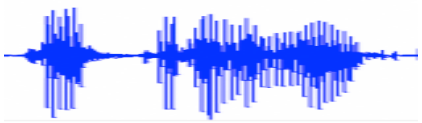
MASK: Robust Local Features for Audio Fingerprinting

Xavier Anguera, Antonio Garzón and Tomasz Adamek

Telefonica Research

Outline

- What is audio fingerprinting
- MASK proposal
- Experiments
- Conclusions

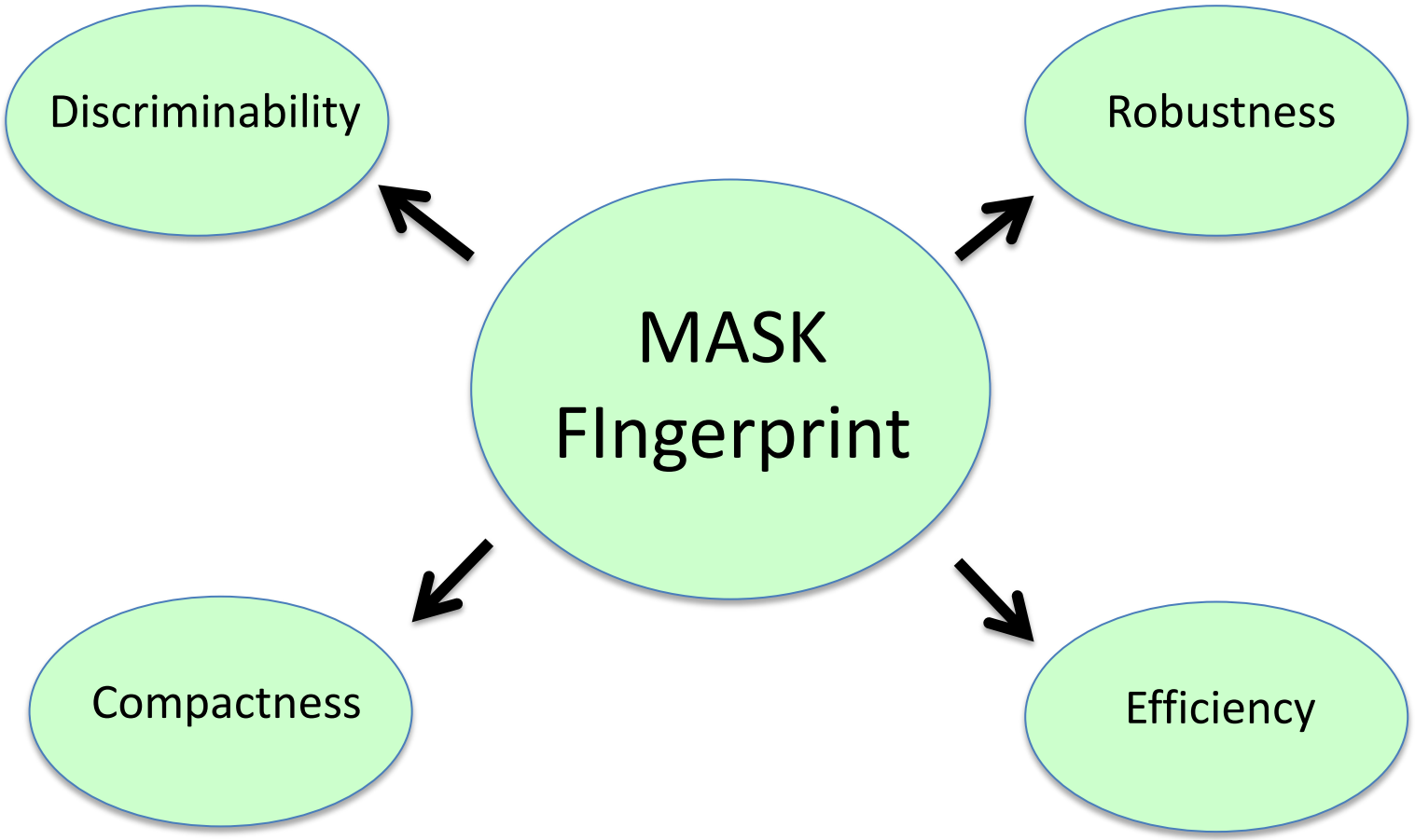


?

watermarking

fingerprinting

What makes a good audio fingerprint?



MASK == **M**asked **A**udio **S**pectral **K**eypoints

Considered prior art

- Avery Wang, “An industrial strength audio search algorithm,” in Proc. ISMIR, 2003.

- Jaap Haitsma and Antonius Kalker, “A highly robust audio fingerprinting system,” in Proc. ISMIR, 2002.

- Shumeet Baluja and Michele Covell, “Waveprint: Efficient wavelet-based audio fingerprinting”, Proc. Pattern Recognition 41 (2008)

Time-to-frequency transformation

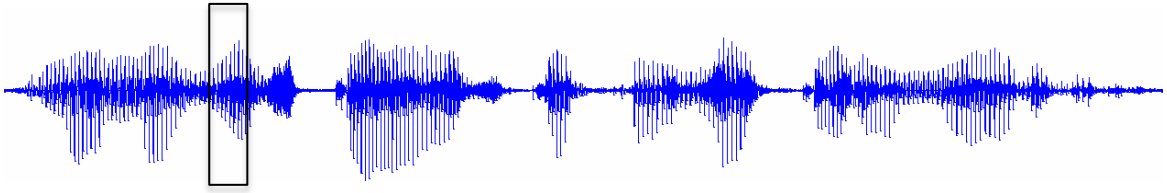
Salient spectral points search

Local mask application

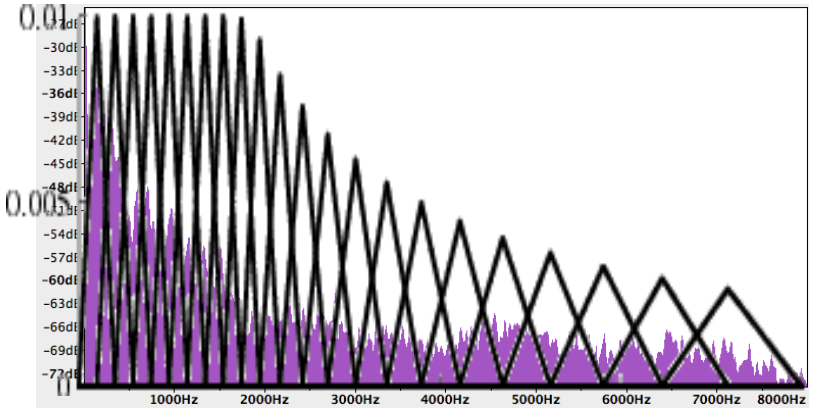
MASK fingerprint encoding and storage

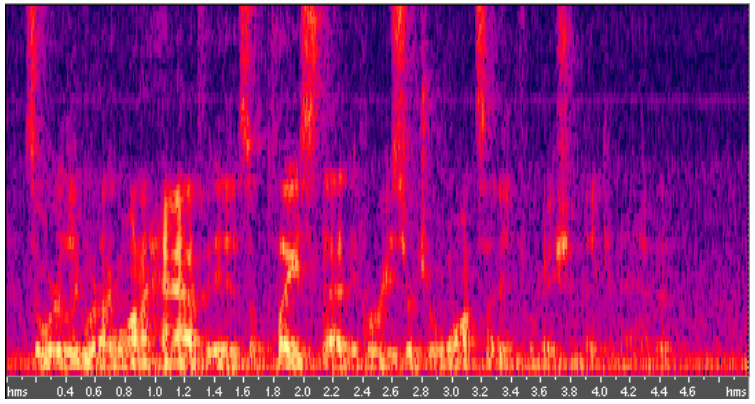
0110101000110110
0001101110110111
0001001001111010
0000101111101010
1111011000100110
0010010100110101

10ms, 100ms Hamming window



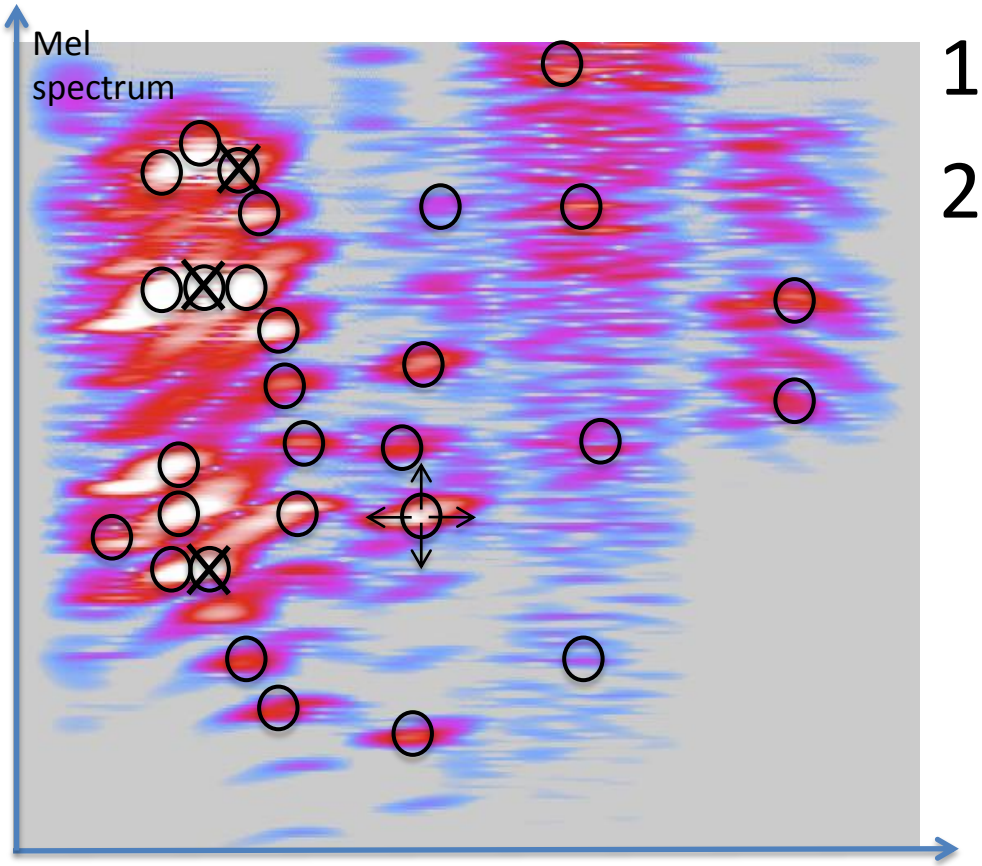
FFT 1024, bandwidth limited to 300-3KHz





18 or 34 MEL-spectrum bands

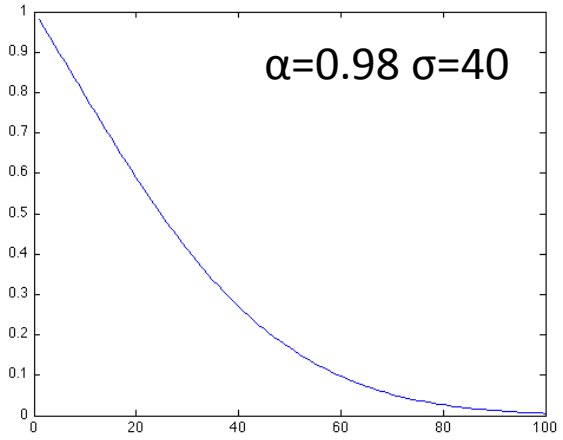
Selection of salient spectral points



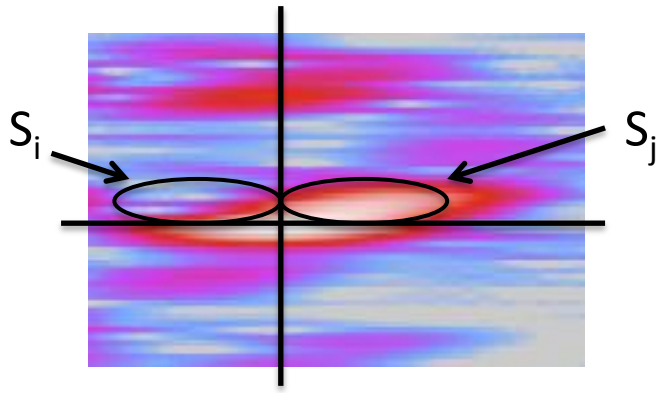
("beautiful", Timit-female)

1. Detect all maximas
2. Trim to the desired number

$$Thr[n] = \alpha^{\Delta t} E[n-1] \exp - \frac{(\Delta t)^2}{2 * \sigma^2}$$



Spectral masking around salient points

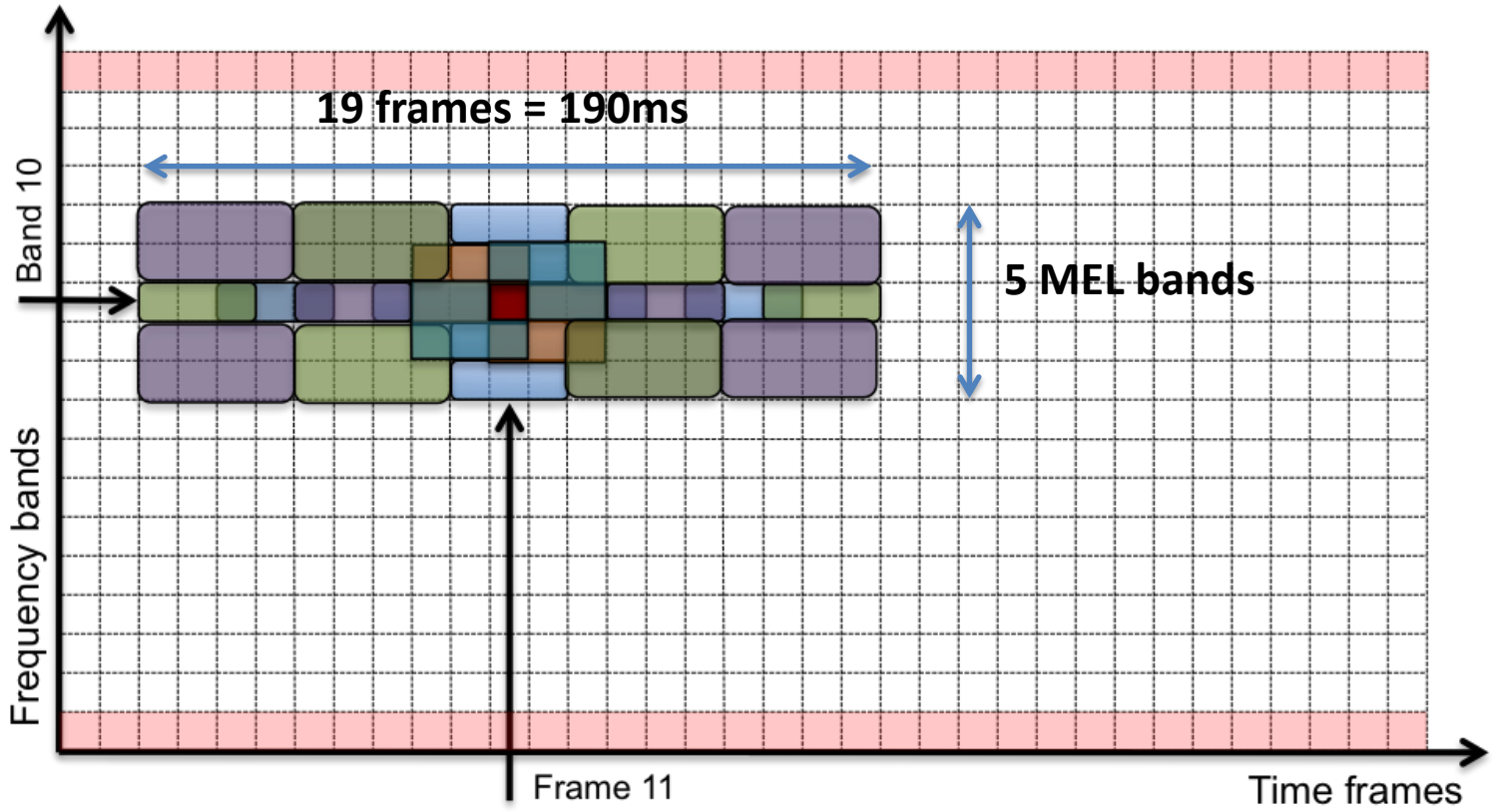


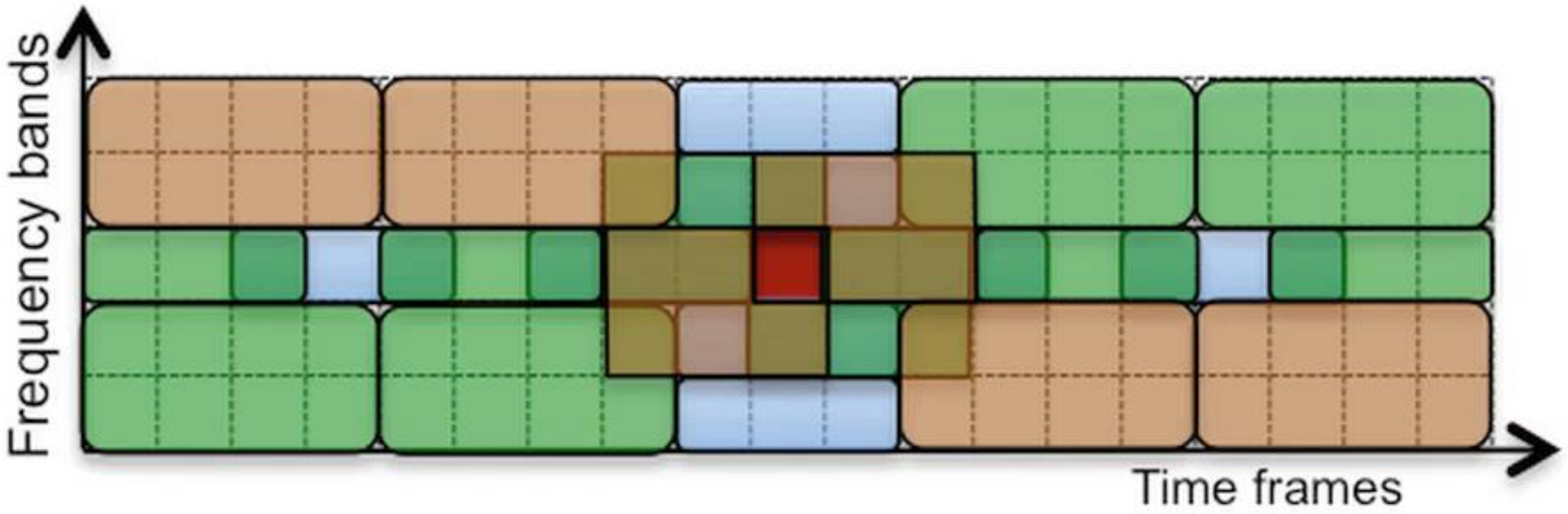
Spectral Regions

- Include one or several time-frequency values
- Overlaps are allowed
- The number of comparisons defines the size of the fingerprint
- Designed manually (for now)

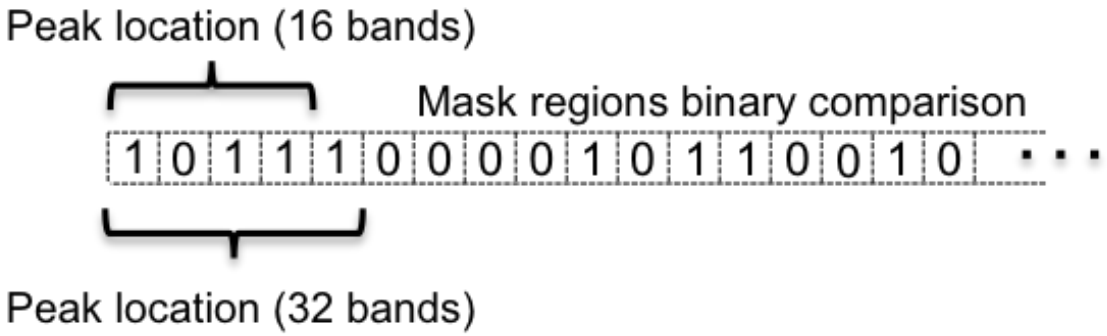
$$\begin{cases} S_i > S_j \rightarrow b[n] = 1 \\ otherwise \rightarrow b[n] = 0 \end{cases}$$

Current MASK regions



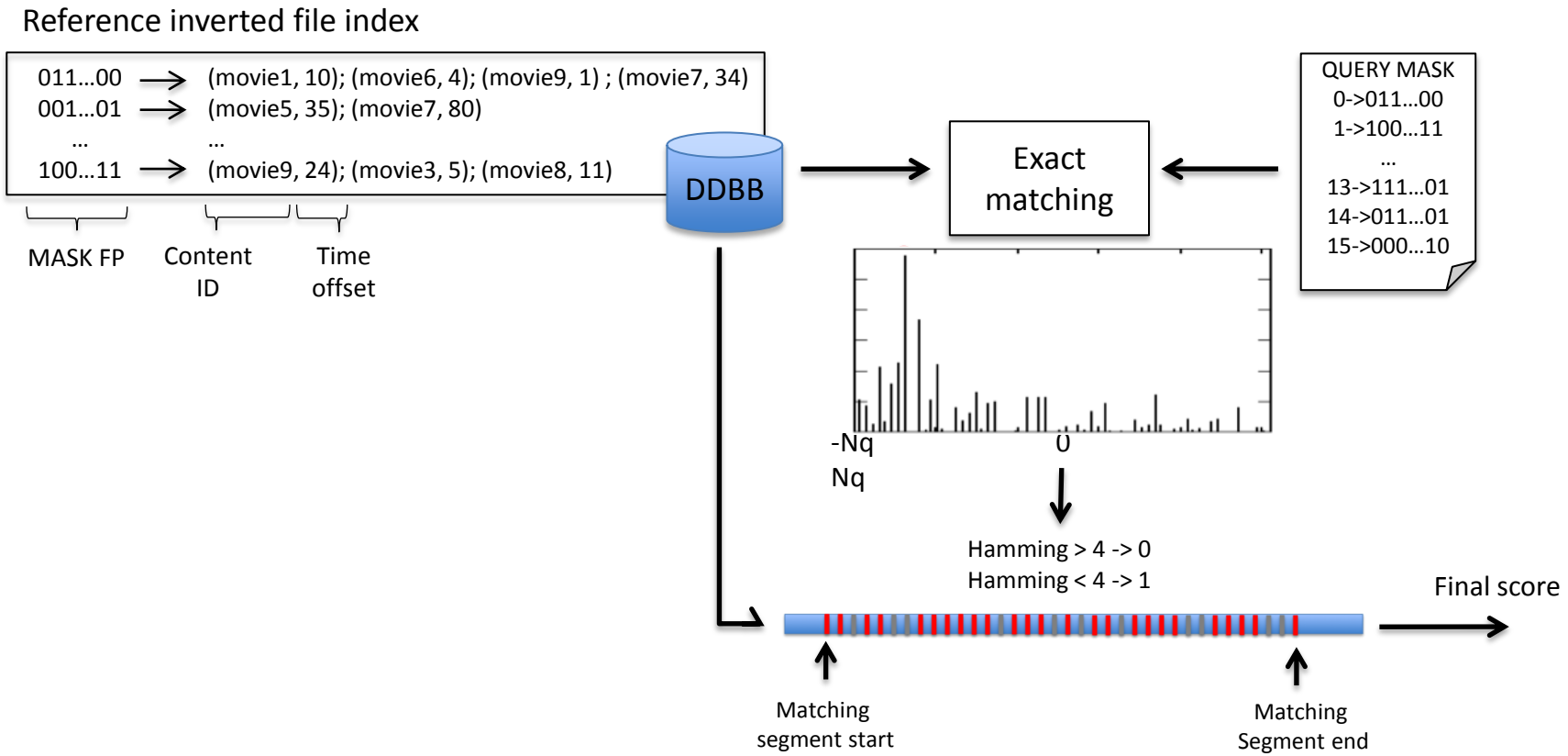

Frequency bands
Time frames

Fingerprint encoding



- 4-5 bits for the MEL band where the maxima is located
- 22 Bits obtained from spectral regions comparison

Indexing and retrieval



Experimental section

database

NIST-TRECVID 2010-2011 data for video-copy detection

- 400h reference videos
- 1400 audio queries per year (201 unique videos X 7)
- 7 audio transformations
  - original
  - MP3 compression
  - MP3 compression + multiband companding
  - Bandwidth limit (500-3K) + single-band companding
  - Mixed with speech
  - Mixed with speech + multiband companding
  - Mixed with speech + bandwidth limit + monoband companding

Metric & baseline

- normalized detection cost rate (NDCR) in balanced profile

$$NDCR_{BALANCED} = P_{MISS} + 200 R_{FA}$$

- Compare results with a similar fingerprint to the Philips fingerprint

Jaap Haitsma and Antonius Kalker, “A highly robust audio fingerprinting system,” in Proc. International Symposium on Music Information Retrieval (ISMIR), 2002.

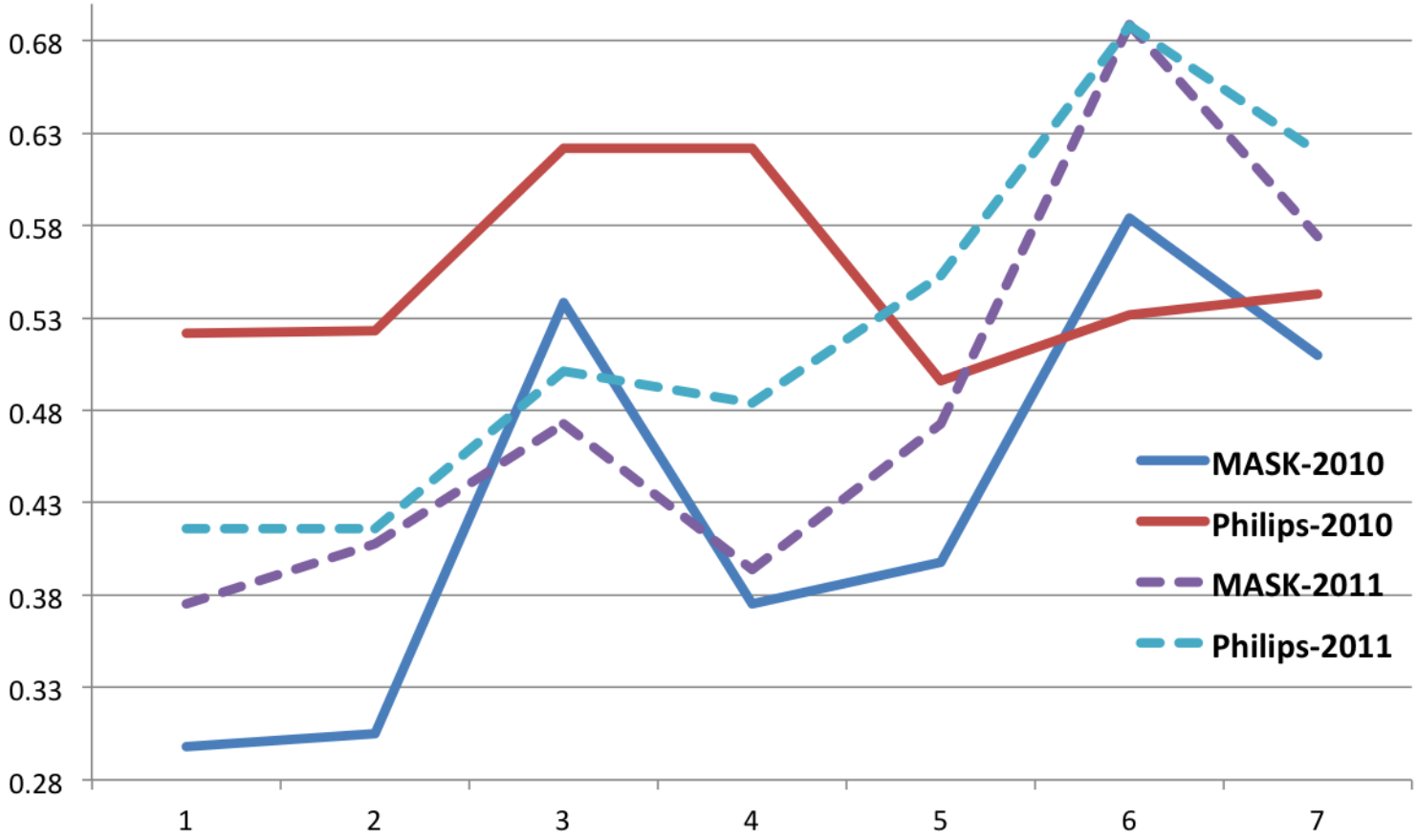
Table 2. Comparison of Minimum NDCR scores

system	# results	dataset	Min. NDCR	% improve.
Philips	1	2010	0.55	–
MASK			**0.43**	21.8%
Philips	20	2010	1.03	–
MASK			0.79	23.3%
Philips	1	2011	0.53	–
MASK			**0.48**	9.4%
Philips	20	2011	0.96	–
MASK			0.82	14.5%

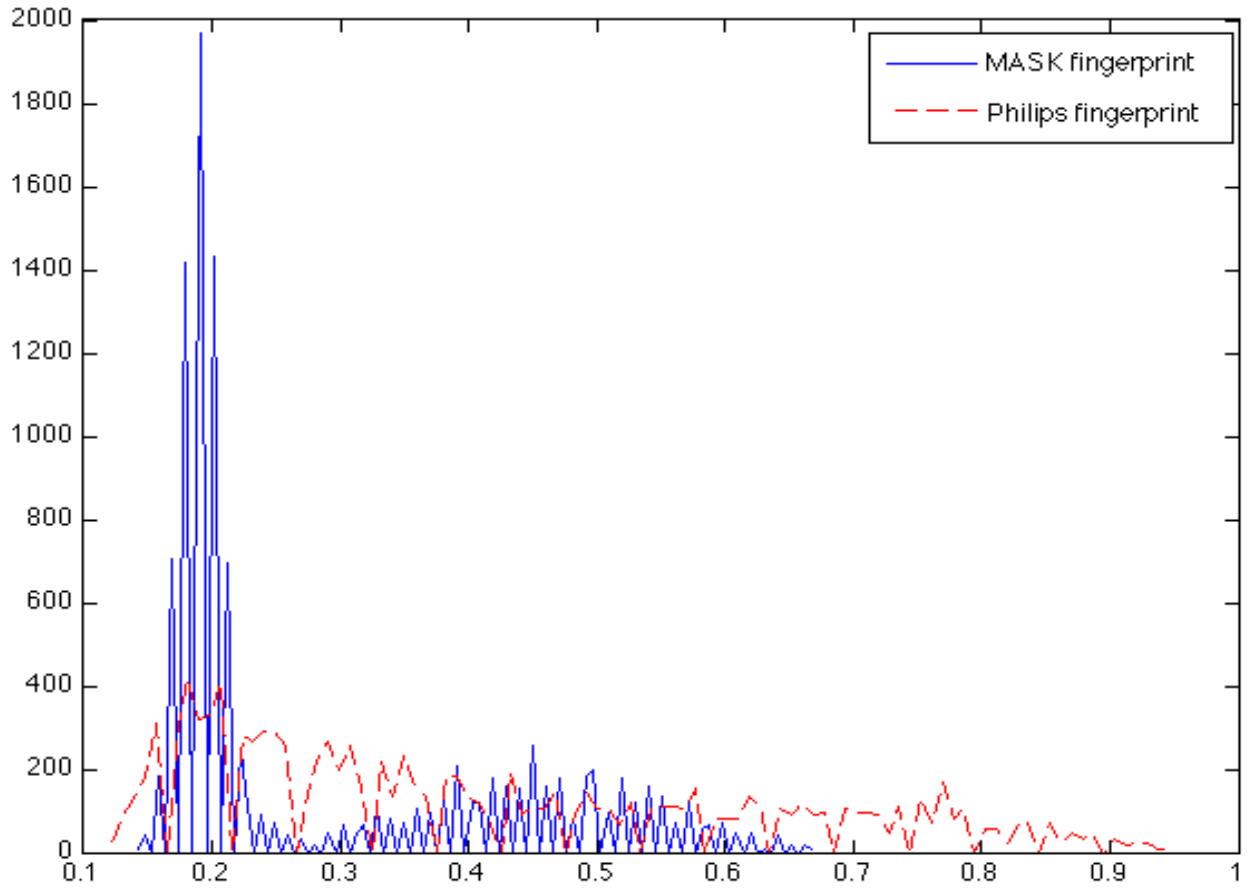
Table 3. Comparison of Actual NDCR scores

system	# results	dataset	Thr. std.	Act. NDCR	% improve.
Philips	1	2010	0.11	0.60	–
MASK			0.03	**0.44**	26.6%
Philips	20	2010	0.12	1.19	–
MASK			0.04	0.91	23.5%
Philips	1	2011	0.08	0.57	–
MASK			0.03	**0.50**	12.2%
Philips	20	2011	0.09	1.18	–
MASK			0.06	1.02	13.5%

Comparison per transformation



Scores histogram



Conclusions

- A novel binary fingerprint is proposed to improve on some shortcomings from well reputed prior art.
- We show that we can extract the FP and use it for VCD with excellent results