CBMI 2024

21st International Conference on Content-based Multimedia Indexing,

September 18-20, Reykjavik, Iceland

Invariant Audio Prints for Music Indexing and Alignment

Rémi Mignot[1], Geoffroy Peeters[2]

[1] STMS Lab – IRCAM, Sorbonne Université, CNRS (UMR-9912), Paris, France

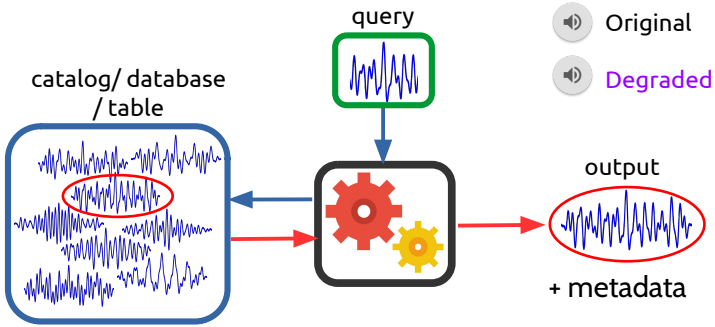
[2] LTCI - Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

test

Introduction

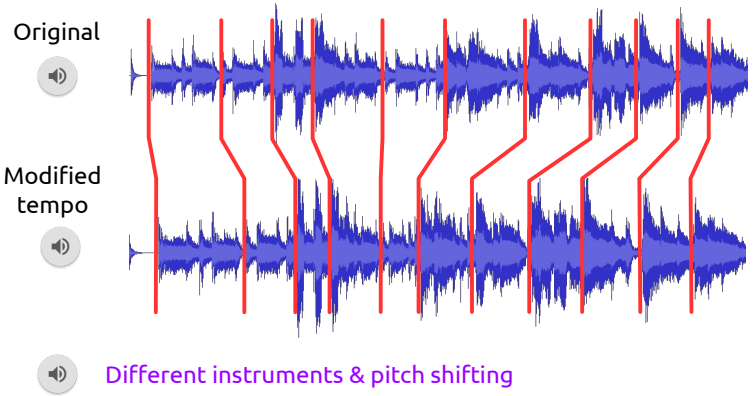
- Audio ***Indexing***

  *Find the "**reference song**" from a **music catalog** based on the signal content of a given **audio excerpt***



- Audio-to-audio ***Alignment***

  *Search the **time mapping** between **two occurrences** of the same music (covers e.g.)*



- <u>Robustness</u>: find a solution which still works when the query is ***transformed / degraded***

  → *time stretching, pitch shifting, noise addition, distortion, audio effects, and different instruments (for alignment)*

- <u>Remark</u>: we use the same approach for both tasks.

Method overview

- Process chain

  Derivation of codes that are: ✓ ***robust to transformations / degradations*** and

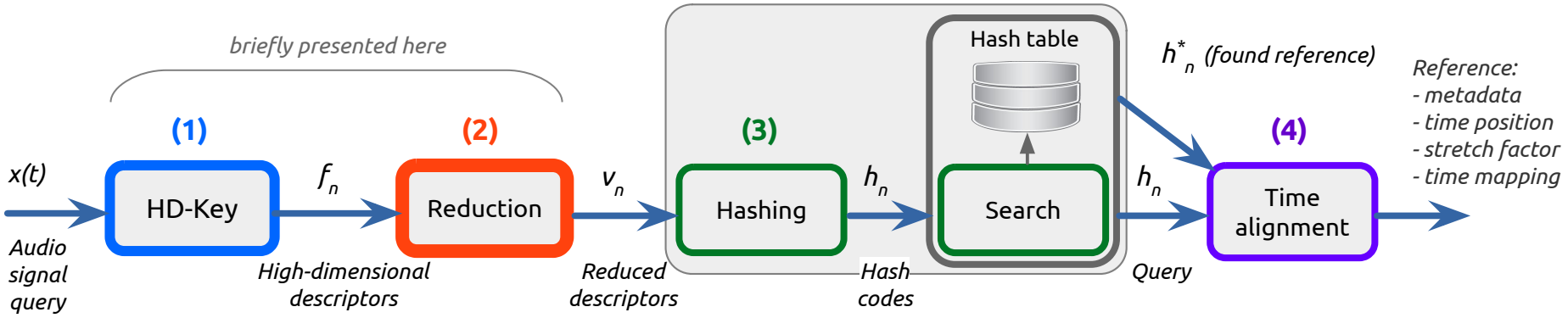
  ✓ ***relevant to the musical content*** (unlike spectrogram *peak-pairs* methods)

(1) High-dimensional audio keys (1056) → *design of **audio descriptors** robust to some transformations,*

(2) Dimension reduction (40) → ***learning** of a linear projection robust to degradations,*

(3) Hashing → *hash codes tolerant to bit corruption (LSH-based),*

(4) Time alignment → *DTW-based alignment to estimate the time mapping.*

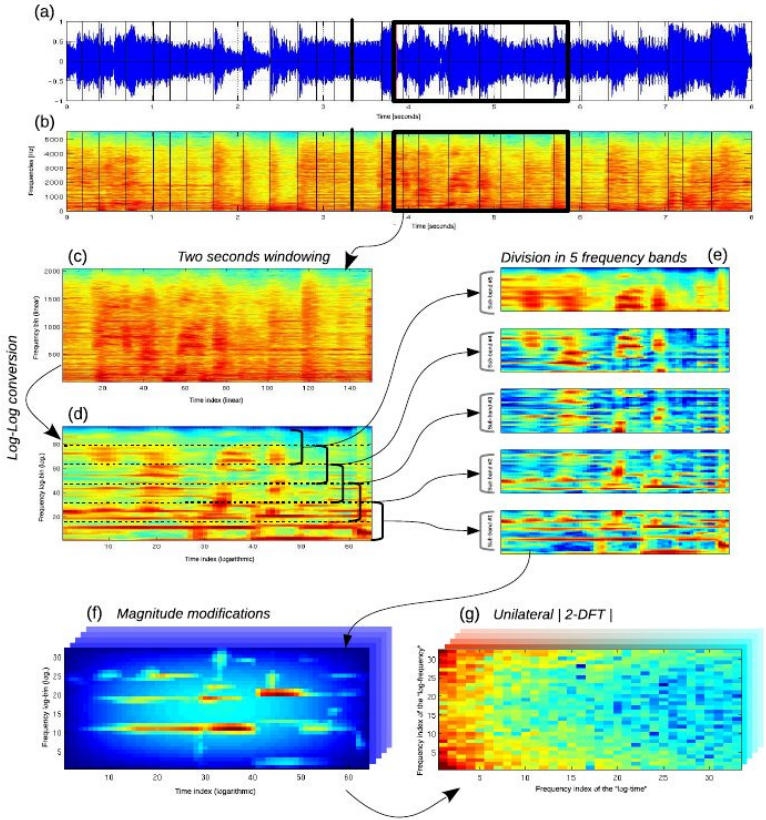


(1) High-dimensional Audio Keys

- Audio descriptors
  - ✔ *relevant to the musical content*
    → *inspired by audio classification (modulation spectrum).*
  - ✔ *robust to transformations* ***by design*** :
    - Manipulations of sub-spectrograms
      - *Log. scale* of *frequencies* and *time* (d),
      - Frequency *band* splitting (e),
      - *Amplitude transformation* (f)
      - Magnitude of 2D-*Discrete Fourier Transform* (g).
    - Based on properties of:
      - Logarithmic function,
      - Shift invariance of | DFT |,
      - Amplitude change,

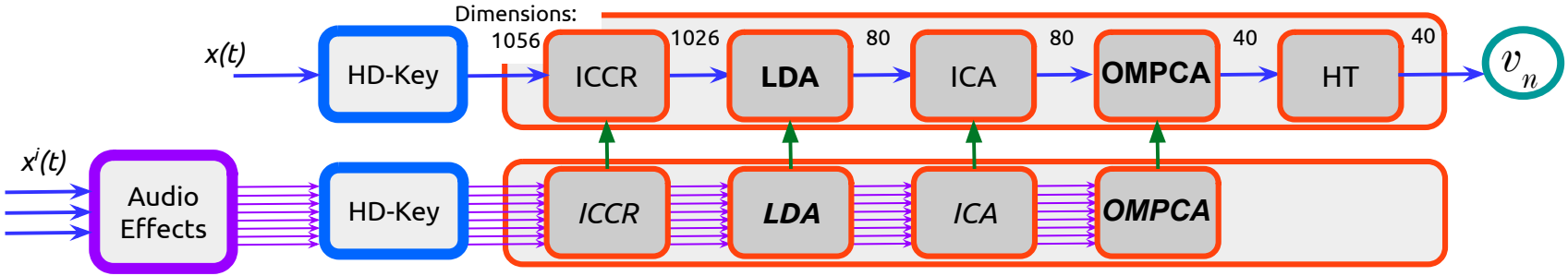
      → The descriptors are *robust by design* to:
      Pitch and time changes, and noise, filtering,

→ dimension 1056…



(2) Robust dimensional reduction

- Reduction of the dimensions
  - Chain of ***linear transformations*** :
    - *Discriminant* analyses, or *Independent Component* Analyses, and O*rthogonal* projections,
  - Dimension reduction 1056 → 40,
    with output variables v_n with **properties**:
    - *centered*, *normalized*, and ***mutually uncorrelated***,
    - ***robust*** to transformations/degradations, and
    - ***discriminant*** to the original signal.
  - Training dataset:
    → many *transformed versions* of music examples (~*data augmentation*).



Two experiments *(see the paper)*

- *Segmentation* and *indexation* of **music medleys**
  → with time and pitch changes + degradations,
  and within a reference catalog of ~40 000 songs.

- *Audio-to-audio time alignment* of synthesized **MIDI covers**
  → with time varying tempo, pitch shifting,
  changed instruments and removed drums.
  → use of local distances computed for the derived audio codes.

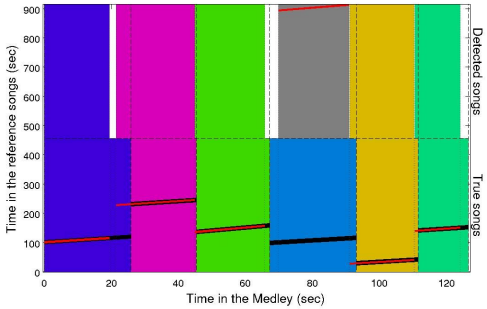
➢ ***See the paper for all the quantitative results***

- Overall conclusions:

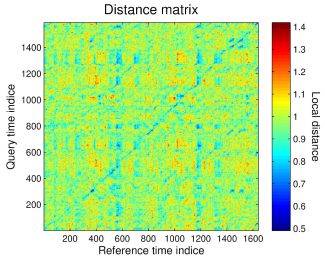
  ✔ Audio indexing:
  not as good as other approaches for some degradations (e.g. noise),
  but still robust, especially for pitch and time changes.

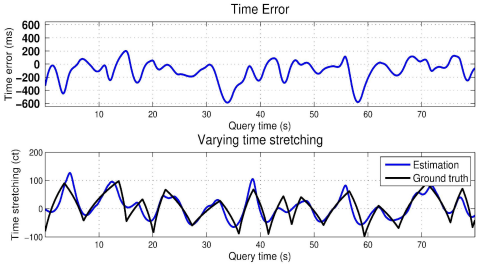
  ✔ Audio-to-audio time alignment:
  The results prove that the derived audio codes are quite
  robust to transformations and
  they are representative to the musical content,
  even with different instruments.

Original medley
Transformed medley
+ realigned medley (right channel)



Original synthesized MIDI song
Transformed MIDI song
+ realigned MIDI (right channel)





Bonus experiment

- Time alignment of an acoustic guitar cover of *Little Wing* (Jimi Hendrix).

Original recording
“Little Wing” (Jimi Hendrix)

Tempo: ~70 BPM

Acoustic guitar + voice cover
“Little Wing” by Corey Heuvel

Tempo: ~60 BPM
with accelerations/decelerations,
some longer transitions,
quite different scores
but respected structure (at the beginning)

Bonus experiment

- Time alignment of an acoustic guitar cover of *Little Wing* (Jimi Hendrix).

1. Processing: Time alignment between the two recordings based on the derived audio prints and DTW.
2. Realignment: The original recording of Hendrix is then realigned to the cover, and inserted into the video.

Remark: longer transitions, e.g. between the 2nd verse and the solo, at 1:40.
→ *the original is strongly stretched*

left channel: unchanged cover sound
right channel: synchronized original recording

all channels: synchronized original recording

THANK YOU

For more details:

- Questions ?
- Read the paper #107:
  - *Rémi Mignot & Geoffroy Peeters,*
  - *“Invariant Audio Prints for Music Indexing and Alignment”*
- See the poster this afternoon.