On the use of sparse time-relative auditory codes for music

Pierre-Antoine Manzagol, Thierry Bertin-Mahieux, Douglas Eck

June 25^{th} 2008

What's wrong with conventional features?

- time-frequency tradeoff
- sensitive to arbitrary alignment of blocks with signal

Sparse Auditory codes

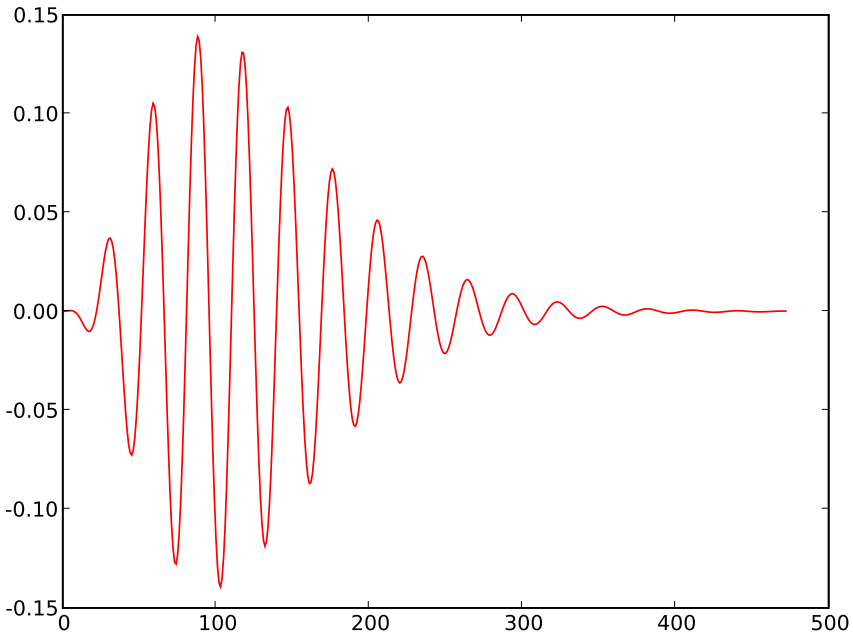
The signal x is modelled as a linear superposition of time shiftable kernels ϕ that are placed at precise time locations τ with a given scaling coefficient s. Formally:

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} s_i^m \phi_m(t - \tau_i^m) + \epsilon(t),$$

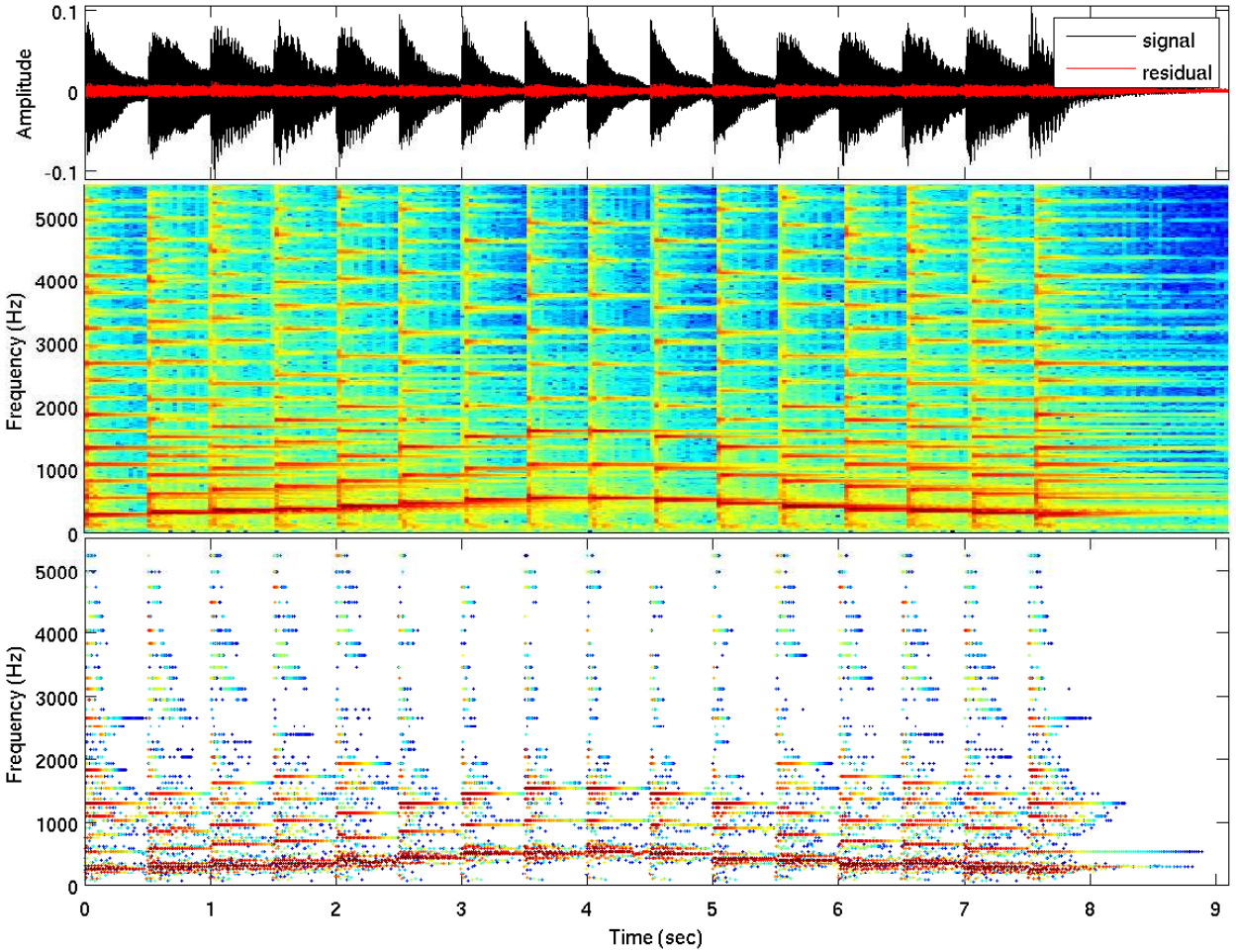
where m runs over the different kernels, i over the different instances of a given kernel and ϵ represents additive noise. The length of the kernels is variable.

Gammatones

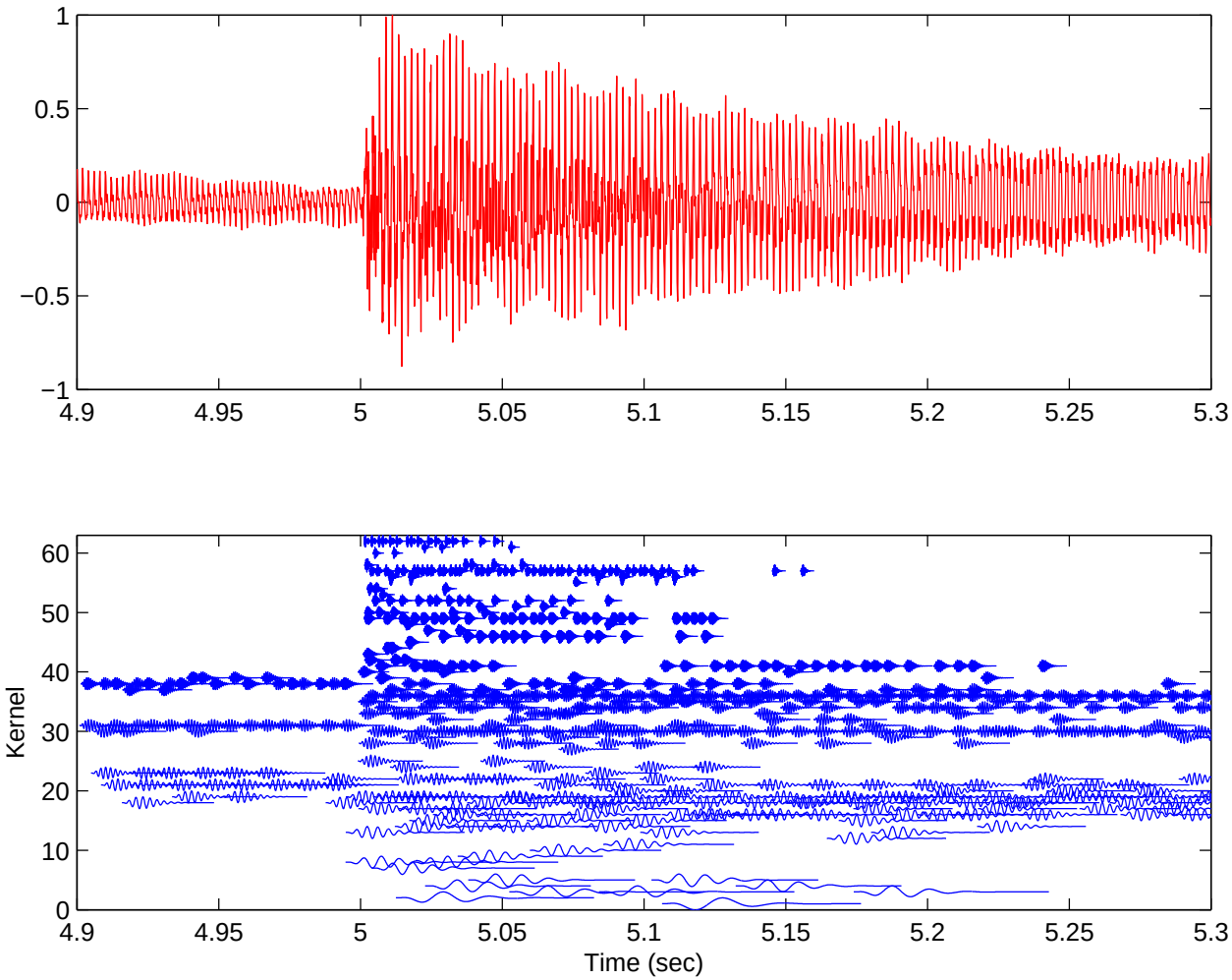
Gammatone kernels are set according to an equivalent rectangular band (ERB) filter bank cochlear model using Slaney's auditory toolbox for Matlab [5], as done in [6].



C-major scale piano



Close-up of piano sounding an A4



Sparse time-relative auditory codes

- **Sparse**: notion that only a small number of things (out of a large number of possible things) are going on at the same time.
- **Time-relative**: the events are located at a precise point in time.
- **Auditory**: biologically validated from what goes on in the cochlea.
- **Efficient**
- **No time-frequency trade-off or sensitivity to arbitrary alignment of signal**

Extra

- Various encoding algorithms.
- Possibility of either using predefined kernels (say Gammatones) or of learning kernels.

Contributions

- First use of these codes on complex western commercial music.
- Used the codes as naive input for genre classification (as good as MFCCs)
- Learn some kernels which are qualitatively different, but somewhat disappointing.

Encoding Tzanetakis

# of kernels	10^3	10^4	2.5×10^4	5×10^4
32	2.82 (0.97)	8.57 (2.17)	12.69 (2.87)	16.60 (2.88)
64	3.02 (1.03)	9.07 (2.29)	13.44 (2.98)	17.43 (2,70)
128	3.08 (1.04)	9.24 (2.31)	13.76 (2.99)	17.76 (2.52)

Table: Mean (standard deviation) of the SNR (dB) obtained over 100 songs (ten from each genre), as a function of the number of Gammatones in the codebook and of the maximum number of spikes allowed to reach a maximal value of 20 dB SNR.

- 64 Gammatones, ERB 20Hz-Nyquist, 20db SNR
- About 1 hour to encode a 30 second segment!!!

Encoding Tzanetakis - basic genre statistics

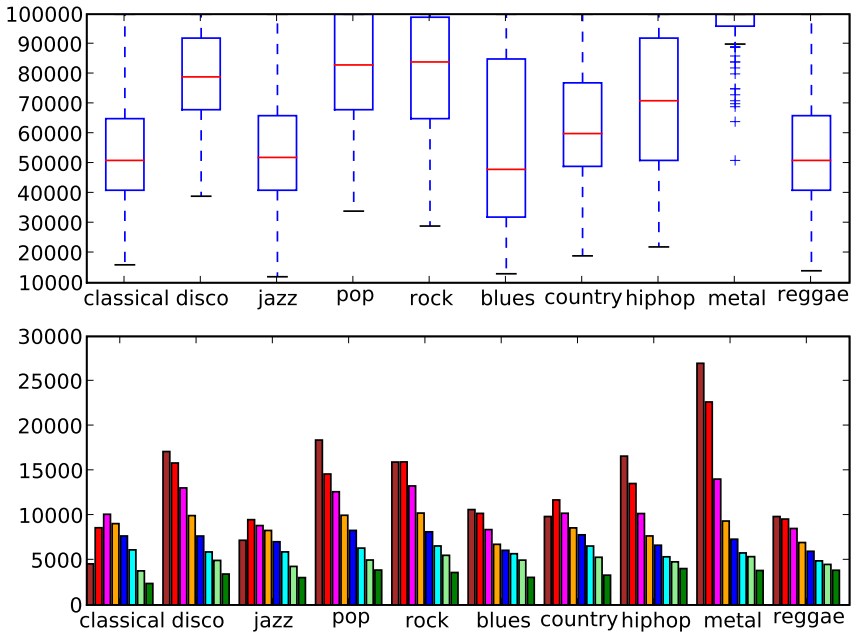
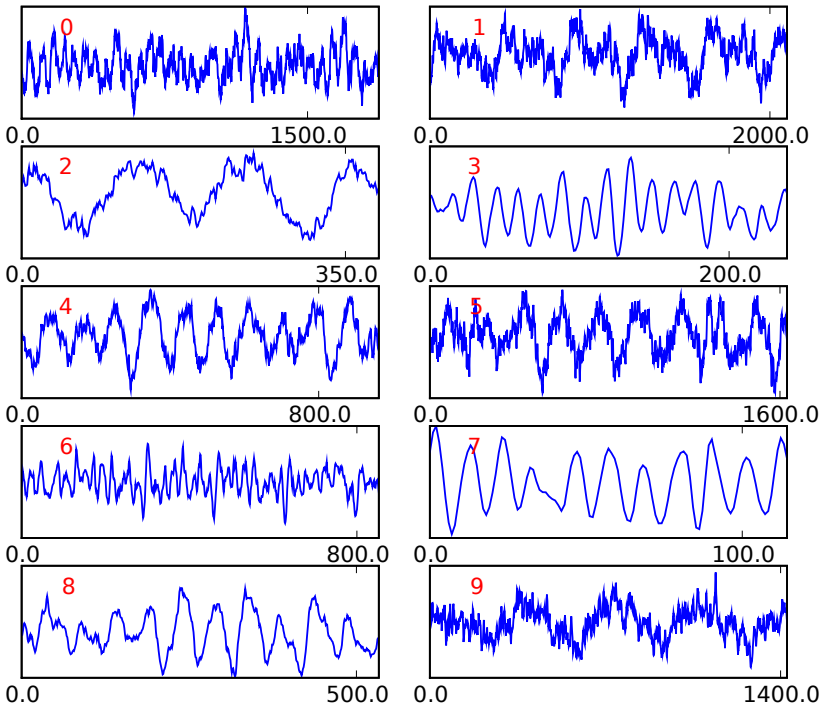


Figure: **Top:** box-and-whisker plot of the number of spikes (out of a maximum of 10^5) required to encode songs from various genres up to a SNR of 20dB. The red line represents the median and the box extends from the lower quartile to the upper one. **Bottom:** Bar plot of the mean number of kernels used per song depending on the genre. The 64 kernels are split into eight bins going from the lowest frequency Gammatones (left) to the highest ones (right).

Genre Recognition on Tzanetakis

- Features: basic statistics on the spikegrams for windows of 5 seconds.
- Results similar to those obtained with MFCCs.

Learned kernels



Learned kernels - results

kernel	10^3	10^4	2.5×10^4	5×10^4
gammatones	2.82 (0.97)	8.57 (2.17)	12.69 (2.87)	16.60 (2.88)
learned	1.86 (0.57)	6.06 (1.49)	8.86 (2.00)	11.53 (2,37)

Table: Mean (standard deviation) of the SNR (dB) obtained over 100 songs (ten from each genre) with either 32 Gammatones or 32 learned kernels in the codebook.

- Encoding not as good.
- Genre recognition results not as good.

Future work

Three specific issues that will need to be dealt with:

- How to put the features in a usable form?
- Learning kernels.
- Computational issue.

References

T. Blumensath and M. Davies.

Sparse and shift-invariant representations of music.
IEEE Transactions on Speech and Audio Processing, 2006.

G. Davis, S. Mallat, and M. Avellaneda.

Adaptive greedy approximations.
Constructive approximation, 13(1):57–98, 2004.

S.G. Mallat and Zhifeng Zhang.

Matching pursuits with time-frequency dictionaries.
Signal Processing, IEEE Transactions on, 41(12):3397–3415, December 1993.

M. Plumbley, S. Abdallah, T. Blumensath, and M. Davies.

Sparse representations of polyphonic music.
Signal Processing, 86(3):417–431, March 2006.

M. Slaney.

Auditory toolbox, 1998.
(Tech. Rep. No. 1998-010). Palo Alto, CA:Interval Research Corporation.

E. Smith and M. S. Lewicki.

Efficient coding of time-relative structure using spikes.
Neural Computation, 17(1):19–45, 2005.

E. Smith and M. S. Lewicki.

Efficient auditory coding.
Nature, 439:978–982, February 2006.

Questions?