



Audio identification using sinusoidal modeling and application to jingle detection



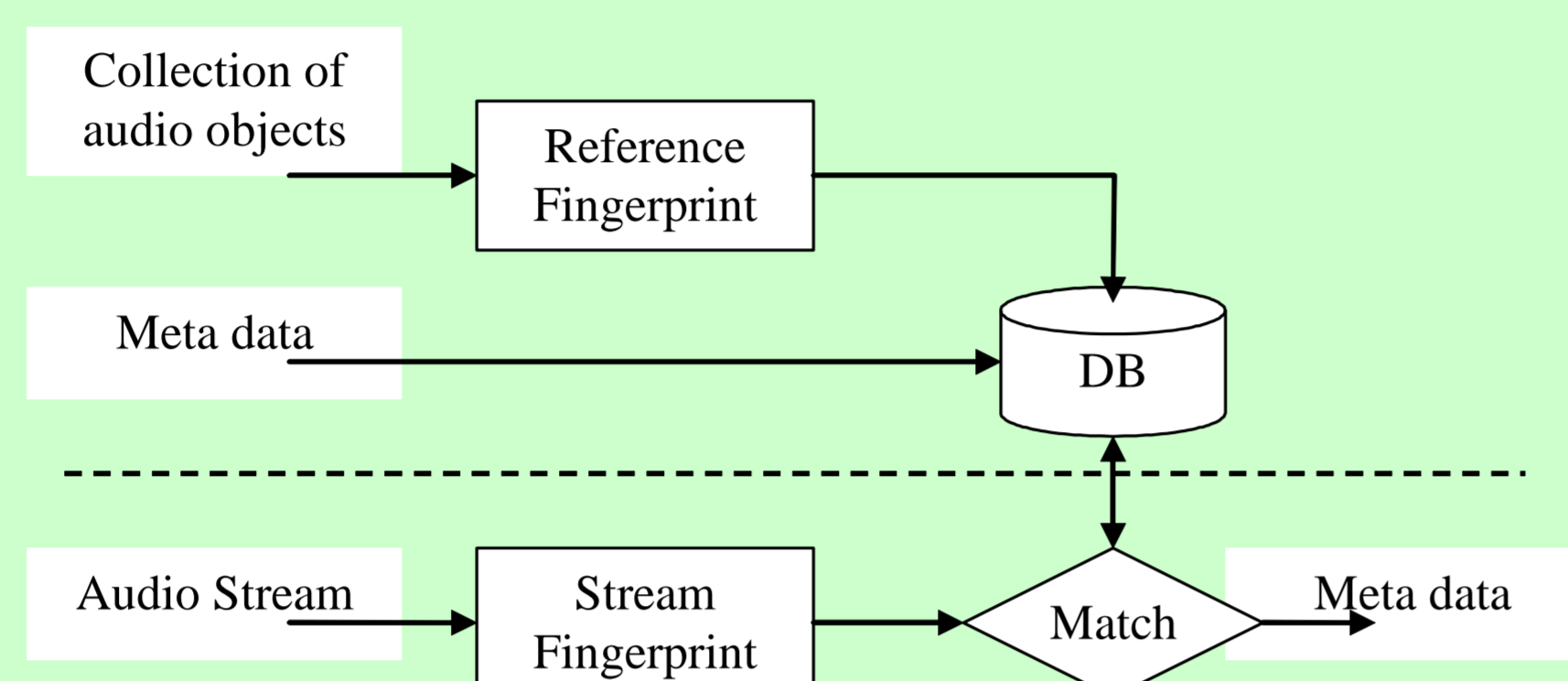
Michaël Betser, Patrice Collen, Jean-Bernard Rault
France Télécom R&D, Rennes

For further information, please contact: michael.betser@orange-ft.com.

Abstract

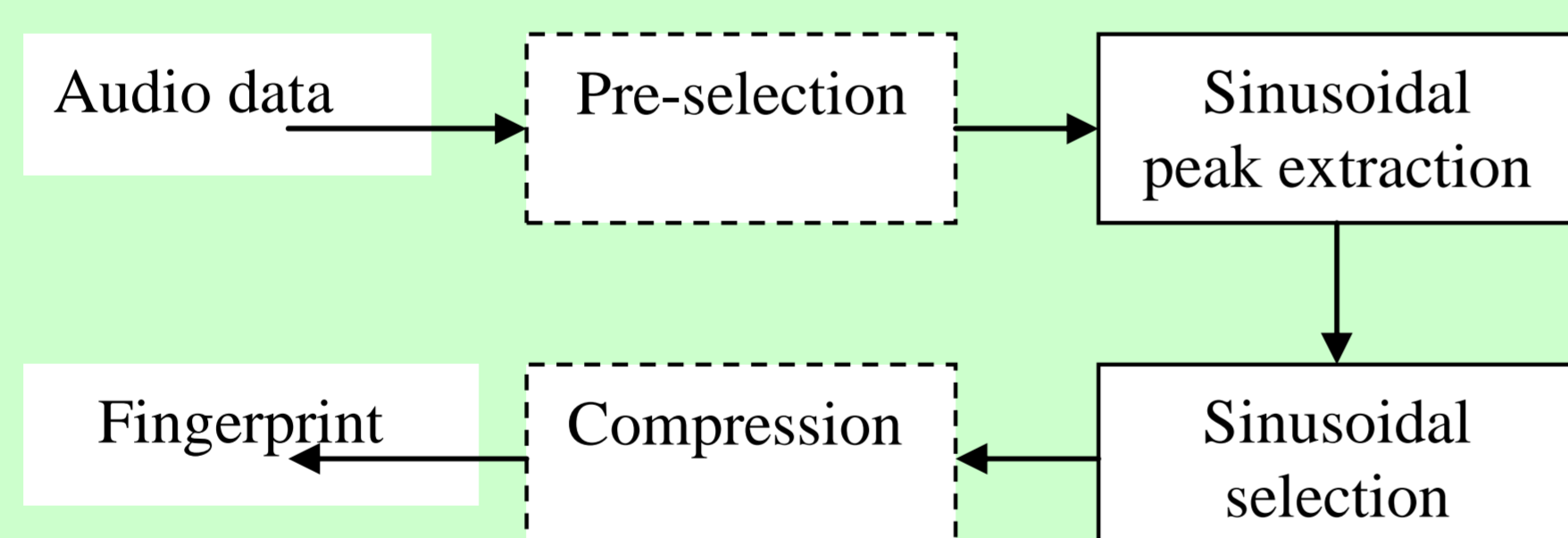
This article presents a new descriptor dedicated to Audio Identification (audioID), based on sinusoidal modeling. The core idea is an appropriate selection of the sinusoidal components of the signal to be detected. This new descriptor is robust against usual distortions found in audioID tasks. It has several advantages compared to classical subband-based descriptors including an increased robustness to additive noise, especially non-random noise such as additional speech, and a robust detection of short audio events. This descriptor is compared to a classical subband-based feature for a jingle detection task on broadcast radio. It is shown that the new introduced descriptor greatly improves the performance in terms of recall/precision.

Fingerprint system



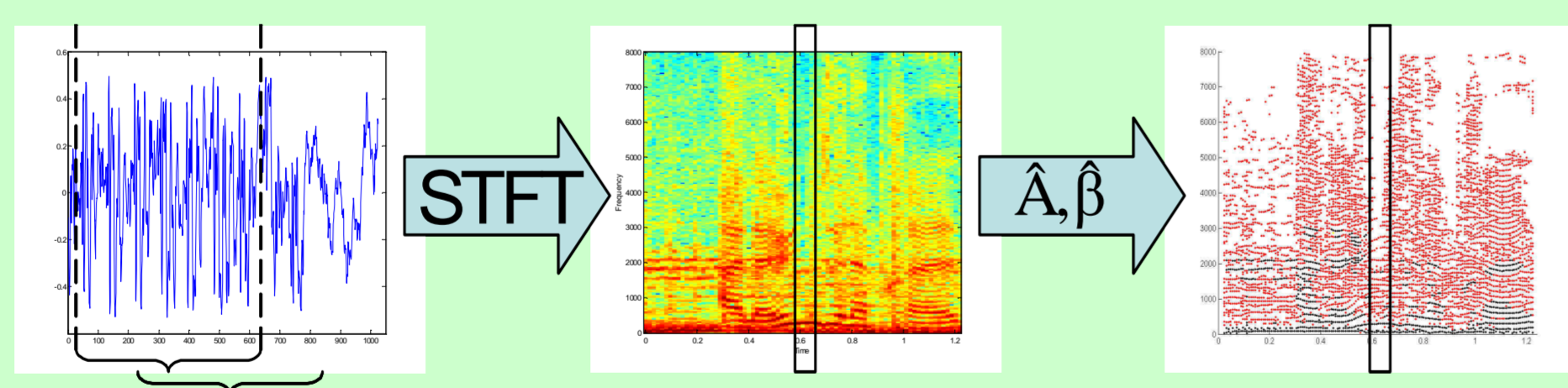
- Applications:
 - research of information about a document,
 - document detection for structuring purposes,
 - document detection for broadcast control
- Possible alterations:
 - additive noise
 - canal distortions
 - speed change
 - equalization, sub-sampling
 - compression
 - system granularity
- Limitations of subband based Fingerprint system:
 - most of the information is carried by the predominant sinusoidal components.
 - use of low energy portions of the audio signal.

Sinusoidal fingerprint creation



- Preprocessing : band-filtering (300-4000Hz), spectral whitening

- Sinusoidal analysis:



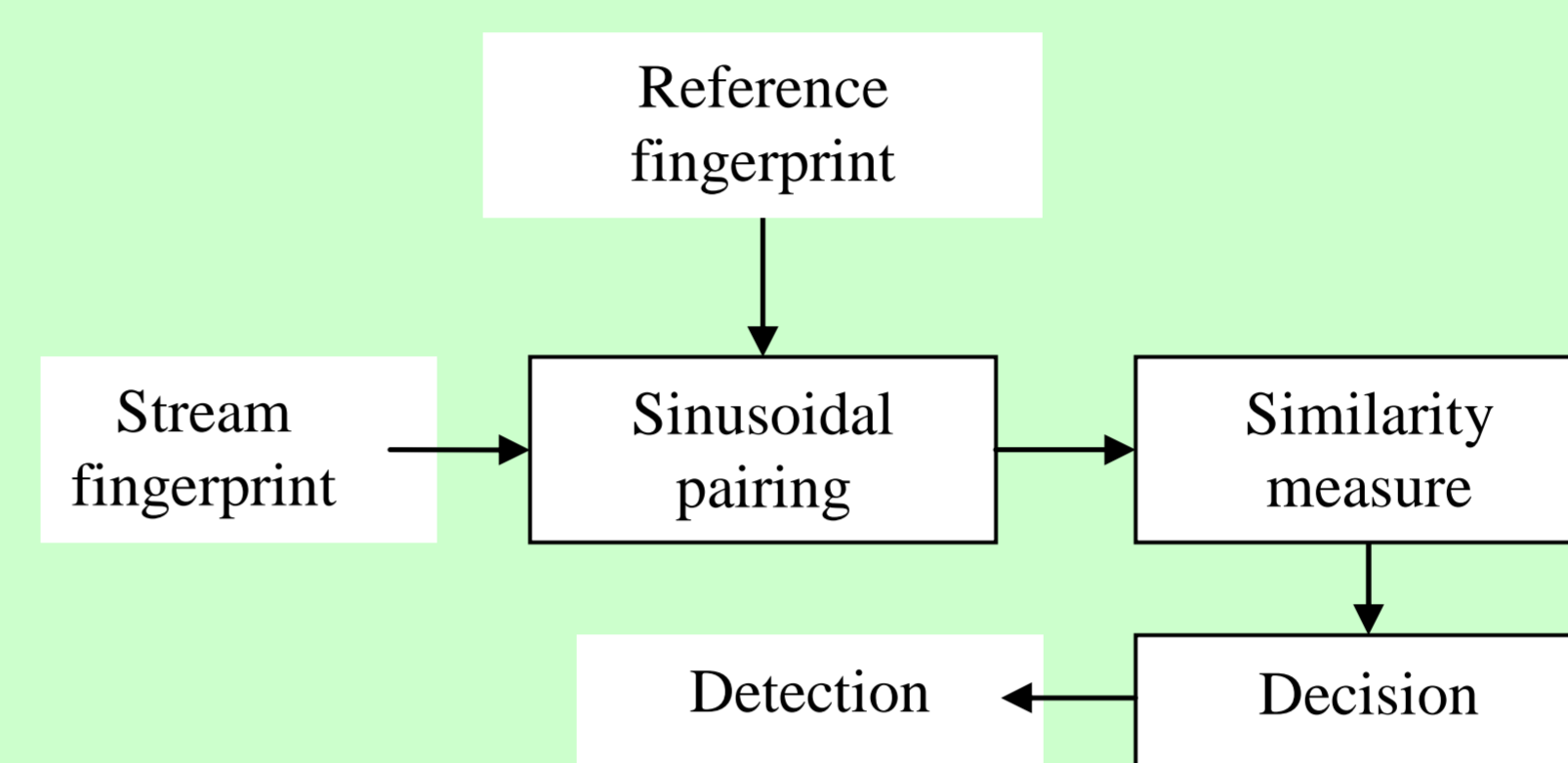
- Reference sinusoidal selection :
 - stability relatively to the frequency estimation method
 - selection of the K most energetic sinusoids per frame (in average)
 - selection is done on mid-term signal blocs (~30 s)

- Stream sinusoidal selection :
 - Q most energetic sinusoids per frame ($Q \gg K$)

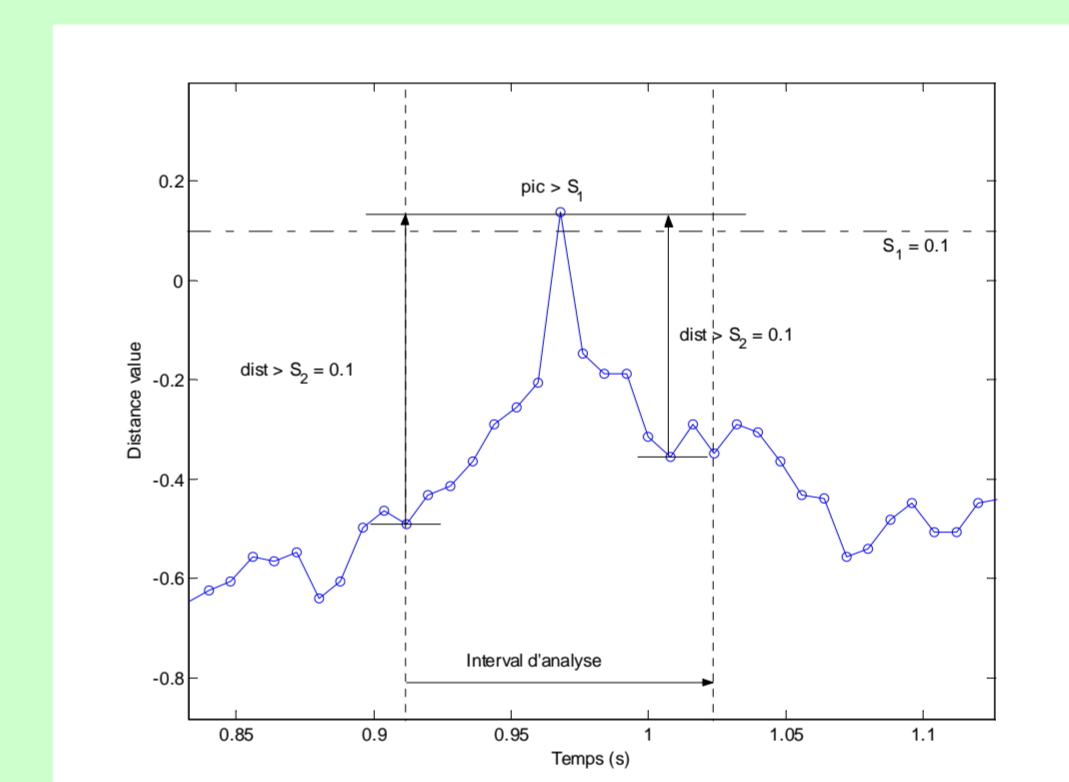
- Remarks:
 - Stream peak selection can be simpler
 - More peaks are kept to take into account additive sinusoidal sounds and equalization

- Compression:
 - Keeping frequency only
 - 16 bit frequency coding (1Hz precision)

Sinusoidal fingerprint comparison



- Comparison is done by block:
 - 1 block of size M of the stream
 - all possible blocks of size M in the references
- Sinusoidal pairing:
 - pairing is done frame by frame
 - each sinusoid of the reference is paired with a sinusoid of the stream using a frequency tolerance of a few Hz
- Similarity measure:
 - number of reference peaks correctly found
 - normalization by the number of peak per second in the reference
- Decision:
 - Looking for sharp peaks in the similarity curves
 - Use of two thresholds: absolute and relative



Evaluation

- 30 France Info jingles detection task
- Corpus: 18 hours of France info + 48 hours of other radios
- Alteration: AM, AM+MP3, AM+Speech
- Comparison with subband-based algo (Haitsma et al. 2001)
- Detection based on one second blocks
- Results:
 - no false alarm
 - occurrence recall

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

- duration recall

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

Conclusion

- Better modeling of the signal
- Reliable recognition of 1s sounds
- Robust to additive sinusoidal noise, increased robustness to other noises