Informed Spectral Analysis for Under Determined Audio Source Separation

Fourer Dominique

LaBRI - Université Bordeaux I

17 Novembre 2011

Plan

Source Separation Problem

Observation model

Instantaneous discrete sound mixture signal :

$$x[n] = \sum_{k=1}^{K} s_k[n] + r[n] \tag{1}$$

with $r[n]$ is a residual noise signal.

Monaural sound mixture

- The number K of sources present in the mixture is greater than the number of observation (under determined configuration).
- No orthogonality assumption (sources may overlap in time and frequency.)

State of the Art

Purpose of presented work

Recover each $s_k[n]$ signals from $x[n]$ with the minimal distortion (in the less squared-error sense).

Existing Approaches for audio under determined source separation

- Model-based inference : estimation of source signal parameters using prior information (e.g. harmonic model, sinusoidal modelling, GMM, ...).
- Unsupervised learning : non-parametric approach that attempts to extract signal characteristics from data. (e.g. ICA, NMF, sparse coding)
- Psychoacoustically motivated methods : organization of psychoacoustic cues (e.g. CASA)

Sinusoidal Modelling

Source decomposition using the stationary model for the analysis of a local frame :

$$s_k[n] = \sum_{l=1}^{L} a_l \cos\left(\omega_l n + \phi_l\right) \qquad (2)$$

where $(a, \omega, \phi) \in \mathbb{R}^3$ denotes respectively the amplitude, frequency and phase.

Why sinusoidal modelling ?

- Sparse representation of musical signal (efficient for low bit rate coding MPEG4-SSC/HILN),
- a and ω are perceptual parameters,
- allows efficient sound transformation. (e.g. time-stretching, transposition)

Sinusoidal Modelling

Parameters estimation

- Reassignment method [Kodera,Villedary & Gendrin 1976] achieves to reach Cramèr Rao lower Bound (CRB).
- Generalized derivative method for non-stationary model [Marchand & Depalle 2008]

$$\hat{\omega}(t, \omega_k) = \omega_k - \underbrace{\Im\left(\frac{S_{w'}(t, \omega_k)}{S_w(t, \omega_k)}\right)}_{-\Delta_\omega}.$$

$$\hat{a}(t, \omega_k) = \left|\frac{S_w(t,\omega_k)}{W(\Delta_\omega)}\right| \; \hat{\phi} = \angle\left(\frac{S_w[t,\omega_k]}{W(\Delta_\omega)}\right)$$

- S_w is the STFT of s using window w
- $w' = \frac{\partial w(t)}{\partial t}$ and $W = \mathcal{F}(w)$

Sinusoidal Modelling : Theoretical bounds

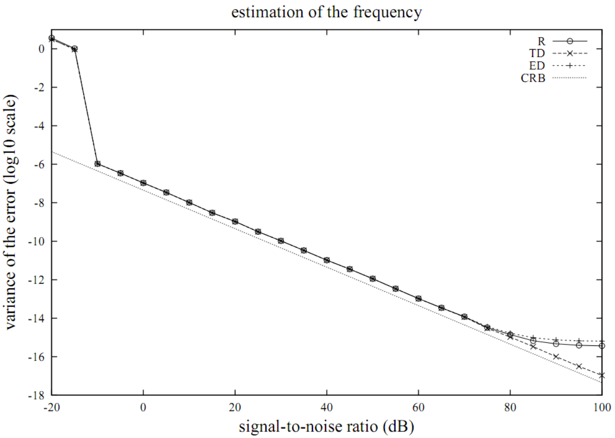


FIG.: Frequency estimation

Sinusoidal Modelling : Theoretical bounds

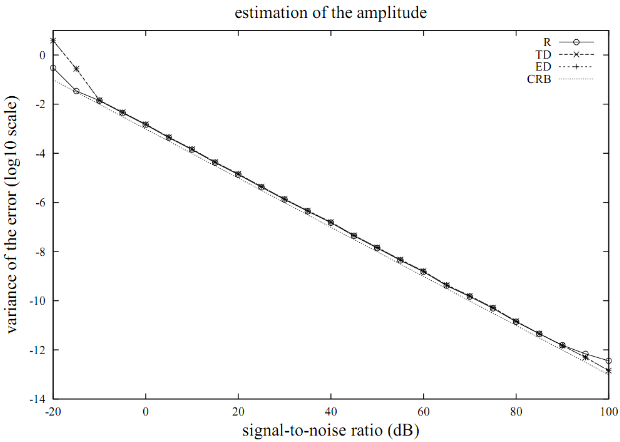


FIG.: Amplitude estimation

Sinusoidal Modelling : Theoretical bounds

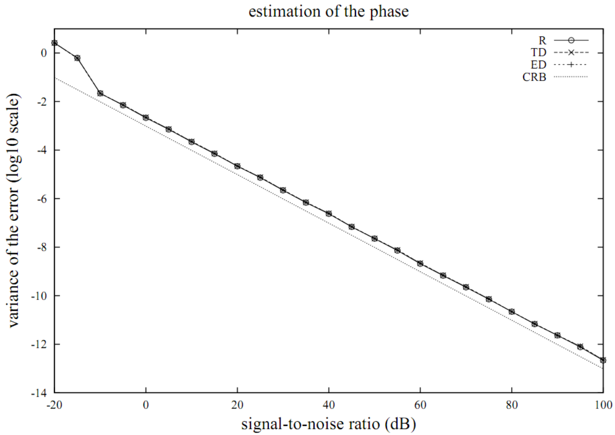


FIG.: Phase estimation

Plan

Informed Source Separation

Why ?

- Classic estimators have theoretical limitations (CRB),
- high-quality is required by demanding applications,
- the original separated audio sources signals are available during the mixing process (recording studio).

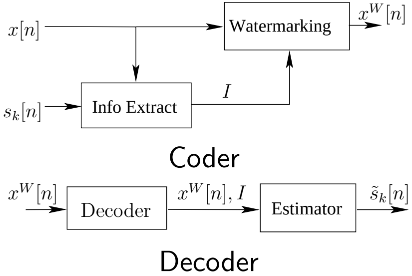
Motivation

- Derive new informed-estimators that combine classic estimation with extra-information,
- optimize the rate-distortion ratio of these estimators,
- hide extra-information into the signal itself (Watermarking)*.

Approach comparison

Classic estimation

Informed estimation



Informed Source Separation

State of the art

Classic source separation combined with side extra information :

- Spectral envelope + clustering + Spatial filtering [Gorlow, Marchand 2011],
- magnitude spectrogram compression + Wiener filtering + [Liuktus & al. 2011],
- spectral envelope + Wiener filtering [Parvaix, Girin 2009],

*E*stimate $\tilde{s}_k$ without prior knowledge about signal model or parameters.

Model based informed analysis

Motivation

- find the minimal amount of extra-information necessary to reach a fixed target precision from any classic estimator,
- allow a bit-per-bit scalable quality control,
- generalize the informed-analysis approach to allow a theoretical study.

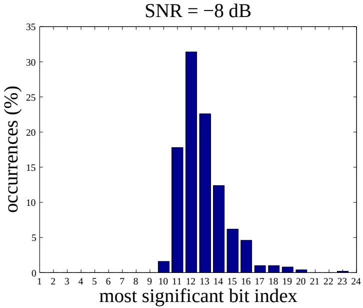
Intuition

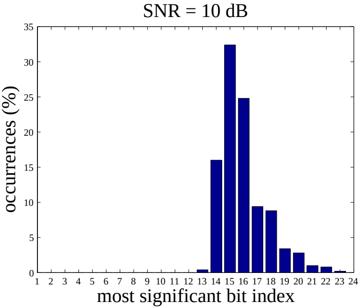
A classic estimator can be combined with the minimal complementary information required to systematically correct errors and reach a target quality.

Observation

Experimentation

Histogram of the Most Significant Bit (fixed point binary representation) of the estimation error committed with the reassignment method for frequency estimation of a sinusoid combined with a white noise.





Informed spectral analysis principle

Scalar case

- Let be $p \in [0, 1)$ a real parameter and $\hat{p}$ its estimation (obtained by a classic spectral analysis).
- Let be $\mathcal{C}_d$ the d-length fixed point binary coding application and $\mathcal{D} = \mathcal{C}^{-1}$. Let $C = C_1, C_2, ..., C_d$ denote $\mathcal{C}_d(p)$
- Let be $I_\sigma = \text{msb}\left(\mathcal{C}\left(|p - \hat{p}|\right)\right)$ for a significant number of occurrences over $\hat{p}$. (I_σ is the upper bound of the CI of the estimator).

$$\mathcal{C}_d(p) = \underbrace{C_1, C_2, \ldots, C_{I_\sigma - 1}}_{\text{reliable part}}, \underbrace{C_{I_\sigma}, \ldots, C_d}_{\text{unreliable part}}. \tag{3}$$

How to correct errors ? 1/2

Proposal of solution

Substitution of the unreliable part of $\mathcal{C}(\hat{p})$ with the exact bit values extracted from $\mathcal{C}(p)$

$$
\begin{array}{lll}
p: & 0.4032 & \mathcal{C}(p): \quad 0\ 1\ 1\ 0 \quad \mathbf{0\ 1\ 1\ 1\ 0} \quad 0\ 1\ 1\ 1 \\
\hat{p}: & 0.3831\ (|\epsilon| = 0.0201) & \mathcal{C}(\hat{p}): \quad 0\ 1\ 1\ 0 \quad \mathbf{0\ 0\ 1\ 0\ 0} \quad 0\ 0\ 1\ 0 \\
\tilde{p}: & 0.4026\ (|\epsilon| = 0.006) & \mathcal{C}(\tilde{p}): \quad 0\ 1\ 1\ 0 \quad \mathbf{0\ 1\ 1\ 1\ 0} \quad 0\ 0\ 1\ 0
\end{array}
$$

$$|\tilde{p} - p| \leq |\hat{p} - p|$$

How to correct errors 2/2

Problem : What happens when ...

p :	0.2776	$\mathcal{C}(p)$:	0 1 0 0	**0 1 1 1**	0 0 0 1 0
$\hat{p}$:	0.2473 ($\|\epsilon\| = 0.0303$)	$\mathcal{C}(\hat{p})$:	0 0 1 1	**1 1 1 1**	0 1 0 1 0
$\tilde{p}$:	0.2161 ($\|\epsilon\| = 0.0615$)	$\mathcal{C}(\tilde{p})$:	0 0 1 1	**0 1 1 1**	0 1 0 1 0

$$|\tilde{p} - p| > |\hat{p} - p|$$

Substitution may increase the error.

Solution

We transmit the $l_\sigma - 1$ bit for verification

p :	0.2776	$\mathcal{C}(p)$:	0 1 0 **0** **0 1 1 1** 0 0 0 1 0
$\hat{p}$:	0.2473 ($\lvert\epsilon\rvert = 0.0303$)	$\mathcal{C}(\hat{p})$:	0 0 1 **1** **1 1 1 1** 0 1 0 1 0

- If l_σ is exact then we have $p - 2^{-l_\sigma} \leq \hat{p} \leq p + 2^{-l_\sigma}$
- When $\mathcal{C}(p)[l_\sigma - 1] \neq \mathcal{C}(\hat{p})[l_\sigma - 1]$ we check 2 cases :

$\tilde{p}^+$:	0.2786 ($\lvert\epsilon\rvert = 0.0010$)	$\mathcal{C}(\tilde{p}^+)$:	0 1 0 0 **0 1 1 1** 0 1 0 1 0
$\tilde{p}^-$:	0.1536 ($\lvert\epsilon\rvert = 0.1240$)	$\mathcal{C}(\tilde{p}^-)$:	0 0 1 0 **0 1 1 1** 0 1 0 1 0

- The best informed value verifies $\hat{p} - 2^{-l_\sigma} \leq \tilde{p} \leq \hat{p} + 2^{-l_\sigma}$

Results

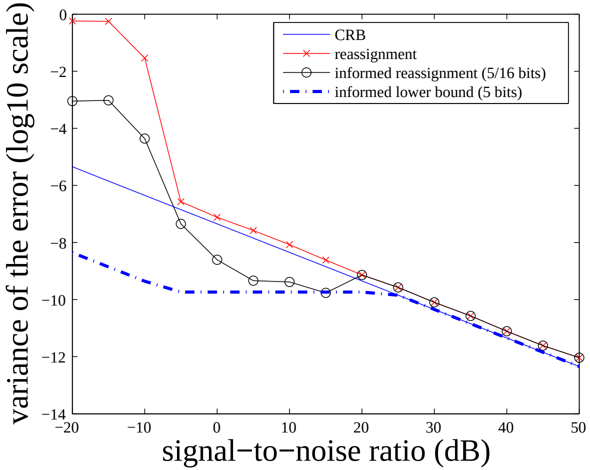


FIG.: Comparison with CRB for frequency estimation with $d = 16$

Results

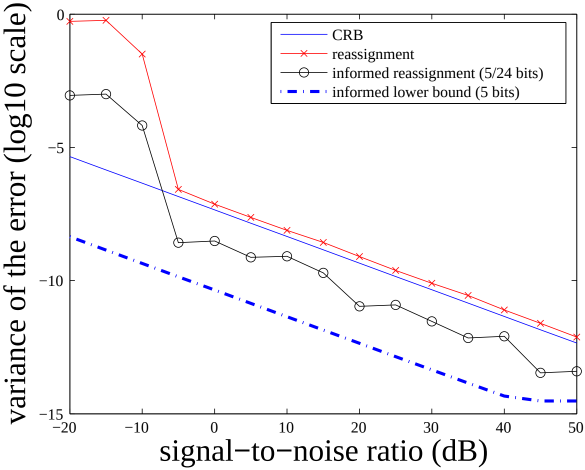


FIG.: Comparison with CRB for frequency estimation with $d = 24$

Generalization to $P \in \mathbb{R}^3$ for sinusoidal model

- Parameter is a vector of $\mathbb{R}^3$: $P = (a, \omega, \phi)$
- Coding application : $\mathcal{C}_d(P) = \mathcal{C}_e(a), \mathcal{C}_f(\omega), \mathcal{C}_g(\phi)$ with $d = e + f + g$

Distortion measure

Weighted squared error between synthesized signals.

$$\mathcal{D}(P, \hat{P}) = \sum_{n=1}^{N} w[n] \left| a \cos\left(\omega n + \phi\right) - \hat{a} \cos\left(\hat{\omega} n + \hat{\phi}\right) \right|^2 \qquad (4)$$

Vector quantization problem

How to :

- find the minimal d for a given distortion measure,
- find e, f, and g that minimize $\mathcal{D}(P, \hat{P})$ (bit allocation),
- taking advantage of dependence between parameters (e.g. It is useless to allocate bit to ω and ϕ if $a \approx 0$).

Solution

Entropy Constrained Unrestricted Spherical Quantization (ECUSQ)[Korten, Jeusen & Heusdens 2007] :

- Define distortion as a function of entropy H_t,
- define a variable-length quantizer that minimize $\mathcal{D}(H_t)$,

ECUSQ

Using high-rate assumption, $\mathcal{D}$ can be expressed as a function of error over each component :

$$\mathcal{D}(\tilde{a}, \Delta_a, \Delta_\omega, \Delta_\phi) \approx \frac{||w||^2}{24} \left(\Delta_a^2 + \tilde{a}^2(\Delta_\phi^2 + \sigma_w^2 \Delta_\omega^2)\right)$$

Let be $f_{A,\Omega,\Phi}(a, \omega, \phi)$ the joint probability density of P and g the quantizer point density.
Thus we can express overall average distortion :

$$\bar{\mathcal{D}} = \frac{||w||^2}{24} \int\int\int f_{A,\Omega,\Phi}(a, \omega, \phi) \left(g_A^{-2}(a, \omega, \phi)\right.$$
$$\left. + \tilde{a} \left(g_\Phi^{-2}(a, \omega, \phi) + \sigma_w^2 g_\Omega^{-2}(a, \omega, \phi)\right)\right) da d\omega d\phi$$

with $\sigma_w = \frac{1}{||w||^2} \sum_{n=0}^{N-1} w[n]^2 n^2$

ECUSQ

Finally we have to minimize using entropy constraint using Lagrangian multiplier :

$$\nu = \bar{\mathcal{D}} + \lambda h(A, \Omega, \Phi) \tag{5}$$

We obtain :

$g_A(a, \omega, \phi) = \sqrt{\frac{||w||^2}{12\lambda \log_2(e)}}$
$g_\Phi(\tilde{a}, \omega, \phi) = \tilde{a} g_A(a, \omega, \phi)$
$g_\Omega(\tilde{a}, \omega, \phi) = \tilde{a} \sigma_w g_A(a, \omega, \phi)$

with $\lambda = \frac{||w||^2 \exp\left(\log(2)(-\frac{2}{3}(H_t) - 2b(A) - \log 2(\sigma_w))\right)}{12 \log_2(e)}$

ECUSQ

Distortion Rate Function

Average distortion as a function of the entropy (with high-rate assumption) :

$$D_{\text{ECUSQ}} = \frac{||w||^2}{8} 2^{-(2/3)(H_t - h(A,\Omega,\Phi)) - 2b(A) - log_2(\sigma_w)} \tag{6}$$

with $b(A) = \int f_A(a) \log_2(a) da$

ECUSQ

Quantizer point density functions

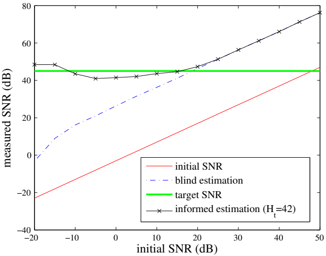
$$g_A(a,\omega,\phi) = 2^{(1/3)\tilde{H}_t - 2b(A) - \log_2(\sigma_w)}, \tag{7}$$
$$g_\Phi(\tilde{a},\omega,\phi) = \tilde{a} g_A(a,\omega,\phi), \tag{8}$$
$$g_\Omega(\tilde{a},\omega,\phi) = \tilde{a}\sigma_w g_A(a,\omega,\phi), \tag{9}$$

Notices

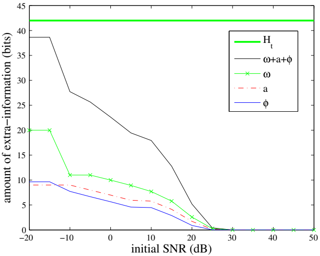
- Quantization steps are given by $\Delta = g^{-1}$,
- Optimal quantizers for ω and ϕ are linearly dependent on $\tilde{a}$,
- the bit allocation function $b_{a,\omega,\phi}$ is computed from $\lceil \log_2(g) \rceil$

Simulation for $P \in \mathbb{R}^3$

$$\text{SNR}^{\text{target}} = 45dB \Rightarrow H_t \approx 42bits$$



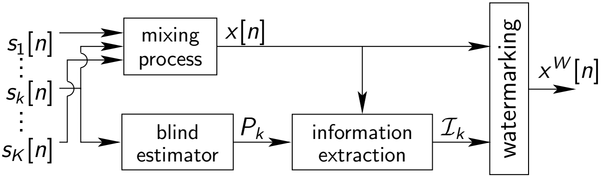
SNR



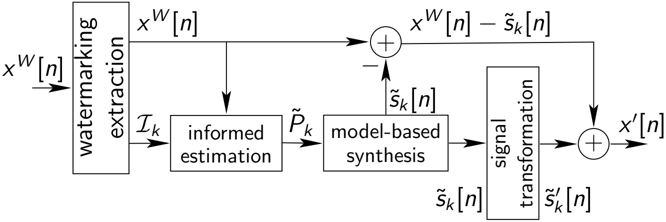
extra information bit rate

Plan

Method Overview



(a) Coder



(b) Decoder

Method summary : coder

input : $s_k[n]$: isolated source signals
output : $x^W[n]$: watermarked mixture

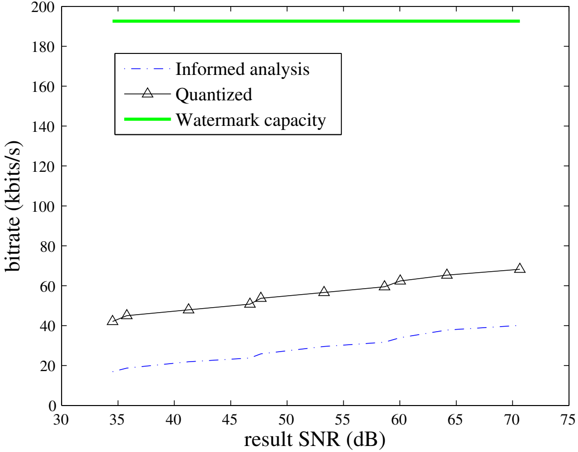
- Estimate $P_{k,l}$ from $s_k[n]$ using reassignment method.
- Compute $b_{a,\omega,\phi}$ from $P_{k,l}$ using the ECUSQ.
- Compute binary mask$[n]$
- Estimate $I_{\sigma,k,l}$ and $\mathcal{I}_{k,l}$ from $\hat{P}_{k,l}$ using the informed spectral analysis method with simulated mixing process combined with watermark.
- Compute $x^W[n]$ using QIM-based watermarking containing mask$[n]$, $I_{\sigma,k,l}$ and $\mathcal{I}_{k,l}$.

Method summary : decoder

input : $x^W[n]$: watermarked mixture
output : $\tilde{s}_k[n]$, $\tilde{P}_{k,l}$: isolated source signals and parameters

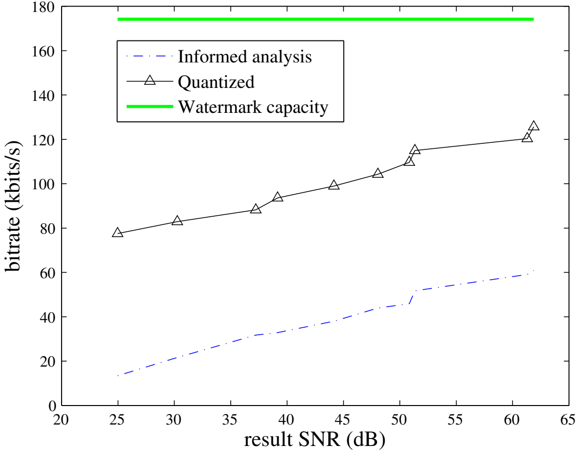
- Recover mask$[n]$, $I_{\sigma,k,l}$ and $\mathcal{I}_{k,l}$ from watermark extraction from $x^W[n]$ and ECUSQ for $b_{a,\omega,\phi}$ computation.
- Estimate $\hat{P}_{k,l}$ using mask$[n]$ and reassignment method.
- Compute $\tilde{P}_{k,l}$ with $I_{\sigma,k,l}$ and $\mathcal{I}_{k,l}$ using the informed spectral analysis.
- Synthesize $\tilde{s}_k[n]$ from $\tilde{P}_{k,l}$.

Results with real sounds



guitar

Results with real sounds



voice

Plan

Happy ending

Conclusion

We have proposed method for informed-analysis of sounds mixture using a quality constraint.

Future work

- theoretical study and comparison with Shannon Lower Bound,
- applications to other audio signal models and estimators,
- optimization of mask[n] computation using prior knowledge about sound structure.