Presenter: Anders Skovsgaard

Audio Identification using Sinusoidal Modeling and Application to Jingle Detection

Michaël Betser **Patrice Collen** **Jean-Bernard Rault**

France Télécom R&D
4 rue du clos courtel, 35510 Cesson-Sévigné, France
e-mail: firstname.lastname@orange-ftgroup.com

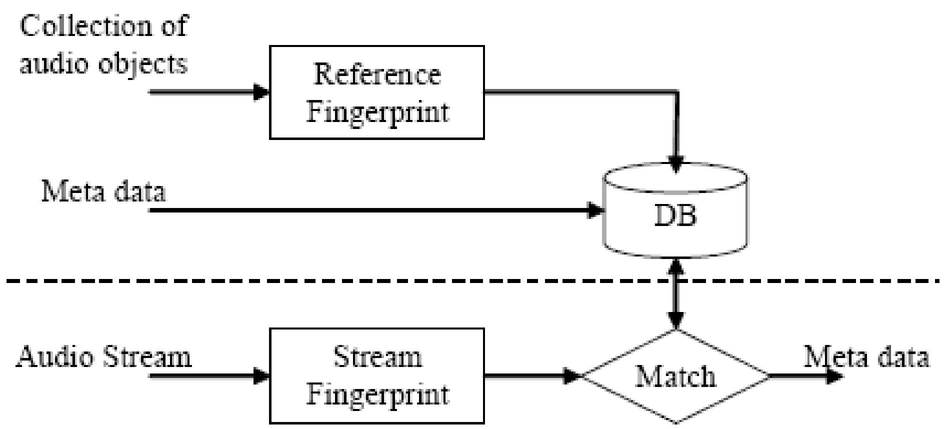
Outline

- Motivation
- Classical Fingerprint systems
  - Classical analysis scheme
  - Fourier transformation
  - The limitations of Classical Fingerprint
- The proposed solution: Sinusoidal Fingerprint
  - Four step model
  - Fingerprint comparison
  - Jingle detection
- Related work
- Evaluation

Motivation

- Music with additional speech is hard to recognize.
- Most audio identification systems aim at real music not e.g. radio.
- Detecting noisy jingles from radio stations.

Classical analysis scheme

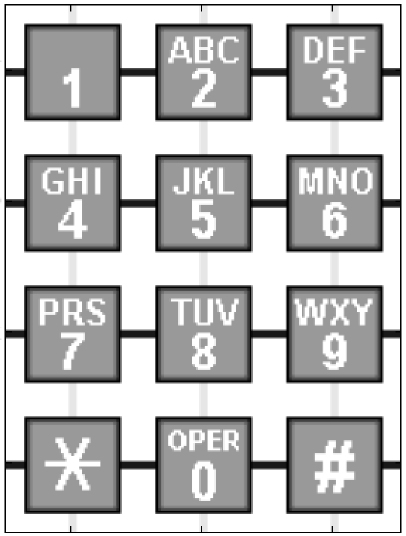


Often no distinction between Stream Fingerprint and Reference Fingerprint generation

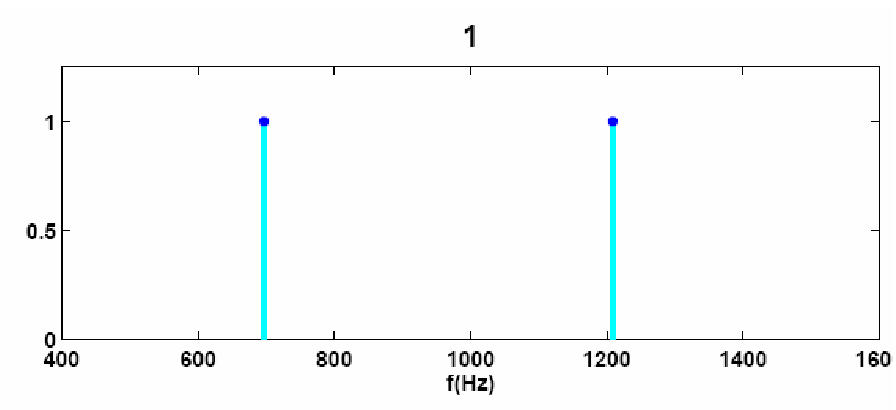
Fourier transformation

- Used by the classical an the proposed solution.
- Separates a waveform into sinusoids of different frequency.

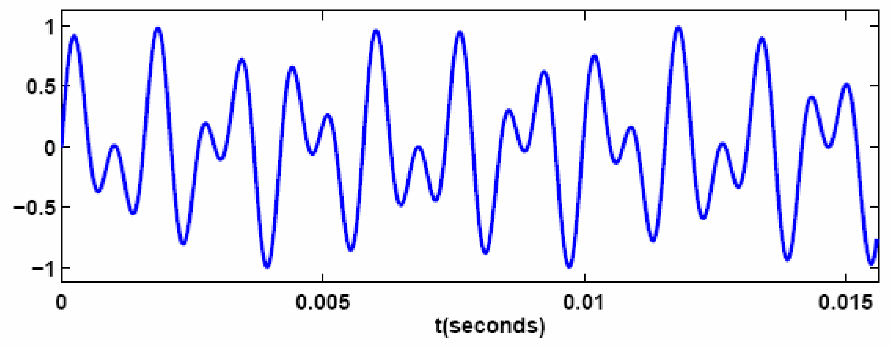
- Simple example:



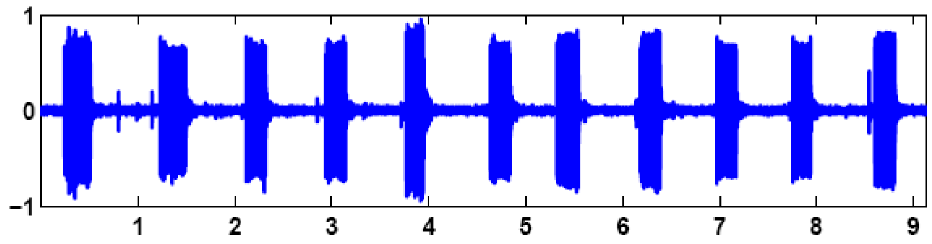
Fourier transformation (example)



Frequencies generated by “1” button

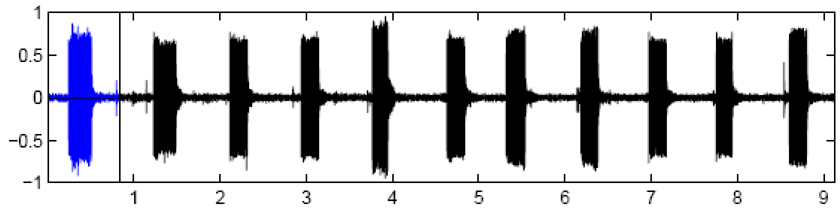


The signal obtained by averaging the sine with the frequencies.

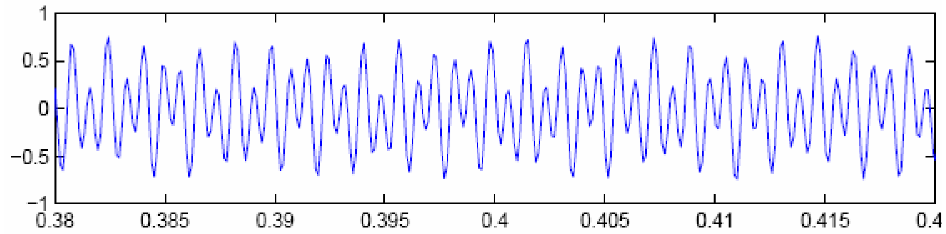


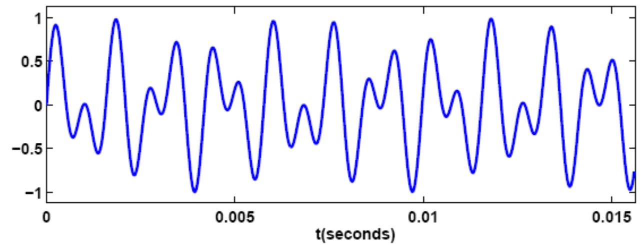
Recording of 11-digit number. Notice the noise between the numbers.

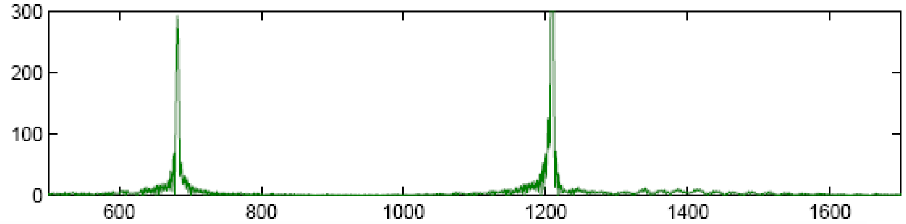
Fourier transformation (example)

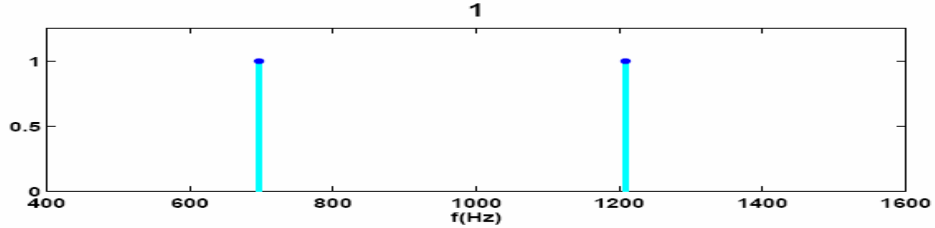


Isolating one number.









After Fourier transformation

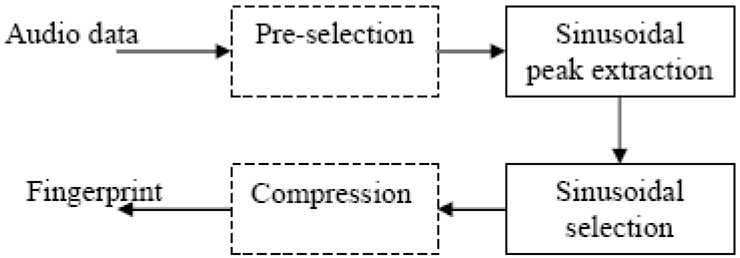
The limitations of Classical Fingerprint

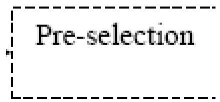
- The paper claims that only the predominant sinusoidal components should be used. (based on experiments)
- Existing systems only partially take this into account.

The proposed solution:

Sinusoidal Fingerprint

- Four step model

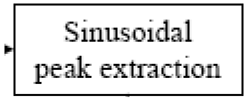




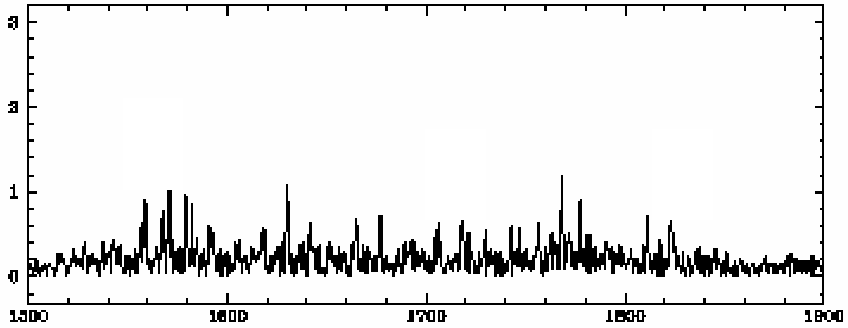
Low-pass filter – cut at 4 kHz

Only keep the frequency (discards e.g. amplitude) coded in 16 bits for each peak selected.

Four step model

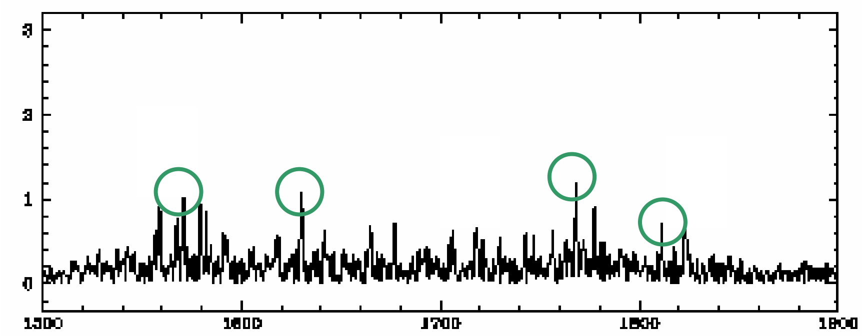


Decompose to sinusoids (Fourier) with a set of parameters used in "peak selection" (including amplitude, phase and frequency).
Frequency spectrum:



Four step model

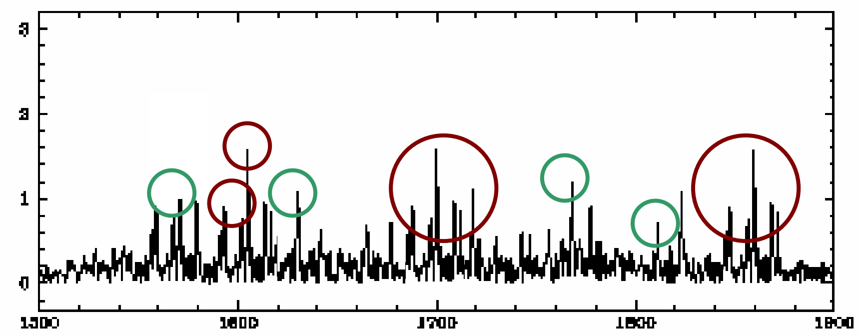
Select the predominant and stable peaks.
The “stream peak selection” should contain more peaks than “reference peak selection” (maybe strong noise)



Fingerprint comparison

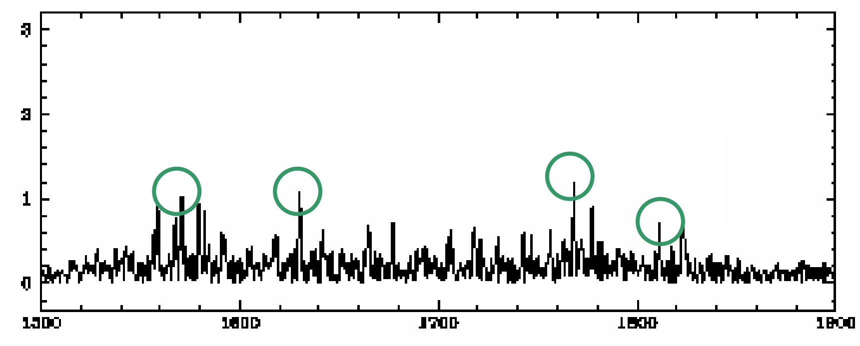
Check frame by frame for each reference audio if there is a frequency match

Stream audio



Noise
(e.g. speech)

Reference audio



Jingle detection

	AM	AM+MP3	AM+SP
Sinusoidal	97	95	83
HKO	89	85	67

Occurrence recall comparison in percent

	AM	AM+MP3	AM+SP
Sinusoidal	79	68	53
HKO	60	57	34

Duration recall comparison in percent

Occurrence of a block of 1 second.

Duration not as good as occurrence because e.g. a block with speech is not recognizes. Also shorter jingles is limiting.

Related work

- We use a "codebook" of frequencies. It is calculated by clustering frequencies of sample data (e.g. 20 songs recorded from microphone).
- A vector with 16 frequencies representing 62.5 ms is created and represented by a symbol from BASE64.
- Result:
  mmmTTcbJ0008ipiNvG33TTTCCCCTTT333
- The database problem:
  Find best similar substring on-the-fly.
  (e.g.. mATmbJ00)

Evaluation

- Bad parts
  - Missing details (3.2. “with a set of parameters including”)
  - Claim that their solution is the best based on one experimental article (2.2.)
  - Suspect it to be extremely slow when comparing. Stream fingerprint has huge overhead of peaks in order to work with random noise.
  - Many “magic” numbers. (3.3. “M superior to a hundred”, “Q should be greater”, 5.2 “Tf should be slightly higher”)

Evaluation

- Good parts
  - They have implemented and tested it in the real world.
  - Clear idea of the paper.
  - Many references to related work.

Presenter: Anders Skovsgaard

A Review of Algorithms for Audio Fingerprinting

Pedro Cano and Eloi Batlle
Universitat Pompeu Fabra
Barcelona, Spain
Email: {pedro.cano, eloi.batlle}@iua.upf.es

Ton Kalker and Jaap Haitsma
Philips Research Eindhoven
Eindhoven, The Netherlands
Email: ton.kalker@ieee.org, jaap.haitsma@philips.com

Outline

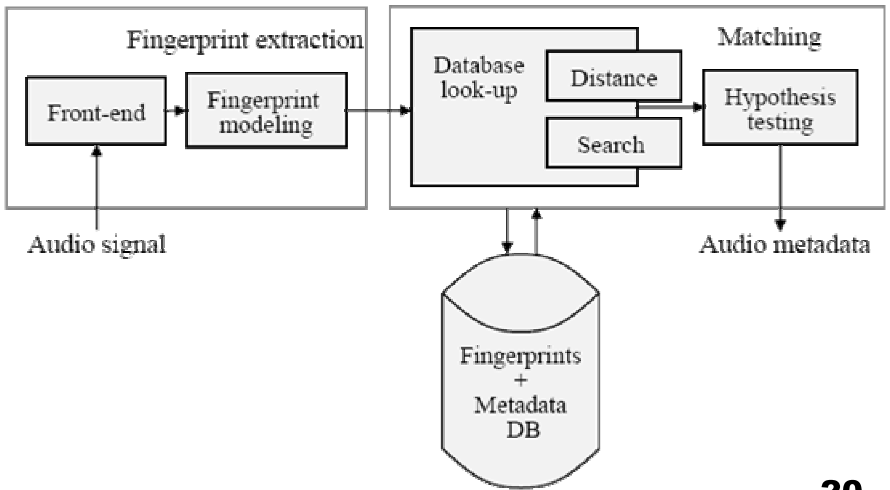
- Motivation
- General Fingerprint Framework
  - Bit matching vs. Content-based Audio Identification
  - Front-End of the Framework
  - Fingerprint Models
  - Searching
- Evaluation

Motivation

- Several ways of recognizing audio and generating fingerprint.
- Provide an overview of the different techniques.

General Fingerprint Framework

- Bit matching
  - E.g. hash methods (MD5). Efficient but extremely fragile.
  - Works only with the bits – not content.
- Content-Based Audio Identification
  - Works at the audio level.
  - Robust to random noise

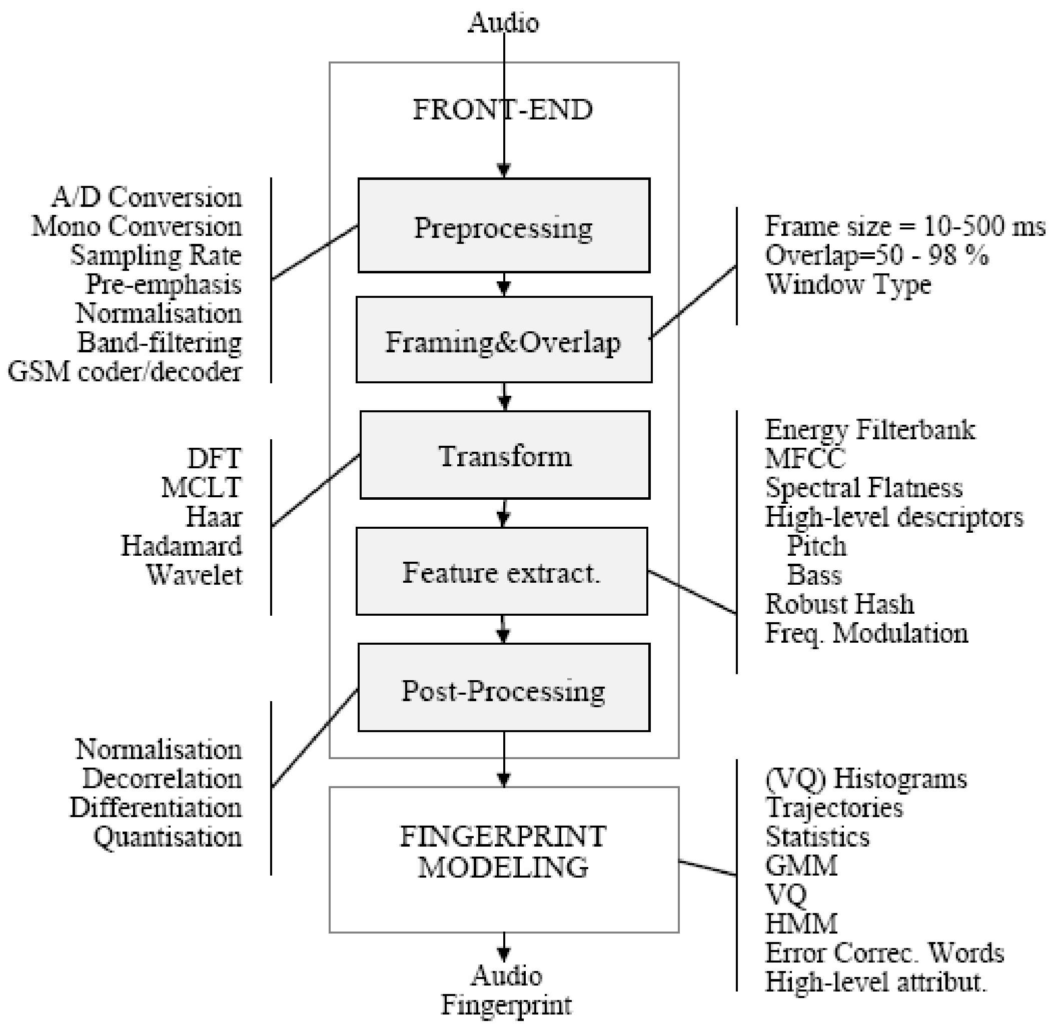


Front-End of the Framework

Preprocessing:
Digitalize
Simulate the channel
GSM coder/decoder

Framing&Overlap:
Divide into frames where signal is stationary.
Overlapping if frame size is larger than variation velocity.

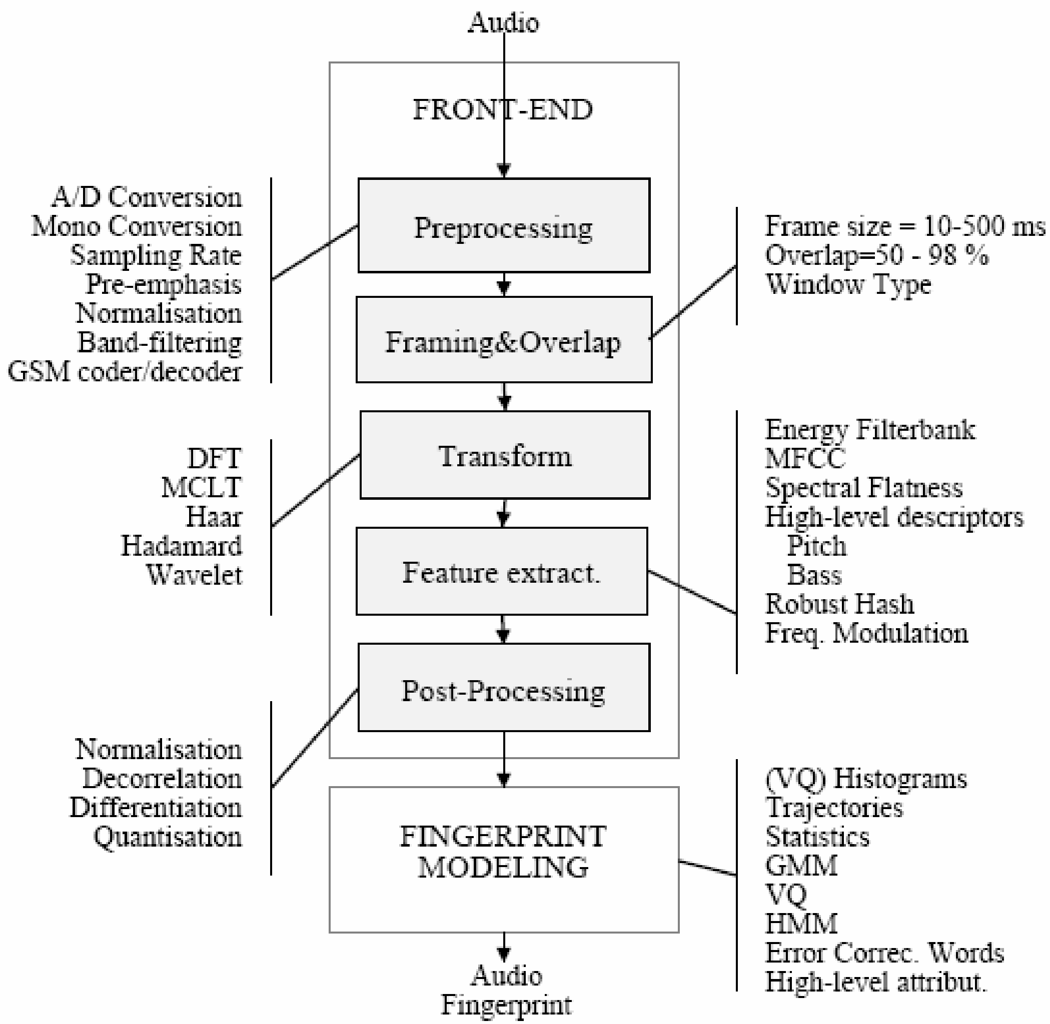
Transform:
Transform to frequency domain.



Front-End of the Framework

Feature extraction:
Divide into bands and extract most meaningful with regard to the human auditory system.

Post-Processing:
Reduce memory requirements, remove distortions



Fingerprint Models

- Fingerprint can be based on the complete or partial lengths of the song.
- Remove redundancies (vectors with same frequencies).
- Use average frequency spectrum, beat per minute.
- Compacting a sequence of vectors to a single mean vector.

Searching

- Using distance techniques. E.g. Hamming distance:

  1011101 and **1001001** is 2.

  2143896 and **2233796** is 3.

- Spatial Access Methods (multidimensional vectors).

Evaluation

- Bad parts
  - Several misspellings (“distorions”, “fingeprint”, “and son on”, “represention”).
  - Many concepts introduced in short article = superficial and assumes comprehensive DSP knowledge.

Evaluation

- Good parts
  - Covers many different techniques.
  - Framework is clear (figures) and the descriptions comes in natural order.