Doping audio for enhanced audio processing

Gaël Mahé

Université Paris Descartes

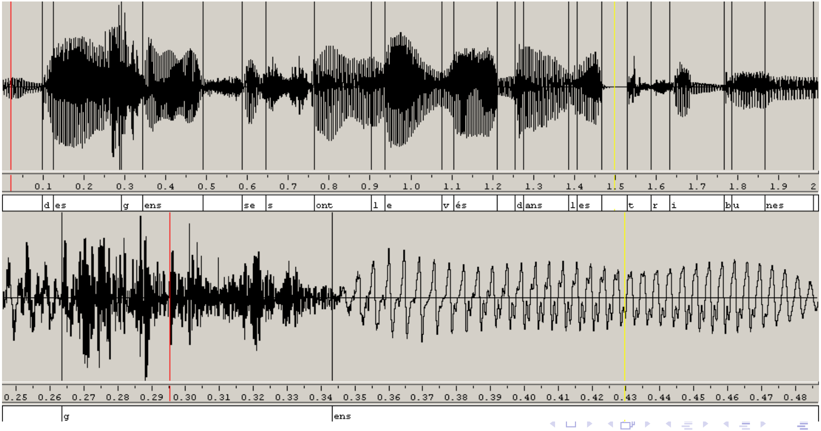
March 2016

Plan

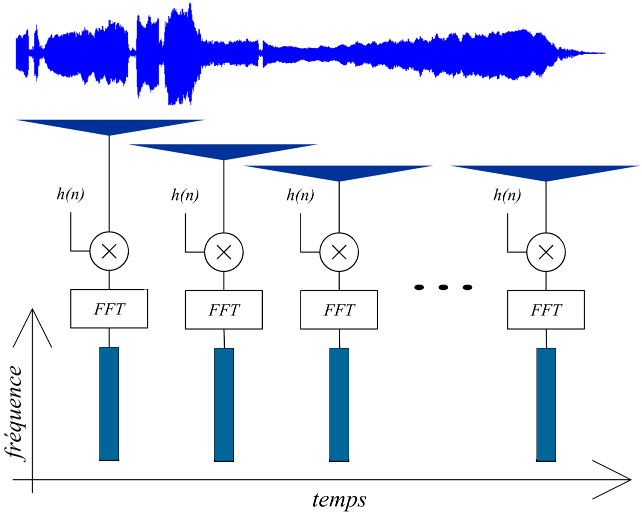
Doping audio for enhanced audio processing

Local stationarity

- Audio signals are **non-stationary**
- but can be seen as **locally** stationary (on 10 to 100 ms).

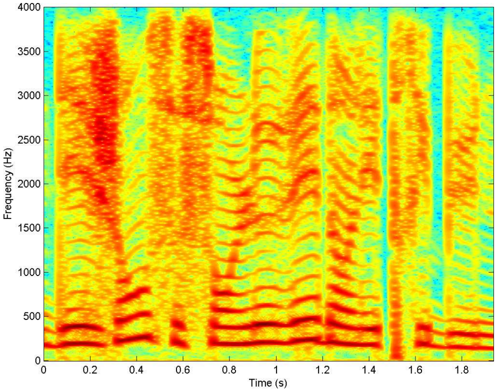


Time-frequency analysis with STFT (1)



Time-frequency analysis with STFT (2)

Spectrogram of 2s of speech:



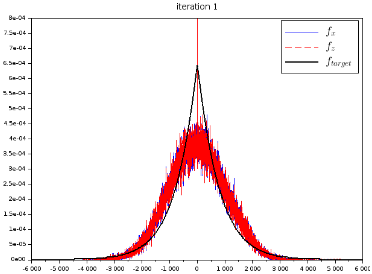
Histogram

Time samples $x(n)$ and time-frequency samples $|X(t,f)|$ have generalized Gaussian PDF:

$$f(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-\left(\frac{|x-\mu|}{\alpha}\right)^{\beta}\right), \quad \text{with:}$$

- α = scale factor (related to variance)
- β = shape parameter

Example : histogram (time samples) of 10s of piano at 32 kHz:

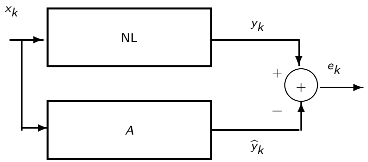


Doping audio for enhanced audio processing

Non-linear audio systems identification

Identification of a non-linear (NL) audio system:

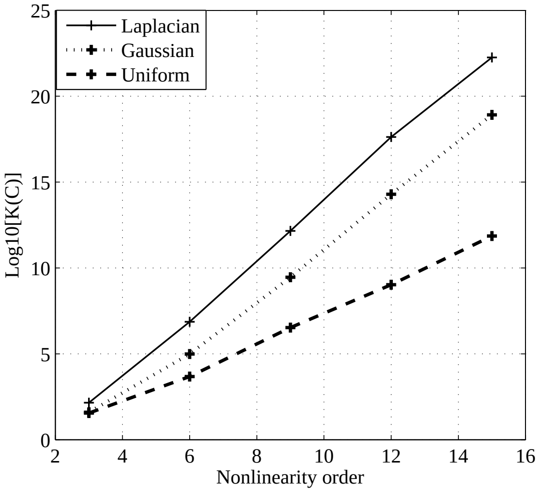
- amplificator, loudspeaker, microphone...
- NL results from electrical, mechanical and acoustical effects
- Modelization : polynomial model (without memory) / Volterra filters (with memory)



Identification performance rely on the condition number of $\mathbf{C}_\mathbf{x} = E[X_k X_k^T]$:

- X_k consists of all $x_k^{m_1} x_{k-1}^{m_2} \dots x_{k-M+1}^{m_M}$ / $m_1 + m_2 + \dots + m_M \leq \mathrm{NL\ order}$
- $K(\mathbf{C}_\mathbf{x}) = \log_{10}\left(|\lambda_{max}|/|\lambda_{min}|\right)$,
  where λ_{max} and λ_{min} = max and min eigenvalues of $\mathbf{C}_\mathbf{x}$.

Memoryless non-linear audio systems identification

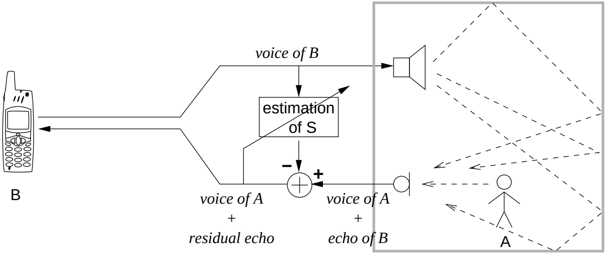


How to enhance the condition numbers ?

- Modify the distribution of the audio signal
- Or orthonormalize:
  - Gram-Schmidt algorithm, whatever the distribution
  - simpler alternative: use Hermite polynomials if the signal is Gaussian

Conclusion: we would like the signal be more Gaussian.

Acoustic echo cancellation (AEC)



- Identification of S through minimization of the variance of the residual echo, when A is not speaking.
- adaptive algorithms: stochastic gradient descent and extensions
- Convergence speed depends on signal whiteness...
  audio is not white!
- Performance in steady state depend on signal stationarity... audio is not stationary!

Source separation

Principles:

- n source signals $S_1 \dots S_n$
- mixture matrix A of dimensions $p \times n$
- p mixtures $X_1 \dots X_p$

$$X = AS, \quad \text{with } S = [S_1 \dots S_n]^\top \text{ et } X = [X_1 \dots X_p]^\top$$

Goal: estimate S from X without knowing A

Difficulty depends on

- $p \geq n$?
- Do we know some properties of the sources? (distributions, moments, bounds, parametric model...)
- Instantaneous or convolutive mixture?

Separation of a determined ($p \geq n$) instantaneous mixture

- **Darmois' Theorem**: If the sources are independent and at most one of them is Gaussian, then any $Y = BX$ such that the Y_i are independants is equal to S up to permutations and changes of scales.

$\rightarrow$ Separation algorithms based on Y_i **independance maximization**

- Performance depend on the sources distributions and on the knowledge (or not) of them.

We would like to control the distribution of the sources, and possibly make them non-Gaussian.

Underdetermined source separation

When number of sensors $<$ number of sources,
Sparse Components Analysis (SCA) methods, relying on:

- assumption of sparse sources: few non-zero coefficients or heavy tail distribution (small shape parameter in the case of a GG distribution)
- even better: jointly sparse sources,
  i.e. with rare ovelapping in a given space of represention.

Again, we would like to control the distribution of the sources, and possibly make them sparse.

Summary

- Signal processing base algorithms often rely on (or work better with) strong assumptions:
  - stationarity
  - whiteness
  - specific distribution
- ...although audio signals verify none of them!

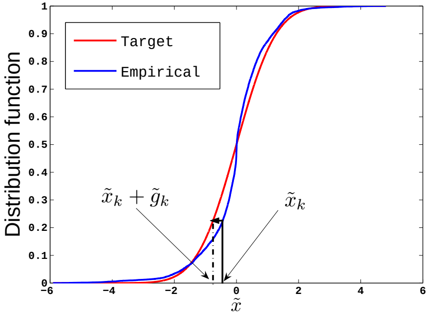
▶ **Doping watermarking** = inaudibly force properties of the audio signals to make them fullfill algorithms requirements.

Doping audio for enhanced audio processing

Ex: Gaussianization (1)

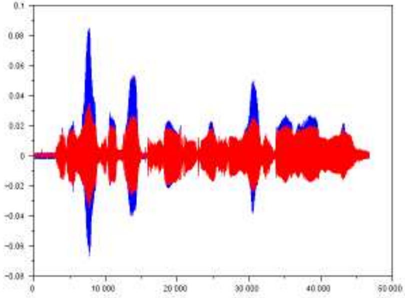
- Empirical cumulative distribution function (CDF): $F_X^{emp}(x_k) = \Pr[X \leq x_k] = |\{X \leq x_k\}|/N$
- Target CDF: F^{target}, Gaussian with same mean and variance
- Histogram equalization as in image processing: add g_k to each x_k, so that:

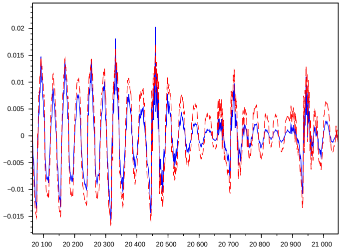
$$F^{target}(x_k + g_k) = F_X^{emp}(x_k)$$



Ex: Gaussianization (2)

Gaussianization of 3s of speech sampled at 16kHz :





Speech Gaussianized speech

Formalization

- Histogram and cumulative histogram of a digital time séquence x:

$$\forall\, k \in \mathbb{Z},\ \begin{cases} f_x(k) = |\{x(n)\,|\,x(n) = k\}| \\ F_x(k) = \sum_{i=-\infty}^{k} f_x(i) = |\{x(n)\,|\,x(n) \leq k\}|, \end{cases} \quad (1)$$

- Same formulas with a real time-frequency representation $X(m, f)$.
- Let f_{target} the target histogram.
  Find a transformation $x \rightarrow z$ so that :

$$\begin{cases} f_z \simeq f_{target} & (2) \\ \text{distortion}(x, z) \text{ is inaudible} & (3) \end{cases}$$

  - (2) $\Leftrightarrow$ min $\mathrm{d}(f_z, f_{target})$,
    where d denotes a PDF dissimilarity measure.
  - (3) depends on the transformation.

Doping audio for enhanced audio processing

The problem

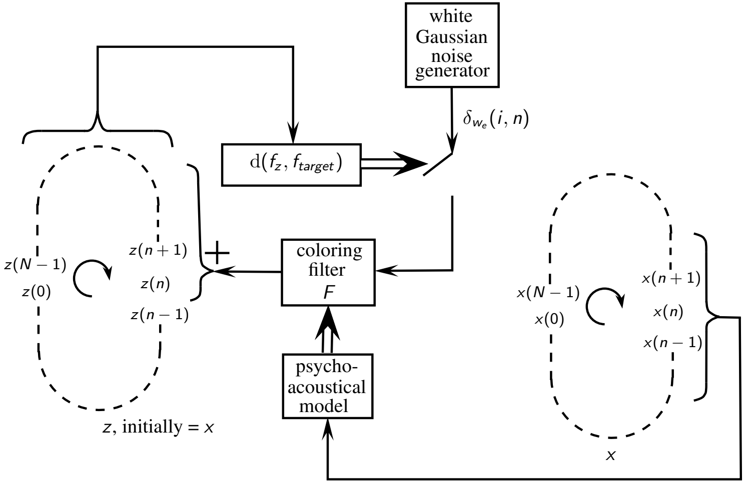
- Additive transformation $x \rightarrow z = x + w$ so that:

$$\begin{cases} \min \mathrm{d}(f_z, f_{target}) & (4) \\ \gamma_w(m, \nu) < \gamma_{mask}(m\nu) \quad \forall \, \text{frame} \, m, & (5) \end{cases}$$

- Difficulties:
  1. Histogram optimization on the whole signal
     vs local constraints (one masking threshold per frame)
  2. Histogram of the time-domain samples
     vs constraints in the frequency domain

$\rightarrow$ Ad hoc heuristic: add iteratively low level noises contributing to (4), with DSP parallel to $\gamma_{mask}(m, \nu)$

Perceptual controlled histogram reshaping



How to respect the perceptual constraint?

- For q iterations, $z = x + w$,
  with $w = f * w_e$,
  where $w_e(n) = \sum_{i=1}^{q} \delta(i,n)\delta w_e(i,n)$
  where $\delta(i,n) = 0 \text{ or } 1$ according to decision of adding $\delta w_e(i,n)$
- Since the $\delta w_e(i,n)$ are independent
  and under assumption that $\delta(i,n)$ are independent,
  w_e is a white noise.
- The constraint $\gamma_w(\nu) < \gamma_{mask}(\nu)$ becomes:

$$|F(\nu)|^2 \sigma_{w_e} < \gamma_{mask}(\nu),$$

i.e.:

$$\begin{cases} |F(\nu)|^2 = \gamma_{mask}(\nu) \\ \sigma_{w_e} < 1 \quad \text{controlled by choosing } q \text{ and } (\sigma_i)_{1 \le i \le q} \end{cases} \tag{6}$$

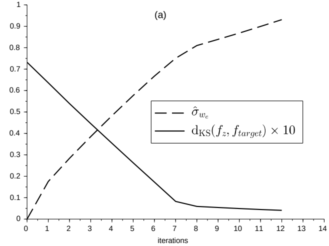
Application to “sparsification” (1)

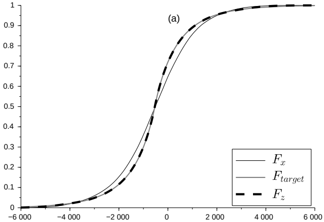
Experimental conditions:

- Mono-instrumental signals, 10 to 15s, sampling frequency 32kHz
- Filter F with finite impulse response of length 50, appromixating the MPEG-1 masking threshold
- Target histogram : generalized Gaussian with shape parameter divided by 2

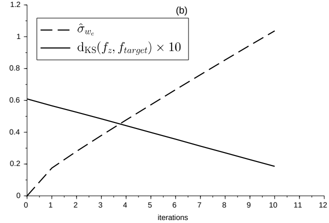
Application to "sparsification" (2)

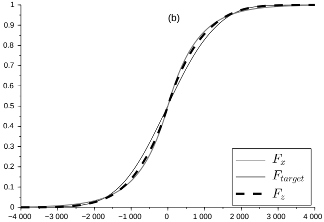
- Bass, shape parameter 1.6 → 0.8, final ODG = -0.85 :





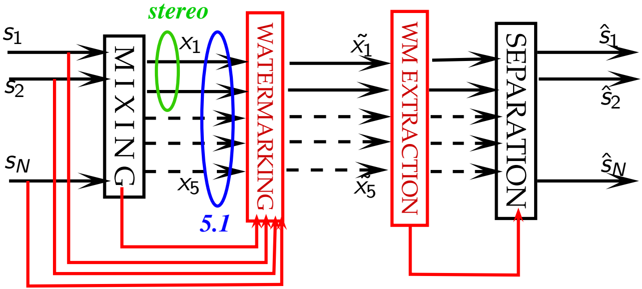
- Piano, shape parameter 2.1 → 1.05, final ODG = -1.1 :



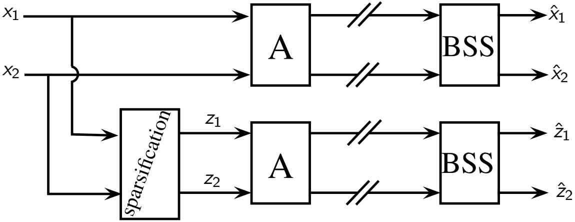


Application of the application

Informed Source Separation:



Experiment



- Mixture matrix:

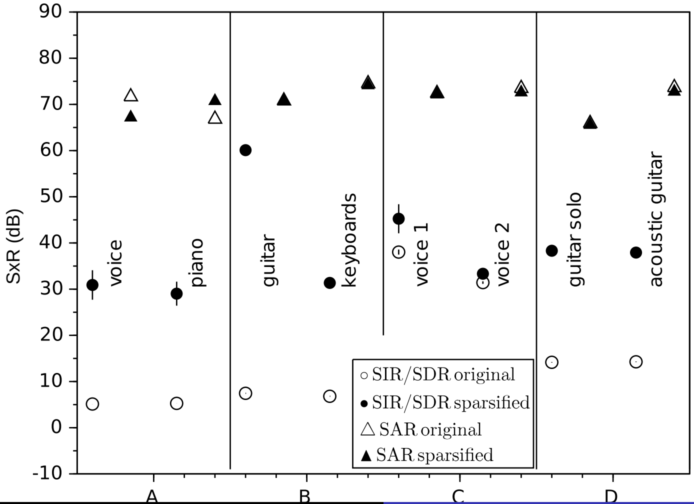
$$\mathbf{A} = \begin{pmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{pmatrix}$$

- BSS : separation using algorithm FastICA

Sparsification results

Mixture	instruments	shape parameter	ODG mix
A	voice piano	$1.8 \rightarrow 1.6$ $2.1 \rightarrow 1.6$	-0.5
B	guitar keyboards	3 $1.9 \rightarrow 1.3$	-0.3
C	voice 1 voice 2	$1 \rightarrow 0.9$ $1 \rightarrow 0.9$	-0.6
D	guitar solo guitar acoustic	$1.5 \rightarrow 1.5$ $1.4 \rightarrow 1.1$	-0.8

Separation results



Doping audio for enhanced audio processing

Goals

- Let $|S(m, f)|$ the spectrogram computed by STFT
- Let assume a generalized Gaussian PDF of shape parameter β
- Target CDF: F_{target} generalized Gaussian with $\beta' < \beta$
- A transformation $S \to \tilde{S}$
  so that $|\tilde{S}(m, f)| = F_{target}^{-1}\left(F_{emp}(|S(m, f)|)\right)$
  would reach the PDF target but not fullfill the inaudibility constraint.

▶ Progressive transformation under perceptual constraint, using an iterative algorithm.
G. Mahé *et al.*, "Perceptually controlled doping for audio source separation", EURASIP J. on Advances in Signal Processing, march 2014

Algorithm

After initializing $\tilde{S}$ to S, process iteratively the spectrogram:

- If $F_{|\tilde{S}|} < F_{target}$ on $I = [10^{-\Delta/20}|\tilde{S}(m,f)|; |\tilde{S}(m,f)|[$, then reduce $|\tilde{S}(m,f)|_{\mathrm{dB}}$ of Δ (in dB)
- If $F_{|\tilde{S}|} > F_{target}$ on $I = [|\tilde{S}(m,f)|; 10^{\Delta/20}|\tilde{S}(m,f)|[$, then increase $|\tilde{S}(m,f)|_{\mathrm{dB}}$ of Δ (in dB)

Hence,

- $|F_{|\tilde{S}|} - F_{target}|$ decreases on interval I.
- which reduces $\mathrm{d}(f_{|\tilde{S}|}, f_{target})$.

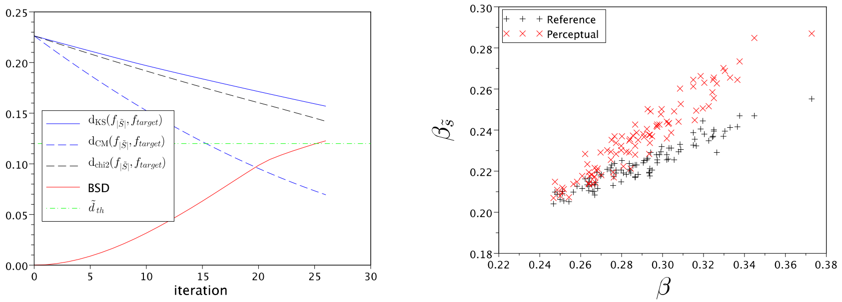
The process stops when $\mathrm{distortion}(s, \tilde{s})$ is audible.

Perceptual aspects

- Transformation = filtering by $H(m,f) = \frac{|\tilde{S}(m,f)|}{|S(m,f)|}$
- Each $H_{\mathrm{dB}}(m,f)$ cannot differ much from its neighbors:
  - horizontally: max. difference max $\Delta_{max}^{time}(f) \rightarrow$ avoid musical noise
  - vertically: max. difference $\Delta_{max}^{freq}(f) \rightarrow$ avoid robotic sound
- Which measure of distortion $\mathrm{distortion}(s,\tilde{s})$?
  - Transform spectra in "loudness spectra"
    = spectra in sones on a Bark frequency scale
  - Bark Spectral Distortion:

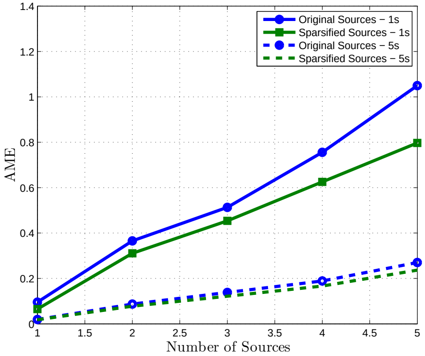
$$BSD(m) = \frac{\sum_{b=1}^{N_b} \left(S_s(m,b) - S_{\tilde{s}}(m,b)\right)^2}{\sum_{b=1}^{N_b} S_s(m,b)^2}$$

Results : sparsity



Results : source separation in a stereo mixture

- Angular mean error (AME)
  when estimating the columns of the mixture matrix



- source separation: interferences and distortions reduced of 1 to 2dB only, compared to without sparsification

Another approach: ℓ_0-sparsity

- **SCA needs jointly sparse sources**,
  i.e. with rare ovelapping in the space of represention
- **Idea: zeroing time-frequency bins below the masking threshold**

▶ $\searrow$ probability of having more than 2 active sources per time-frequency bin (m, f)

▶ ▶ simple determined case: 2 mixtures of 2 sources

- P. Balazs *et al.*, IEEE Trans. on Audio, Speech and Lang. Processing, 2010 :
  Eliminates 36% of the STFT spectrogram, $F_e = 16$ kHz
  But sparsity lost through OLA synthesis.
- J. Pinel and L. Girin, Proc. AES Conf., 2011 :
  Eliminates 75% of the MDCT spectrogram, $F_e = 44,1$ kHz
  Sparsity conserved across synthesis/analysis operations.

The problem...

... is to transmit information of activity :
For each source, binary matrix indicating if the source is active.
$\rightarrow$ bitrate 44,1 kbit/s per source... $=$ compressed audio bitrate!

Conclusion

Is it watermarking?

- Not in the classical sense (no explicit message embedded)
- But properties embedding trough an imperceptible alteration of the sound
  $\simeq$ watermarking

More classical applications possible:

- Properties forcing as a fragile watermark (integrity proof)
- Create patterns of short-term histograms
  as symbols for message embedding