Extraction from Polyphonic Music Signals

Gaël Richard
Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, France

With help from J. Salamon, E. Gomez, D. Ellis, J-L Durrieu, B. Fuentes, A. Ozerov, A. Liutkus

International Workshop on Acoustic Signal Enhancement (IWAENNC 2014)
Sept. 11th, 2014
Audio recordings

- What is an audio recording?
Audio recordings

What is an audio recording?

- It is composed of audio objects or sources…
  - piano, drums, guitar, …

- … Which are mixed together into a mixture (i.e. the audio recording) which is possibly multichannel (stereo is the most common for music)
Audio recordings

■ What is an audio recording?

- It is composed of *audio objects* or *sources*...

  piano  drums  guitar  ....  (stop)

- .... Which are mixed together into a *mixture* (i.e. the audio recording) which is possibly multichannel (stereo is the most common for music)

■ In most cases only the mixture is available which limits *Active Listening* capabilities ...
Applications

What could we do if we had the separated audio objects?

- Active listening
- Karaoke
- Remixing
- Music information retrieval
  - Cover song detection,
  - Music transcription (audio-to-midi, instrument recognition,...)
- Audio Classification
- ....
What is « melody extraction » ?

- Also termed
  - Audio melody extraction
  - Predominant melody extraction/estimation
  - Predominant fundamental frequency estimation

- The aim: to obtain a sequence of frequency values representing the pitch of the dominant melodic line

The problem more precisely …

- **Definition:**

  \[ \text{melody line} = \text{sequence of } f_0 \text{ (fundamental frequency values)} \] of the lead instrument or voice of a polyphonic music audio signal.

**Polyphonic music audio signal:**

- a music audio signal where two or more notes can sound simultaneously
Some difficulties of the problem

- « Voicing detection »: Determining when the leading voice is active

- « Polyphony »: Presence of multiple concurrent instruments

- « Harmonicity »: The notes of each instrument are often harmonically related

- « Mixing effects »: Presence of sound effects (reverberation, dynamic compression,...)

- « Melody tracking »: Associate the different fundamental frequencies obtained to the melody line
Some difficulties of the problem

- Top: Spectrogram of the music signals (in dB)
- Bottom: extracted melody [Salomon2012] (blue) and ground truth (red).

From "Blind" melody extraction to Informed melody extraction

- Strictly speaking, "Blind" melody extraction is only done using the audio mixture.
  - In practice, some (limited) priors or assumptions are used, e.g.:
    - Harminicity of the lead instrument
    - Production model of the lead instrument

- Informed melody extraction
  - Side information is transmitted to the extraction module
  - Extraction is done using the mixture and the side information
Informed melody extraction

User-guided melody extraction

User

Mix → Melody extraction algorithm → Estimated melody

Melody extraction algorithm

Side information

User-guided melody extraction
From “Blind” to Informed

Blind
Independent component analysis

Supervised, weakly-informed
“e.g., the singer is known”

Informed
“e.g., a humming of the melody to be extracted is known”
Content

- Introduction
- (« blind ») Main melody extraction
  - Salient based approaches
  - Source-separation-based approaches
  - Alternative approaches
  - Evaluation
- Informed main melody extraction
- Conclusion
From monopitch estimation to Main melody extraction

• … a task similar to monophonic pitch extraction..

• Classically, monophonic pitch extraction estimates a sequence \( \hat{f}_{\text{mono}} \) of pitch values as:

\[
\hat{f}_{\text{mono}} = \arg\max_f \sum_\tau S_x(f_\tau, \tau) + C(f)
\]

- Sequence of pitch values
- Temporal constraints
- Function indicating the likelihood Of the pitch candidates at each time trame \( \tau \)
From monopitch estimation to Main melody extraction

- In melody extraction

\[ y(t) = x(t) + n(t) \]

- But « noise » includes other periodic signals, potentially harmonically related to the melody
- The melody may not be always active or be the dominant source…
From monopitch estimation to Main melody extraction

- Two main directions for main melody estimation

  - **Salience-based** approaches, using a modified pitch salience function calculated over the mixed signal.
    \[
    \hat{f}_{sal} = \arg \max_f \sum_{\tau} S'_y(f, \tau) + C'(f)
    \]

  - **Source-separation** approaches using an estimation of the separated leading voice component \(\hat{x}(t)\)
    \[
    \hat{f}_{sep} = \arg \max_f \sum_{\tau} S_{\hat{x}}(f, \tau) + C''(f)
    \]
Salience-based approaches

- Overview
Source separation approaches

- Overview

Source Separation
- Source/filter model
- Harmonic/percussive separation
- Repeating structure separation

Estimated model parameters

Melody

Analysis

$\rightarrow f_0$
Salience based approaches
Saliency-based approaches

- **Pre-processing**
  - Use of a band-pass-pass filter (Goto2004)

- Use of an equal-loudness filter (Salomon2012)
  - 10th order infinite impulse response (IIR) filter cascaded with a 2nd order Butterworth high pass filter, (Robinson,2013)

---


Salience-based approaches

- **Spectral Transform**
  - Typically based on Short-time Fourier Transform
  - Others: CQT (Cancela2004), Multi-resolution FFT (Dressler2006), Exploitation of Perceptual principles (Richard2013)


Salience-based approaches

- **Spectral Peak Processing**
  - **Objective:**
    - Removing peaks which are not related to the lead voice (for ex. based on sinusoidality criteria [Goto2004])
    - Or Reducing the influence of timbre (spectral envelope whitening, see for exemple [Cancela2006])
    - Or Computing instaneous frequencies (see for example Dressler2006)

---


Salience-based approaches

**Salience Function**

- Provides an estimate of the salience of each possible pitch value over time
- Many approaches …
  - Obtained as a weighted sum of the amplitude of harmonic frequencies …
  - Use of tone models (Goto2004, Marolt2004)
  - Use of summary autocorrelation (Paiva2006),…

Salience-based approaches

- **Salience Function**
  - Presence of ghost notes …
    - Use of « tricks » to reduce these « octave errors »
      - Ex: spectral smoothness,
    - Most errors are practically removed by the tracking stage

Salience-based approaches

- **Melody selection and tracking**
  - Tracking using:
    - Clustering (Marolt2004),
    - heuristic-based tracking agents (Dressler2006, Goto2004),
    - HMM (Ryynanen, Yeh2012), or
    - Dynamic Programming (Rao2010, Hsu2010)
Salience-based approaches

- **Voicing detection**
  - **Aim:**
    - To determine when the melody is present.
    - Usually applied at the very end…
    - Possible approaches: using a pre-frame fixed or dynamic threshold on the salience function
    - Using a « silence » state in HMM based approaches
Source separation based approaches
Source separation approaches

Numerous strategies exist:

- Exploiting prior information of the singing voice component (e.g. a source/filter model) [Durrieu2010]
- Exploiting Harmonic / Percussive separation (singing voice is a temporally variable harmonic component) [Ono2010]
- Exploiting the repeating structure of the background (and the on-repeating nature of the singing voice component) [Rafii2013, Liutkus2012]
An example of singing voice separation Source using Non-Negative Matrix factorization

- General principle:
  - The sources are recovered by filtering the mixtures

\[
\hat{s} \quad = \quad \mathcal{F} \quad \begin{cases} x \quad , \quad \Theta \end{cases} \quad \text{sources} \quad \text{filtering technique} \quad \text{mixtures} \quad \text{parameters}
\]
A popular model for audio source separation: NMF

- NMF = Non-negative Matrix Factorization

\[ WH \approx V \]

Original spectrogram

"Templates or Atoms"

"Activations"
A popular model for audio source separation: NMF

- How the template matrix W and activation matrix H are obtained [Lee&al. 1999]?

- Minimization of

\[ D(V|\hat{V} = WH) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn}) \]

- Typical distances and divergences used:

| Euclidean          | \( d_{EUC}(a|b) = (a - b)^2 \) |
|--------------------|----------------------------------|
| Kullback-Leibler divergence | \( d_{KL}(a|b) = a \log\left(\frac{a}{b}\right) - a + b \) |
| Itakura-Saito divergence      | \( d_{IS}(a|b) = \frac{a}{b} - \log\left(\frac{a}{b}\right) - 1 \) |

\( \beta \)-divergence

\[ d_{\beta}(a|b) = \begin{cases} \frac{1}{\beta(\beta-1)} (a^{\beta} + (\beta - 1)b^{\beta} - \beta ab^{\beta-1}) & \beta \in \mathbb{R} \setminus \{0, 1\} \\ a \log \frac{a}{b} + (b - a) & \beta = 1 \\ \frac{a}{b} - \log \frac{a}{b} - 1 & \beta = 0 \end{cases} \]
A popular model for audio source separation: NMF

How the template matrix $W$ and activation matrix $H$ are obtained [Lee&al. 1999]?

- In general, the cost function is not convex in $(W,H)$. However, it is separately convex in $W$ and $H$ (for Euclidean and Kullback-Leibler divergence).

- The solution is iteratively obtained by means of multiplicative update rules:

  For example with the Euclidean distance:

  \[
  \begin{align*}
  H &\leftarrow H \otimes \frac{W^T V}{W^T(WH)} \\
  W &\leftarrow W \otimes \frac{V H^T}{(WH)H^T}
  \end{align*}
  \]
A popular model for audio source separation: NMF

- NMF does not necessarily provide a semantically meaningful decomposition in the absence of “constraints”

Templates correspond to musical notes

- Templates are built from half of each note and are less semantically meaningful
- Activations are less sparse
- Templates grouping for source recovery
An example of model-based constraints for main melody separation using NMF

- **The model:** \( \text{Audio} = \text{Voice} + \text{Music} \)
  - The voice \( \text{Voice} \) follows a source filter production model: \( \text{Voice} = \text{Source} \times \text{Filter} \)
  - Each component (Voice and Music) is represented by separate NMF

\[
S_{\text{Audio}} = (W^{F_0} H^{F_0}) \odot (W^{\phi} H^{\phi}) + (W^M H^M)
\]

- Spectrogram of the input audio signal
- Spectrogram of the singing voice
- Spectrogram of the background music
An example of model-based constraints for main melody separation using NMF

Illustration of the source/filter model with NMF

An example of model-based constraints for main melody separation using NMF

- **Exemple of Blind leading voice extraction** [Durrieu&al.2011]

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Backgrounds</th>
<th>Leading voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singing voice</td>
<td><img src="1" alt="Sound" /></td>
<td><img src="2" alt="Sound" /></td>
<td><img src="3" alt="Sound" /></td>
</tr>
<tr>
<td>Trumpet</td>
<td><img src="1" alt="Sound" /></td>
<td><img src="2" alt="Sound" /></td>
<td><img src="3" alt="Sound" /></td>
</tr>
</tbody>
</table>

Evaluation
Evaluation: several measures

- **Voicing recall rate**: Proportion of frames labeled as melody frames in the ground truth and that are estimated as melody frames by the algorithm.

- **Voicing false alarm rate**: Proportion of frames labeled as non-melody in the ground truth and that are estimated as melody frames by the algorithm.
Evaluation: several measures

- **Raw pitch accuracy**: Proportion of melody frames in the ground truth for which the pitch estimation is considered correct (i.e. within half a semi-tone)

- **Raw chroma accuracy**: same as raw pitch accuracy but without counting octave errors

- **Overall accuracy**: combines pitch accuracy and voicing detection accuracy
Evaluation: are we improving?

- Evolution of the best overall accuracy result over the years (on 6 MIREX collections)
Alternatives approaches

• Using machine learning approaches (Poliner)
• Using the repetitive structure of the music (and non-repeating structure of singing voice) [Rafii2013, Liutkus2012]

• Combining source separation (SS) and salience-based (SB) approaches:
  − SB can bring prior information for SS based approaches
  − And SS can bring a « lead voice enhanced » spectrogram for SB approaches
  − Towards informed methods..

Informed Source Separation
Towards Informed melody extraction …

- Significant performance gain can be obtained by using better prior information....

- **Informed melody extraction**
  - Side information is transmitted to the extraction module
  - Extraction is done using the mixture and the side information

---


Informed melody extraction

For example: User-guided melody extraction
User-guided source separation

Main approaches

Example-based approaches

- User-guided source separation
  - User examples
  - Example-guided source separation
  - Estimated sources

Annotation-based approaches

- Annotation-guided source separation
  - User annotations
  - Estimated sources

Mixture

Feedback
User-guided source separation
Interactive time-pitch annotation-informed separation

- The user paints the parts corresponding to the melody in the GUI

- Algorithm is re-run but with many zero values in the initial decomposition for the melody part

- Several iterations are possible

User-guided source separation

Example-based approaches

Separation by humming

- Demonstration video [Smaragdis & Mysore 2009]


http://www.cs.illinois.edu/~paris/demos/al/user-guide.mp4
Conclusion

- Steady improvement in melody extraction in the last decade…
- Mainly targeted to singing voice melodies …

**Challenges:**

- Going from Singing voice to instrument music
- Target higher polyphony (5+ music sources)
- Target songs with backing vocals
- Improving the voicing detection
- Public access to larger annotated databases …
Additional References